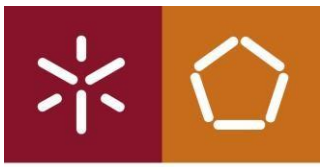

Universidade do Minho
Escola de Engenharia

Ana Catarina Vieira Ribeiro

Previsão dos fatores de risco e caracterização de doentes internados nos cuidados intensivos

Dissertação de Mestrado
Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Trabalho realizado sob a orientação de:
Professor Doutor Carlos Filipe Portela (Professor Auxiliar Convidado DSI)
Professor Doutor Manuel Filipe Vieira Torres dos Santos (Professor Associado do DSI)

Outubro, 2016

Agradecimentos

O desenvolvimento desta dissertação, bem como o seu término, foram momentos definitivos no meu percurso académico, tendo sido constituída por diversos altos e baixos que foram ultrapassados não só por mim, mas principalmente com a ajuda de diversos intervenientes.

Em primeiro lugar um agradecimento em especial ao meu orientador Professor Doutor Carlos Filipe Portela por todo o suporte e paciência ao longo destes meses e pela total disponibilidade, bem como pelo profissionalismo exemplar no exercício da sua orientação. Um agradecimento também ao Centro Hospitalar do Porto pelo fornecimento dos dados que, sem os quais, não teria sido possível realizar esta dissertação.

Aos meus pais que foram os principais fomentadores do meu ensino, que trabalharam muito e se esforçaram mais ainda para me possibilitar a formação num curso extraordinário que me trará, com certeza, um futuro brilhante.

Ao meu namorado e melhor amigo que nunca me deixou desistir em momento algum, tendo sido a principal fonte de inspiração e o meu porto de abrigo para a realização não só desta dissertação, mas também de todo o meu percurso académico e pessoal.

Aos meus amigos, os verdadeiros, que sabem perfeitamente quem são, que me acompanharam deste o primeiro momento e desde então nunca mais me abandonaram. Um muito obrigada por todos os momentos memoráveis que me proporcionaram, momentos os quais foram também uma fonte de inspiração e de força para o término desta dissertação.

Por fim, mas não menos importante, quero deixar um último agradecimento à Universidade do Minho e ao Departamento de Sistemas de Informação pela instituição fantástica que me permitiu crescer a nível pessoal e académico e me permitirá um futuro na área dos Sistemas de Informação (SI) o qual pretendo abraçar com todo o empenho.

Resumo

A Medicina Intensiva (MI) é uma das áreas mais críticas da Medicina. A sua característica multidisciplinar torna-a muito abrangente, reunindo todo o tipo de profissionais de saúde, bem como um local com equipamentos e condições especiais, denominadas Unidades de Cuidados Intensivos (UCI). Tendo em conta o seu ambiente crítico torna-se evidente a necessidade de prever admissões às UCI, pois, para além de constituírem custos adicionais para as instituições e ocuparem recursos desnecessariamente, admissões não planeadas são arriscadas para os doentes que se encontram debilitados. Ao longo dos anos os Sistemas de Informação (SI) têm acompanhando o desenvolvimento da Medicina, tornando-se instrumentos imprescindíveis para o tratamento de doentes, sobretudo através dos Sistemas de Apoio à Decisão (SAD) que apresentam as informações pertinentes sobre os doentes, sem necessidade análise manual de dados. Deste modo, a utilização de SAD na Medicina é crucial, principalmente na MI, em que as decisões têm, muito frequentemente, de ser tomadas com celeridade sempre no melhor interesse do doente. Um SAD pode ser constituído por diferentes técnicas, como é o caso do Data Mining (DM). A presente dissertação envolve descoberta de conhecimento em bases de dados extraídas a partir do sistema de apoio à decisão INTCare, localizado no Centro Hospitalar do Porto (CHP). Foi utilizado um conjunto de técnicas de DM, nomeadamente *Clustering* e Classificação, tendo por base diferentes algoritmos e métricas de avaliação. Assim foram descobertos padrões naturais nos dados, nomeadamente através da formação de dois grupos de características (*Clusters*) dos doentes internados em UCI e identificando os atributos mais críticos nestes *Clusters*. Além disso, foram obtidas previsões com cerca de 97% de capacidade de acertar nos doentes internados (sensibilidade) e que, apesar de criar demasiados Falsos Positivos (63% de especificidade), permitiu obter modelos que permitam que os médicos possam agir de forma proactiva e preventiva, tendo sido esta uma das principais motivações desta dissertação. A presente dissertação serviu para aumentar o número de estudos que aplicam técnicas de DM em MI, particularmente para realização de previsão de internamentos em UCI. Deste modo, contribui-se com conhecimento para a comunidade científica não só de DM, mas também para a Medicina, de modo a potenciar o processo de tomada de decisão médica e na procura pela melhoria dos serviços prestados aos doentes.

Palavras-Chave

Data Mining, Medicina Intensiva, Sistemas de Apoio à Decisão Clínica, Sistema INTCare, Clustering, e Classificação.

Abstract

Intensive Medicine is one of the most critical areas of medicine. Its multidisciplinary feature makes it a very wide area that gathers all kinds of health professionals as well as a place with special equipment and conditions known as Intensive Care Unit. Having in account its critical environment it becomes evident the need to forecast Intensive Care Unit admissions because, besides being additional costs for institutions and occupy resources unnecessarily, unplanned admissions are risky for patients who are debilitated. Over the years, Information Systems are accompanying the development of medicine and have become essentials instruments for the treatment of patients especially using Clinical Systems Decision Support that have relevant information about patients without the need to manually analyse clinical data. Therefore, the use of DSS is crucial in medicine, particularly in the IM in which decisions must very often be taken speedily always in the best interest of the patient. This Decision Support Systems may be constituted by different techniques such as Data Mining (DM). This dissertation involves knowledge discovery in databases extracted from the Clinical Decision Support System being used in Centro Hospital do Porto (CHP) and named INTCare System. It was used a set of DM and rating techniques including clustering and classification which are based on different algorithms and evaluation metrics. Thereby, natural patterns were discovered in the data particularly through the formation of two groups of characteristics (clusters) of patients admitted to Intensive Care Unit and through the identification the most critical attributes in these clusters. Moreover, it was obtained predictions with approximately 97% of ability to get properly forecast admissions to Intensive Care Unit (Sensitivity) and despite creating too many false positives (63% specificity) it also created models that allow doctors to act proactively and preventively which is one of the main motivations of this dissertation. This dissertation served to increase the number of studies that apply DM techniques in Intensive Medicine particularly for performing predictions of admissions to Intensive Care Units. Thus, knowledge was created for the scientific community not only of DM, but also of medicine in order to promote the process of clinical decision-making and to improve services rendered to patients.

Keywords

Data Mining, Intensive Medicine, Clinical Decision Support Systems, INTCare System, Clustering and Classification.

Índice

Índice de Figuras

Índice de Tabelas

Índice de ilustrações

Lista de Abreviaturas, Siglas e Acrónimos

AAG – Asma Aguda Grave

APACHE – *Acute Physiology and Chronic Health Evaluation*

ATA – Artroplastia Total da Anca

BD – Base de dados

BI – *Business Intelligence*

CHP – Centro Hospitalar do Porto

CI – Cuidados Intensivos

DCBA – Descoberta de Conhecimento na Base de Dados

DM – *Data Mining*

DPOC – Doença Pulmonar Obstrutiva Crónica

EH – Emergência Hospitalar

EHR – *Electronic Health Record*

EMR – *Electronic Medical Record*

EPR – *Electronic Patient Record*

HTA – Hipertensão Arterial

LOD – *Logistic Organ Dysfunction*

MI – Medicina Intensiva

MEWS – *Medical Early Warning Score*

ML – *Machine Learning*

MODS – *Multiple Organ Dysfunction Score*

MPM – *Mortality Probability Model*

OSF – *Organ System Failure*

SAD – Sistemas de Apoio à Decisão

SADC – Sistemas de Apoio à Decisão Clínico

SAPS – *Simplified Acute Physiology Score*

SI – Sistemas de Informação

SOFA – *Sequential Organ Failure Assessement*

SWIFT – *Stability and Workload Index for Transfer*

TI – Tecnologias de Informação

UCI – Unidade de Cuidados Intensivos

UCII – Unidade de Cuidados Intensivos Intermédios

UCIP – Unidade de Cuidados Intensivos Pediátricos

WBS - *Work Breakdown Structure*

Capítulo I – Introdução

1.1. Enquadramento

As organizações de saúde todos os dias geram e armazenam grandes volumes de dados. Ao longo dos anos tem-se percebido que a evolução da tecnologia tem um impacto na forma como estes volumes de dados são tratados. Nos últimos anos têm aparecido diversas formas de automatizar todo o processo de gestão de dados (Milovic and Milovic, 2012).

O surgimento de técnicas como Data Mining (DM) e a sua aplicação na saúde tem possibilitado a melhoria dos serviços prestados aos doentes e, consequentemente salvar vidas que, sem recurso a estes métodos, não seria possível. Estas técnicas têm sido implementadas na saúde para responder não só às necessidades dos doentes, mas também às dos médicos, que necessitam de auxílio nas suas atividades diárias (Milovic and Milovic, 2012).

DM é o processo de descoberta de conhecimento e padrões a partir de dados armazenados. Este processo é a fusão de diferentes disciplinas que vão desde Inteligência Artificial (IA) e matemática estatística, até à visualização de informação (Han, 2009).

Segundo Fayyad (Fayyad et al., 1996), o DM faz parte do processo de Descoberta de Conhecimento na Bases de Dados (DCBD), em que são aplicados a análise de dados e algoritmos de descoberta, sendo possível encontrar conhecimento nos dados, previamente desconhecimento.

Nos últimos 70 anos surgiu e tem vindo a ser desenvolvida uma área da medicina que tem possibilitado salvar a vida de doentes críticos, em risco de vida – Medicina Intensiva (MI) (Saúde, 2003).

A MI aplica um conjunto de técnicas e procedimentos que visam a monitorização, diagnóstico e tratamento de doentes, 24 horas por dia. Esta é uma área multidisciplinar, visto envolver várias outras áreas da medicina, como a cardiologia, a cirurgia, entre outras (Saúde, 2003), e está focada no tratamento de falências orgânicas através do suporte aos seis sistemas orgânicos de um doente (Silva et al., 2008).

Para a prática deste tipo de medicina existem as Unidades de Cuidados Intensivos (UCI), que são os locais preparados com os equipamentos e recursos humanos necessários para tratar doentes críticos (Saúde, 2003).

Os Sistemas de Informação (SI) são equipamentos vitais nas UCI pois, dão suporte a diversas atividades no seu dia-a-dia. Começaram por ser simples registos digitais de dados (Haux, 2006) e foram evoluindo até sistemas inteligentes de previsão de eventos críticos ou Sistemas de Apoio à Decisão (SAD) que permitem tomar decisões mais eficazes e que podem ditar a sobrevivência de um doente (Gago et al., 2005).

Com o aumento da população mundial, aumentou também as necessidades de cuidados de saúde das pessoas. Para prestar os serviços na melhor forma possível é necessário gerir eficientemente os recursos utilizados nas UCI. Estes recursos são referentes tanto a camas e equipamentos, como a profissionais de saúde (médicos, auxiliares de saúde, entre outros) (Caldeira et al., 2010).

De forma a gerir todos estes aspetos é necessária uma seleção correta dos doentes que realmente necessitam de cuidados intensivos (Medicine, 1999).

Um dos problemas mais comuns nas UCI está relacionado com as admissões e readmissões não planeadas. Estas representam um custo acrescido para as UCI e podem ser originadas por decisões mal tomadas pelos médicos (Veloso et al., 2015).

Para a resolução dos problemas descritos anteriormente a utilização de técnicas de DM é uma das opções. Através da sua característica preditiva, existem projetos que permitem antecipar eventos críticos num doente (Silva et al., 2008), readmissões (Braga et al., 2014; Veloso et al., 2015; Veloso et al., 2014), entre outros. No entanto, ainda não existem uma aplicação semelhante nas admissões às UCI que permita prever, ou de algum modo caraterizar, os doentes que poderão ser internados em cuidados intensivos através de técnicas de DM. Os projetos existentes para previsão de admissão a UCI são baseados em análises estatísticas.

Existem alguns modelos criados que permitem agrupar os doentes em grupos de risco, com características que indiquem uma posterior admissão à UCI. Deste modo, os médicos podem verificar em que grupo se inserem os seus doentes e estar cientes da possível evolução da sua saúde e, assim, ter um planeamento de possíveis admissões (Goldman et al., 1996; Tsai et al., 2014). Ter um planeamento de possíveis admissões à UCI e caracterização dos fatores de risco, torna-se vantajoso para o planeamento e alocação eficiente de recursos e, ainda, permite que os médicos, doentes e familiares possam estar cientes dos riscos associados à condição de saúde dos doentes (Van den Bosch et al., 2012).

Este projeto de dissertação enquadra-se na utilização de técnicas de DM para confirmar e identificar a existência de padrões nos dados, que permitam prever se um determinado doente contém características que indiquem que poderá ser internado numa UCI e prever qual a probabilidade de ele ser admitido numa UCI. Enquadra-se ainda no projeto de investigação INTCare e os modelos serão desenvolvidos utilizando dados reais provenientes da Unidade de Cuidados Intensivos do Centro Hospitalar do Porto (CHP) (Braga et al. 2014, Marins et al. 2013, Portela et al. 2013, Portela et al. 2014a). O CHP é constituído por 3 instituições: Hospital de Santo António, Hospital Joaquim Urbano e Centro Materno-Infantil do Norte. O CHP presta serviços de cuidados de saúde e desenvolve atividades de investigação e ensino. É constituído por diversos departamentos e diversas áreas da medicina, incluindo a Medicina Intensiva. A UCI está dividida em 3 unidades: UCI I, UCI II e UCI Intermédia (UCII).

O projeto INTCare, em utilização no CHP, é um sistema inteligente de apoio à decisão que permite melhorar os serviços prestados na UCI sugerindo tratamentos e procedimentos adequados, suportando o processo de tomada de decisão clínico em tempo real (Portela et al., 2014a). Este sistema obtém dados de diferentes fontes da UCI e, utilizando técnicas de BI e DM, é extraído conhecimento que permite fazer previsões do desenvolvimento das condições de saúde e verificar outros parâmetros relevantes para o tratamento dos doentes (Gago et al., 2005; Portela et al., 2014a; Veloso et al., 2014).

O motivo para a escolha deste tema deve-se ao facto de este unir duas áreas de interesse: a medicina, nomeadamente a Medicina Intensiva e DM

1.2. Objetivos

Esta dissertação tem como principal finalidade dar resposta às seguintes questões de investigação:

1. "É possível encontrar fatores de risco que permitam caracterizar os doentes internados na UCI?"
2. "É possível prever qual a probabilidade de um doente ser admitido às UCI?"

Para obter resposta às questões de investigação foram definidos os seguintes objetivos:

- Criar padrões tipo de doentes internados na UCI;
- Identificar fatores de riscos de doentes internados na UCI;
- Utilizar técnicas de DM para caracterizar doentes internados na UCI;

- Utilizar técnicas de DM para criar modelos de previsão de internamentos em UCI;

A partir dos objetivos definidos, são esperados os seguintes resultados:

- Conjunto de padrões que permitam definir o doente tipo da UCI;
- Fatores de riscos associados ao internamento em CI;
- Grupo de características dos doentes internados em CI;
- Modelos com boas capacidades de previsão de internamentos em CI;
- Novo conhecimento útil na área dos Sistemas de Informação (SI) e CI.

1.3. Materiais e métodos

Foram utilizados dois métodos, nomeadamente *Design Science Research* como método de investigação e *Cross Industrial Standard Process for Data Mining*, como método de Data Mining, as quais foram combinadas para se adequarem à presente dissertação. De seguida são apresentadas cada um dos métodos, bem como o modo como foram fundidos.

1.3.1. Design Science Research

Design Science Research (DSR) é um método de investigação, sendo definido por um conjunto de técnicas e perspetivas para realizar um trabalho de investigação em sistemas de informação, com a criação de novo conhecimento através do desenvolvimento de artefactos (Peffers et al. 2007).

Peffers (Peffers et al. 2007) define DSR como um processo que contém seis fases e está ilustrado na figura 1.

Figura 1 - Fases do Design Science Research (adaptada de (Peffers et al. 2007))

1. Identificação do problema: O DSR começa pela identificação do problema, onde é apresentado a questão da pesquisa e é justificado o motivo do trabalho. É também elaborada a revisão de literatura para compreender o estado do problema e a importância da sua solução;

2. Identificação dos objetivos da solução: nesta fase são definidos todos os objetivos exequíveis que tem de ser cumpridos para chegar ao artefacto projetado;
3. Desenvolvimento: após os pontos anteriores serem finalizados, inicia-se o desenvolvimento do artefacto de acordo com a projeção anterior, construindo uma solução viável para o problema a resolver;
4. Demonstração: no final da projeção do artefacto, este é aplicado a um contexto para que a sua funcionalidade seja testada. Os testes podem ser feitos através de experiências, simulações, casos de estudo, entre outros;
5. Avaliação: o artefacto é analisado e avaliado para verificar se vai totalmente de acordo com os objetivos estabelecidos e se vai de encontro às condições necessárias para a sua validação. A intenção é analisar a eficácia e eficiência do artefacto;
6. Comunicação: todo o trabalho é documentado e apresentado em artigos científicos, conferencias, entre outros.

1.3.2. Cross Industrial Standard Process for Data Mining

CRISP-DM (*Cross Industrial Standard Process for Data Mining*) foi desenvolvido em 1996, é um modelo de processo *standard* hierárquico, que é constituído por diversas fases com quatro níveis de abstração: fase, tarefa genérica, tarefa especializada e instância de processos (Chapman et al. 2000). A figura 2 representa o ciclo das fases do método CRISP-DM.

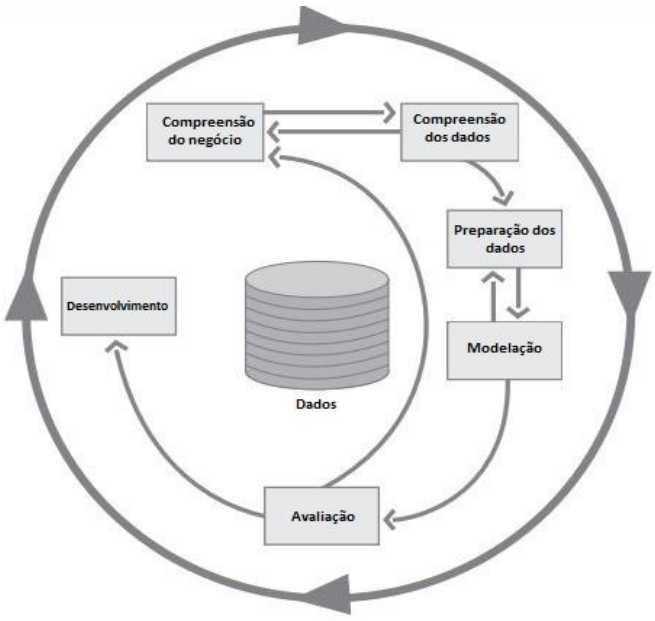

Figura 2 - Fases do CRISP-DM (adaptada de (Chapman et al. 2000)

1. Compreensão do negócio: compreender os objetivos e requisitos do projeto, convertendo-os na definição do problema de DM e construindo um planeamento preliminar para ir de encontro a esses objetivos;
2. Compreensão dos dados: inicia-se com a recolha dos dados ao qual se aplica o seu estudo de modo a compreender melhor e identificar as possíveis falhas e falta de qualidade nos dados. Também é possível retirar as primeiras conclusões em relação a informação "escondida";
3. Preparação dos dados: cobre todo o conjunto de atividades de construção do *dataset* final. Inclui a limpeza e correção dos problemas previamente encontrados e formatação dos dados, para os preparar para modelação na ferramenta de DM;
4. Modelação: Seleção e aplicação de várias técnicas de modelação e de requisitos dos dados. Por vezes, será necessário voltar à fase de preparação dos dados para formatar os dados de diferentes formas;
5. Avaliação: revisão e avaliação do modelo e conclusões. É necessário também rever todos os passos anteriores para certificar que não existe nenhuma falha. Verifica-se se existe algum problema do negócio que não foi suficientemente abordado. No final desta fase, deverá ser possível verificar o resultado do processo de DM.
6. Implementação: apesar de já ter sido criado conhecimento, é necessária a criação de *reports* ou outras formas de apresentar os resultados de forma legível pelo utilizador final. Os resultados podem ser implementados, por exemplo, no sistema de apoio à decisão da organização.

Na tabela 1 estão organizadas as fases deste método, juntamente com as suas tarefas.

Tabela 1 - Fases e subfases do CRISP-DM (adaptada de (Chapman et al. 2000))

Compreensão do negócio	Compreensão dos dados	Preparação dos dados	Modelação	Avaliação	Implementação
Determinar os objetivos do negócio	Coletar dados iniciais	Selecionar dados	Selecionar técnicas de modelação	Avaliar resultados	Planear a implementação
Avaliar a situação	Descrever dados	Limpar dados	Gerar testes	Rever o processo	Monitorização e manutenção do plano

Compreensão do negócio	Compreensão dos dados	Preparação dos dados	Modelação	Avaliação	Implementação
Determinar os objetivos do Data Mining	Explorar dados	Construir dados	Construir modelos	Determinar os próximos passos	Produzir relatório final
Produzir o plano do projeto	Verificar a qualidade dos dados	Integrar dados	Avaliar modelo		Rever o projeto
		Formatar dados			

1.3.3. Métodos combinados

Devido à necessidade de utilizar ambas os métodos descritos nas duas secções anteriores, foi necessário fazer uma fusão de métodos para poder construir um plano de trabalho baseado nas mesmas.

Deste modo, foram observadas e analisadas todas as fases do DSR e CRISP-DM e surgiu o cruzamento de fases representado na tabela 2 que é composta por cinco fases. Por exemplo, faz sentido que a Fase 1 da Junção dos métodos incorpore a primeira fase do CRISP-DM, designada por Compreensão do negócio, juntamente com a primeira e segunda fase do DSR, designadas Identificação do problema e Identificação dos objetivos da solução, devido às suas características comuns.

Tabela 2 - Combinação de DSR e CRISP-DM

			Junção dos métodos				
			Fase 1	Fase 2	Fase 3	Fase 4	Fase 5
Métodos	CRISP-DM	Compreensão do negócio	X				
		Compreensão dos dados		X			
		Preparação dos dados		X			
		Modelação		X			
		Avaliação			X		

			Junção dos métodos				
			Fase 1	Fase 2	Fase 3	Fase 4	Fase 5
		Implementação				X	
	DSR	Identificação do problema	X				
		Identificação dos objetivos da solução	X				
		Desenvolvimento		X			
		Demonstração		X			
		Avaliação			X		
		Comunicação					X

Na fase 1, foi realizada uma pesquisa sobre SI e Medicina, de modo a compreender o negócio em questão, bem como a área em que a dissertação se insere. Esta é importante tanto para a fase seguinte como para compreender quais os problemas a solucionar e para identificar os objetivos a cumprir.

A fase 2 envolveu a exploração dos dados de forma a compreender qual a ligação das tabelas, o seu propósito e qual a relevância dos atributos. Os dados foram ainda analisados na procura de erros e omissões, que foram posteriormente corrigidos, tendo sido os registos em branco o tipo de erro mais encontrado. Estes dados foram então submetidos a técnicas de DM para modelação de dados, nomeadamente *Clustering* e Classificação. Para o *Clustering* foi utilizado algoritmo *k-means* e para a Classificação foram utilizados *Naïve Bayes, Support Vector Machine, Decision Trees* e *Generalized Linear Model*.

Na fase 3 foram utilizadas diversas métricas para avaliação dos modelos: *Davies-Bouldin Index, Silhouete, Inter-cluster distance* e *Distance to Centroids* para o *Clustering*. Para a Classificação foram utilizados Acuidade, Sensibilidade, Especificidade e Taxa de erro.

A fase 4 não foi realizada devido ao facto de esta dissertação ter sido realizada ainda no âmbito académico, não havendo implementação dos modelos. No entanto, importa referir que

esta fase será realizada após conclusão da dissertação e os modelos resultantes serão então incorporados no sistema INTCare do CHP.

Por ultimo, a fase 5 envolveu a escrita do presente documento e de todos os artefactos relacionados, nomeadamente a pré-dissertação, bem como 3 artigos (*Patients' Admissions in Intensive Care Units – A clustering overview; Predicting Patients admission in Intensive Care Units using Data Mining; A study about Patients' Admissions in Intensive Care Units).*

1.3.4. Ferramentas utilizadas

As ferramentas empregues neste projeto foram *Oracle SQL Developer*, onde está armazenada a base de dados utilizada, permitindo executar a análise de dados e, ainda, detém uma componente de *Data Mining* (DM), abarcando diversos algoritmos. É ainda utilizado *Microsoft Excel*, também para análise das tabelas. Para a aplicação das técnicas de DM foram utilizados o Orange e o RapidMiner. Uma breve descrição de cada uma destas ferramentas, bem como a sua utilização no âmbito desta dissertação estão expostas de forma sucinta na tabela 3.

