Universidade do Minho
Escola de Engenharia

Eva Alexandra Pereira da Silva

Plataforma de *Business Intelligence* para o Estudo de Infeção Nosocomial

outubro de 2014

Universidade do Minho
Escola de Engenharia

Eva Alexandra Pereira da Silva

Plataforma de *Business Intelligence* para o Estudo de Infeção Nosocomial

Dissertação de Mestrado
Mestrado Integrado em Engenharia Biomédica
Ramo de Informática Médica

Trabalho efetuado sob a orientação do
Professor Doutor António Carlos da Silva Abelha
e supervisão da
Mestre Luciana Almeida Cardoso

outubro de 2014

Agradecimentos

Ao professor doutor António Abelha agradeço a disponibilidade para a orientação deste trabalho, bem como todo o apoio ao longo do desenvolvimento do mesmo.

À mestre Luciana Cardoso agradeço a ajuda e as sugestões sempre oportunas.

Ao professor doutor José Machado e ao doutor Filipe Portela agradeço a partilha de conhecimentos e toda a ajuda prestada ao longo da elaboração deste projeto.

Ao Serviço de Sistemas de Informação do Centro Hospitalar do Porto agradeço a oportunidade de realização deste trabalho, bem como as informações e meios disponibilizados.

À minha família agradeço o apoio e ajuda prestados nos momentos certos.

À Eliana agradeço a companhia durante as horas de trabalho e a partilha de ideias e conhecimentos.

Resumo

O controlo e a prevenção de infeções nosocomiais são essenciais para a redução de custos, bem como para a melhoria dos cuidados prestados numa instituição de saúde. Por outro lado, o tratamento de dados que permitam compreender, caracterizar e monitorizar as infeções possibilita um controlo e uma prevenção mais eficaz das mesmas. Sendo um método automatizado e eficiente para o tratamento de dados, a tecnologia de *Business Intelligence* permite a extração de informação importante para gerar conhecimento que pode auxiliar o processo de tomada de decisão dos profissionais de saúde.

O principal objetivo deste trabalho é o desenvolvimento e implementação de uma plataforma de *Business Intelligence* que permita o estudo da incidência de infeção nosocomial nas Unidades de Medicina do Centro Hospitalar do Porto. Este estudo é feito através da apresentação de um conjunto de indicadores clínicos (informações importantes extraídas dos dados referentes a infeções nosocomiais) que ajudam a analisar e caracterizar estas infeções. Por conseguinte, depois de identificados os indicadores relevantes, torna-se pertinente desenvolver um sistema que permita tratar os dados, extrair os indicadores destes e apresentá-los, de forma atrativa, na plataforma. Por sua vez, a plataforma facilita a análise das informações que disponibiliza, apoiando a tomada de decisões, nomeadamente através da identificação dos principais fatores de risco. Assim, o sistema atua como um Sistema de Apoio à Decisão Clínica, podendo auxiliar no controlo e prevenção destas infeções.

Pretende-se ainda estudar a aplicabilidade da tecnologia de *Data Mining* na criação de modelos de classificação capazes de prever a ocorrência de infeções nosocomiais, na presença de determinados fatores de risco.

O conhecimento obtido com a análise dos indicadores e as previsões efetuadas pode possibilitar a diminuição da incidência de infeção nosocomial e, consequentemente, a redução dos custos associados à sua ocorrência, bem como o aumento da segurança e do bem-estar dos doentes, ao permitir a tomada de decisões mais fundamentadas. A aplicação de *Business Intelligence* na área da saúde contribui para melhorar não só o fluxo de trabalho diário nas unidades de saúde, como também a qualidade dos cuidados prestados.

Abstract

The control and prevention of nosocomial infections are essential to reduce costs and improve the care delivered in healthcare institutions. On the other hand, the treatment of data that allow to understand, characterize and monitor the infections enables a more effective control and prevention of them. Being an automated and efficient method to treat data, the Business Intelligence technology allows to retrieve important information to create knowledge that can assist the decision making process of the healthcare professionals.

The main goal of this work is to develop and implement a Business Intelligence platform for the study of nosocomial infection incidence in the Medicine Units of *Centro Hospitalar do Porto*. This study is made through the presentation of a set of clinical indicators (important informations extracted from the nosocomial infection data) that help to analyse and characterize these infections. Therefore, after identifying the relevant indicators for this study, it becomes pertinent to develop a system that treats data, extracts the indicators from data and presents the indicators, in an attractive way, on the platform. The platform makes the analysis of the information derived from processed data easier, assisting the decision making, namely in the identification of the main risk factors. Thus, the system acts as a Clinical Decision Support System, helping in the control and prevention of these infections.

This work also intends to study the applicability of Data Mining technology in the creation of classification models, capable to predict the occurrence of nosocomial infections, in the presence of certain risk factors.