Tabela 3 - Ferramentas utilizadas

Ferramenta	Descrição
Microsoft Excel 2016	Programa integrado na suite de produtos *Microsoft Office* 2016. É um editor de folhas de cálculo com diversas funcionalidades ao nível da construção de gráficos. Nesta dissertação foi útil para construção de alguns gráficos, visualização mais simplificada devido à sua interface simples e intuitiva, bem como para a exploração dos erros das tabelas da base de dados do CHP
Oracle SQL Developer	Ambiente de desenvolvimento de bases de dados da Oracle que permite a gestão e interação com bases de dados. No âmbito desta dissertação, foi responsável por conter a base de dados extraídos do CHP, tendo permitido o tratamento dos dados antes de serem utilizados no processo de DM. Foi ainda a ferramenta utilizada para realizar a parte da Classificação desta dissertação
Orange	Ferramenta *open source* para análise e visualização gráfica de dados, DM e *Machine Learning* (ML), apresentando uma interface extremamente simples e intuitiva. Foi utilizada para a construção

Ferramenta	Descrição
	de cenários e aplicação de técnicas de *Clustering* e realização de diversos testes.
RapidMiner	Ambiente integrado para ML, DM, Text Mining, análise de dados e de negócio. No âmbito desta dissertação, esta ferramenta foi utilizada para a aplicação de técnicas de *Clustering* nos dados, bem como a utilização do *Davies-Bouldin Index*.

1.4. Estrutura do documento

O presente documento está organizado através das seguintes secções:

- Introdução: é feito o enquadramento do projeto de dissertação, é apresentado o âmbito da realização deste projeto e a motivação para tal. São, ainda, identificados os objetivos e os resultados esperados;
- Capítulo II – Revisão de Literatura: reúne as informações necessárias sobre projetos e investigações existentes nas áreas envolventes à do presente projeto de dissertação. Permite definir os conceitos relacionados com o tema e compreender o que o projeto de dissertação irá acrescentar ao que já existe na área. Está dividida em 4 grandes secções: a primeira é a contextualização, em que é explicado como foi conduzido o processo de produção da revisão de literatura; a segunda é o tema da medicina, onde é feita uma breve história da medicina e o papel dos SI na evolução, ao longo do tempo. De seguida, é descrito o processo clínico eletrónico. É ainda abordado o tema da MI, sendo apresentados dados sobre admissões, métodos de triagem e scores médicos utilizados. Ainda dentro da secção da medicina, existe ainda a subsecção do DM na saúde, onde são descritos e relacionados os conceitos de DM, DCBD e SAD, sendo direcionados para a saúde; na secção seguinte é descrito o projeto INTCare; na penúltima secção, estão reunidos os sistemas de previsão de internamentos nas UCI; a secção final é destinada a conclusões retiradas a partir da revisão de literatura;
- Capítulo III – Fase 1: reúne as atividades de compreensão do negócio, bem como a identificação do problema e os objetivos da dissertação;
- Capítulo IV – Fase 2: compreende a fase de compreensão dos dados, onde é apresentada uma descrição e análise dos dados, sendo que posteriormente foram identificados os erros presentes nos dados, para posterior correção. Após a sua correção, são aplicadas as técnicas de DM às duas abordagens previamente definidas e são expostos os resultados;

- Capítulo V – Fase 3: compreende a avaliação e discussão dos resultados obtidos;
- Capítulo VI – Fase 4 e 5: estas fases, que incluem a implementação dos modelos de DM e a descrição do motivo da não realização desta fase, por se tratar de um trabalho no âmbito académico, bem como de que forma foram desenvolvidos os artefactos associados à presente dissertação;
- Capítulo VIII – Conclusões: são apresentados os resultados obtidos, bem como trabalho futuro e ainda contém algumas considerações finais para concluir o trabalho realizado, bem como são apresentadas as principais dificuldades encontradas ao longo da realização da dissertação;
- Referências: lista de referências bibliográficas utilizadas ao longo da elaboração do presente documento;
- Anexos: reúne um conjunto de elementos (imagens, tabelas, etc.) que complementam o trabalho apresentado ao longo do documento.

Capítulo II – Revisão de literatura

2.1. Contextualização da revisão de literatura

O *Capítulo II - Revisão de literatura* reúne o universo de conceitos envolvidos na temática a ser abordada nesta dissertação, bem como trabalhos já existentes e a situação atual, de forma a introduzir o tema e analisar o estado da arte.

Para a escrita desta revisão de literatura iniciou-se uma pesquisa e posterior leitura de artigos publicados no âmbito do projeto INTCare, no qual esta dissertação está inserida. Posteriormente, foi feita uma pesquisa bibliográfica de todo o tipo de conteúdo científico importante centrado nos conceitos chave desta dissertação. Foram utilizados artigos científicos, artigos de conferências, livros, decretos-lei e informações provenientes dos *sites web* de instituições relacionadas com a medicina e a medicina intensiva. Foi ainda utilizada bibliografia de bases de dados digitais: *GoogleScholar, ScienceDirect, IEEE* e *Springer*. Para pesquisa de informações das instituições envolvidas, foi utilizado o motor de busca Google.

A pesquisa foi realizada utilizando um conjunto de conceitos em português e inglês. Os principais conceitos utilizados foram os seguintes (em português os termos foram traduzidos):

- *Data Mining;*
- *Intensive care/medicine;*
- *Intensive care units;*
- *Predictive data mining in healthcare;*
- *Decision Support Systems in healthcare;*
- *Admission to intensive care;*
- *Knowledge discovery in databases ;*
- *Medical information Systems.*

A escolha de bibliografia lida integralmente baseou-se nos seguintes passos:

1. Leitura do *abstract.;*
2. Quando o primeiro passo não era suficiente para a escolha, foi lida a conclusão e a introdução;
3. Para tirar algumas dúvidas sobre a relevância da bibliografia em análise, foram observadas partes do desenvolvimento dos documentos;

4. Após estes passos, se a bibliografia não suscitou interesse, foi excluída.

Após a pesquisa bibliográfica, o passo seguinte foi organizá-la por assunto abordado, sendo posteriormente lida e analisada. A bibliografia escolhida teve por base o ano, o número de citações, o autor envolvido, bem como a relevância do seu tema.

Ao longo da escrita da revisão de literatura e à medida que era necessário complementar informação, foram feitas mais pesquisas ocasionais dentro dos parâmetros, anteriormente descritos. No final da leitura da bibliografia, foi construída uma estrutura do documento, que foi alterada sempre que necessário, dando origem à seguinte organização:

O presente capítulo designado *Capítulo II – Revisão de literatura* está dividido essencialmente em quatro secções. A primeira é a contextualização, que corresponde à descrição de todo o processo de pesquisa e seleção de bibliográfica, estabelecendo parâmetros e conceitos principais utilizados nesta pesquisa que estarão presentes e serão definidos ao longo do documento.

O capítulo seguinte designado *A Medicina* começa com uma pequena definição e introdução histórica sobre esta ciência. É descrita a forma como a medicina e os métodos de diagnóstico se foram desenvolvendo com o passar dos anos, com base no que vários autores escreveram. Também dentro deste desenvolvimento temporal, é referido de que forma a introdução de Sistemas de Informação (SI) e as Tecnologias de Informação (TI), no domínio da saúde, contribuíram para uma melhoria dos serviços prestados. De seguida, após esta introdução de SI e TI na medicina, foi descrito processo clínico eletrónico e definidas as tecnologias utilizadas.

Sendo que esta dissertação está relacionada com a Medicina Intensiva, foram apresentados os conceitos e um pequeno contexto histórico do seu desenvolvimento, incluindo a criação das Unidades de Cuidados Intensivos e, também os tópicos diretamente relacionados com o tema da dissertação, nomeadamente critérios de admissão ou triagem nos cuidados intensivos. Nesta secção são também apresentados os modelos existentes para decidir se um doente deve ou não ser internado nos cuidados intensivos (modelos de triagem). Ainda dentro da secção da Medicina, existe uma parte que descreve alguns dos scores médicos utilizados que permitem a caracterização de diversas condições dos doentes.

De seguida, existe uma outra secção denominada Data Mining na saúde que serve essencialmente para, primeiro apresentar este tema, bem como abordagens utilizadas para o aplicar, principalmente na Medicina Intensiva. É também aqui que é apresentado o processo de

descoberta de conhecimento nas bases de dados, também um conceito importante a presente dissertação. Por fim, para finalizar esta secção, são apresentados os sistemas de apoio à decisão e de que forma estes podem ser aplicados na Medicina, nomeadamente na área da medicina intensiva.

Na secção seguinte, surge o sistema INTCare, onde é apresentado e descrito o seu funcionamento e objetivo, de forma a compreender como um Sistema de Informação inteligente pode ser indispensável na medicina. Serve também para referir em que fase de desenvolvimento se encontra, neste momento.

Existe ainda uma secção direcionada para os sistemas de previsão de internamentos em UCI encontrados na literatura.

O último capítulo é a Conclusão que sumariza e reúne comentários sobre o documento redigido, comparando tudo o que foi abordado nesta revisão de literatura com o trabalho a ser realizado na próxima fase.

2.2. A Medicina

Segundo a definição do dicionário de Oxford, a medicina é a ciência ou prática de diagnósticos, tratamentos e prevenção de doenças[1]. Não é possível datar a origem desta ciência visto que já existe desde pré-história e tem evoluindo até à medicina que conhecemos hoje.

Até ao século XVIII, os locais onde eram prestados serviços hospitalares, eram a casa dos doentes ou dos médicos. Os hospitais eram utilizados sobretudo por pessoas sem condições financeiras, ou por pessoas que não tivessem ninguém que pudesse cuidar deles em casa (Reiser 1981).

2.2.1. A medicina e os sistemas de informação

Segundo Reiser (Reiser 1981), no início do século XVII ainda não existia qualquer tipo de tecnologia direcionada para a medicina. O diagnóstico feito aos doentes baseava-se em três técnicas: aquilo que o doente podia descrever em palavras, a observação do médico ou examinação do corpo e o comportamento e aparência do doente. Durante o século seguinte também não existiram grandes desenvolvimentos relacionados com técnicas e tecnologias de

[1] http://www.oxforddictionaries.com/definition/english/medicine

diagnóstico. No entanto, começaram-se a fazer cada vez mais registos em papel sobre doenças conhecidas e a organizar as informações que existiam.

Desde a sua origem, os Sistemas de Informação (SI) têm tido um papel fundamental na otimização e automatização de processos em todas as organizações. Especialmente na área da saúde, o rápido desenvolvimento da tecnologia permitiu dar resposta a vários problemas que, sem ela, não teriam uma solução simples (Braga et al. 2014).

O objetivo da introdução de sistemas de informação na medicina foi e continua a ser a melhoria da qualidade de cuidados prestados aos doentes. Precisamente por esse motivo, estes sistemas são centralizados nos doentes e na melhoria das suas condições de saúde (Haux 2006).

Torna-se evidente que sem o acesso a informações importantes, não seria possível tomar decisões sem existirem consequências para os doentes que, por vezes, podiam ser fatais.

Ao longo dos anos e assim que se começou a desenvolver tecnologia direcionada à saúde, iniciou-se uma passagem de registos baseados em papel para uma tecnologia baseada em computadores, em que os dados começaram a ser registados digitalmente (Haux 2006).

Haux (Haux 2006) considera que existe uma vantagem e uma desvantagem para esta mudança. A vantagem é a oportunidade da utilização dos dados dos doentes para melhorar a tomada de decisão. A desvantagem é a alta complexidade tecnológica que envolve a construção de tais sistemas.

No início do século XIX, começaram a ser criadas as primeiras tecnologias baseadas em computadores para auxiliar a medicina (Reiser 1981). Nos anos 60, a tecnologia já existente era centrada em pequenas funções de alguns departamentos especiais, como radiologia ou laboratórios. Mais tarde, depois dos anos 70, estes sistemas foram alargados para os hospitais, funcionando como um todo dentro dessas organizações (Haux 2006).

Inicialmente, os sistemas de informação eram utilizados pelos médicos ou pelos administrativos, passando, mais tarde, também a serem acedidos por enfermeiros e outros profissionais de saúde. Atualmente, os doentes e familiares podem utilizar estes sistemas para tirarem as suas dúvidas. Além disso, permitiu que os dados fossem utilizados mais facilmente para investigação, dando origem ao desenvolvimento de novas tecnologias e novas descobertas (Haux 2006).

A partir dos anos 90 já era possível incluir imagens nos dados armazenados, como radiologias (Haux 2006) e, consequentemente, o volume de dados adquiridos eletronicamente aumentaram exponencialmente (Hanson Iii and Marshall 2001).

Haux (Haux 2006) dá o exemplo de um simples centro médico universitário em que a cada ano são guardados cerca de 40 000 documentos de dados clínicos de doentes. Armazenando-os em papel, o seu volume anual seria de 1 500 metros. Caso fossem armazenados digitalmente, ocupariam por volta dos 5 *terabytes*.

Hoje em dia existe uma dominância dos sistemas baseados em computadores, mas ainda existem registos em papel, principalmente por motivos legais ou facilidade de uso (Haux 2006). Haux (Haux 2006) considera esta coexistência prejudicial pois dá origem a mais custos e maior dificuldade em ter acesso à informação relevante.

No final do século XIX começaram a ser criados os primeiros sistemas baseados em regras com abordagem *top-down*, que têm o objetivo de simular o pensamento humano, em que se inicia com um problema complexo que é dividido em várias partes até se chegar a problemas mais pequenos e mais fáceis de resolver. Como exemplo deste tipo de sistemas está o ONCOCIN e o MYCIN (Hanson Iii and Marshall 2001).

Com base no sistema MYCIN, um programa de consultoria baseada em regras que recomendava medicações para doenças infeciosas, surgiu o GUIDON, um sistema com interface gráfica, que permitia investigar uma base de conhecimento durante o diagnóstico de um doente, para suportar a decisão (Richer and Clancey 1985).

O CASNET (*Casual-Association Network*) era também um sistema que representava conhecimento médico e suportava a tomada de decisão. Utilizava modelos estatísticos, reconhecimento de padrões e algoritmos clínicos para diagnóstico. Mostrava um grande nível de capacidade quando confrontado com doenças complexas (Kulikowski and Weiss 1982).

Neste caso, a Medicina Intensiva é uma das áreas da Medicina que mais beneficia da utilização e dos avanços das tecnologias e sistemas de informação (Braga et al. 2014).

Todos os dias, os Médicos deparam-se com situações críticas onde devem ser tomadas decisões demasiado importantes e que podem determinar a sobrevivência de um doente. Estes desafios diários não podem ser tomados sem serem conhecidos os dados dos doentes e, por

vezes, não há tempo para serem estudados e analisados todos os dados devido à urgência da situação (Braga et al. 2014, Gago et al. 2005).

Deste modo, os sistemas de informação, que são o suporte das Unidades de Cuidados Intensivos para o processo de tomada de decisão, devem estar adaptados ao ambiente crítico que os envolve, utilizando agentes inteligentes que agem em tempo real e prestam auxílio aos médicos (Gago et al. 2005).

Um exemplo de um sistema para a medicina intensiva que utiliza agentes inteligentes é o sistema INTCare que será descrito, posteriormente, na secção 2.3.3.

2.2.2. Processo clínico eletrónico

A tomada de decisão clinica é um processo que envolve precedentes legais, filosóficos e médicos. Os próprios doentes têm direito a participar neste processo, caso seja verificada a capacidade para tal. Este processo pode ser centralizado no doente, sendo considerado as suas expectativas, ou centralizados na sua condição de saúde, em que é considerado o prognóstico associado ao doente e a potencialidade de cura da opção de tratamento (Levenson 2010, Braga et al. 2014).

Segundo Portela (Portela et al. 2014a), o processo de tomada de decisão numa UCI é a chave para salvar vidas, portanto o mesmo deve estar baseado em informação válida e útil sobre o estado dos doentes.

Sendo o contexto hospitalar um processo contínuo, onde a todo o momento estão a ser geradas novas informações, é necessário geri-las de forma a tornarem-se úteis noutras situações (Hannan 1996). De acordo com Bates et al. (Bates et al. 2001), o aumento do volume de dados sugere que o erro humano em qualquer área da medicina é muito comum.

A procura de dados sobre os doentes era um processo moroso e nem sempre os médicos obtinham toda a informação que necessitavam, ou a mesma não era apresentada de uma forma legível (Jensen and Bossen 2016).

Os registos físicos existentes eram desvantajosos para o apoio à decisão na medida em que um médico tinha de ter o registo físico na sua posse e a análise só poderia ser feita por quem o possuísse (Hannan 1996).

Com o objetivo de substituir os registos baseados em papel onde eram efetuados os registos clínicos dos doentes, foram criados os *Electronic Health Record, Electronic Medical Record, Electronic Patient Record,* em português, Processo Clínico Eletrónico. Estas tecnologias oferecem formas de apresentar os dados dos doentes, para que os médicos os possam consultar e analisar mais rapidamente (Jensen and Bossen 2016).

Um *Electronic Health Record* (EHR) armazena eletronicamente toda a informação de um doente, relativamente ao passado clínico e ao presente. O EHR é utilizado pelos diferentes utilizadores dos serviços de saúde, no qual podem ser consultadas informações atualizadas sobre a saúde dos doentes (Kirch 2008).

Para acederem à aplicação os utilizadores apenas precisam de ter acesso a um computador com o sistema EHR instalado.

Como vantagens o EHR apresenta a disponibilidade dos dados, a facilidade em extrair conhecimento, o acesso fácil e facilita a partilha de informação com os diferentes especialistas. No entanto, também apresenta desvantagens como os custos de implementação elevados, a necessidade de formar os utilizadores e a necessidade de empregar técnicos de Tecnologias de Informação (TI) para a manutenção deste sistema (Kirch 2008).

O *Electronic Medical Record* (EMR) também é utilizado como sinónimo de EHR e é considerado o sistema anterior a EHR e EPR. Este sistema é constituído pela documentação digital, inserida por uma única pessoa de cada serviço (Kirch 2008).

Segundo Hannan (Hannan 1996), um EMR consiste no armazenamento de todos os dados e informações clínicas em formato eletrónico, com ferramentas de suporte ao conhecimento e de processamento de informação.

O EPR inclui todos os dados clínicos de um doente e combina os dados do EHR de cada doente. Este só pode ser acedido num único sistema (Kirch 2008).

Apesar da utilidade destes sistemas, apenas serão úteis se a informação for apresentada de forma clara e organizada, ou seja, a interface do utilizador tem de ser simples e a informação tem de estar organizada da melhor forma de modo a proporcionar uma leitura e análise rápida (Jensen and Bossen 2016).

Muitos médicos não confiam totalmente nestes sistemas, principalmente quando estão a tratar doentes com historiais muito complexos e complicados. Temem que a informação apresentada não seja suficiente ou não esteja inteiramente correta (Jensen and Bossen 2016).

Segundo Jensen (Jensen and Bossen 2016), é importante que, para a implementação destes sistemas, os médicos definem um padrão para a interface e para a forma como os documentos / informações são apresentados.

Com o avanço das TI e dos SI, estes sistemas evoluíram de uma simples apresentação dos dados dos doentes, para sistemas complexos que permitem fazer previsão de readmissões, falhas de órgãos vitais, eventos críticos, entre outros (Veloso et al. 2015, Braga et al. 2014, Portela et al. 2014a, Ramon et al. 2007, Guiza Grandas et al. 2006, Silva et al. 2008).

2.2.3. Medicina Intensiva

A medicina intensiva (MI) surgiu durante o século XX e o seu contributo tem sido cada vez mais de relevante, pois tem possibilitado de salvar a vida de muitas pessoas que, de outro modo, não teriam como sobreviver (Penedo et al. 2013).

Existem diversas opiniões que envolvem o aparecimento desta área da medicina, mas não é possível determinar uma data, nem um local. De entre as várias opiniões, o surgimento da MI durante guerras é a mais comum. Com o objetivo de tratar os soldados feridos durante as batalhas na guerra da Crimeia e na 2ª Guerra Mundial, foram criadas áreas de recuperação de doentes em risco de vida. Segundo Weil (Weil and Tang 2011), esta é uma das origens da Medicina Intensiva mais apoiada.

A Medicina Intensiva é composta por diversas outras áreas da saúde que se reúnem com o objetivo de prever, diagnosticar e tratar doentes que apresentam condições de saúde críticas, onde há a possibilidade de ocorrerem falhas nos órgãos vitais e que necessitem de ser acompanhados continuamente. Importante salientar o facto de estes doentes terem condições de saúde reversíveis (Saúde 2003).

Através da evolução da medicina intensiva obteve-se novos conhecimentos nas áreas de fisiologia, patologia e terapêutica, modificando a capacidade de cuidados prestados, melhorando bastante a capacidade de tratamento de situações clínicas e doenças que, de outro modo, seriam fatais. Estas melhorias tiveram um grande impacto na qualidade de vida de doentes com doenças graves (Penedo et al. 2013).

2.2.3.1. Unidades de Cuidados Intensivos

Para o tratamento dos doentes em condições críticas de saúde que necessitem de cuidados especiais foram criadas as Unidades de Cuidados Intensivos (UCI). Este tipo de local especializado foi evoluindo ao longo da história, sendo originalmente criado com o objetivo de tratar os feridos durantes a guerras (Weil and Tang 2011).

Estas unidades são responsáveis pelos doentes internados, havendo uma monitorização contínua (24 horas por dia) do estado do doente (ex. sinais vitais e ventilação) e um acompanhamento por médicos de diferentes especialidades que estão preparados para intervir em qualquer situação crítica (Saúde 2003). As UCI são consideradas um ambiente crítico onde a maioria dos doentes se encontram em coma, e a recuperar de situações graves de saúde (Portela et al. 2014a).

Até ao fim da primeira metade do século XX, as UCI não foram reconhecidas oficialmente, nem tinham qualquer equipamento especializado como hoje-em-dia. No entanto, com os avanços tecnológicos verificados a partir dos anos 50, a medicina intensiva começou a dispersar-se e a progredir até existirem as UCI atuais (Weil and Tang 2011).

Atualmente, os doentes internados nos cuidados intensivos podem manter as funções do organismo através de vários aparelhos de suporte de vida, como a ventilação mecânica, hemodialise, entre outros (Saúde 2003).

Para além destes tratamentos, as UCI são também locais onde se fazem investigação, com o objetivo de obter mais conhecimentos sobre doenças e estados de saúde críticos, de forma a acumular e sintetizar informações importantes para melhorar a resposta a situações adversas (Saúde 2003).

Bersten e Soni (Bersten and Soni 2013) distinguem os dois tipos de doentes que devem ser admitidos a internamento numa UCI:

- Doentes que necessitem de monitorização e tratamento, devido ao facto de alguma função vital se encontrar em risco de falha;
- Doentes com falhas orgânicas, mas com hipóteses de recuperação.

Segundo Ramon (Ramon et al. 2007) 30% dos doentes ficam internados numa UCI por um período de tempo mais longo, aumentando o seu risco de morte pois ficam mais sensíveis a doenças e a infeções, devido ao seu estado de saúde debilitado.

De acordo com estudos do Ministério da Saúde (Penedo et al. 2013), com o aumento da esperança média de vida, prevê-se que as necessidades hospitalares dos doentes também aumentem, devido à complexidade das doenças que surgem. Deste modo, torna-se essencial a investigação contínua na área, tanto a nível da medicina como a nível tecnológico.

A questão que surge geralmente dentro deste tema é: terão as Unidades de Cuidados Intensivos capacidade de no futuro tratar o número crescente de doentes que necessitam destes cuidados?

2.2.3.2. Admissão a UCI

De forma a prestar melhores cuidados de saúde aos doentes é necessário, em primeiro lugar, ter uma forma eficiente de os admitir corretamente aos cuidados intensivos. São sugeridos vários modelos de triagem de doentes para auxiliar a decisão de se deve ou não internar um doente numa UCI (Medicine 1999).

Segundo a Administração Central do Sistema de Saúde (ACSS 2012), um doente é considerado internado quando é admitido a uma unidade hospitalar, ocupando uma cama, com permanência superior a 24 horas. No entanto, esta definição não se aplica aos Cuidados Intensivos, visto que, por exemplo, um doente pode necessitar apenas de monitorização contínua durante algumas horas após, por exemplo, a realização de uma cirurgia ou de um procedimento clínico.