The knowledge, achieved by the analysis of the indicators and the predictions accomplished, may allow the decrease of nosocomial infection incidence. Consequently, it may allow the reduction of the costs associated with its occurrence, as well as the increase of patients' safety and well-being, by allowing a more reasoned decision making process. The application of Business Intelligence to the healthcare sector contributes to improve not only the daily workflow of healthcare units, but also the quality of the healthcare delivered.

Conteúdo

Índice de Figuras

Índice de Tabelas

Acrónimos e Siglas

AIDA Agência para a Integração, Difusão e Arquivo de Informação Médica e Clínica. 10, 54, 97

BI *Business Intelligence*. 4–8, 14–20, 24, 25, 33–39, 41–43, 47–50, 54, 64, 69, 74, 75, 77, 78, 87, 90, 93–97

CCI Comissão de Controlo de Infeção. 5, 47
CDE *Community Dashboard Editor*. 44, 64, 90
CHP Centro Hospitalar do Porto. 5, 6, 10, 40, 47–53, 55, 74, 79, 89, 92–97
CRISP-DM *CRoss-Industry Standard Process for Data Mining*. 31–33, 78, 81, 95

DCBD Descoberta de Conhecimento em Bases de Dados. 30, 31, 78
DM *Data Mining*. 6, 7, 17, 20, 29–33, 40–43, 55, 77–90, 92, 95–97
DSA *Data Staging Area*. 18
DW *Data Warehouse*. 16–22, 25, 29, 31, 33–36, 38, 40, 41, 49, 50, 54, 56, 67, 75, 117

ETL *Extract Transform Load*. 16–20, 25, 34, 42, 43, 50, 54, 64, 75, 97

FN Falso Negativo. 84, 85
FP Falso Positivo. 84, 85

HIV *Human Immunodeficiency Virus*. 52, 79
HOLAP *Hibrid On-Line Analytical Processing*. 29
HTML *HyperText Markup Language*. 44

KPIs *Key Performance Indicators*. 37, 38, 117, *Glossário: Key Performance Indicators*

MDX *Multidimensional Expressions*. 44, 45, 69
MOLAP *Multidimensional On-Line Analytical Processing*. 28, 29

OLAP *On-Line Analytical Processing*. 16, 17, 20, 25, 26, 28–30, 42, 44, 45, 55, 64, 66–69, 75, 97
OMS *Organização Mundial de Saúde*. 2

PCE Processo Clínico Eletrónico. 10, 37, 54
PL/SQL *Procedural Language/Structured Query Language*. 40, 50, 64

ROLAP *Relational On-Line Analytical Processing*. 29, 44

SAD Sistemas de Apoio à Decisão. 14, 15
SADC Sistemas de Apoio à Decisão Clínica. 1, 6, 75, 77, 87, 95, 96
SAM Sistema de Apoio ao Médico. 10
SAPE Sistema de Apoio às Práticas de Enfermagem. 10
SI Sistemas de Informação. 10–12, 14, 38
SIH Sistemas de Informação Hospitalar. 9–12, 36, 37
SONHO Sistema de Gestão de Doentes Hospitalares. 10
SQL *Structured Query Language*. 29, 33, 36, 40, 44, 65

TI Tecnologias de Informação. 1, 7, 9–11, 13, 14, 17, 39

VN Verdadeiro Negativo. 84, 85
VP Verdadeiro Positivo. 84, 85

XML *eXtended Markup Language*. 44, 56, 64

Capítulo 1

Introdução

Atualmente, devido a questões financeiras e à elevada competitividade entre organizações, é imprescindível que uma instituição de saúde utilize adequadamente os seus recursos, prestando serviços de alta qualidade. De modo a atingir estes objetivos é necessário tomar decisões fundamentadas, rápidas e de qualidade.

Com os avanços tecnológicos, a implementação de Tecnologias de Informação (TI) na área da saúde tem aumentado significativamente, sendo responsável pelo elevado volume de dados recolhidos atualmente. Estes dados, quando apropriadamente explorados, podem revelar informações muito importantes sobre os processos que decorrem na organização, permitindo identificar eventuais problemas e oportunidades de melhoria no ambiente hospitalar. Assim, estes dados são indispensáveis para fundamentar as decisões e fazer com que estas sejam mais acertadas.

Devido à importância atribuída à tomada de decisão em contexto hospitalar torna-se então fundamental a existência de ferramentas computacionais capazes de auxiliar e facilitar este processo. Neste sentido, nos últimos anos tem-se intensificado a exploração de dados clínicos recorrendo a essas ferramentas com o objetivo de criar informação e, consequentemente, conhecimento, de uma forma rápida e automática. Assim, surgiram os Sistemas de Apoio à Decisão Clínica (SADC), sistemas computorizados desenvolvidos para facilitar e melhorar a tomada de decisão clínica. Através do auxílio a

este processo, estes sistemas permitem melhorar a qualidade dos cuidados de saúde prestados, bem como reduzir erros médicos e custos nas instituições de saúde [1].

1.1 Enquadramento e Motivação

Uma infeção nosocomial ou infeção adquirida em ambiente hospitalar é uma infeção contraída numa instituição de saúde, ou seja, que não estava presente ou em fase de incubação no momento de admissão do doente. Esta pode ocorrer durante as 48 horas após a admissão do doente, durante 30 dias após uma cirurgia ou durante 3 dias após a alta, abrangendo assim as infeções adquiridas na instituição mas que só são detetadas após a alta [2–4]. As infeções nosocomiais também incluem infeções ocupacionais apresentadas pelos profissionais que trabalham na instituição de saúde [4].

Segundo a *Organização Mundial de Saúde* (OMS) [4], nos países em desenvolvimento, em cada 100 doentes hospitalizados, 10 contraem uma infeção nosocomial. No caso dos países desenvolvidos estas infeções afetam 7 em cada 100 doentes hospitalizados. Além disso, todos os anos mais de 4 milhões de doentes são afetados por infeções nosocomiais na Europa e 1.7 milhões nos Estados Unidos da América [4].

Existem vários fatores que contribuem para a aquisição de uma infeção nosocomial, como por exemplo a idade e o estado imunitário do doente, o seu tempo de permanência na instituição de saúde, os procedimentos médicos a que foi submetido, a utilização de antibióticos, os métodos de diagnóstico utilizados, entre outros. O ambiente hospitalar apresenta também muitos focos de infeção, objetos ou ambientes nos quais os microrganismos podem sobreviver ou multiplicar-se, tais como as instalações, os aparelhos invasivos ou equipamentos utilizados, ou até os doentes, profissionais de saúde e visitantes [3, 5]. Desta forma, nas instituições de saúde, qualquer falha nas práticas de controlo e prevenção de infeção, em combinação com o estado imunitário debilitado do doente, poderá facilmente resultar na aquisição de uma infeção nosocomial [5]. Por conseguinte, as unidades de cuidados intensivos, devido ao estado imunitário dos doentes internados nestas unidades e

também aos procedimentos invasivos efetuados, apresentam uma probabilidade de ocorrência destas infeções significativamente superior [2, 3].

As infeções nosocomiais têm muito impacto na mortalidade e morbilidade dos doentes de uma instituição de saúde. Além disso, um doente com uma infeção adquirida em ambiente hospitalar permanece mais tempo na instituição, podendo mesmo ser necessário o seu reinternamento, resultando, consequentemente, em custos adicionais para a organização [2, 3, 5]. Estas infeções são também um fator muito importante para avaliar a qualidade dos cuidados prestados. Portanto, o seu controlo e a sua prevenção são fundamentais, permitindo a redução de custos para a organização, a diminuição do risco de infeção por parte dos profissionais de saúde, doentes ou pessoas que visitam a instituição, bem como a redução do desconforto e do sofrimento dos doentes.

A figura 1.1 resume os principais fatores que podem contribuir para a ocorrência de infeções nosocomiais, quer ao nível das características do doente e do seu estado de saúde, quer ao nível do ambiente hospitalar ou ainda ao nível das medidas de combate à infeção. Esta figura apresenta ainda as principais consequências destas infeções.

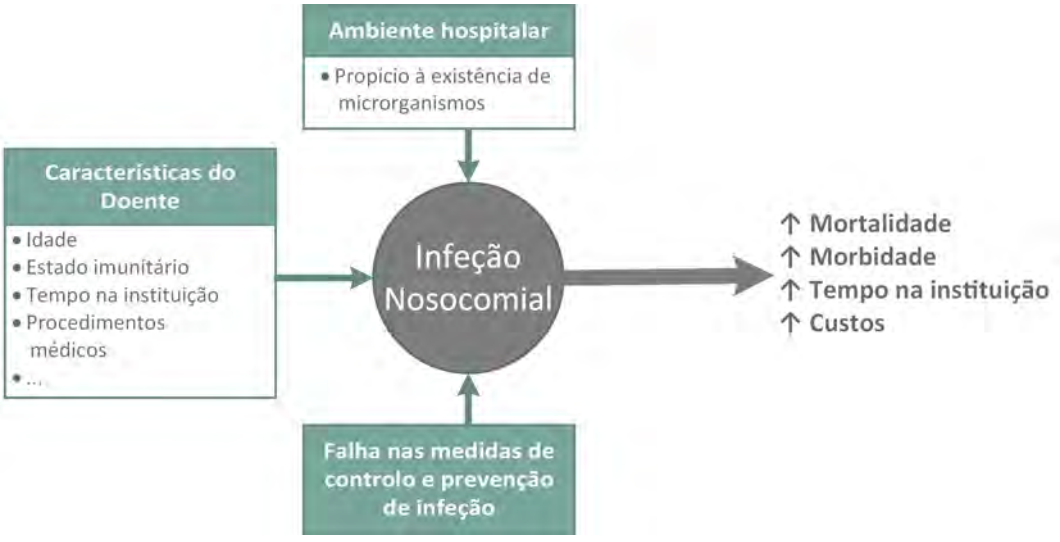

Figura 1.1: Fatores que podem contribuir para a ocorrência de infeções nosocomiais e consequências das mesmas.