Vários autores (Medicine 1999, Caldeira et al. 2010) sugerem a identificação de dois tipos de doentes: os de alto risco e os de baixo risco. Os primeiros são aqueles que estão demasiado doentes para beneficiarem de cuidados intensivos e não teriam qualquer melhoria no seu estado de saúde. Os segundos são aqueles que não estão doentes o suficiente para necessitarem de cuidados intensivos e que podem ser tratados em outras unidades hospitalares. É ainda sugerida a hospitalização de doentes de baixo risco em Unidades de Cuidados Intensivos Intermédias, onde terão outro tipo de tratamento, sem ocuparem as UCI e os seus recursos (Medicine 1999, Arabi et al. 2004).

A má seleção de doentes a serem internados nas UCI provoca a ocupação de camas e outros recursos, como médicos e enfermeiros, que poderiam ser utilizados corretamente com outros doentes (Arabi et al. 2004). Além disso, a estadia prolongada numa UCI expõe os doentes a infeções e inflamações que complicam a sua condição de saúde (Arabi et al. 2004, Ramon et al.

2007). Portanto, é extremamente necessária a existência de algum critério de triagem ou de caracterização de doentes indicados para admissão aos cuidados intensivos. Esta necessidade vai de encontro aos objetivos definidos nesta dissertação.

Os doentes podem ser admitidos de urgência por necessitarem de cuidados especiais que permitam manter as funções dos seus órgãos, ou porque necessitam apenas de serem monitorizados de forma contínua, sendo medicados de acordo com a sua condição clínica (Ramon et al. 2007).

Um dos modelos o apresentado pelos autores (Medicine 1999, Arabi et al. 2004), designado Modelo de Diagnóstico tem por base a identificação do tipo de doença/condição que o doente apresenta, dentro dos tópicos listados abaixo:

- Sistema cardíaco
- Sistema pulmonar
- Sistema neurológico
- Overdose
- Problemas gastrointestinais
- Sistema endócrino
- Cirúrgico (pós operação)

Existe ainda o Modelo de Parâmetros Objetivos em que são analisados parâmetros, como os sinais vitais, valores de exames laboratoriais, resultados de radiologia, ultrassonografia, tomografia, eletrocardiograma ou outras descobertas físicas detetadas pelos médicos (como dilatação anormal das pupilas, percentagem de queimaduras no corpo, entre outros) (Medicine 1999).

Em 2013 foi realizado um estudo (Orsini et al. 2013) no *New York City Health and Hospital Corporation* (HHC) contemplando 165 doentes admitidos a Unidades de Cuidados Intensivos. Neste estudo foram utilizados dados do departamento de emergência, enfermaria, Unidade de cuidados Pós-Anestesia e cirurgia, em que 43% dos doentes foram internados na UCI e 18,3% acabaram por falecer após internamento.

Dos doentes admitidos na UCI, 64,8% são provenientes do serviço de emergência hospitalar, 26,8% das enfermarias, 5,6% das Unidades de cuidados pós-anestesia e os restantes 2,8% eram provenientes da cirurgia. Este estudo apresenta também as principais causas de admissão nos

cuidados intensivos: 28,2% apresentavam Sépsis, 23,9% tinham falhas respiratórias e 11,3% tinham hemorragias gastrointestinais (Orsini et al. 2013).

Quase 90% dos doentes recusados aos cuidados intensivos foram considerados "demasiado bem" para beneficiarem do internamento. No entanto, 6 destes doentes faleceram posteriormente a serem recusados nas UCI (Orsini et al. 2013).

Os médicos nem sempre seguem os critérios de triagem definidos porque, tendo em conta o facto de todos os doentes serem diferentes, por vezes é necessário analisar os seus casos individualmente.

Após o estudo, os profissionais de saúde começaram a prestar mais atenção aos doentes que apresentem doenças cardiorrespiratórias e Sépsis, devido ao facto de estes terem maior probabilidade de serem internados nas UCI (Orsini et al. 2013).

Existem ainda vários autores que apresentam o Modelo de Priorização como um método eficaz de triagem. O método consiste em categorizar os doentes de acordo com os 4 graus de prioridade de internamento (Caldeira et al. 2010, Medicine 1999):

- Prioridade 1: doentes instáveis, em estado crítico que necessitam de monitorização 24 horas por dia e de tratamentos intensivos;
- Prioridade 2: doentes que requerem monitorização e podem necessitar de alguma intervenção por parte dos médicos;
- Prioridade 3: doentes instáveis, com hipóteses de cura diminuídas devido a outra doença ou agravamento do estado de saúde;
- Prioridade 4: doentes que não devem ser internados por terem condições irreversíveis ou por não estarem doentes o suficiente para necessitarem de cuidados intensivos.

No Brasil foi realizado um estudo (Caldeira et al. 2010) para analisar as percentagens relacionadas com a admissão nos cuidados intensivos, em que foram incluídos dados de 359 doentes, dos quais, 66,6% foram admitidos aos UCI. Neste caso, a principal causa de admissão foi a Sépsis. No local deste estudo, foi utilizado o modelo de priorização e verificou-se que os cuidados intensivos eram mais eficazes em doentes com prioridade 1 ou 2, sendo que os restantes graus de prioridade tinham maior taxa de mortalidade pois não estavam no local correto para serem tratados.

Um estudo publicado em 2013 (Mullins, Goyal and Pines 2013), realizado entre 2008 e 2009 em UCI dos Estado Unidos revela que houve 50% de aumento de admissões a cuidados intensivos, em que as duas principais razões de admissão foram dores no peito e dificuldades respiratórios. Em média, os doentes esperaram 5 horas no serviço de Emergência antes de serem encaminhados para a UCI.

Com o agravamento da situação económica mundial, o planeamento detalhado dos recursos (humanos, financeiros e materiais) torna-se imperativo na gestão de uma UCI. Os custos hospitalares agravam-se também, sendo necessário gerir todos os custos e recursos envolvidos.

No caso de uma readmissão não planeada nas UCI são utilizados recursos, provavelmente já alocados no planeamento, que trazem problemas de custos acrescidos e sobrecarga dos recursos disponíveis. Estas readmissões são vistas como um indicador de qualidade do serviço de saúde prestado e está diretamente relacionada com a má decisão por parte dos médicos em dar alta aos doentes (Veloso et al. 2015, Braga et al. 2014, Veloso et al. 2014).

Um doente é considerado readmito se for admitido à mesma unidade hospitalar, com os mesmos sintomas, até 30 dias após a alta (ACSS 2012).

No âmbito do projeto INTCare, forma realizados vários estudos para construir um modelo que permitisse prever readmissões de doentes nos cuidados intensivos. Por exemplo, através da utilização de técnicas de Clustering é possível obter informação sobre o tipo de doente que os médicos devem ter mais atenção antes de lhes dar alta, devido à probabilidade de poder ser readmitido (Veloso et al. 2015, Braga et al. 2014).

2.2.3.3. Scores médicos

Com o objetivo de ajudar os profissionais de saúde em diversas tarefas como a caracterização de um doente e a severidade da sua condição de saúde, surgiram os scores médicos, à cerca de 30 anos atrás (Bouch and Thompson 2008).

Dentro do contexto da medicina intensiva, os scores de severidade existentes permitem melhorar a qualidade dos cuidados intensivos (Silva et al. 2008), permitindo a quantificação do nível de severidade de doença e da probabilidade de morte (Bouch and Thompson 2008).

O score APACHE (*Acute Physiology and Chronic Health Evaluation*) surgiu em 1981 e utiliza vários parâmetros (variáveis fisiológicas, sinais vitais, urina, score neurológico (escala de *glasgow*),

idade e condição mórbida) que podem ter impacto na razão de alta do doente (melhoria ou morte). Por volta de 1985 surgiu o APACHE II que estima o risco de morte, baseando-se no pior valor nas primeiras horas de admissão, quantificando a severidade de doenças. Em 1991, foi criado o APACHE III, semelhante ao anterior, mas com mais variáveis. Este score foi criado com o objetivo de melhorar o poder estatístico utilizado e prever a alta, individualmente, e quais os fatores que a influenciam (Saleh et al. 2015).

O SAPS (*Simplified Acute Physiology Score*) surgiu em 1984, utiliza 14 variáveis fisiológicas e o seu desvio-padrão, utilizando os dados das primeiras 24 horas de admissão. Mais tarde surgiu o SAPS II, que calcula o score de severidade, utilizando o pior valor de 17 variáveis, durante as primeiras 24 horas (Saleh et al. 2015)

Ainda existe o SOFA (*Sequential Organ Failure Assessment*) que classifica o grau de falha de 6 sistemas orgânicos (respiratório, neurológico, cardiovascular, hepático, renal e hematológico) numa escala de 0 a 4. Este é calculado 24 horas após admissão (Bouch and Thompson 2008).

Para além dos scores referidos anteriormente existem ainda muitos outros como: *Mortality Probability Model* (MPM) (Silva et al. 2008, Bouch and Thompson 2008), *Multiple Organ Dysfunction* Score (MODS) (Silva et al. 2008, Bouch and Thompson 2008), *Logistic Organ Dysfunction* (LOD) (Silva et al. 2008), *Organ System Failure* (OSF) (Bouch and Thompson 2008) e MEWS (*Medical Early Warning Score) (Subbe et al. 2001)*.

Existe ainda um score médico utilizado nos cuidados intensivos, designado *Stability and Workload Index for Transfer* (SWIFT) que prevê a probabilidade de um doente ser readmitido numa UCI (Veloso et al. 2015).

2.3. Data Mining na saúde

2.3.1. Data Mining e abordagens

Inicialmente, a análise estatística era aplicada aos dados que eram recolhidos manualmente. Quando os dados começaram a ser armazenados automaticamente nas bases de dados, sentiu-se a necessidade de criar técnicas que os analisassem também de forma automática. Esta necessidade tem evoluindo com o passar dos anos, assim como as técnicas e ferramentas associadas (Asghar and Iqbal 2009).

O Data Mining (DM) surgiu, segundo Turban (Turban, Sharda and Delen 2011), por volta dos anos 80, quando se utilizava a análise estatística e a inteligência artificial para a análise de dados. No caso da medicina, o Data Mining começou a ser utilizado para identificar melhores formas de diagnóstico, tratamento de doenças e descoberta de melhores medicamentos.

Segundo Han (Han 2009), o conceito e técnicas de Descoberta de Conhecimento na Base de Dados, surgiram pela primeira vez em 1989 no *Workshop on Knowledge Discovery from Data* e o conceito de DM em 1995 em *The First International Conference on Knowledge Discovery and Data Mining*. E desde então têm-se realizado inúmeras conferencias e são publicadas revistas sobre o tema.

A figura 3 descreve a relação entre dados, informação e conhecimento.

Figura 3 - Pirâmide do conhecimento (adaptada de (Zaitseva et al. 2015))

Analisando a figura 3, uma grande quantidade de dados pode ser transformada em informação através de aplicação de um contexto. A informação pode ser transformada em conhecimento através da sua análise e agregação que permitem descobrir padrões que representem o conhecimento sobre o seu contexto. Nesta última fase, normalmente é aplicado o DM (Zaitseva et al. 2015).

O termo DM era utilizado para descrever o processo de descoberta de padrões nos dados. Este termo foi evoluindo ao mesmo tempo que eram desenvolvidas novas ferramentas e *softwares* (Turban et al. 2011).

Segundo Jing (Jing 2009), o objetivo principal de DM é a aplicação de métodos e algoritmos para descobrir e extrair padrões a partir dos dados armazenados nas bases de dados, em que anteriormente esses dados foram pré-processados.

Turban (Turban et al. 2011) afirma que esta é uma forma de desenvolver *Business Intelligence* (BI) a partir de dados recolhidos e armazenados numa organização.

DM é uma área que reúne várias outras disciplinas e tecnologias, como gestão de dados, aprendizagem computacional, inteligência artificial, estatística, visualização da informação, reconhecimento de padrões, entre outras, e pode ser aplicada a qualquer setor (financeiro, medicina, gestão...) (Fayyad, Piatetsky-Shapiro and Smyth 1996).

Este processo pretende obter dados de alto nível a partir de dados de baixo nível e não responde a questões nem descobre soluções automaticamente, sem orientação. É necessário formular as questões certas, pois o resultado vai depender da forma como a questão foi colocada (Turban et al. 2011).

A realização de um processo de descoberta de conhecimento com DM pode ter dois grandes objetivos: verificação e descoberta. Na verificação, o sistema verifica apenas a hipótese dos utilizadores. Na descoberta, o sistema encontra os padrões automaticamente. Este último pode ser dividido em previsão, em que o sistema procura padrões nos dados que permita fazer uma previsão do futuro, e em descrição, em que o sistema encontra os padrões e apresenta-os de forma legível por humanos (Fayyad et al. 1996). Estes padrões podem ser encontrados sob diversas formas: regras de negócio, tendências, correlações ou modelos de previsão (Turban et al. 2011).

Quando se fala em dados, a intenção é referir coleções de factos resultantes de experiências e observações, que podem estar criados de diversas formas: Categóricos (Nominal ou Ordinal) ou Numéricos (Intervalo ou Rácio) e, ainda, podem existir em formato de data/horas, texto não estruturado, áudio, entre outros (Turban et al. 2011)

Os métodos mais utilizados de data mining são: Classificação, Regressão e *Clustering* (Fayyad et al. 1996, Turban et al. 2011):

- Classificação: os dados são organizados em classes pré-definidas. Não é possível estabelecer uma fronteira exata que separe as classes;
- Regressão: para determinar a relação entre a variável dependente (*target*) e variáveis de valor real, para prever futuras ocorrências ou valores futuros;
- *Clustering*: identifica grupos ou categorias naturais dentro dos dados.

Dentro dos métodos existem vários algoritmos associados como, árvores de decisão, *k-means, naive bayes*, *Support Vector Machines* entre outros (Turban et al. 2011).

Como foi dito anteriormente, o DM também é utilizado na medicina. Cada vez mais frequentemente, esta técnica ajuda a salvar vidas, na medida em que permite prever eventos críticos que possam ser fatais para os doentes (Berner 2007).

No caso da medicina intensiva é crucial detetar problemas com antecedência ou saber informações importantes a tempo de auxiliar tomadas de decisão urgentes. Antes da utilização de DM ou de qualquer outro tipo de análise computacional dos dados, os médicos analisavam todos os dados dos doentes para tentar tirar conclusões e previsões. Estes dados podem ser descritos com mais de 250 parâmetros diferentes e o ser humano apenas consegue analisar por volta de 7 (Guiza Grandas et al. 2006).

Assim, o papel de DM na medicina intensiva torna-se poderoso e indispensável para a análise dos dados.

Guiza et al. (Guiza Grandas et al. 2006) enumera algumas dos desafios de DM quando aplicados à medicina intensiva. Começa por dar atenção ao facto de existirem fontes heterógenas de dados, em que estas podem ser, por exemplo, dados históricos, dados dos equipamentos de monitorização, dados demográficos, sendo difícil fazer um uso apropriado dos mesmos. De seguida, é direcionada a atenção para as diferentes características de cada ser humano, sendo importante, para além da análise global dos dados, analisar o doente individualmente. Quando lidamos com dados clínicos, o volume de dados armazenados pode ser muito grande. Assim, é também importante existir um sistema de gestão de dados. Estes dados devem ser tratados pois podem estar incorretos, incompletos ou escritos de várias maneiras diferentes. Por último, os dados devem ser interpretáveis para poderem ser analisados facilmente.

2.3.2. Descoberta de conhecimento de bases de dados

Segundo Frawley (Frawley, Piatetsly-Shapiro and Matheus 1992), em 1992 foi estimado que a quantidade de informação armazenada a nível mundial duplicava a cada 20 meses. Uma simples operação como uma chamada telefónica, ou pagamento com cartão de crédito e outras simples atividades do quotidiano, contribuem para este aumento significativo.

Antes dos avanços da tecnologia que usufruímos neste século, grande parte dos dados das organizações eram simplesmente armazenados nas bases de dados sem serem analisados (Fayyad et al. 1996). Frawley (Frawley et al. 1992) define base de dados como uma coleção

logicamente integrada de dados guardados e organizados em um ou mais ficheiros, com o objetivo de facilitar o armazenamento, modificação e obtenção da informação.

Segundo Zaitseva et al. (Zaitseva et al. 2015), se tivermos informações suficientes sobre um determinado tema, podemos identificar factos gerais que o caracterizam. Esses factos são denominados conhecimento.

O ser humano não tem capacidade de analisar muitos dados ao mesmo tempo, nem de encontrar padrões subtis no meio de um grande volume de dados armazenados em bruto. No entanto, começou-se a perceber que o tratamento e a análise destes dados forneciam vantagem competitiva às organizações e permitiam aumentar a eficiência, pois era possível extrair conhecimento útil. Este conhecimento pode ser beneficamente utilizado no processo de tomada de decisão de qualquer organização, sendo, hoje-em-dia, indispensável na área da saúde (Portela et al. 2014a).

Segundo Fayyad (Fayyad et al. 1996), o avanço da tecnologia foi crucial para compensar a lacuna da capacidade de processamento de dados por parte do ser humano. As técnicas computacionais que foram surgindo permitiram o desenvolvimento de ferramentas que analisassem as bases de dados de forma mais rápida e mais eficiente.

Frawley (Frawley et al. 1992) considera que a combinação de interesses das organizações e de investigadores deu origem a uma procura por ferramentas e técnicas de descoberta de conhecimento nas bases de dados.

Devido à falha que existia na compreensão dos dados organizacionais armazenados, surgiu, por volta do seculo XX, o conceito e técnicas de Descoberta de Conhecimento na Base de Dados (DCBD). Percebeu-se que seria um recurso valioso para as organizações se estes dados fossem mais estudados e não, simplesmente, armazenados (Frawley et al. 1992, Fayyad et al. 1996) .

A definição de Frawley (Frawley et al. 1992) para descoberta de conhecimento é a extração não trivial de informação implícita, desconhecida e potencialmente útil, nos dados armazenados.

Segundo Fayyad (Fayyad et al. 1996), o processo de DCBD consiste em métodos e técnicas computacionais desenvolvidas para tirar conclusões estratégicas, descobrindo padrões nos dados armazenados. Zaitseva (Zaitseva et al. 2015) afirma que o objetivo da DCBD é

transformar grandes volumes de dados em informação e conhecimento útil, apresentando-a de forma a ser facilmente interpretada.

Na figura 4 está representado os passos do processo de descoberta de conhecimento nos dados ilustrado por Fayyad.

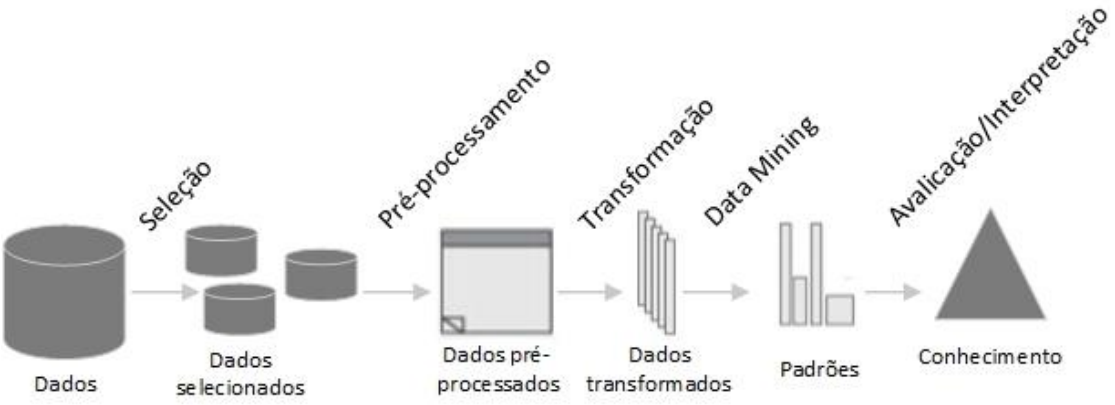

Figura 4 - Processo DCBD (Adaptada de(Fayyad et al. 1996))

O processo de DCBD é constituído por cinco fases (Fayyad et al. 1996):

1. Seleção: selecionar ou criar o conjunto de dados necessário, ou focar-se num conjunto de variáveis, que podem ser extraídos de diferentes fontes de dados;
2. Pré-processamento: limpeza e pré-processamento dos dados selecionados para correção de erros de ortografia e preenchimento adequado dos registos vazios;
3. Transformação: encontrar as características úteis para representar os dados de acordo com objetivo;
4. Data Mining: procura e análise de padrões de interesse estratégico nos dados;
5. Interpretação/Avaliação: verificar os resultados, analisar se os padrões são de interesse ou não, se estes contêm conhecimento útil e avaliar o seu grau de precisão. Após a finalização das fases todas, o conhecimento extraído pode ser documentado e reportado às partes interessadas ou incorporado noutro sistema como um Sistema de Apoio à Decisão.

Segundo Zaitsevsa et al. (Zaitseva et al. 2015), um dos principais problemas da medicina atual é o processamento e a análise da grande quantidade de dados que são criados a partir dos sistemas clínicos eletrónicos. Assim, surge a importância da DCBD, necessitando de processos automáticos ou semiautomáticos que transformem os dados e os analisem de forma a suportar a tomada de decisão clínica.

2.3.3. Sistemas de Apoio à Decisão

Segundo Power (Power 2008), não é possível estabelecer uma cronologia exata sobre a história dos Sistemas de Apoio à Decisão (SAD), nem para qualquer outra tecnologia. Os primeiros SAD a serem implementados surgiram nos finais dos anos 60, devido ao avanço da computação. Os investigadores começaram a estudar as formas que permitiam que os computadores auxiliassem os utilizadores aquando a tomada de decisão.

Com os avanços das TI nascem novas formas de suportar a decisão, existindo atualmente SAD para qualquer setor da indústria. Deste modo, também o termo SAD ganha diferentes definições por parte dos diversos autores da área (Power 2008).

Para Turban (Turban et al. 2011), SAD é um tema abrangente que descreve qualquer sistema computacional que suporta a tomada de decisão numa organização. Este autor afirma que estes sistemas necessitam, em primeiro lugar, de dados para ser possível resolver os problemas. Na figura 5 está ilustrada a arquitetura de alto nível de um Sistema de Apoio à Decisão:

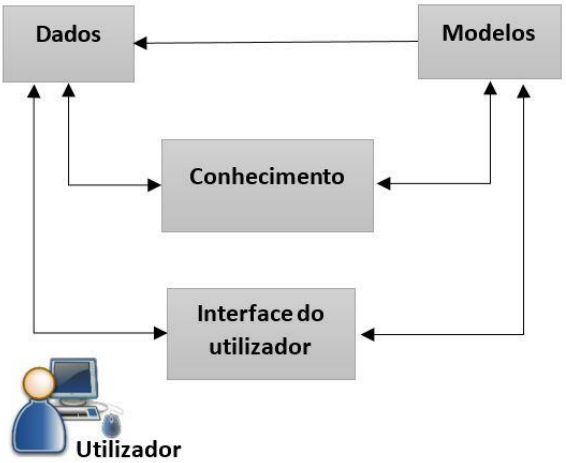

Figura 5 - Arquitetura de alto nível de um SAD (adaptada de (Turban et al. 2011))

De acordo com a arquitetura de um SAD (figura 5), os dados são manipulados por modelos construídos. Quase todos os sistemas têm a componente do conhecimento. Os utilizadores interagem com o sistema através da interface do utilizador (Turban et al. 2011).

Mann e Watson (Mann and Watson 1984) definem SAD como um sistema interativo que fornece ao utilizador acesso a modelos de decisão e dados para suportar a tomada de decisão semiestruturada e não-estruturada.

Segundo a definição de Keen e Scott (Keen and Morton 1978), este sistema junta recursos intelectuais de indivíduos com as capacidades do computador melhorar a qualidade de decisão. É um sistema de suporte baseado em computadores que lida com problemas semiestruturados.

Berner (Berner 2007) descreve um SAD em 3 componentes básicos: a base de conhecimento, o controlo de inferência e um mecanismo de comunicação ao utilizador. A base do conhecimento contém informação compilada quase sempre na forma de *IF-THEN*. O mecanismo de inferência contém regras aplicadas à base do conhecimento que, neste caso, contém os dados dos doentes. Por último, o mecanismo de comunicação serve para mostrar ao utilizador os resultados, para dar apoio nas decisões.

Segundo alguns autores (Power 2008, Turban et al. 2011), SAD pode ser dividido em cinco categorias:

- *Communication-driven*: os sistemas são construídos através de comunicação e tecnologias de suporte à decisão;
- *Data-driven*: analisam grandes volumes de dados dos sistemas das organizações e suportam a decisão, permitindo que os utilizadores extraiam a informação útil retirada desses dados;
- *Document-driven*: integram vários tipos de tecnologias de processamento e armazenamento de dados para fornecer análise de documentos;
- *Knowledge-driven*: contêm conhecimento sobre um domínio em particular que permite resolver os problemas dentro desse domínio;
- *Model-driven*: utilizam modelos para realizar questões "WHAT-IF" e outros tipos de análises e permite acesso ao utilizador para controlar esses modelos.

No caso da medicina, existem os Sistemas de Apoio à Decisão Clínicos (SADC), que são definidos por Berner como sistemas computacionais desenhados para auxiliar a tomada de decisão dos profissionais de saúde no momento em que a decisão é feita. Este tipo de sistema foi criado para diminuir os erros médicos e aumentar a segurança e bem-estar dos doentes. Segundo Berner, os primeiros SADC foram criados apenas para mostrarem os dados dos doentes (Berner 2007).

O Centro Hospitalar de São João no Porto implementou uma plataforma *Business Intelligence* para suporte à decisão capaz de recolher, armazenar e incorporar dados clínicos dos doentes. Chama-se VITAL (Vigilância, Monitorização e Alerta) e já venceu vários prémios, incluindo o *Microsoft Health Innovation Awards 2014*. (São João 2014)

Outro SAD presente na medicina, neste caso, na medicina intensiva, é o Guardian, que utiliza agentes autónomos e vários algoritmos, dando soluções aos profissionais de saúde, como terapias a aplicar, diagnósticos das condições dos doentes, entre outros, sendo utilizado em tempo real (Larssan and Hayes-Roth 1998).

Existe ainda o Micromedex[2], um sistema inteligente em tempo real, que monitoriza os doentes para identificar quais estão em risco e quais os candidatos a intervenções. Além disso, apresenta os resultados em *dashboards*, utilizando dados de diversas fontes e auxilia os médicos no seu trabalho diário. Esta tecnologia pode ser acedida remotamente através de computadores ou dispositivos móveis, a qualquer hora.

Na secção 2.5. está descrito outro Sistema de Apoio à Decisão Clínico para a medicina intensiva, designado INTCare, utilizado atualmente no Centro Hospitalar do Porto e que permite prever diversas condições de saúde e eventos críticos.

A utilização destes SADC tem mostrado melhorias no tratamento dos doentes e nos custos de tratamento (Berner 2007). Assim sendo, percebe-se a importância do desenvolvimento e da sofisticação destes sistemas e da sua implementação na medicina e, neste caso, na medicina intensiva. Sendo esta uma área onde os doentes se encontram em condições de saúde muito críticas e onde muitas decisões têm de ser tomadas rapidamente com o objetivo de salvar a vida de pessoas, é necessária a utilização destes sistemas de suporte à decisão em tempo real, que apresentem soluções rápidas e eficazes aos profissionais de saúde.

2.4. Sistema INTCare

O sistema INTCare é um SAD com características *pervasive* desenvolvido para a área da Medicina Intensiva. Atualmente encontra-se instalado na Unidade de Cuidados Intensivos do Centro Hospitalar do Porto (CHP), Porto, Portugal. Este sistema encontra-se em constante desenvolvimento e testes e realiza o processo de Descoberta de Conhecimento em Bases de Dados automaticamente e em tempo real (Braga et al. 2014, Marins et al. 2013, Portela et al. 2013, Portela et al. 2014a).

O objetivo da criação deste sistema foi o de modificar o tempo de resposta de reativa para a proactiva (Braga et al. 2014) e melhorar a qualidade de tratamento dos doentes, sugerindo tratamentos, terapias e procedimentos adequados e, ainda, permite prever as condições clínicas

[2] http://micromedex.com/360-care-insights

do doente nas próximas horas. Com o INTCare é possível melhorar a base de conhecimento na qual os profissionais de saúde se baseiam para tomar decisões (Veloso et al. 2014).

Este sistema permite apresentar a informação dos doentes em qualquer lugar e a qualquer hora, para auxiliar o processo de tomada de decisão (Braga et al. 2014, Marins et al. 2013, Portela et al. 2013, Portela et al. 2014a, Veloso et al. 2015):

- Dados clínicos dos doentes (sinais vitais, resultados de laboratório, balanço hídrico);
- Eventos críticos (SPo2, frequência cardíaca, temperatura, pressão sanguínea, produção de urina);
- Scores médicos de UCI (SAPS II, SAPS III, Glasgow, SOFA, MEWS e TISS);
- Probabilidade de falência orgânica (cardiovascular, respiratório, renal, hepático, hematológico e neurológico);
- Probabilidade de readmissão;
- Probabilidade de alta clínica;
- Arritmias cardíacas e eventos críticos em pressão sanguínea;
- Barotrauma e extubação;
- Tempo de estadia.

Inicialmente os dados eram obtidos de forma manual. A maioria dos dados eram registados em papel e gravados manualmente na Base de Dados (Portela et al. 2014a). Após a implementação deste sistema, os dados passaram a ser recolhidos e armazenados através dos agentes inteligentes e dos procedimentos automáticos (Veloso et al. 2014). A automatização de processos através da utilização de agentes inteligentes permite a recolha dos dados e realização do processamento (ETL) automaticamente e em tempo-real (Portela et al. 2014c).

A figura 6 apresenta a arquitetura do sistema INTCare que está dividido em 4 subsistemas: aquisição de dados, gestão do conhecimento, inferência e interface do utilizador (Braga et al. 2014, Portela et al. 2013, Veloso et al. 2015).

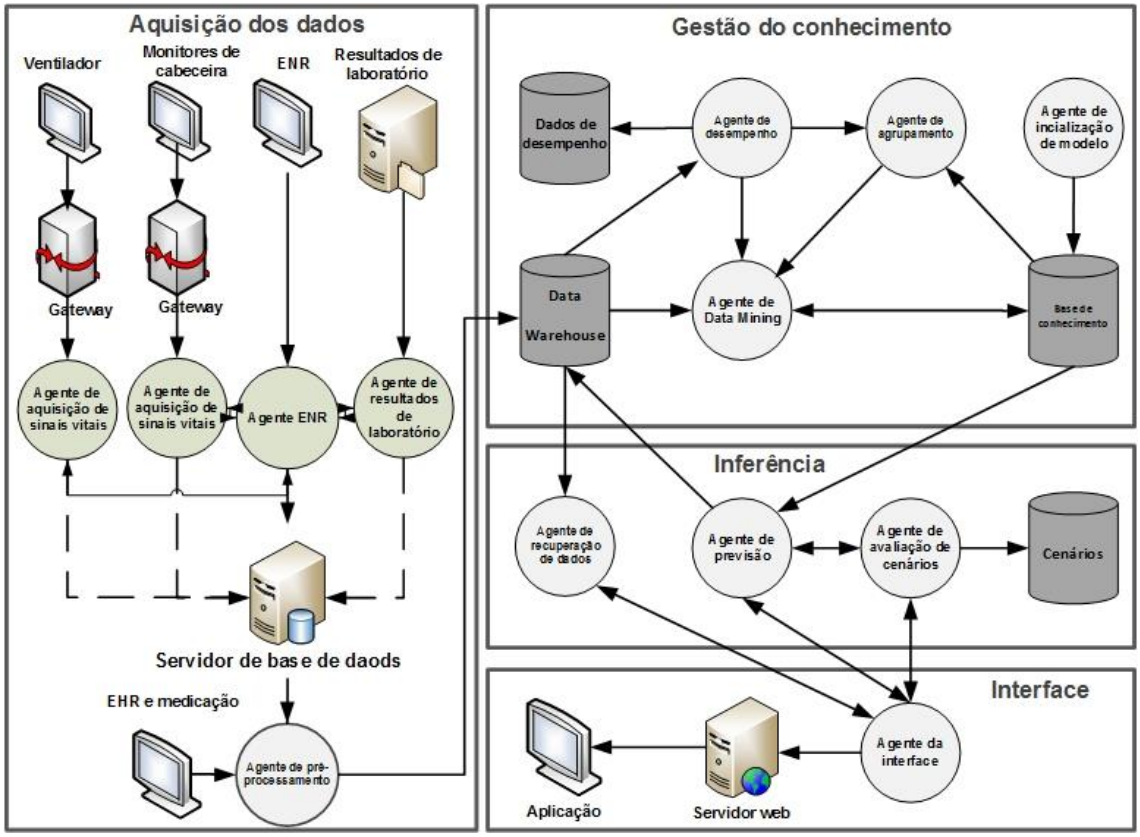

Figura 6 - Sistema INTCare (adaptado de (Marins et al. 2013)

A recolha de dados é realizada a partir de 5 fontes: monitores de cabeceira (sinais vitais e ventiladores), Processo Clínico Eletrónico, Folha de Enfermagem Eletrónica, laboratório e medicação (Portela et al. 2013, Portela et al. 2014a, Portela et al. 2014c). Estes dados são armazenados na base de dados e enviados para o sistema INTCare após pré-processamento dos dados (Gago et al. 2005).

Na gestão do conhecimento, os dados passam pelo processo de ETL (*Extract, Transform, Load*) e são armazenados num Data Warehouse. Estes são utilizados pelos agentes de Data Mining (DM) que criam novos modelos para a base de conhecimento. O agente de inicialização de modelo preenche também a base de conhecimento com modelos do treino *offline*. O agente de desempenho acede ao Data Warehouse para realizar atualizações e o agente de agrupamento pretende melhorar os sistemas de previsão, combinando vários modelos (Gago et al. 2005).

Na fase de inferência, os dados solicitados pela interface são recuperados do Data Warehouse, através do agente de recuperação de dados. O agente de previsão aplica modelos armazenados na base do conhecimento e no Data Warehouse, para responder às questões. Para isso, são criados cenários avaliados pelo agente de avaliação de cenários (Gago et al. 2005).

Por último, na interface, o agente da interface recebe os resultados obtidos e permite que sejam consultados pelos profissionais de saúde através do sistema INTCare e da Folha de Enfermagem Eletrónica, após estes serem enviados para um servidor web que permite a sua disponibilização em qualquer lugar e a qualquer hora (Gago et al. 2005). Os dados podem ser consultados através de qualquer dispositivo móvel com acesso à internet (Portela et al. 2014a).

Este sistema chama, ainda, a atenção para a importância da criação de sistemas e ambientes inteligentes (AmI) capazes de estabelecer interação entre humanos, máquinas e o ambiente envolvente. No caso da saúde, este tipo de sistema combina a informação sobre as condições dos doentes (Marins et al. 2013).

Recentemente, os avanços do projeto INTCare permitiram a identificação de eventos críticos para todas as variáveis monitorizadas (sinais vitais, ventiladores):

- Deteta valores críticos, utilizando uma tabela com valores mínimos e máximos;
- Deteta variações nos atributos utilizando a análise de uma curva de valores.

Deste modo, é possível compreender rapidamente a condição do doente, visto que o sistema envia alertas assim que ocorrer alguma variação nos valores (Portela et al. 2015a).

Um dos eventos críticos que é previstos é a arritmia cardíaca – ritmo cardíaco fora do normal – um dos principais problemas verificados nos doentes. Este evento crítico deve ser previsto com antecedência pois aumenta o risco de outros ataques ou insuficiência cardíaca. Utilizando o *Support Vector Machines* foram obtidos resultados com sensibilidade 95% e precisão de 85%, o que mostra capacidade de o sistema prever futuros eventos críticos e auxiliarem os médicos a tomar decisões (Portela et al. 2014b).

É, também, possível prever eventos críticos relacionados com pressão sanguínea. A utilização de modelos de DM possibilita o cálculo da probabilidade de um doente ter alguma complicação relacionada com pressão sanguínea, apresentando sensibilidade dos resultados de cerca de 90 a 95%. Estes resultados são cruciais para prever estes eventos críticos que podem dar origem a ataques cardíacos, falhas orgânicas e outros. (Portela et al. 2015b).

Este tipo de computação inteligente está inserida no conceito de Ambiente Inteligente (AmI) e permite o sucesso da interpretação da informação extraída de sensores e pode-se adaptar às necessidades do utilizador, sendo possível ser implementado noutros serviços hospitalares (Portela et al. 2015a).

Assim, é possível ter um sistema *pervasive*, porque o utilizador não precisa de estar preocupado com a informação gerada devido ao facto de este estar preparado para o alertar em situações anormais. Esta característica é vantajosa para os médicos pois é demasiado difícil analisar e comprar todos os valores em tempo real (Portela et al. 2015a). Em termos gerais os médicos preferem ser alertados para o pior cenário de DM de modo a poderem prevenir a sua ocorrência. Assim é preferível que seja previsto um evento que não ocorra, do que não ser previsto e ocorrer (Portela et al. 2015b).

O desenvolvimento deste projeto tem sido uma mais-valia para a medicina intensiva devido ao facto de corrigir muitas lacunas nas unidades de saúde e melhorar claramente o tratamento prestado aos doentes. Nota para a utilização de técnicas de DM para previsões de eventos críticos que não conseguem ser previstos apenas através do conhecimento dos médicos.

A informação retirada das fontes do sistema INTCare podem ser representadas em diversas formas como gráficos, tabelas e texto. Atualmente encontra-se em fase de implementação uma nova forma de visualização desta informação, em que os dados dos doentes são apresentados tendo por base uma linha temporal (Braga et al. 2015).

Esta forma de representação dos dados apresenta os eventos críticos do doente de forma cronológica, sendo possível verificar em que data ocorreu, a sua duração, a frequência e período de ocorrência. Permite, ainda, que os médicos possam ver a evolução da saúde do doente ao longo do tempo, ou seja, funciona como uma estruturação de dados clínicos históricos, criando novo conhecimento em tempo real. Por último, esta linha temporal apresenta as previsões de agravamento da condição de saúde dos doentes (Braga et al. 2015).

Apesar das funções indispensáveis do sistema INTCare e de outros sistemas inteligentes utilizados na área da saúde, não foi encontrado nenhum sistema na revisão de literatura que esteja direcionado para a previsão de uma futura admissão em unidades de cuidados intensivos. Neste sentido, este projeto de dissertação procurou desenvolver modelos de DM utilizando os dados recolhidos pelo sistema INTCare. Ao nível da arquitetura do sistema INTCare, este projeto interfere e apresenta resultados na fase 2 e 3: Gestão do Conhecimento e Inferência

2.5. Sistemas de previsão de internamentos em UCI

Um dos projetos existentes, realizado em Nova Iorque, tem como finalidade a triagem de doentes de alto risco de internamento em UCI após uma Artroplastia Total da Anca (ATA). Consiste

num modelo de identificação de riscos pré-operatório, que permite prever se um doente terá maior probabilidade de ser admitidos a UCI após a operação (Kamath et al. 2013).

Os principais objetivos deste projeto são a redução de admissões não planeadas, de complicações graves e da mortalidade no hospital (Kamath et al. 2013). Para a construção deste modelo foi utilizada a análise estatística em 175 doentes submetidos a ATA, dando origem a 5 fatores de risco: idade superior a 75 anos, revisão da cirurgia, depuração da creatinina inferior a 60mL/MIN, antecedentes de ataques do miocárdios e índice me massa corporal superior a 35kg/m^2 (Kamath et al. 2013).

Foram utilizados testes de *Pearson Chi-square e T-student*, para analisar dados demográficos, taxa de altas de doentes com admissão não planeada às UCI, morbidade e mortalidade. Estes dados foram calculados com um processador SPSS (Kamath et al. 2013).

Este modelo permitiu diminuir as admissões não planeadas de 7,1% para 2,2%, o que mostra uma importância da caracterização de doentes e identificação de fatores que podem prever complicações posteriores às operações. (Kamath et al. 2013).

Um outro exemplo é o estudo de controlo de caso realizado na Holanda a crianças que foram admitidas a Unidades de Cuidados Intensivos Pediátricos (UCIP) devido a Asma Aguda Grave (AAG). O grupo de controlo são as crianças com asma que não foram admitidas a UCIP (Van den Bosch et al. 2012).

Foram utilizadas 25 variáveis candidatas a fatores de risco que foram analisadas através de análise estatística. Foi utlizado o teste *Mantel-Haenszel* para identificar variáveis associadas à asma e a necessidade de hospitalização, através de *softwares* de estatística (SPSS e SAS) (Van den Bosch et al. 2012).

Inicialmente foram identificados 14 fatores de riscos, mas, após uma análise de regressão logística condicional, sobraram apenas 4 fatores de risco: fumador ativo ou passivo, alergias, hospitalizações anteriores devido a asma e casa sem condições higiénicas

Bosch et al. (Van den Bosch et al. 2012) consideram importante que os médicos, pais e doentes estejam cientes deste fatores de riscos para prevenir complicações na saúde dos doentes e diminuir as admissões não planeadas nas UCIP.

Num outro estudo foram utilizados dados provenientes de 7 hospitais da Califórnia, referentes a 10 682 doentes com dor aguda no peito (dados entre 1984 e 1986) para serem identificados

potenciais fatores que possam permitir a previsão do desenvolvimento de outras complicações consequentes que requerem admissões a UCI. Para o conjunto de dados de validação foram utilizados dados de 4676 doentes de um dos hospitais. Os dados foram submetidos a análise estatística, utilizando testes como *Chi-square, T-student, Proportional-Hazards* e *Mantel-Haenszel* (Goldman et al. 1996).

No final, os fatores de riscos identificados que devem ser considerados nestes casos de dor aguda no peito são: ondas Q no eletrocardiograma que indiquem enfarte do miocárdio, aumento da segmentação ST e outras mudanças eletrocardiográficas significativas. A partir destes fatores, foram identificados 4 grupos dentro do conjunto de dados de validação (Goldman et al. 1996).

Esta identificação do tipo de doentes e dos fatores de riscos associados à dor aguda no peito permite a previsão de complicações na saúde do doente que levem a uma admissão não planeada nas UCI. Estes aspetos também são importantes para decisões referentes à hospitalização do doente, como por exemplo, o nível de cuidados que o doente necessita (Goldman et al. 1996).

Um outro estudo realizado em Itália, em 1 297 doentes após uma grande recessão pulmonar verificada entre 2000 e 2006, permitiu a construção de um sistema de *scores* para prever a admissão a UCI de doentes com complicações após recessão pulmonar. As variáveis utilizadas foram analisadas através da regressão logística *stepwise* e avaliadas com análise *bootstrap*. Foram identificados 3 grupos de classes de risco a partir dos *scores* das suas variáveis. Através deste sistema podem ser determinadas as necessidades pós-operatórias dos doentes, que podem beneficiar de uma admissão direta às UCI em vez de irem para outras unidades hospitalares (Brunelli et al. 2008).

Um estudo realizado na França permite estratificar doentes com Pneumonia em 4 categorias de risco de admissão à UCI dentro de 3 dias após a sua apresentação numa emergência hospitalar (EH). As categorias de 1 a 2 são de risco moderado, apesar de ser necessária alguma monitorização, não necessitam de ser admitidos a UCI. Nas categorias de 3 a 4, os doentes devem ser transferidos para as UCI, posteriormente à sua entrada como EH (Labarère et al. 2012).

Este estudo mostra ainda a necessidade uma previsão de internamentos em UCI, pois os doentes que são admitidos sem estar planeado têm maior taxa de morte devido ao seu estado de saúde debilitado (Labarère et al. 2012).

Segundo Labarère et al. (Labarère et al. 2012), existem vários indexes de avaliação de severidade de doença como *Pneumonia Severity Index* (PSI) e o CURB-65 (*Confusion, Urea,*

Respiratory, Blood pressure e age => 65), que permitem identificar doentes de baixo risco, mas não são uteis para a previsão de internamentos. Para isso criaram o Ri*sk of Early Admission to Intensive Care Unit* (REA-UCI).

Este projeto utiliza 11 parâmetros clínicos, laboratoriais e de raio-x, que são utilizados para análise estatística através de testes como *Wilcoxon, Chi-square, Fischer Exact*, entre outros (Labarère et al. 2012).

O REA-UCI é comparado a outras ferramentas como o PSI e o CURB-65, apresentado melhores valores em termos de previsão de admissões na UCI. No entanto, segundo os autores (Labarère et al. 2012), o REA-UCI não demonstra vantagem de precisão sobre os outros modelos.

Um último projeto encontrado dentro do âmbito da previsão de admissões na UCI, é um estudo de controlo de caso direcionado a doentes com Sépsis, realizado na China. O objetivo deste projeto era verificar se os conceitos de Sépsis, *Predisposition, Insult, Response and Organ Dysfunction* PIRO), pode ser utilizado para prever o risco de transferências não planeadas para UCI e, deste modo, diminuir a morte no hospital e otimizar os recursos (Tsai et al. 2014).

Foram utilizados dados de 313 doentes que tiveram admissões na UCI não planeadas dentro de 48 horas após terem chegado à EH e, para ao grupo de controlo, foram utilizados dados de 736 doentes aleatórios que não foram admitidos. Dois terços do total dos doentes constituem um grupo de derivação que foi utilizado para construir o modelo PIRO. O terço restante é o grupo de validação (Tsai et al. 2014).

Neste caso também foi utilizada a ferramenta SPSS para realizar a análise estatística e foram identificados 16 fatores de risco para prever a detioração de doentes dentro de 48 horas após EH (Tsai et al. 2014).

Segundo o Tsai et al (Tsai et al. 2014), a monitorização contínua poderia ser uma ajuda para verificar sinais de detioração, mas isso implicaria a utilização de recursos que poderiam estar a ser aplicados de forma mais eficiente noutros doentes. Os autores (Tsai et al. 2014) pretendem, ainda, juntar este modelo com os sistemas *Electronic Health Record* (EHR) para, posteriormente, construir um Sistema de Apoio à Decisão (SAD).

Na pesquisa realizada não foram encontrados sistemas de previsão de admissão na UCI que utilizassem técnicas de Data Mining (DM). No entanto, existem sistemas, como é o caso do

INTCare (Braga et al. 2014, Veloso et al. 2014, Veloso et al. 2015), que possuem funções de previsão de readmissões a UCI, através de técnicas de DM.

Existe um score médico designado *Medical Early Warning Score* (MEWS) que pode ajudar a identificar doentes em situação de risco que poderão necessitar de admissão numa UCI e a severidade de doença (Subbe et al. 2001). No entanto, este sistema de *score*, por si só, não é o mais eficiente para previsão de internamentos, pois só utiliza dados de sinais vitais (Tsai et al. 2014).

2.6. Conclusão da revisão de literatura

A medicina é uma área com necessidade de constante desenvolvimento e evolução, não só em termos científicos, como também em termos tecnológicos. Sendo os SI uma componente importante do quotidiano das instituições de saúde, o estudo contínuo de novas formas de suportar o processo de tomada de decisão clinico torna-se, claramente, benéfico, não só para os médicos, mas também para os doentes.

Na revisão de literatura apresentada, foi possível verificar que a relação existente entre a medicina e os SI já é longa e ainda tem muito para evoluir. Muitas vezes os SI conseguem resolver problemas que os médicos só por si não conseguem.

No caso das técnicas de DM, também se verifica que a sua utilização para a medicina é bastante vantajosa. É, assim, possível agrupar o conhecimento médico e aplicá-lo de forma mais rápida e mais eficaz.

Principalmente na MI, a utilização de técnicas de DM é indispensável, devido ao facto de uma UCI ser um ambiente crítico e ser necessário tomar decisões que, por vezes, podem determinar a sobrevivência de um doente. Estas técnicas possibilitam que os médicos obtenham a informação relevante e correta que necessitam para basear as decisões, sem precisarem de desperdiçar tempo a analisá-la manualmente.

Um dos aspetos importantes do DM na MI é a previsão de admissões ou readmissões. Quando estas não estão planeadas, as UCI têm de reajustar os seus recursos, o que, por vezes, pode prejudicar outros doentes e são, ainda, custos adicionais para a instituição de saúde. Através da pesquisa feita, é possível concluir que existem alguns sistemas de previsão de internamentos baseados em análise estatística. O único sistema encontrado que utiliza técnicas de DM é direcionado para a previsão de readmissões.

Pode-se, ainda, verificar que o DM e as suas previsões são aplicadas a situações / áreas específicas dentro da MI, ou a doenças/condições de saúde, devido ao facto de a MI ser composta por diversas disciplinas. Deste modo, será selecionada uma área dentro da MI para aplicar as previsões de internamentos e caracterizar os doentes internados em UCI.

Torna-se, assim, óbvia a importância da existência de um sistema que permita caracterizar doentes internados nas UCI e prever admissões a UCI.

Este sistema enquadra-se na área dos SAD, pois será um sistema que suportará o processo de tomada de decisão clínico, apresentando características de doentes que poderão ser internados nas UCI. Apresentará as informações relevantes para que os médicos não necessitem de analisar os dados manualmente, permitindo que estes se possam basear no sistema para tomar decisões em relação aos seus doentes. De realçar que os modelos a desenvolver não substituirão o conhecimento clínico, mas serão uma importante ajuda no processo de decisão.

Através de técnicas de DM serão identificados grupos de características de doentes que têm maior probabilidade de ser internados em UCI. Caso o doente se enquadre em algum destes grupos, os profissionais de saúde terão acesso a essa informação, podendo utilizá-la como base para a sua decisão.

Deste modo, o projeto de DM a desenvolver irá fornecer aos médicos uma base de informação que permita prever o internamento dos seus doentes nas UCI. Será, ainda benéfico, no sentido de melhorar o planeamento de internamentos e de recursos utilizados, melhorando o serviço prestado aos doentes.

Capítulo III – Fase 1

Segundo os médicos o internamento que advém de admissões não planeadas nas Unidades de Cuidados Intensivos (UCI) contém bastantes riscos para os doentes que se encontram mais debilitados. Estes internamentos podem agravar as doenças ou até mesmo provocar morte. Além disto, prejudica a gestão dos hospitais, ocupando recursos adicionais. Por sua vez, a utilização não planeada de recursos pode, ainda, afetar negativamente outros doentes, alterando os seus planos de tratamento, ou adiando cirurgias. Estes são uns dos principais problemas que as UCI encontram no seu dia-a-dia e é de extrema relevância criar métodos de planear com antecedência os internamentos.

Deste modo, surge a oportunidade de utilizar técnicas de Data Mining (DM) para realizar estas previsões, baseadas nos dados existentes dos doentes. Esta previsão pode ser feita após a descoberta de padrões nos dados, que permita criar grupos de doentes internados em UCI, com características semelhantes.

Esta dissertação tem como principais objetivos de negócio caracterizar doentes internados em UCI e verificar a existência de padrões que permitam prever se um doente poderá necessitar de CI no futuro.

Os objetivos de DM são criar modelos úteis, capazes de identificar grupos de doentes internados em UCI com características semelhantes e capazes de prever internamentos em UCI.

A solução passará por modelos de DM com utilidade para o INTCare do Centro Hospitalar do Porto (CHP) que permita auxiliar os médicos aquando a tomada de decisão, apresentando informação útil. Esta informação útil passa por apurar se as características do doente estão presentes em algum dos grupos identificados, com maior probabilidade de ser internado em UCI. Deste modo, os médicos poderão dar mais atenção aos doentes que poderão necessitar de Cuidados Intensivos (CI) e prevenir o internamento dos mesmos.

Capítulo IV – Fase 2

No presente capítulo estão apresentadas todas as atividades associadas à fase 2 e 3, incluindo, não só a exploração, tratamento e modelação de dados, como também a avaliação dos resultados obtidos.

4.1. Recolha e descrição dos dados

Esta fase foi iniciada com a extração e recolha de dados e a sua posterior exploração/descrição. Como já foi referido anteriormente, os dados utilizados para a realização desta dissertação foram recolhidos na Unidade de Cuidados Intensivos (UCI) do Centro Hospitalar do Porto (CHP), estando compreendidos entre os anos 2010 e 2016

Os dados dos doentes internados na UCI foram filtrados e estão organizados em 3 tabelas de uma base de dados. As tabelas designam-se: DOC_ADMISSAO, UCI_INTERNADO_ANT_MV, e UCI_INTERNADOS_ADMISSAO_ALL.

A primeira tabela da base de dados (DOC_ADMISSAO) contém 7 174 linhas, com 12 atributos, a segunda tabela (UCI_INTERNADOS_ANT_MV) contém 5 270 linhas, com 8 atributos e a terceira (UCI_INTERNADOS_ADMISSAO_ALL) contém 2 730 linhas, com 53 atributos.

De forma a compreender a razão de existências dos atributos, foi feita a descrição dos atributos de cada tabela da base de dados, que está organizada nas tabelas 3, 4 e 5. Foram ainda apresentados nas tabelas 6, 7 e 8 os atributos pertencentes a cada tabela da base de dados, sendo indicado o tipo de atributo e o intervalo de valores, se existente. Importa referir que todas as tabelas contêm o atributo NPROCESSO e EPISODIO, em que o primeiro é um número único que identifica o doente e o segundo é o número associado a cada episódio de admissão. Para além destes, contêm também um atributo DATACRIACAO que indica o dia em que os doentes foram internados ou admitidos, quer na UCI, quer no hospital. Relativamente à tabela DOC_ADMISSAO e UCI_INTERNADOS_ANT_MV, existem mais atributos em comum, como é o caso de SEXO, que indica o género dos doentes e o atributo DATAN, que indica a sua data de nascimento. Quanto à tabela DOC_ADMISSAO e UCI_INTERNADOS_ADMISSAO_ALL, existe em comum o atributo SERVICO, que indica qual o serviço hospitalar onde o doente foi admitido/internado.

As tabelas da base de dados estão interligadas através dos atributos NPROCESSO e EPISODIO, que permitem aceder a todos os dados de um doente, bem como as situações em que este esteve presente no hospital e na UCI.

Tabela 4 - Descrição dos atributos da tabela DOC_ADMISSAO

Nome do atributo	Descrição
MODULO	Tipo de episódio (internamento, consulta, etc.)
ESP	Código de identificação de especialidade
DESP	Nome da especialidade
DATAALTERACAO	Última data de alteração do documento
ESTADO	Estado do documento de admissão
ORDEM	Versão do documento de admissão

Tabela 5 - Descrição dos atributos da tabela UCI_INTERNADO_ANT_MV

Nome do atributo	Descrição
NUM_CAMA	Número da cama ocupada pelo doente
SEXO	Sexo do doente
DATA_SAIDA	Data de saída do doente da UCI
N_SEQUENCIAL	Número sequencial de admissão

Tabela 6 - Descrição dos atributos da tabela UCI_INTERNADOS_ADMISSAO_ALL

Nome do atributo	Descrição
URGENCIA	Verifica se o doente esteve neste serviço hospitalar
SALAEMERGENCIA	Verifica se o doente esteve neste serviço hospitalar
OBS	Verifica se o doente esteve neste serviço hospitalar
BLOCOOPERATORIO	Verifica se o doente esteve neste serviço hospitalar
ENFERMARIA	Verifica se o doente esteve neste serviço hospitalar
OUTRAUCI	Verifica se o doente esteve neste serviço hospitalar
UNIDADEINTERMEDIA	Verifica se o doente esteve neste serviço hospitalar
DIRECTA	Verifica se o doente esteve neste serviço hospitalar
OUTROHOSPITAL	Verifica se o doente esteve neste serviço hospitalar
OUTRASITUACAO	Verifica se o doente esteve neste serviço hospitalar

Nome do atributo	Descrição
TIPOINTERNAMENTO	Verifica se o doente foi internado de urgência ou se foi planeado
TIPOINTERNAMENTOCIRURGIA	Verifica se o doente foi internado com ou sem cirurgia
GLASGOW_HOSPITAL	Classifica o doente de acordo com o *Glasgow Coma Scale* aquando entrada no hospital
GLASGOW_SERVICO	Classifica o doente de acordo com o *Glasgow Coma Scale* aquando a entrada no serviço hospitalar
SURDO	Indica se o doente é surdo
MUDO	Indica se o doente é mudo
CEGO	Indica se o doente é cedo
ALERGIAS	Indica se o doente tem alergias
VIAAEREA	Indica se o doente tem problemas respiratórios
PACEMAKER	Indica se o doente possui pacemaker
DEFICIENCIAFISICA	Indica se o doente possui alguma deficiência física
DEFICIENCIAPSIQUICA	Indica se o doente possui alguma deficiência mental
SEQUELASAVC	Indica se o doente sofreu algum AVC
ALCOOLISMO	Indica se o doente é alcoólico
TOXICODEPENDENTEIV	Indica se o doente é toxicodependente
ATRBH	Atributo hospitalar H
ATRBS	Atributo hospitalar S
INSUFHEPATICACRONICA	Indica se o doente tem alguma insuficiência hepática crónica
INSUFRENALCRONICA	Indica se o doente tem alguma insuficiência renal crónica
INSUFCARDIACA	Indica se o doente tem alguma insuficiência cardíaca
INSUFRESPIRATORIACRONICA	Indica se o doente tem alguma insuficiência respiratória crónica
DPOC	Indica se o doente sofre de Doença Pulmonar Obstrutiva Crónica (DPOC)
DOENCAHEMATOLOGICA	Indica se o doente sofre de alguma doença hematológica

Nome do atributo	Descrição
NEOPLASIANAOMETASTIZADA	Indica se o doente sofre de algum tumor não metastizado
NEOPLASIAMETASTIZADA	Indica se o doente sofre de algum tumor metastizado
CORTICOTERAPIADELONGADURACAO	Indica se o doente passou por corticoterapia de longa duração
QUIMIOTERAPIA	Indica se o doente realizou algum tratamento de quimioterapia
RADIOTERAPIA	Indica se o doente realizou algum tratamento de radioterapia
OUTROSTRATUMUNOSSUPRESSORES	Indica se o doente realizou outros tratamentos
TRANSPLANTADO	Indica se o doente sofreu algum transplante
HTAEMTRATAMENTO	Indica se o doente está em tratamento de Hipertensão Arterial (HTA)
SEQUELASAVC2	Indica se o doente sofreu algum Acidente Vascular Cerebral (AVC)
DIABETESINSULINATRATADA	Indica se o doente sofre de diabetes com valores de insulina tratada
DIABETESINSULINANAOTRATADA	Indica se o doente sofre de diabetes com valores de insulina não-tratada (diabetes tipo 2)
VENTILACAOMECANICA	Indica se o doente está sob ventilação mecânica
FARMACOSVASOATIVOS	Indica se o doente tem algum medicamento vasoativo
DATACRIACAO	Indica a data de criação do registo na UCI
INSUFCARDIAC	Indica se o doente sofre de alguma insuficiência cardíaca

Tabela 7 - Características dos atributos da tabela DOC_ADMISSAO

Nome do atributo	Tipo de atributo	Intervalo de valores
NPROCESSO	Número	-
EPISODIO	Número	-
MODULO	Texto	Internamento
DATAN	Data	-

SEXO	Texto	1; 2 Em que 1 é Homem e 2 é Mulher
SERVICO	Texto	-
ESP	Texto	-
DESP	Texto	-
DATACRIACAO	Data	[10.07.30; 16.02.14]
DATAALTERACAO	Data	[10.11.22; 16.02.18]
ESTADO	Texto	0; 1; 2; 9
ORDEM	Número	1

Tabela 8 - Características dos atributos da tabela UCI_INTERNADO_ANT_MV

Nome do atributo	Tipo de atributo	Intervalo de valores
NUM_CAMA	Texto	[001; 012]
NPROCESSO	Número	-
EPISODIO	Número	-
SEXO	Texto	1; 2
DATAN	Data	-
DATA_ENTRADA	Data	[06.06.30; 16.01.11]
DATA_SAIDA	Data	[06.06.30; 16.01.14]
N_SEQUECIAL	Número	-

Tabela 9 - Características dos atributos da tabela UCI_INTERNADOS_ADMISSAO_ALL

Nome do atributo	Tipo de atributo	Intervalo de valores
NPROCESSO	Texto	-
EPISODIO	Texto	-
SERVICO	Texto	UCI
URGENCIAS	Texto	Falso; Verdadeiro;
SALAEMERGENCIA	Texto	Falso; Verdadeiro;
OBS	Texto	Falso; Verdadeiro;
BLOCOOPERATORIO	Texto	Falso; Verdadeiro;
ENFERMARIA	Texto	Falso; Verdadeiro;
OUTRAUCI	Texto	Falso; Verdadeiro;
UNIDADEINTERMEDIA	Texto	Falso; Verdadeiro;
DIRECTA	Texto	Falso; Verdadeiro;

Nome do atributo	Tipo de atributo	Intervalo de valores
OUTROHOSPITAL	Texto	Falso; Verdadeiro;
OUTRASITUACAO	Texto	Falso; Verdadeiro;
TIPOINTERNAMENTO	Texto	Urgente; Programado
TIPOINTERNAMENTOCIRURGIA	Texto	Com cirurgia: Sem cirurgia
GLASGOW_HOSPITAL	Texto	-
GLASGOW_SERVICO	Texto	-
SURDO	Texto	Falso; Verdadeiro;
MUDO	Texto	Falso; Verdadeiro;
CEGO	Texto	Falso; Verdadeiro;
ALERGIAS	Texto	Falso; Verdadeiro;
VIAAEREA	Texto	Falso; Verdadeiro;
PACEMAKER	Texto	Falso; Verdadeiro;
DEFICIENCIAFISICA	Texto	Falso; Verdadeiro;
DEFICIENCIAPSIQUICA	Texto	Falso; Verdadeiro;
SEQUELASAVC	Texto	-
ALCOOLISMO	Texto	Falso; Verdadeiro;
TOXICODEPENDENTEIV	Texto	Falso; Verdadeiro;
ATRBH	Texto	Falso; Verdadeiro;
ATRBS	Texto	Falso; Verdadeiro;
INSUFHEPATICACRONICA	Texto	Falso; Verdadeiro;
INSUFRENALCRONICA	Texto	Falso; Verdadeiro;
INSUFCARDIACA	Texto	Falso; Verdadeiro;
INSUFRESPIRATORIACRONICA	Texto	Falso; Verdadeiro;
DPOC	Texto	Falso; Verdadeiro;
DOENCAHEMATOLOGICA	Texto	-
NEOPLASIANAOMETASTIZADA	Texto	-
NEOPLASIAMETASTIZADA	Texto	Não; Sim
CORTICOTERAPIADELONGADURACAO	Texto	Falso; Verdadeiro;
QUIMIOTERAPIA	Texto	Falso; Verdadeiro;
RADIOTERAPIA	Texto	Falso; Verdadeiro;
OUTROSTRATUMUNOSSUPRESSORES	Texto	Falso; Verdadeiro;

Nome do atributo	Tipo de atributo	Intervalo de valores
TRANSPLANTADO	Texto	Falso; Verdadeiro;
HTAEMTRATAMENTO	Texto	Falso; Verdadeiro;
SEQUELASAVC2	Texto	Falso; Verdadeiro;
DIABETESINSULINATRATADA	Texto	Falso; Verdadeiro;
DIABETESINSULINANAOTRATADA	Texto	Falso; Verdadeiro;
VENTILACAOMECANICA	Texto	-
FARMACOSVASOATIVOS	Texto	-
DATACRIACAO	Data	[10.10.19; 16.01.25]
SALAEMERGENCIA_SCORE	Número	-
RISCO	Texto	-
INSUFCARDIAC	Texto	-

4.2. Explorar os dados

Os dados extraídos da UCI do CHP são na maioria referentes a Homens. No caso da tabela DOC_ADMISSAO, os dados são de 4 370 homens (cerca de 60,9%) e de 2 804 mulheres (cerca de 39,1%), como é possível verificar na ilustração 1.

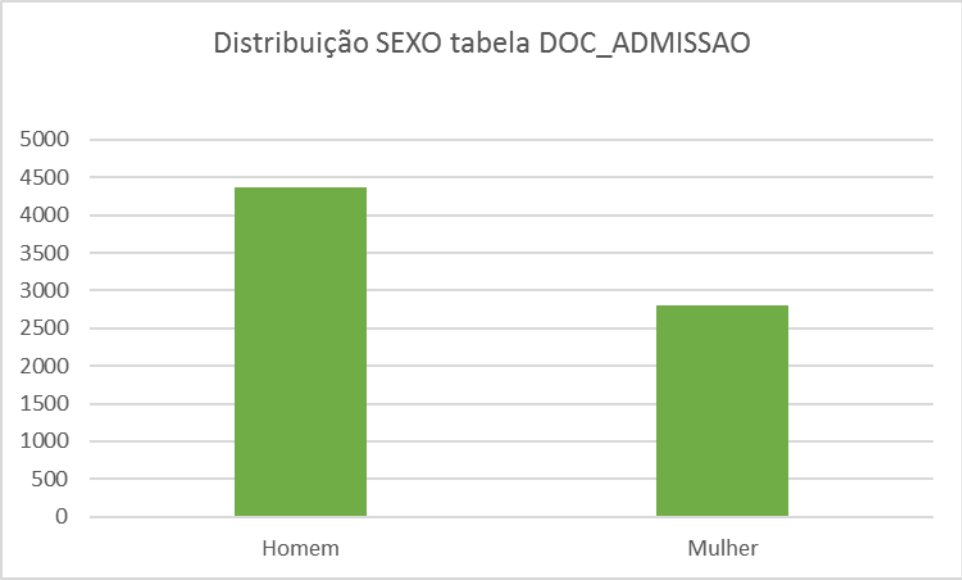

Ilustração 1 - Distribuição dos valores do atributo SEXO da tabela DOC_ADMISSAO

No caso da tabela UCI_INTERNADO_ANT_MV, os dados são de 3 257 homens e de 2013 mulheres, como é possível verificar na ilustração 2.

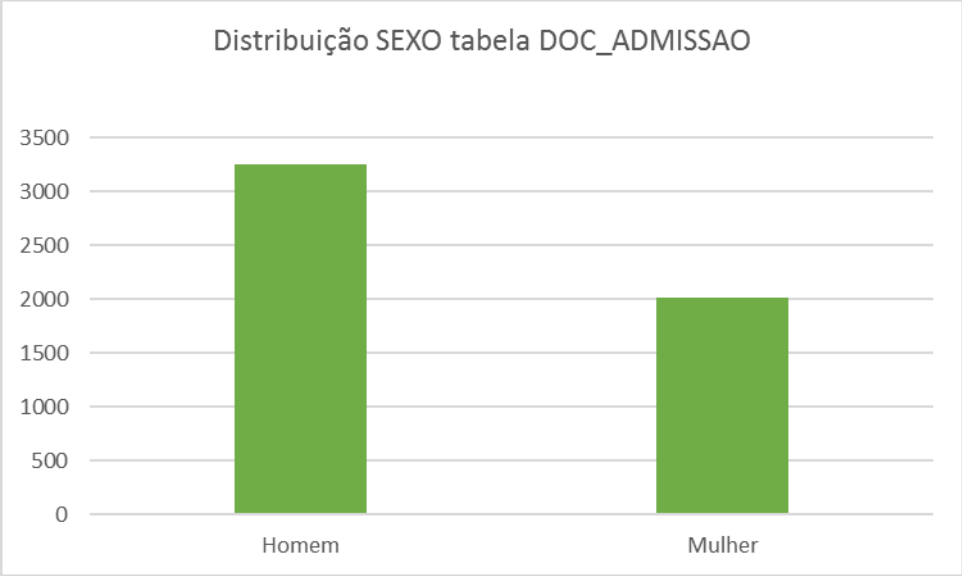

Ilustração 2 - Distribuição dos valores do atributo SEXO da tabela UCI_INTERNADO_ANT_MV

Após apurar que o sexo masculino é o mais frequente nos dados, foram verificadas quais as especialidades que os doentes estiveram antes de serem internados nas UCI com mais registos de mulheres do que de homens no documento de admissão:

- Cirurgia2, sendo que as Mulheres constituem 58,6% dos registos associados ao atributo;
- Cirurgia plástica, sendo que as Mulheres constituem 100% dos registos associados ao atributo;
- Medicina B, sendo que as Mulheres constituem 53% dos registos associados ao atributo;
- Neurocirurgia, sendo que as mulheres constituem 54,7% dos registos associados ao atributo;
- Obstetrícia, sendo que as Mulheres constituem 100% dos registos associados ao atributo;
- Pneumologia, sendo que as Mulheres constituem 60% dos registos associados ao atributo;
- Maternidade Júlio Dinis – CMIN, sendo que as Mulheres constituem 100% dos registos associados ao atributo.

É possível, ainda, através da análise aos dados, aferir que SCI é o serviço com mais ocorrências por parte dos doentes, seguido do "Unidade Interna de Medicina C" (UIMC) e do "Unidade de Transplante Hepático e Pancreático" (UTHP). Por sua vez, os serviços "Outras UCI", "Dermatologia", "IADC", "Maternidade Júlio Dinis – CMIN" (MJDNP) são os que têm menos ocorrências. Existem, ainda, 1 946 doentes admitidos num serviço hospitalar e que têm alta no mesmo dia.

Foi também realizada uma análise estatística das idades antes e depois da passagem pela UCI. Como é possível verificar na tabela 9, as métricas utilizadas (Idade máxima, mínima, média, moda e mediana) não possuem grande discrepância de valores. No caso da moda e da idade mínima, estas não diferem (75 anos e 13 anos, respetivamente).

Tabela 10 - Análise dos valores da idade no documento de admissão

Medida	Documento de admissão	Admissão/Internamento em UCI
Máxima	96	101
Mínima	13	13
Média	62,19	59,07
Moda	75	75
Mediana	64	60

Dos doentes internados na UCI, 667 ficaram internados por apenas algumas horas, não completando 24 horas de internamento, sendo 0 o número de dias mínimo de internamento presente nos dados e constituindo cerca de 12,7% dos registos. Em relação aos doentes que ficaram mais dias internados, o número máximo de dias de internamento é de 894, provavelmente de algum doente em coma. 1 281 doentes ficaram internados por 1 dia, sendo este o número de dias de internamento mais frequente, constituindo cerca de 24,3% dos registos. A média de dias de internamento são cerca de 5 dias.

Em relação ao trajeto pelo hospital dos doentes que foram internados na UCI, o local por onde estes doentes mais transitaram foi o bloco operatório, com 1 411 registos (cerca de 74,7%), seguido pela enfermaria, com 265 registos (cerca de 14%). Pela sala de observação, apenas passaram 5 doentes (0,3%). Além disso, a maioria dos doentes foram internados de urgência, existindo apenas 761 doentes que tinham internamento programado (40,3%).

Dos doentes presentes no documento de admissão, cerca de 6 211 doentes (86,6%) foram internados posteriormente na UCI.

4.3. Verificar qualidade

Para concluir a compreensão dos dados, é imprescindível explorá-los com o objetivo de verificar a sua qualidade, procurando erros ou omissões de informação (campos em branco). É verificada a integridade e exatidão dos dados, apurando, também, a coerência dos atributos e do seu relacionamento. Por fim, é apresentado um conjunto de possíveis soluções para correção dos

erros. Esta fase antecede a preparação dos dados, onde são selecionadas as soluções mais adequadas. Importa também referir que a tabela com os atributos, erros encontrados e hipóteses de correção encontram-se no anexo IV, sendo de seguida apresentado de forma sucinta quais os principais erros encontrados em cada tabela, bem como de que forma foram corrigidos.

- DOC_ADMISSAO

Após uma análise da primeira tabela na procura de incoerências, foram encontrados erros de dois tipos: valores todos iguais, o que torna os atributos dispensáveis do processo de DM e, portanto, o atributo pode ser eliminado e valores em branco que, quando não é possível preenchê-los, devem ser eliminadas as respetivas linhas. No entanto, 7 dos 12 atributos não apresentavam qualquer tipo de erros, tendo sido, portanto, inalterados.

- UCI_INTERNADO_ANT_MV

Nesta tabela não foi encontrado nenhum erro, nem relativamente a valores em brancos, nem valores repetidos. Sendo assim, a referida tabela não sofreu nenhum tipo de alterações.

- UCI_INTERNADOS_ALL

Por fim, relativamente aos erros encontrados na última tabela, foram encontrados 4 atributos em que os valores estavam todos em branco, ou seja, o atributo passa a ser dispensável do processo de DM, pois não apresentava nenhum contributo, como também é o caso do atributo SERVICO, no qual os valores são todos iguais (INT). O atributo NPROCESSO, EPISODIO e DATACRIACAO são os únicos que não apresentam qualquer tipo de erro. Existe ainda um erro na inserção dos dados para o atributo GLASGOW_HOSPITAL e GLASGOW_SERVICO. Este atributo serve para classificar o nível de consciência de um doente baseando-se num teste ocular, verbal e motor, sendo que a pontuação total se encontra no intervalo dos valores numéricos de 3 a 15. No entanto, em algumas linhas encontram-se valores como “sedada”, “sedado”, ou “sedada com Propofol”. Deste modo, estas linhas também foram eliminadas pois não contribuíam positivamente para o processo de DM. Por fim, o erro mais encontrado nos atributos foi o aparecimento de algumas linhas em branco, que, mais uma vez, foram eliminadas do *dataset* submetido ao processo de DM.

4.4. Preparar os dados

Para a tabela DOC_ADMISSAO foi feito o cálculo da idade à data de criação do documento de admissão, a partir da data de nascimento. Alem disso, foi calculado o número de dias que os

doentes estiveram admitidos em unidades hospitalares. Nota para os valores que são "0" significam que o doente esteve apenas admitido por algumas horas no mesmo dia.

Para a tabela UCI_INTERNADO_ANT_MV foi, de igual modo, calculada a idade e o número de dias que os doentes estiveram internados na UCI. Tal como noutra tabela, os valores que são "0" significam que o doente esteve apenas internado por algumas horas no mesmo dia.

Foi ainda criada o atributo RISCO com os atributos [0; 1] que identifica doentes que têm características que permitem considerá-los doentes de risco, devido à gravidade das doenças que apresentam. Deste modo, se o doente não for de risco, é atribuído o valor 0, caso contrário, é atribuído o valor 1. O conjunto de doenças consideradas de risco foi definido pelo INTCare do CHP.

Os atributos considerados de risco são atribuídos com o valor 1 caso se verifique pelo menos um dos seguintes atributos:

- INSUFHEPATICACRONICA
- INSUFRENALCRONICA
- INSUFCARDIACA
- INSUFRESPIRATORIACRONICA
- TRANSPLANTADO
- HTAEMTRATAMENTO
- SEQUELASAVC2

Com base na análise da qualidade dos dados, é possível verificar que nem todos são relevantes para o estudo e, portanto, serão excluídos para o processo de DM. Esses atributos estão apresentados na tabela 11, bem como a respetiva fundamentação.

Tabela 11 - Atributos excluídos no processo de DM

Atributo	Solução
MODULO ORDEM SERVICO	Todos os valores são iguais, não acrescentando informação relevante para o estudo
NPROCESSO EPISODIO DATAN DATACRIACAO DATAALTERACAO	O atributo não apresenta nenhuma informação relevante para o estudo

Atributo	Solução
NUM_CAMA N_SEQUENCIAL ESTADO ORDEM	
SEQUELASAVC VENTILAÇÃOMECANICA FARMACOSVASOATIVOS INSUFCARDIAC	Os valores encontram-se em branco, não acrescentando informações relevantes para o estudo
NEOPLASIANAOMETASTIZADA NEOPLASTIAMETASTIZADA	Devido à quantidade de nulos, os dois atributos são fundidos para dar origem ao atributo NEOPLASTIA
URGENCIA SALAEMERGENCIA OBS BLOCOOPERATORIO ENFERMARIA OUTRAUCI UNIDADEINTERMEDIA DIRECTA OUTROHOSPITAL OUTRASITUACAO	Foi criada o atributo PROVENIENCIA com o nome dos atributos eliminados onde estes eram “Verdadeiro”

Todos os outros atributos da base de dados foram empregues no estudo visto que contribuem para o alcance do objetivo da dissertação.

Em relação aos nulos, foram eliminados todos os registos da tabela UCI_INTERNADOS_ADMISSAO_ALL que apresentaram os atributos dos doentes como NULL. A tabela foi filtrada de forma a suprimir todos os valores em branco.

Para além disso, os registos do atributo GLASGOW_HOSPITAL e GLASGOW_SERVICO foram modificados de forma a indicarem apenas o valor atribuído que que se encontra nos valores de 3 a 15.

Para inserção dos dados no RapidMiner e no Orange, tiveram de ser convertidos para dados numéricos. Todos os atributos que tivessem valores *True* ou *False* foram convertidos para 1 ou 0, respetivamente.

Foi criado ainda um atributo denominado DIASADMISSAO que apresenta qual o número total de dias que um doente esteve internado/admitido, desde a sua data de entrada até à data de saída. Na fase de modelação, mais propriamente, na aplicação de técnicas de classificação, este atributo deu origem a duas abordagens relativamente à criação de classes de dias e que estão descritas numa fase posterior desta dissertação.

Por fim, no atributo PROVENIENCIA os valores foram convertidos de acordo com a tabela 12.

Tabela 12 - Valores convertidos no atributo PROVENIENCIA

Valor atributo	Valor convertido
URGENCIA	1
SALAEMERGENCIA	2
OBS	3
BLOCOOPERATORIO	4
ENFERMARIA	5
OUTRAUCI	6
UNIDADEINTERMEDIA	7
DIRECTA	8
OUTROHOSPITAL	9
OUTRASITUACAO	10

4.5. Modelação dos dados

Após aplicadas todas as alterações na fase de preparação dos dados, foram, então, utilizadas as ferramentas selecionadas para aplicação das técnicas de DM. A modelação dos dados foi realizada em duas partes: primeiro foi feito o *Clustering* e na segunda parte foi feita a Previsão utilizando a Classificação.

As técnicas de *Clustering* e Classificação foram aplicadas à tabela com os dados dos doentes internados na UCI (UCI_INTERNADOS_ADMISSAO_ALL). Isto deve-se ao facto de, na tabela de documento de admissão os testes efetuados permitiram verificar que não acrescentava valor ao projeto, nem os primeiros resultados foram suficientes para avançar com a modelação.

Relativamente à tabela UCI_INTERNADO_ANT_MV, foi rapidamente excluída dado que não continha dados que contribuíssem para o processo de DM.

A modelação dos dados iniciou-se a construção de cenários para *Clustering*, em que foi necessário agrupar os atributos para apurar a existência de padrões que possibilitassem criar grupos (*Clusters*) de doentes e categoriza-los.

Para realização da modelação, foi necessária a construção de diferentes cenários através dos atributos da tabela UCI_INTERNADOS_ADMISSAO_ALL. Estes cenários foram utilizados tanto na parte de *Clustering* como na Classificação, tendo sido construídos mais cinco cenários na segunda parte. Assim, o primeiro cenário criado foi o *Case Mix*, que abarca todas as variáveis do *dataset*. Os cenários seguintes foram concebidos a partir de testes. Isto é, primeiro foram agrupados atributos relacionados com scores médicos atribuídos aos doentes. Os cenários subsequentes foram constituídos para verificar que preponderância tem a inclusão ou exclusão das variáveis. Foram também criados cenários com atributos dos diversos tipos de insuficiências e de internamento e cirurgia. No final obtiveram-se 10 cenários com os diferentes grupos de variáveis que posteriormente foram submetidos aos algoritmos de *Clustering* e Classificação.

As variáveis incluídas para construção dos cenários foram as seguintes:

- VZ: TIPOINTERNAMENTO
- VAA: TIPOINTERNAMENTOCIRURGIA
- VA: GLASGOW_HOSPITAL
- VB: GLASGOW_SERVICO
- VC: SURDO
- VD: MUDO
- VE: CEGO
- VF: ALERGIAS
- VG: VIAAEREA
- VH: PACEMAKER
- VI: DEFICIENCIAFISICA
- VJ: DEFICIENCIAPSIQUICA
- VK: ALCOOLISMO
- VL: TOXICODEPENDENTEIV
- VM: ATRBH

- VN: ATRBS
- VO: INSUFHEPATICACRONICA
- VP: INSUFRENALCRONICA
- VQ: INSUFCARDIACA
- VR: INSUFRESPIRATORIACRONICA
- VS: DPOC
- VT: DOENCAHEMATOLOGICA
- VU: CORTICOTERAPIADELONGADURACAO
- VV: HTAEMTRATAMENTO
- VAE: QUIMIOTERAPIA
- VAF: RADIOTERAPIA
- VAG: OUTROSTRATUMUNOSSUPRESSORES
- VAB: TRANSPLANTADO
- VW: SEQUELASAVC2
- VX: DIABETESINSULINATRATADA
- VX: DIABETESINSULINANAOTRATADA
- VAC: RISCO
- VAD: NEOPLASIA
- VY: PROVENIENCIA

Estes atributos foram então organizados em diferentes cenários listados de seguida:

- S1 = {Todos os atributos};
- S2 = {VZ; VAA; VA; VB; VAC; VY};
- S3 = {VA; VY; VZ; VW; VAB};
- S4 = {VAA; VZ; VAB; VAC};
- S5 = {VO; VP; VQ; VR; VY};
- S6 = {VAC; VA; VY};
- S7 = {VO; VP; VQ; VR; VAA; VAB; VZ; VAC; VY};
- S8 = {VZ; VAA; VAB};
- S9 = {VAA; VAB; VZ; VM; VN; VT};
- S10 = {VAD; VAE; VAF; VAG; VZ}.

Importa salientar, que foi definido que, caso os resultados provenientes dos cenários construídos inicialmente para o *Clustering* não fossem satisfatórios, seriam realizados mais testes a outros cenários para confirmar esses resultados, tendo sido por isso criados mais cinco cenários, dando um total de 15:

- S11 = {VX, VT, VV, VZ};
- S12 = {VQ, VH, VW, VZ};
- S13 = {VZ, VG, VF, VR, VS};
- S14 = {VC, VD, VE, VZ};
- S15 = {VZ, VAA}.

4.5.1. Aplicação de técnicas de Clustering

Foi então iniciada a construção dos 10 cenários na ferramenta Orange, aos quais foi aplicado o algoritmo K-Means. Este simples processo está ilustrado na figura 7, onde é utilizada a tabela UCI_INTERNADOS_ADMISSAO_ALL com os dados tratados, nos quais são selecionados os atributos correspondente a cada cenário (neste exemplo, o cenário 4), tendo sido aplicado ao algoritmo K-Means. Os resultados podem ser visualizados através de gráficos de distribuição, bem como através de uma tabela que associa cada registo aos clusters formados.

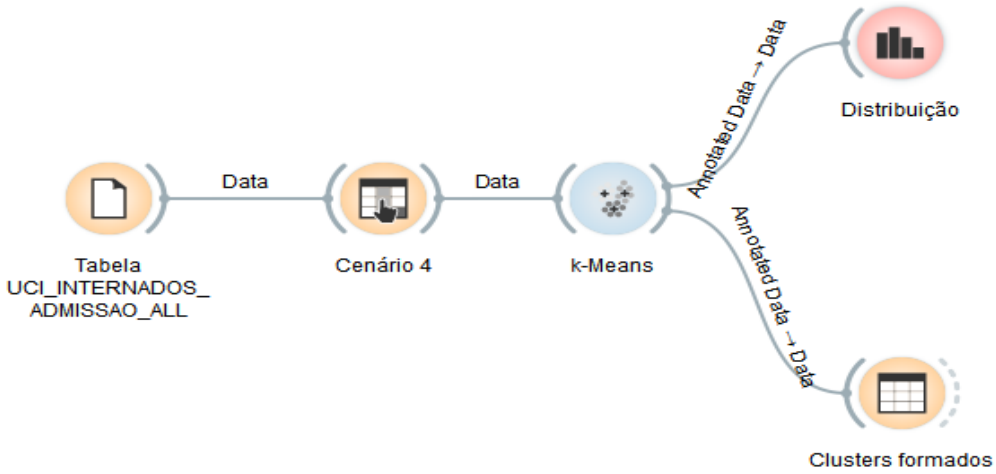

Figura 7 - Aplicação de Clustering no Orange

Como já foi suprarreferido, foi executado o algoritmo K-Means, tendo sido possível realizar testes a partir de três tipos de técnicas de *Scoring*:

- *Silhouete*: valores encontram-se no intervalo [-1; 1], sendo 1 o melhor valor;

- *Inter-cluster distance*: quanto menor o valor, melhor;
- *Distance to centroids*: quanto menor o valor, melhor.

Assim, no Orange, foram gerados 30 modelos:

Técnicas = {Silhouete, Inter-Cluster Distance, Distance to Centroids}

DMM = {*10 Scenarios, 3 Techniques, 1 Sampling Method, 1 Representation Method, 1 Target*}

Na tabela 13 é possível verificar os três melhores resultados dos testes realizados aos cenários. Para verificação dos resultados para todos os cenários, pode ser consultada uma tabela no anexo V. A tabela apresenta os atributos de cada cenário, bem como o pior e o melhor resultado para cada teste. A titulo de exemplo, o cenário 8, constituído por 3 atributos, teve como pior resultado do *Silhouete* "0,64" e como melhor resultado "1". Como já foi referido, neste tipo de teste o resultado ótimo é 1, o que significa que o melhor modelo tem os resultados mais próximo deste valor.

Tabela 13 - Resultados dos três melhores cenários com a aplicação do algoritmo k-means na ferramenta Orange

Cenário	Atributos	Teste	Pior	Melhor
Cenário 4	TIPOINTERNAMENTOCIRURGIA TRANSPLANTADO TIPOINTERNAMENTO RISCO	Silhouete	0,42	0,81
		Inter-cluster distance	1,8	1,4
		Distance to centroids	0,93	0,07
Cenário 8	TIPOINTERNAMENTOCIRURGIA TRANSPLANTADO TIPOINTERNAMENTO	Silhouete	0,64	1
		Inter-cluster distance	1,6	1,5
		Distance to centroids	0,43	6,2e-17
Cenário 10	TIPOINTERNAMENTO NEOPLASIA QUIMIOTERAPIA RADIOTERAPIA OUTROSTRATIMUNOSSUPRESSORES	Silhouete	0,87	0,93
		Inter-cluster distance	1,7	1,4
		Distance to centroids	0,12	0,04

Como é possível observar, o melhor cenário é o cenário (8), sendo aquele que apresenta melhores valores. Comparando os cenários entre si, é facilmente analisado que o cenário (8) difere do cenário (4) apenas por excluir o atributo RISCO. Através desta exclusão foi verificado que o atributo RISCO influência negativamente os resultados, sendo, por isso, desnecessária a sua inclusão no cenário (8). Ao contrário do atributo RISCO, o atributo TIPOINTERNAMENTO possui um impacto positivo nos resultados, tendo sido incluída em diferentes cenários. Isto deve-se ao

facto deste atributo permitir uma melhor separação dos valores e uma clara distribuição entre os clusters formados. Ainda assim, os melhores cenários são o (4), (8) e, ainda, o cenário (10), que inclui os atributos de doentes com neoplasia e respetivos tratamentos. Este cenário, apresentou 0,93 para o teste *Silhouete,* 1,4 para *Inter-cluster Distance* e 0,04 para *Distance to Centroids*.

O mesmo processo foi realizado no RapidMiner, como mostra a figura 8, onde foi selecionada a tabela UCI_INTERNADOS_ADMISSAO_ALL, bem como os atributos para o cenário pretendido, seguido da aplicação de *Clustering*, onde foram realizados testes para diferentes números de clusters (2, 5 e 10), tendo-se demonstrado que o número de clusters mais apropriado era 2. De seguida, existe ainda um componente "Performance" que permite selecionar o teste a realizar (*Davies-Bouldin Index*).

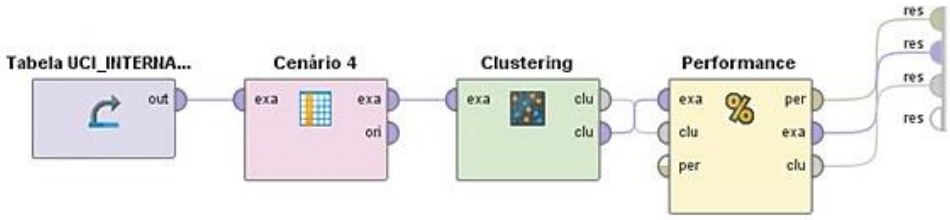

Figura 8 - Aplicação de *Clustering* no RapidMiner

Assim, no RapidMiner foram gerados 30 modelos (nota para que o cálculo dos modelos neste caso terá de ser multiplicado por três devido ao facto de terem sido realizados os testes três vezes com diferentes números de clusters (2, 5 e 10):

Técnicas = {*Davies-Bouldin Index*}

DMM = {*10 Scenarios, 1 Technique, 1 Sampling Method, 1 Representation Method, 1 Target* }

Foi aplicado o *Davies-Bouldin Index* nos cenários criados previamente. Na tabela 14 estão expostos os resultados obtidos para cada cenário, tendo sido então efetuados testes para 10, 5 e 2 *clusters*.

Nota para o Davies-Bouldin Index no RapidMiner que apresenta maioritariamente valores negativos. A justificação para esta situação passa pelo facto de quando a densidade dos dados é baixa, os valores são automaticamente assinalados como negativos. Assim sendo, quanto mais negativo for o valor, melhor é o resultado.

Tabela 14 - Resultados de cada cenário com a aplicação do algoritmo *Davies-Bouldin Index* na ferramenta RapidMiner

Cenário	Teste	10 clusters	5 clusters	2 clusters
S1	Avg within centroid distance	-3.56	-5.49	-12.41
	Davies Bouldin	-1.24	-0.87	-0.74
S2	Avg within centroid distance	-2.93	-4.36	-11.27
	Davies Bouldin	∞	-0.8	-0.7
S3	Avg within centroid distance	-0.880	-1.63	-5.19
	Davies Bouldin	∞	-0.63	-0.56
S4	Avg within centroid distance	-0.03	-0.19	-0.5
	Davies Bouldin	∞	-0.72	-1.35
S5	Avg within centroid distance	-0.22	-0.5	-1.47
	Davies Bouldin	∞	: -0.77	-0.42
S6	Avg within centroid distance	-1.038	-1.97	-5.15
	Davies Bouldin	∞	-0.96	-0.56
S7	Avg within centroid distance	-0.592	-0.91	: -2.20
	Davies Bouldin	∞	-1.06	-0.56
S8	Avg within centroid distance	-0.023	-0.03	-0.25
	Davies Bouldin	∞	∞	-0.65
S9	Avg within centroid distance	-0.072	-0.08	-0.31
	Davies Bouldin	∞	-0.75	-0.73
S10	Avg within centroid distance	-0.058	-0.06	-0.12
	Davies Bouldin	∞	∞	-0.37

Como é possível verificar após a análise da tabela anterior, o melhor cenário do RapidMiner difere do melhor cenário do Orange. No entanto, neste caso, o cenário (4) apresenta melhores resultados aquando a formação de 2 *clusters*.

Para os três melhores cenários (4, 8 e 10) foram analisados os valores possíveis de cada atribuo que estão presentes nos dois clusters originados. Na tabela 15 é possível observar esses valores, sendo possível observar que a variável TIPOADMISSAO tem os seus dois valores claramente divididos pelo cluster 1 e o cluster 2 nestes três cenários. Mais uma vez se observa que este atributo tem uma influência bastante positiva na construção de *clusters*.

Tabela 15 - Distribuição dos valores por cada cluster dos três melhores cenários

Cenário	Atributo	Cluster 1	Cluster 2
Cenário 4	TIPOINTERNAMENTOCIRURGIA	0 e 1	0 e 1
	TRANSPLANTADO	0 e 1	0 e 1
	TIPOINTERNAMENTO	**1**	**0**
	RISCO	0 e 1	0 e 1
Cenário 8	TIPOINTERNAMENTO	**0**	**1**
	TIPOINTERNAMENTOCIRURGIA	0 e 1	0 e 1
	TRANSPLANTADO	0 e 1	0 e 1
Cenário 10	TIPOINTERNAMENTO	**0**	**1**
	NEOPLASIA	0 e 1	0 e 1
	QUIMIOTERAPIA	0 e 1	0 e 1
	RADIOTERAPIA	0 e 1	0 e 1
	OUTROSTRATIMUNOSSUPRESSORES	0 e 1	0 e 1

Na ilustração 3 é facilmente observável a separação dos valores entre os dois *Clusters*, para o atributo TIPOADMISSAO. Este atributo é constituído pelo valor "0" e o valor "1" que se dividiram totalmente pelos dois clusters, sendo que o valor "0" está presente apenas no *Cluster* 2 (604 vezes) e o valor "1" está presente apenas no *Cluster* 1 (1 255 vezes).

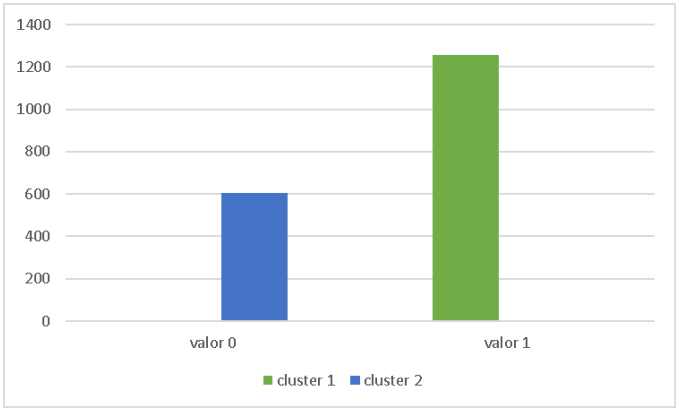

Ilustração 3 - Distribuição dos valores entre os clusters no atributo TIPOADMISSAO

Para observar estes valores, está presente na tabela 16 a distribuição dos valores pelos dois clusters, para os atributos dos três melhores cenários (4, 8 e 10). Aqui se verifica que o atributo TIPOINTERNAMENTO permite uma total distribuição dos valores pelos dois *Clusters*. Para os cenários apresentados, este atributo distribui o valor "0" para o *Cluster* 2 e o valor "1" para o

Cluster 1. Isto significa que este atributo permite dividir claramente os doentes em dois grupos de características comuns, sendo então um dos melhores atributos para criação de *Clusters*.

Tabela 16 - Distribuição dos valores de cada variável dos 3 melhores cenários pelos dois clusters

Cenário	Atributo	Valor	Cluster 1	Cluster 2
Cenário 4	TIPOINTERNAMENTOCIRURGIA	0	42,87%	4,64%
		1	57,13%	95,36%
	TRANSPLANTADO	0	90,84%	95,20%
		1	9,16%	4,80%
	TIPOINTERNAMENTO	0	0,00%	100,00%
		1	100,00%	0,00%
	RISCO	0	52,75%	50,99%
		1	47,25%	49,01%
Cenário 8	TIPOINTERNAMENTO	0	0,00%	100,00%
		1	100,00%	0,00%
	TIPOINTERNAMENTOCIRURGIA	0	42,87%	4,64%
		1	57,13%	95,36%
	TRANSPLANTADO	0	90,84%	95,20%
		1	9,16%	4,80%
Cenário 10	TIPOINTERNAMENTO	0	0,00%	100,00%
		1	100,00%	0,00%
	NEOPLASIA	0	97,13%	92,22%
		1	2,87%	7,78%
	QUIMIOTERAPIA	0	97,77%	94,70%
		1	2,23%	5,30%
	RADIOTERAPIA	0	99,04%	97,68%
		1	0,96%	2,32%
	OUTROSTRATIMUNOSSUPRESSORES	0	96,10%	96,85%
		1	3,90%	3,15%

4.5.2. Aplicação de técnicas de Classificação

Seguidamente ao de *Clustering*, procedeu-se à Classificação de modo a verificar a possibilidade de realizar previsões de admissões/internamentos nas UCI. Os cenários criados inicialmente foram então introduzidos no Oracle SQL Developer. A figura 7 constitui um exemplo da construção dos modelos (neste caso para o cenário 1). É escolhida a fonte de dados, onde são selecionados os atributos pretendidos, previamente tratados e preparados para este processo. Esta fonte de dados é então ligada ao componente de classificação. Este, por sua vez, permite selecionar quais os algoritmos que serão empregues na modelação, bem como qual o *target* pretendido.

Figura 9 - Construção do modelo para o cenário 1

Para esta fase, foi realizado o processo de classificação dos cenários duas vezes, tendo sido utilizadas duas abordagens ao problema, utilizando como target os dias de admissão/internamento. Por outras palavras, na primeira abordagem foi selecionado um *target* baseado em 2 classes:

- Classe 1: doentes admitidos em menos de 24 horas;
- Classe 2: doentes internados por mais de 24 horas.

Para a segunda abordagem, foi selecionado um *target* baseado em 4 classes;

- Classe 1: doentes admitidos em menos de 24 horas;
- Classe 2: doentes internados num período superior a 24 horas até 3 dias;
- Classe 3: doentes internados num período superior a 3 dias até 5 dias;Classe 4: doentes internados num período superior a 5 dias.

Importa referir que foi feita uma distinção entre os conceitos de **doentes admitidos** e **doentes internados**, em que o primeiro corresponde aos doentes em que o período de hospitalização não foi superior a 24 horas e os segundos, foram internados num período superior a 24 horas.

Os algoritmos selecionados para a classificação foram SVM (*Support Vector Machine*), GLM (*Generalized Linear Model*), NB (*Naive Bayes*) e DT (*Decision Trees*). No entanto, para a

classificação com o target constituído por 4 classes, não foi possível utilizado GLM, pois não é aplicável a previsões com mais de 2 classes, sendo, portanto, utilizada na segunda classificação.

Após a construção de cada cenário e a aplicação dos algoritmos, foram obtidos diversos resultados, nomeadamente a matriz confusão, curva ROC, LIFT e Profit. A matriz confusão é constituída por:

- Verdadeiros Positivos (Verdadeiro Positivo -VP),
- Verdadeiros Negativos (Verdadeiro Negativo - VN),
- Falsos Positivos (Falso Positivo - FP),
- Falsos Negativos (Falso Negativo - FN).

A título de exemplo, é, de seguida apresentada na tabela 17, a matriz confusão para o S1 para a primeira abordagem, extraída a partir da ferramenta utilizada. De modo sucinto, o Verdadeiro Positivo e o Verdadeiro Negativo indicam qual o número de valores que foram previstos corretamente e, neste caso, 151 valores "1" e 19 dos valores "2" foram previstos corretamente. Relativamente aos Falsos Positivos, estes também indicam o número de vezes que um valor foi previsto incorretamente, quando o resultado é previsto incorretamente como positivo, mas o seu valor real é negativo, sendo que neste caso são 557 valores incorretamente previstos. Por fim, os Falsos Negativos correspondem a valores previstos incorretamente, em que o valor previsto é negativo e o valor real é positivo, tendo sido neste caso 9 o número de valores incorretamente previstos. No caso da Medicina, os médicos obviamente preferirem modelos altamente capazes de acertar corretamente nas previsões, mas os Falsos Positivos são também adequados dado que os médicos preferem que os modelos prevejam valores positivos e na realidade sejam negativos, do que ao contrario, dado que pode pôr em risco a vida dos seus doentes. Para verificar todas as matrizes confusão obtidas para a primeira abordagem, bem como o cálculo das métricas, pode ser consultado o anexo VI.

Tabela 17 - Matriz confusão para o cenário 1 na primeira abordagem

	1	2
1	151	9
2	557	19

Para a primeira abordagem foi realizado o cálculo de 4 métricas, nomeadamente Acuidade, Sensibilidade, Especificidade e Taxa de Erro;

$$\text{Acuidade} = \frac{VP+VN}{VP+VN+FP+FN} \times 100$$

$$\text{Sensibilidade: } \frac{VP}{VP+FN} \times 100\ (\%)$$

$$\text{Especificidade: } \frac{VN}{VN+FP} \times 100\ (\%)$$

$$\text{Taxa de erro total: } 100 - Acuidade\ (\%)$$

Assim sendo, utilizando o exemplo do cenário 1, são obtidos os seguintes valores com as métricas anteriores:

- Acuidade = $\frac{151+19}{151+19+557+9} * 100$ = 23,1%
- Sensibilidade = $\frac{151}{151+9} * 100 = 94{,}4\%$
- Especificidade = $\frac{19}{19+557} * 100 = 3{,}3\%$
- Taxa de erro total = $100 - 23{,}1 = 76{,}9\%$

Para a segunda abordagem apenas foi possível o cálculo da acuidade, dado que o target inclui quatro classes. Mais uma vez, a título de exemplo, é apresentada na tabela 18, a matriz confusão para o S1 para a segunda abordagem, sendo possível consultado no anexo VII todas as matrizes confusão desta abordagem.

Tabela 18 - Matriz confusão do cenário 1 na segunda abordagem

	1	2	3	4
1	0	49	0	111
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Além da construção de novos cenários, procedeu-se ao *Sampling* dos dados, de modo a equilibrar os valores do target. Este passo foi importante dado que o desequilíbrio ou a dominância de um valor sob outro, provoca que a minoria dos dados apresente um maior número de erros. O *Sampling* dos dados procedeu-se com a replicação da minoria dos dados, as vezes que fossem necessárias até chegar a um equilíbrio com a maioria. Por exemplo, na abordagem dois, dado que o target tinha apenas 20,8% de valores "1", esses dados foram replicados, de forma integral, quatro vezes, de modo a balancear a sua distribuição.

Assim sendo, iniciou-se a modelação da primeira abordagem, através da qual foram criados 120 modelos:

- DMM = {15 cenários, 4 técnicas de classificação, 2 métodos de *sampling*, 1 *Target*}

Como exemplo, pode ser utilizado o cenário 15 para demonstrar como é calculado o número de modelos criados:

- S15 = {VZ, VAA};
- Técnicas de Classificação = {NB, DT, GLM, SV};
- Métodos de *Sampling* = {2};
- Target = {Internamento}.

Com estas informações foi possível obter 8 modelos:

- DMM = {1 cenário * 4 técnicas de classificação * 2 métodos de *sampling* * 1 *target*}

Dado a grande quantidade de modelos criados, são aqui apresentados os três melhores resultados, cujos foram obtidos através do cenário (3), (6) e (11). A escolha dos melhores resultados foi feita de acordo com a preferência que os médicos têm por modelos com maior sensibilidade de prever uma possível admissão na UCI. Os modelos que apresentar a maior sensibilidade são considerados melhores, pois oferecem a oportunidade de agir proactivamente, de modo a evitar admissões, o que pode ser crucial na sobrevivência de um doente, este tipo de ação pode levar à criação de falsos positivos. Mais concretamente, os médicos preferem modelos com Sensibilidade superior a 85 e Especificidade superior a 60 o que, neste caso, apenas acontece nos cenários (3) e (6).

De um modo geral, fazendo uma análise dos resultados para cada um dos algoritmos aplicados, o algoritmo NB e DT apresentaram os melhores resultados (os resultados dos melhores cenários foram iguais nestes dois algoritmos). Na tabela 19 estão expostos os resultados dos três melhores cenários (3, 6 e 11), para os algoritmos NB e DT (dado que são iguais).

Tabela 19 - Resultados da primeira abordagem

Cenários	Medida	Resultados
3 *e* 6	Acuidade	58.22%
	Especificidade	62.96%
	Sensibilidade	97.40%
	Taxa de erro	41,78%
11	Acuidade	57.77%
	Especificidade	55.41%

Cenários	Medida	Resultados
	Sensibilidade	95.71%
	Taxa de erro	42,23%

Fazendo uma análise mais específica dos melhores resultados de cada medida, é possível verificar que para a Acuidade obteve 58,22%, para os cenários (3), (6) e (7), com os algoritmos NB e DT. Para a Especificidade, os cenários (3) e (6) apresentaram os melhores valores com 62,96% para os algoritmos NB e DT. Para a Sensibilidade, 5 cenários (10,12, 13, 14 e 15) obtiveram 100%, para o algoritmo NB e 3 cenários (3, 6 e 7) com 97,40% para o algoritmo DT. Por fim, a taxa de erros total, os melhores cenários foram 41,78% para o cenário (3), (6) e (7), para os algoritmos DT e NB.

Assim sendo, de um modo geral, para esta abordagem o melhor cenário foi o cenário 3, que é direcionado para doentes que tiveram algum AVC e sofreram algum transplante, incluindo ainda, os atributos da escala de Glasgow, qual a proveniência dos doentes e o tipo de admissão.

Importa também salientar, através da observação da tabela 19 que a Sensibilidade teve o maior valor para os cenários (3) e (6), o que significa que o modelo é capaz de prever quase todos os targets de valor "1". Por outro lado, o valor baixo que foi obtido da Especificidade significa que, apesar do modelo ser capaz de prever um maior número de valores "1", é também responsável por o maior número de Falsos Positivos. No entanto, os resultados são satisfatórios pois podem permitir que os médicos ajam de forma preventiva e possam diminuir a taxa de admissões e a taxa de mortalidade e, consequentemente, melhorar o serviço oferecido aos doentes.

A segunda abordagem foi realizada com o objetivo de verificar quais os dias de admissão que eram mais facilmente previstos, tendo sido organizados em quatro classes. Deste modo foram originados 120 modelos:

- DMM = {15 cenários, 4 técnicas de classificação, 2 métodos de *sampling*, 1 Target}

Como exemplo, pode ser utilizado o cenário 15 para demonstrar como é calculado o número de modelos criados:

- S15 = {VZ, VAA};
- Técnicas de Classificação = {NB, DT, GLM, SV};
- Métodos de *Sampling* = {1};
- Target = {Internamento};
- Classes = {1, 2, 3, 4}.

Com estas informações foi possível obter 8 modelos:

- DMM = {1 cenário * 4 técnicas de classificação * 2 métodos de sampling * 1 target}

Assim, são aqui apresentados os melhores resultados na segunda abordagem, cujos três melhores resultados foram obtidos através do cenário (1), (2) e (3). A escolha do melhor cenário foi realizada com base no calcula da Acuidade. Assim sendo, o cenário (1), teve 70,49% de acuidade nos algoritmos NB e DT e 52,96% no algoritmo SVM. O cenário (3), teve 48,92% para os algoritmos NB e DT, bem como o cenário (2). No entanto, o cenário (3) teve um SVM ligeiramente superior, de 21,70% de acuidade e o cenário (2) teve 27,29%. Na tabela 20 estão apresentados os seis melhores resultados de Acuidade obtidos para cada um dos algoritmos aplicados. Como pode ser facilmente observável, para além dos resultados serem ligeiramente mais baixos, estão dotados ainda de uma grande similaridade, o que, possivelmente, deve-se à discrepância de registos tendo em conta o target, constituindo um *dataset* ligeiramente desequilibrado.

Tabela 20 - 6 melhores resultados da segunda abordagem

Cenário	NB	DT	SVM
S1	70,49%	70,49%	52,56%
S3	48,92%	48,92%	21,70%
S2	48,92%	48,92%	21,29%
S7	48,92%	48,92%	17,52%
S12 e S13	48,92%	48,92%	15,50%
S8 e S10	48,92%	48,92%	15,23%

Capítulo V – Fase 3

Após a realização de todas as fases do trabalho, nomeadamente a recolha, análise e tratamento dos dados, bem como modelação e a obtenção de resultados, segue-se a discussão dos resultados, de forma a confrontá-los com os objetivos propostos inicialmente para esta dissertação.

A fase mais morosa da dissertação foi a verificação da qualidade dos dados, dado que foi necessária uma análise aprofundada dos *datasets* de modo a identificar os erros, bem como hipóteses de correção viáveis, dado que foram encontrados um grande número de valores em branco.

Na fase de *Clustering* foram obtidos bons resultados e permitiu sobretudo conhecer quais as variáveis de maior criticidade que definem os doentes internados nas UCI. Assim, os cenários com melhores resultados foram relativos a atributos de tipo de internamento, tipo de internamento por cirurgia e transplantado, bem como atributos relativos a doentes com tumores/cancro e respetivos tratamentos. É possível concluir sobretudo que o atributo tipo de internamento é crucial para agrupar os doentes, mas que o nível de risco prejudica os resultados das técnicas de *Clustering*.

Os cenários utilizados na primeira fase da dissertação foram utilizados para os testes na fase de classificação. No entanto, os testes iniciais não permitiram a obtenção de resultados satisfatórios tendo sido realizado a replicação de dados (*Sampling*), bem como a construção de mais cinco cenários. Além disso, foram realizados testes a duas abordagens, onde são distinguidos os doentes admitidos (menos de 24 horas) de doentes internados (superior a 24 horas). Na primeira abordagem, foram obtidos bons resultados, tendo o cenário 3 os melhores resultados. Neste cenário, o modelo é capaz de acertar cerca de 97% dos doentes que são internados, mas são criados demasiados Falsos Positivos, dado os 63% de Especificidade. No entanto, como já foi referido, os resultados são satisfatórios dado que permite uma estratégia preventiva por parte dos médicos.

A partir da segunda abordagem, direcionada para os dias de internamento (utiliza 4 classes de dias), os resultados foram menos satisfatórios, mas também permitiram uma análise útil para esta dissertação. Dos resultados obtidos importa salientar que o cenário 1, que contém todas as variáveis do *dataset*, obteve os melhores resultados, com cerca de 70% de Acuidade, o que mostra que todas as variáveis são importantes para esse modelo.

Os resultados obtidos permitiram oferecer uma nova visão para a comunidade científica, sendo relevante no domínio dos Sistemas de Apoio à Decisão (SAD), Data Mining (DM) e Medicina

Intensiva (MI), pois poderá ser útil para os profissionais de saúde na prevenção e diminuição de admissões não planeadas nas UCI. Além disso, poderá se uma forma de dar um caminho a novas investigações na área de DM, tendo por base os resultados e conclusões obtidas.

Capítulo VI – Fase 4 e 5

Neste capítulo foi agregada a fase 4 e fase 5. O motivo principal passa pelo facto de, como já foi explicado na secção 1.3.3. Métodos combinados, a fase 4 não ser realizada nesta dissertação devido a esta estar no âmbito académico. No entanto, a utilização e implementação dos modelos resultantes desta dissertação será realizada posteriormente à sua conclusão, sendo integrados no sistema INTCare do Centro Hospitalar do Porto (CHP).

A fase 5 compreende a "Comunicação" do método *Design Science Research*, o qual inclui o desenvolvimento dos diversos artefactos que facilitaram a disseminação do trabalho de dissertação, bem como os resultados obtidos. Para tal, para além do presente documento, foram ainda desenvolvidos e submetidos dois artigos científicos em duas conferências internacionais, os quais foram aceites e posteriormente publicados/apresentados nas respetivas conferências, bem como mais um artigo para ser publicado numa revista cientifica:

- *Patients' Admissions in Intensive Care Units – A clustering overview;*
- *Predicting Patients admission in Intensive Care Units using Data Mining.*
- *A study about Patients' Admissions in Intensive Care Units.*

As informações principais relativas a cada um dos artigos encontra-se no anexo I, anexo II e anexo III, onde, entre outras, se pode examinar o resumo e as palavras-chave associadas aos artigos. A divulgação amplificada do trabalho realizado na comunidade científica confere uma mais-valia não só para esta dissertação, como também para os trabalhos futuros relativamente ao tema.

Capítulo VII – Conclusões

7.1. Considerações

Os Sistemas de Apoio à Decisão Clínicos (SADC) tornaram-se um foco de grande criticidade num ambiente como os da Unidade de Cuidados Intensivos (UCI), dado que permitem a análise aprofundada dos dados que o pensamento humano não tem a capacidade de descobrir padrões em *datasets* de grandes dimensões. Assim, o surgimento de técnicas como *Data Mining* (DM) tem permitido evoluir no que toca a descoberta de conhecimento em bases de dados e, quando aplicado na Medicina, tem permitido até salvar vidas.

Esta dissertação teve como objetivo principal apurar a possibilidade de categorizar e prever internamentos em UCI, utilizando dados de um SADC, que se encontra numa UCI do Centro Hospitalar do Porto (CHP), bem como obter modelos de DM úteis para suportar o processo de tomada de decisão clínico. Considera-se que estes objetivos foram concluídos com sucesso. Deste modo, foi possível responder às duas questões de investigação colocadas no inicio desta dissertação:

1. "É possível encontrar fatores de risco que permitam caracterizar os doentes internados na UCI?"
2. "É possível prever qual a probabilidade de um doente ser admitido às UCI?"

Além disso, obtiveram-se os resultados pretendidos, nomeadamente ao nível da identificação dos fatores de risco e características dos doentes-tipo internados em UCI, bem como foram obtidos bons modelos para verificar a possibilidade de realizar previsões de internamentos. O facto de terem sido cumpridos os objetivos e os resultados esperados, permitiu a criação de conhecimento útil para a comunidade cientifica.

Os dados utilizados nesta dissertação provieram de três tabelas que continham informação sobre doentes admitidos e/ou internados na UCI. Estes foram explorados de modo a compreender a sua necessidade no processo de DM, bem como para aferir erros e omissões de informação que necessitasse de correção. Esta foi uma das fases mais longas pois requereu que fossem verificados todos os tipos de erros possíveis, bem como encontrar soluções para cada um. O erro mais comum eram as linhas com atributos em branco, o que complicou a formação de um *dataset* final, devido se encontrarem em número elevado. Os erros encontrados foram causados sobretudo

por erros humanos aquando a inserção dos dados, sendo, por isso, necessário automatizar o mais possível o processo de registo de dados.

Para auxiliar a execução deste projeto, foram utilizadas diversas tecnologias, nomeadamente *Oracle SQL Developer, Orange* e *RapidMiner* e recorreu-se a técnicas de *Clustering* e Classificação, as quais dividem o trabalho em duas partes. Relativamente aos métodos utilizados, foi aplicado *Design Science Research* (DSR) para investigação e *Cross Industrial Standard Process for Data Mining* (CRISP-DM) para a aplicação de técnicas de DM. Foram ainda construídos 10 cenários para aplicação das técnicas de *Clustering*, tendo sido construídos mais 5 para a Classificação.

A primeira parte, na qual foram aplicadas técnicas de *Clustering*, nomeadamente o algoritmo *k-means*, com testes *Silhouete*, *Distance to Centroids*, *Inter-cluster distance* e *Davies Bouldin Index*, permitiu identificar os atributos mais importantes de um doente internado em UCI. Foi possível concluir que atributos como o tipo de internamento e tipo de internamento por cirurgia são as variáveis mais críticas e que devem ter maior atenção. No caso do tipo de internamento, este é considerado um dos atributos mais importantes dado que existe uma clara distribuição dos seus valores pelos dois clusters formados. No entanto, verificou-se que atributos como o Risco prejudicam os resultados da aplicação das técnicas de *Clustering*. Assim, apesar de não permitir prever internamentos, os modelos resultantes desta primeira parte permitem obter o conjunto de características dos doentes internados em UCI e, assim, ter em atenção estes doentes. Esta parte da dissertação permitiu então verificar a possibilidade de identificar os fatores de risco e caracterizar os doentes-tipo internados em UCI.

Na segunda parte desta dissertação, foi aplicada a Classificação em duas abordagens: na primeira foi utilizada previsão para duas classes (classe 1 para doentes que estiveram numa UCI num período inferior a 24 horas e classe 2 para doentes que foram internados numa UCI num período superior a 24 horas) e na segunda foi utilizada uma previsão para quatro classes (classe 1 para doentes admitidos num período inferior a 24 horas, classe 2 para doentes internados até 3 dias, classe 3 para doentes internados até 5 dias e classe 4 para doentes internados mais de 5 dias). Foram realizados os primeiros testes que não se demonstraram satisfatórios. Assim, foi realizado o *Sampling* dos dados, tendo sido replicados até ser encontrado um equilíbrio dos valores dos *targets*. Os modelos obtidos deste modo permitiram concluir que a primeira abordagem era mais indicada para realizar previsões, mas a segunda abordagem mostrou que todos os atributos do *dataset* são importantes para realizar a previsão com 4 classes. Para aplicar as técnicas de

classificação foram utilizados 4 algoritmos, nomeadamente *Naive Bayes, Decision Trees, Support Vector Machines* e *Generalized Linear Model.*

Para cada modelo foram realizados testes de métricas: Sensibilidade, Acuidade, Especificidade e Taxa de erro. Para a segunda abordagem só foi possível realizado o teste de acuidade devido ao facto de ser constituída por 4 classes. Assim, importa referir que, para a primeira abordagem, os cenários 3 e 6 permitem obter modelos capazes de prever quase todos os targets de valor "1", devido ao elevado valor de sensibilidade. Não obstante, o valor baixo obtido na especificidade significa que os modelos criam também um maior numero de Falsos Positivos. No entanto, os resultados finais foram satisfatórios dado que permitem uma ação preventiva por parte dos médicos, que permitam agir cautelosamente no tratamento dos doentes e, assim, prevenir internamentos em UCI. Relativamente à segunda abordagem, como foi mencionado anteriormente, apesar dos resultados serem inferiormente satisfatórios, é possível concluir que todos os atributos do *dataset* são importantes para a previsão de internamentos, devido ao seu valor mais elevado de acuidade. Deste modo, foi possível verificar a possibilidade de prever internamentos em UCI através da aplicação de técnicas de DM, bem como cumprir o objetivo de obter bons modelos de previsão de internamentos.

O trabalho desenvolvido nesta dissertação deu ainda origem a três publicações cientificas, em que uma será para *proceedings* de uma conferência, outra será para constituir um capitulo de um livro e, por fim, uma das publicações foi estendida para ser publicada numa revista. Importa referir que estas publicações foram revistas por revisores especializados, tendo sido aceites para publicação e serão indexadas pelo *Scopus*.

Deste modo, é possível concluir que os objetivos propostos foram alcançados, oferecendo um contributo útil para a comunidade científica, médica e de DM, podendo ainda ser útil para posteriores investigações. Espera-se, então, que os modelos obtidos desta dissertação possam constituir um suporte para o processo de tomada de decisão clínico e que permita que os médicos passem a agir de forma preventiva, para diminuir o número de internamentos não planeados em UCI.

7.2. Dificuldades encontradas

A primeira grande dificuldade encontrada foi na procura da informação cientifica diretamente relacionada com o tema. Apesar da grande variedade de trabalhos científicos que utilizem DM na MI, não existem projetos que permitam previsões de internamentos. Os trabalhos encontrados

direcionados para estas previsões apenas utilizam Modelos Estatísticos, ficando aquém das intenções desta dissertação.

Na fase de exploração e tratamento dos dados também foram encontradas algumas dificuldades, dado o grande número de colunas e registos em branco. Este fator poderia influenciar negativamente a dissertação, levando a um prolongamento desta fase, tendo sido, por isso, uma das fases mais morosas.

Na fase de modelação, mais propriamente nos modelos de Classificação, os resultados iniciais não se apresentavam satisfatórios, não permitindo totalmente responder às questões de investigação colocadas inicialmente. Para ultrapassar este obstáculo, foram aplicadas técnicas de *Sampling* dos dados, permitindo uma melhoria dos resultados obtidos e uma melhor resposta ao que se pretendia para a dissertação.

7.3. Trabalho futuro

Para trabalho futuro, importa indicar algumas das orientações mais importantes a seguir numa fase posterior:

- Determinar novos atributos e novos dados que possam ser inseridos na modelação para obter novos modelos e melhores resultados;
- Aplicar diferentes técnicas de DM e novos cenários;
- Utilizar mais abordagens ao nível do target, com diferentes tipos de classes.
- Implementar no sistema de suporte à decisão clínico, INTCare

Referências Bibliográficas

ACSS. 2012. Administração Central do Sistema de Saúde, Circular Normativa nº33/2012.

Arabi, Y., S. Venkatesh, S. Haddad, S. A. Malik & A. A. Shimemeri (2004) The characteristics of very short stay ICU admissions and implications for optimizing ICU resource utilization: the Saudi Experience. *International Journal for Quality in Health Care,* 16**,** 149-155.

Asghar, S. & K. Iqbal. 2009. Automated Data Mining Techniques: A Critical Literature Review. In *Information Management and Engineering, 2009. ICIME '09. International Conference on*, 75-79.

Bates, D. W., M. Cohen, L. L. Leape, J. M. Overhage, M. M. Shabot & T. Sheridan (2001) Reducing the Frequency of Errors in Medicine Using Information Technology. *Journal of the American Medical Informatics Association,* 8**,** 299-308.

Berner, E. S. 2007. *Clinical Decision Support Systems: Theory and Practice*. Springer New York.

Bersten, A. D. & N. Soni. 2013. *Oh's Intensive Care Manual*. Elsevier Health Sciences UK.

Bouch, D. C. & J. P. Thompson (2008) Severity scoring systems in the critically ill. *Continuing Education in Anaesthesia, Critical Care & Pain,* 8**,** 181-185.

Braga, A., F. Portela, M. F. Santos, J. Machado, A. Abelha, Á. Silva & F. Rua (2015) Step Towards a Patient Timeline in Intensive Care Units. *Procedia Computer Science,* 64**,** 618-625.

Braga, P., F. Portela, M. F. Santos & F. Rua. 2014. Data Mining to Predict Patient's Readmission in Intensive Care Units. In *6th International Conference on Agents and Artificial Intelligence*, 270-276. França.

Brunelli, A., M. K. Ferguson, G. Rocco, P. Pieretti, W. T. Vigneswaran, N. J. Morgan-Hughes, M. Zanello & M. Salati (2008) A Scoring System Predicting the Risk for Intensive Care Unit Admission for Complications After Major Lung Resection: A Multicenter Analysis. *The Annals of Thoracic Surgery,* 86**,** 213-218.

Caldeira, V., J. Júnior, A. Oliveira, S. Rezende, L. Araújo, M. Santana, C. Amendola & E. Rezende. 2010. Criteria for patient admission to an intensive care unit and related mortality rates. In *Revista da Associação Médica Brasileira*, 528-534. Brasil.

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

Fayyad, U., G. Piatetsky-Shapiro & P. Smyth. 1996. From Data Mining to Knowledge Discovery in Databases. In *American Association for Artificial Intelligence*, 37-54.

Frawley, W. J., G. Piatetsly-Shapiro & C. J. Matheus. 1992. Knowledge Discovery in Databases: An Overview. In *Artificial Intelligence Magazine*.

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

Goldman, L., E. F. Cook, P. A. Johnson, D. A. Brand, G. W. Rouan & T. H. Lee (1996) Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain. *New England Journal of Medicine,* 334**,** 1498-1504.

Guiza Grandas, F., D. Fierens, J. Ramon, H. Blockeel, G. Meyfroidt, M. Bruynooghe & G. Van den Berghe. 2006. Predictive data mining in intensive care. In *Annual Machine Learning Conference of Belgium and The Netherlands*. Bélgica.

Han, J. (2009) Data Mining. - *Encyclopedia of Database Systems***,** 595-598.

Hannan, T. J. 1996. Electronic Health Record. In *Health Informatics: An Overview*. Australia.

Hanson Iii, C. W. & B. E. Marshall (2001) Artificial intelligence applications in the intensive care unit. *Critical care medicine,* 29**,** 427-435.

Haux, R. (2006) Health information systems – past, present, future. *International Journal of Medical Informatics,* 75**,** 268-281.

Jensen, L. G. & C. Bossen (2016) Factors affecting physicians' use of a dedicated overview interface in an electronic health record: The importance of standard information and standard documentation. *International Journal of Medical Informatics,* 87**,** 44-53.

Jing, H. 2009. Advances in Data Mining: History and Future. In *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on*, 634-636.

Kamath, A. F., J. T. Gutsche, Z. N. Kornfield, K. D. Baldwin, L. M. Kosseim & C. L. Israelite (2013) Prospective Study of Unplanned Admission to the Intensive Care Unit after Total Hip Arthroplasty. *The Journal of Arthroplasty,* 28**,** 1345-1348.

Keen, P. G. W. & S. Morton. 1978. Decision Support Systems: An Organizational Perspective Reading: Addison-Wesley.

Kirch, W. 2008. Electronic Health Record. In *Encyclopedia of Public Health,* ed. W. Kirch, 326. Netherlands: Springer.

Kulikowski, C. A. & S. M. Weiss (1982) Representation of Expert Knowledge for Consultation: The CASNET and EXPERT Projects. *Artificial Intelligence in Medicine*.

Labarère, J., P. Schuetz, B. Renaud, Y.-E. Claessens, W. Albrich & B. Mueller (2012) Validation of a Clinical Prediction Model for Early Admission to the Intensive Care Unit of Patients With Pneumonia. *Academic Emergency Medicine,* 19**,** 993 -1003.

Larssan, J. E. & B. Hayes-Roth (1998) Guardian: intelligent autonomous agent for medical monitoring and diagnosis. *Intelligent Systems and their Applications, IEEE,* 13**,** 58-64.

Levenson, S. A. (2010) The health care decision-making process framework *Maryland Medicine,* 11**,** 13-17.

Mann, R. I. & H. J. Watson (1984) A contingency model for user involvement in DSS development. *MIS Q.,* 8**,** 27-38.

Marins, F., L. Cardoso, F. Portela, M. Santos, A. Abelha & J. Machado. 2013. Intelligent Information System to Tracking Patients in Intensive Care Units. In *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction*, 54-61. Springer International Publishing.

Medicine, S. o. C. C. 1999. Guidelines for ICU Admission, Discharge and Triage. 633-638.

Mullins, P. M., M. Goyal & J. M. Pines (2013) National Growth in Intensive Care Unit Admissions From Emergency Departments in the United States from 2002 to 2009. *Academic Emergency Medicine,* 20**,** 479-486.

Orsini, J., A. Butala, N. Ahmad, A. Llosa, R. Prajapati & E. Fishkin (2013) Factors influencing triage decisions in patients referred for ICU admission. *Journal of clinical medicine research,* 5**,** 343-349.

Peffers, K., T. Tuunanen, M. A. Rothenberger & S. Chatterjee (2007) A design science research methodology for information systems research. *Journal of management information systems,* 24**,** 45-77.

Penedo, J., A. Ribeiro, H. Lopes, J. Pimentel, J. Pedrosa, R. Sá & R. Moreno. 2013. Avaliação da situação nacional das unidade de cuidados intensivos - Relatório final. Ministério da saúde.

Portela, F., S. Oliveira, M. Santos, J. Machado & A. Abelha (2015a) A Real-Time Intelligente System for Tracking Patient Condition. *Lectures Notes in Computer Science (LNCS) - Ambient Intelligence for Health (AmIHEALTH 2015),* 9456.

Portela, F., M. Santos, J. Machado, A. Abelha & Á. Silva. 2013. Pervasive and Intelligent Decision Support in Critical Health Care Using Ensembles. In *Information Technology in Bio- and Medical Informatics*, 1-16. Springer Berlin Heidelberg.

Portela, F., M. F. Santos, J. Machado, A. Abelha, F. Rua & Á. Silva. 2015b. Real-time Decision Support using Data Mining to predict Blood Pressure Critical Events in Intensive Medicine

Patients. In *Lecture Notes in Computer Science (LNCS) - Ambient Intelligence for Health (AmIHEALTH2015),* ed. Springer, 77-90

Portela, F., M. F. Santos, J. Machado, A. Abelha, Á. Silva & F. Rua. 2014a. Pervasive and intelligent decision support in Intensive Medicine – the complete picture. In *CAIg - Artigos em livros de atas/Papers in proceedings*, 87-102. Springer.

Portela, F., M. F. Santos, Á. Silva, F. Rua, A. Abelha & J. Machado. 2014b. Preventing Patient Cardiac Arrhythmias by using Data Mining Techniques. In *IEEE Conference on Biomedical Engineering and Sciencies*, 165-170. Sarawak, Malaysia: IEEE.

Portela, F., R. Veloso, M. F. Santos, J. M. Machado, A. Abelha, Á. Silva, F. Rua & S. M. C. Oliveira (2014c) Predict hourly patient discharge probability in intensive care units using data mining.

Power, D. J. 2008. Decision Support Systems: A Historical Review. In *Handbook on Decision Support Systems 1,* ed. Springer.

Ramon, J., D. Fierens, F. Guiza, G. Meyfroidt, H. Blockeel, M. Bruynooghe & G. V. D. Berghe (2007) Mining data from intensive care patients. *Advanced Engineering Informatics,* 21**,** 243 - 256.

Reiser, S. J. 1981. *Medicine and the Reign of Technology*. Cambridge University Press.

Richer, M. H. & W. J. Clancey. 1985. Guidon-Watch: A Graphic Interface for Viewing a Knowledge-Based System. Stanford: Department of computer science, Stanford University.

Saleh, A., M. Ahmed, I. Sultan & A. Abdel-lateif (2015) Comparison of the mortality prediction of different ICU scoring systems (APACHE II and III, SAPS II, and SOFA) in a single-center ICU subpopulation with acute respiratory distress syndrome. *Egyptian Journal of Chest Diseases and Tuberculosis,* 64**,** 843-848.

Saúde, M. d. 2003. Cuidados Intensivos: Recomendações para o seu desenvolvimento.

Silva, Á., P. Cortez, M. F. Santos, L. Gomes & J. Neves (2008) Rating organ failure via adverse events using data mining in the intensive care unit. *Artificial Intelligence in Medicine,* 43**,** 179 - 193.

Subbe, C. P., M. Kruger, P. Rutherford & L. Gemmel (2001) Validation of a modified Early Warning Score in medical admissions. *QJM,* 94**,** 521-526.

São João, C. H. 2014. Clinical Intelligence.

Tsai, J. C.-H., S.-J. Weng, C.-Y. Huang, D. H.-T. Yen & H.-L. Chen (2014) Feasibility of using the predisposition, insult/infection, physiological response, and organ dysfunction concept of sepsis to predict the risk of deterioration and unplanned intensive care unit transfer after emergency department admission. *Journal of the Chinese Medical Association,* 77**,** 133-141.

Turban, E., R. Sharda & D. Delen. 2011. *Decision Support and Business Intelligence Systems*. Prentice Hall.

Van den Bosch, G. E., P. Merkus, C. M. Buysse, A. A. Vaessen-Verberne, W. C. Hop & M. de Hoog (2012) Risk Factors for Pediatric Intensive Care Admission in Children With Acute Asthma. *Respiratory Care,* 57**,** 1391-1397.

Veloso, R., F. Portel, M. F. Santos, A. Abelha, J. Machado, Á. Silva & F. Rua (2015) Categorize readmitted patients in Intensive Medicine by means of Clustering Data Mining. *International Journal of E-Health and Medical Communications (IJEHMC)*.

Veloso, R., F. Portela, M. F. Santos, S. Álvaro, F. Rua, A. Abelha & J. Machado (2014) A Clustering Approach for Predicting Readmissions in Intensive Medicine. *Procedia Technology,* 16**,** 1307 - 1316.

Weil, M. H. & W. Tang (2011) From Intensive Care to Critical Care Medicine: A Historical Perspective. 183**,** 1451-1453.

Zaitseva, E., M. Kvassay, V. Levashenko & J. Kostolny. 2015. Introduction to knowledge discovery in medical databases and use of reliability analysis in data mining. In *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*, 311-320.

Anexo I – Publicação científica: *Patients' Admissions in Intensive Care Units – A Clustering Overview*

Autores: Ana Ribeiro, Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha.

Conferência: *IEEE Conference on Business Informatics - ISA'HEALTH@CBI'2016 - Intelligent Systems and Applications in Healthcare Workshop*

Estado: aceite para publicação.

Abstract— *Intensive Care is one of the most critical areas of medicine. Its multidisciplinary nature makes it a very wide area, requiring all types of healthcare professionals. Given the critical environment of intensive care units, it becomes evident the need to use technology of decision support systems to improve healthcare services and Intensive Care Units management. By discovering the common characteristics of the admitted patients it is possible to improve these outcomes. In this study clustering techniques were applied to data collected from admitted patients in Intensive Care Unit. The best results presented a Silhouette of 1, with a distance to centroids of 6.2e-17 and a Davies-Bouldin index of -0.652.*

Keywords— *Data mining; Clustering; Intensive care; admissions*; INTCare

Anexo II – Publicação científica: *Predicting Patients admission in Intensive Care Units using Data Mining*

Autores: Ana Ribeiro, Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha.

Conferência: *Lecture Notes in Computer Science (LNCS) - Computational Science and Its Applications - MIKE 2016*

Estado: aceite para publicação.

Abstract— *Intensive Care is a critical and multidisciplinary field of medicine with the goal to diagnose and treat patients suffering from critical conditions, where there is the possibility of occurrence of reversible failures in vital organs and/or that need to be monitored continuously. In order to provide better health care to patients it is necessary first of all to have an efficient way to properly admit to intensive care. This paper aims to apply classification data mining techniques to data extracted from an Intensive Care Unit (ICU) targeting patient's admissions. The results obtained in terms of sensitivity were satisfactory (97.4). They were achieved with Naïve Bayes and Decision Trees. These models are a way to support clinical decision support and help physicians to reduce the number of ICU admissions.*

Keywords— *Intensive Care, Data Mining, Classification, Admissions, INTCare System*

Anexo III – Publicação científica: *A study about Patients' Admissions in Intensive Care Units.*

Autores: Ana Ribeiro, Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha.

Revista: *MDPI Information*

Estado: aceite para publicação.

Abstract— *Intensive Care is one of the most critical area of medicine and its composed by all types of healthcare professionals that are prepared to treat all types of conditions/diseases. Due to the life-threatening environment of intensive care units it becomes important to use information technology to create clinical decision support systems to improve healthcare services and Intensive Care Units management. Unplanned and prolonged admissions in ICU are prejudicial to patient's health and implies a rearrangement of ICU resources (beds, doctors, nurses, financial resources, among others). Patients admitted to ICU have common characteristics that when discovered it could improve hospitals outcomes by means of mortality and admissions rates. This study aims to apply clustering techniques to data extracted from admitted patients in Intensive Care Unit in Centro Hospitalar do Porto in Portugal. The best results presented a Silhouette of 1, with a Distance to Centroids of 6.2e-17 and a Davies-Bouldin Index of -0.652, showing that there are some attributes that would clearly separate these patients into two groups (Clusters) like patients with surgery or patients with cancers/tumours.*

Keywords— *Data Mining; Decision Support Systems, Clustering; Intensive Care; Admissions; INTCare System*

Anexo IV: Erros e hipóteses de correção

Na tabela 21, 22 e 23, estão expostos os atributos de cada tabela da base de dados, os seus respetivos erros e hipóteses de correção dos mesmos.

Tabela 21 - Erros e hipóteses correção da tabela DOC_ADMISSAO

Nome do atributo	Erros	Hipóteses de correção
NPROCESSO	Sem erros	-
EPISODIO	Sem erros	-
MODULO	Todos os valores são iguais	Eliminar atributo
DATAN	Sem erros	-
SEXO	Sem erros	-
SERVICO	1 valor em branco	Eliminar linha correspondente
ESP	1 valor em branco	Eliminar linha correspondente
DESP	1 valor em branco	Eliminar linha correspondente
DATACRIACAO	Sem erros	-
DATAALTERACAO	Sem erros	-
ESTADO	Sem erros	-
ORDEM	4 valores em branco 7170 valores todos iguais.	Eliminar atributo

Tabela 22 - Erros e hipóteses de correção da tabela UCI_INTERNADO_ANT_MV

Nome do atributo	Erros	Hipóteses de correção
NUM_CAMA	Sem erros	-
NPROCESSO	Sem erros	-
EPISODIO	Sem erros	-
SEXO	Sem erros	-
DATAN	Sem erros	-
DATA_ENTRADA	Sem erros	-
DATA_SAIDA	Sem erros	-
N_SEQUENCIAL	Sem erros	-

Tabela 23 - Erros e hipóteses de correção da tabela UCI_INTERNADOS_ADMISSAO_ALL

Nome do atributo	Erros	Hipóteses de correção
NPROCESSO	Sem erros	-

Nome do atributo	Erros	Hipóteses de correção
EPISODIO	Sem erros	-
SERVICO	apenas 1 valor existente (INT)	Eliminar atributo
URGENCIA	9 valores em branco	Eliminar linha
SALAEMERGENCIA	9 valores em branco	Eliminar linha
OBS	9 valores em branco	Eliminar linha
BLOCOOPERATORIO	9 valores em branco	Eliminar linha
ENFERMARIA	468 valores em branco	Eliminar linha
OUTRAUCI	468 valores em branco	Eliminar linha
UNIDADEINTERMEDIA	468 valores em branco	Eliminar linha
DIRECTA	468 valores em branco	Eliminar linha
OUTROHOSPITAL	468 valores em branco	Eliminar linha
OUTRASITUACAO	468 valores em branco	Eliminar linha
TIPOINTERNAMENTO	468 valores em branco 24 valores em branco (-escolha-)	Eliminar linha
TIPOINTERNAMENTOCIRURGIA	575 valores em branco 20 valores em branco (-escolha-)	Eliminar linha Substituir "Sem cirúrgica" por "Sem cirurgia"
GLASGOW_HOSPITAL	737 em branco	Eliminar os números dentro dos parenteses
GLASGOW_SERVICO	769 valores em branco 2 valores "SEDADA" 1 valor "sedada com Propofol" 4 valores "sedado"	Eliminar os números dentro dos parenteses
SURDO	468 valores em branco	Eliminar linha
MUDO	468 valores em branco	Eliminar linha
CEGO	468 valores em branco	Eliminar linha
ALERGIAS	468 valores em branco	Eliminar linha

Nome do atributo	Erros	Hipóteses de correção
VIAAEREA	468 valores em branco	Eliminar linha
PACEMAKER	468 valores em branco	Eliminar linha
DEFICIENCIAFISICA	468 valores em branco	Eliminar linha
DEFICIENCIAPSIQUICA	468 valores em branco	Eliminar linha
SEQUELASAVC	Todos os valores em branco	Eliminar atributo
ALCOOLISMO	468 valores em branco	Eliminar linha
TOXICODEPENDENTEIV	468 valores em branco	Eliminar linha
ATRBH	575 valores em branco 56 valores em branco (-escolha-)	Eliminar linha Substituir *Falso* por *False* Substituir *Verdadeiro* por *True*
ATRBS	575 valores em branco 56 valores em branco (-escolha-)	Eliminar linha Substituir *Falso* por *False* Substituir *Verdadeiro* por *True*
INSUFHEPATICACRONICA	468 valores em branco	Eliminar linha
INSUFRENALCRONICA	468 valores em branco	Eliminar linha
INSUFCARDIACA	468 valores em branco	Eliminar linha
INSUFRESPIRATORIACRONICA	468 valores em branco	Eliminar linha
DPOC	468 valores em branco	Eliminar linha
DOENCAHEMATOLOGICA	575 valores em branco 51 valores em branco (-escolha-)	Eliminar linha
NEOPLASIANAOMETASTIZADA	2623 valores em branco	Eliminar linha
NEOPLASIAMETASTIZADA	575 valores em branco 53 valores em branco (-escolha-)	Eliminar linha
CORTICOTERAPIADELONGADURACAO	468 valores em branco	Eliminar linha
QUIMIOTERAPIA	468 valores em branco	Eliminar linha
RADIOTERAPIA	468 valores em branco	Eliminar linha
OUTROSTRATUMUNOSSUPRESSORES	468 valores em branco	Eliminar linha

Nome do atributo	Erros	Hipóteses de correção
TRANSPLANTADO	468 valores em branco	Eliminar linha
HTAEMTRATAMENTO	468 valores em branco	Eliminar linha
SEQUELASAVC2	468 valores em branco	Eliminar linha
DIABETESINSULINATRATADA	468 valores em branco	Eliminar linha
DIABETESINSULINANAOTRATADA	468 valores em branco	Eliminar linha
VENTILACAOMECANICA	Todos os valores em branco	Eliminar atributo
FARMACOSVASOATIVOS	Todos os valores em branco	Eliminar atributo
DATACRIACAO	Sem erros	-
INSUFCARDIAC	Todos os valores em branco	Eliminar atributo

Anexo V – Resultados de *Clustering* no Orange

Tabela 24 - Resultados obtidos da aplicação do algoritmo K-Means no Orange

Cenário	Atributos	Teste	Pior	Melhor
Cenário 1	Todos os atributos	Silhouete	0,22	0,63
		Inter-cluster distance	10,72	8,8
		Distance to centroids	13,48	4,85
Cenário 2	TIPOINTERNAMENTO TIPOINTERNAMENTOCIRURGIA GLASGOW_HOSPITAL GLASGOW_SERVICO RISCO PROVENIENCIA	Silhouete	0,41	0,67
		Inter-cluster distance	10,69	8,93
		Distance to centroids	11,93	3,4
Cenário 3	GLASGOW_HOSPITAL PROVENIENCIA TIPOINTERNAMENTO SEQUELASAVC2 TRANSPLANTADO	Silhouete	0,58	0,72
		Inter-cluster distance	8,827	7,36
		Distance to centroids	5,43	1
Cenário 4	TIPOINTERNAMENTOCIRURGIA TRANSPLANTADO TIPOINTERNAMENTO RISCO	Silhouete	0,42	**0,81**
		Inter-cluster distance	1,8	**1,4**
		Distance to centroids	0,93	**0,07**
Cenário 5	INSUFHEPATICACRONICA INSUFRENALCRONICA INSUFCARDIACA INSUFRESPIRATORIACRONICA PROVENIENCIA	Silhouete	0,62	0,74
		Inter-cluster distance	5,07	3,5
		Distance to centroids	1,5	0,2
Cenário 6	RISCO GLASGOW_HOSPITAL PROVENIENCIA	Silhouete	0,59	0,72
		Inter-cluster distance	8,8	7,07
		Distance to centroids	5,4	0,95
Cenário 7	TIPOINTERNAMENTOCIRURGIA TRANSPLANTADO TIPOINTERNAMENTO RISCO INSUFHEPATICACRONICA INSUFRENALCRONICA INSUFCARDIACA INSUFRESPIRATORIACRONICA	Silhouete	0,55	0,64
		Inter-cluster distance	5,09	3,16
		Distance to centroids	2,4	0,52

Cenário	Atributos	Teste	Pior	Melhor
	PROVENIENCIA			
Cenário 8	TIPOINTERNAMENTOCIRURGIA TRANSPLANTADO TIPOINTERNAMENTO	Silhouete	0,64	**1**
		Inter-cluster distance	1,6	**1,5**
		Distance to centroids	0,43	**6,2e-17**
Cenário 9	TIPOINTERNAMENTOCIRURGIA TRANSPLANTADO TIPOINTERNAMENTO ATRBH ATRBS DOENCAHEMATOLOGICA	Silhouete	0,6	0,9
		Inter-cluster distance	2	1,5
		Distance to centroids	0,53	0,11
Cenário 10	TIPOINTERNAMENTO NEOPLASIA QUIMIOTERAPIA RADIOTERAPIA OUTROSTRATIMUNOSSUPRESSORES	Silhouete	0,87	**0,93**
		Inter-cluster distance	1,7	**1,4**
		Distance to centroids	0,12	**0,04**

Anexo VI: Matrizes confusão Primeira Abordagem para os três melhores cenários

Tabela 25 - Matriz confusão DT do Cenário 3

	1	2
1	749	20
2	542	34

Tabela 26 - Resultados dos testes DT do Cenário 3

Acuidade	58,22%
Especificidade	62,96%
Sensibilidade	97,40%
Erro	41,78%

Tabela 27 - Matriz confusão GLM do Cenário 3

	1	2
1	295	474
2	188	388

Tabela 28 - Resultados dos testes GLM do Cenário 3

Acuidade	50,78%
Especificidade	45,01%
Sensibilidade	38,36%
Erro	49,22%

Tabela 29 - Matriz confusão NB do Cenário 3

	1	2
1	749	20
2	542	34

Tabela 30 - Resultados dos testes NB do Cenário 3

Acuidade	58,22%
Especificidade	63%
Sensibilidade	97%
Erro	41,78%

Tabela 31 - Matriz confusão SVM do Cenário 3

	1	2
1	168	601
2	80	496

Tabela 32 - Resultados dos testes SVM do Cenário 3

Acuidade	49,37%
Especificidade	45,21%
Sensibilidade	21,85%
Erro	50,63%

Tabela 33 - Matriz confusão DT do Cenário 6

	1	2
1	749	20
2	542	34

Tabela 34 - Resultados dos testes DT do Cenário 6

Acuidade	58,22%
Especificidade	62,96%
Sensibilidade	97,40%
Erro	41,78%

Tabela 35 - Matriz confusão GLM do Cenário 6

	1	2
1	220	549
2	140	436

Tabela 36 - Resultados dos testes GLM do Cenário 6

Acuidade	48,77%
Especificidade	44,26%
Sensibilidade	28,61%
Erro	51,23%

Tabela 37 - Matriz confusão NB do Cenário 6

	1	2
1	749	20
2	542	34

Tabela 38 - Resultados dos testes NB do Cenário 6

Acuidade	58,22%
Especificidade	63%
Sensibilidade	97%
Erro	41,78%

Tabela 39 - Matriz confusão SVM do Cenário 6

	1	2
1	136	633
2	51	525

Tabela 40 - Resultados dos testes SVM do Cenário 6

Acuidade	49,14%
Especificidade	45,34%
Sensibilidade	17,69%
Erro	50,86%

Tabela 41 - Matriz confusão DT do Cenário 11

	1	2
1	736	33
2	535	41

Tabela 42 - Resultados dos testes DT do Cenário 11

Acuidade	57,77%
Especificidade	55,41%
Sensibilidade	96%
Erro	42,23%

Tabela 43 - Matriz confusão GLM do Cenário 11

	1	2
1	552	217
2	362	214

Tabela 44 - Resultados dos testes GLM do Cenário 11

Acuidade	56,95%
Especificidade	49,65%
Sensibilidade	71,78%
Erro	43,05%

Tabela 45 - Matriz confusão NB do Cenário 11

	1	2
1	736	33
2	535	41

Tabela 46 - Resultados dos testes NB do Cenário 11

Acuidade	57,77%
Especificidade	55%
Sensibilidade	96%
Erro	42,23%

Tabela 47 - Matriz confusão SVM do Cenário 11

	1	2
1	28	741
2	29	547

Tabela 48 - Resultados dos testes SVM do Cenário 11

Acuidade	42,75%
Especificidade	42,47%
Sensibilidade	3,64%
Erro	57,25%

Anexo VII: Matrizes confusão Segunda Abordagem para os três melhores cenários

Tabela 49 - Matriz Confusão DT do Cenário 1

	1	2	3	4
1	160	0	0	0
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Tabela 50 - Matriz Confusão NB do Cenário 1

	1	2	3	4
1	160	0	0	0
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Tabela 51 - Matriz Confusão SVM do Cenário 1

	1	2	3	4
1	160	0	0	0
2	13	83	26	44
3	3	11	18	14
4	17	148	76	129

Tabela 52 - Matriz Confusão DT do Cenário 2

	1	2	3	4
1	0	49	0	111
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Tabela 53 - Matriz Confusão NB do Cenário 2

	1	2	3	4
1	0	49	0	111
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Tabela 54 - Matriz Confusão SVM do Cenário 2

	1	2	3	4
1	47	41	41	31
2	65	31	40	30
3	22	5	11	8
4	132	77	92	69

Tabela 55 - Matriz Confusão DT do Cenário 3

	1	2	3	4
1	0	49	0	111
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Tabela 56 - Matriz Confusão NB do Cenário 3

	1	2	3	4
1	0	49	0	111
2	0	83	0	83
3	0	20	0	26
4	0	90	0	280

Tabela 57 - Matriz Confusão SVM do Cenário 3

	1	2	3	4
1	51	65	27	17
2	70	46	23	27
3	17	12	7	10
4	96	145	72	57