Segundo Inweregbu *et al.* [2], cerca de um terço das infeções nosocomiais podem ser efetivamente prevenidas e controladas através da implementação de medidas de controlo e prevenção de infeção apropriadas.

Além da implementação destas práticas, uma instituição de saúde deve fazer continuamente a monitorização dos resultados relacionados com os programas de prevenção e controlo de infeções, recorrendo, para isso, à recolha periódica de dados e à análise de indicadores específicos [6]. Estes indicadores são parâmetros que devem ajudar a sumariar e a compreender fatores importantes presentes nos dados, tais como a taxa de infeção nosocomial e os fatores que contribuem para a sua incidência na organização. A análise destes indicadores permite também que os profissionais de saúde responsáveis pelo estudo de infeção identifiquem entre outros: atividades e processos críticos que decorrem no ambiente hospitalar; especialidades clínicas onde a implementação de medidas é essencial para garantir a segurança e o bem-estar dos doentes; áreas prioritárias onde as medidas devem ser executadas com mais urgência. Deste modo, é possível planear e implementar medidas específicas e eficientes para reduzir a taxa de incidência de infeção nosocomial nestas operações críticas e, consequentemente, aumentar a qualidade dos cuidados prestados.

A tecnologia de *Business Intelligence* (BI) pode ser utilizada para gerar e apresentar estes parâmetros, pois permite o tratamento de dados e a análise de informação extraída dos mesmos, de forma fácil e eficiente. Assim, a informação apresentada por um sistema baseado na tecnologia de BI pode ser utilizada para auxiliar e melhorar o processo de tomada de decisão numa instituição de saúde.

O presente trabalho surge da necessidade de monitorizar continuamente o ambiente hospitalar com o intuito de aplicar medidas específicas para diminuir a taxa de incidência de infeção nosocomial. Para fazer essa monitorização através do tratamento de dados clínicos, apresenta-se a possibilidade de recorrer à tecnologia de BI. A motivação deste trabalho está ainda relacionada com a necessidade dos profissionais de saúde responsáveis pelo estudo da infeção nosocomial tomarem decisões rápidas e fundamentadas, a fim de melhorar a produtividade e a eficiência da organização e, simultaneamente,

a qualidade dos cuidados prestados. Pretende-se que este trabalho seja capaz de facilitar a apresentação e análise de informação referente a infeções nosocomiais, de modo a auxiliar estes profissionais no estudo da incidência deste tipo de infeções, bem como na tomada de decisões relacionadas com o controlo e a prevenção das mesmas.

1.2 Objetivos

O objetivo principal do presente trabalho consiste no desenvolvimento de uma plataforma de BI que permita o estudo da incidência de infeção nosocomial nas Unidades de Medicina do Centro Hospitalar do Porto (CHP), um centro hospitalar do norte de Portugal. O estudo realiza-se através apresentação de um conjunto de indicadores capazes de auxiliar na análise e compreensão da incidência da infeção nosocomial no CHP, através da identificação de fatores de risco e outros parâmetros importantes para caracterizar a infeção. Para atingir este objetivo, torna-se necessário realizar um levantamento dos indicadores com interesse para o estudo e desenvolver um sistema de BI para automatizar o tratamento de dados de infeção nosocomial e a extração de informações destes, de forma a gerar conhecimento.

O sistema permite que o utilizador compreenda a incidência de infeção nosocomial e faça a sua monitorização através da plataforma de BI. A plataforma é muito importante, pois as informações que disponibiliza podem ser utilizadas nas decisões relacionadas com o controlo e prevenção de infeções nosocomiais por parte dos profissionais de saúde responsáveis pelo estudo destas infeções. Destes profissionais destacam-se os da Comissão de Controlo de Infeção (CCI) do CHP, o órgão desta instituição responsável pelo estudo, prevenção, deteção e controlo de infeções, bem como pela promoção de ações neste âmbito.

Neste sentido, pretende-se com este projeto:

- desenvolver um sistema baseado em métodos e ferramentas de BI para: automatizar o processo de tratamento dos dados relativos às infeções nosocomiais registadas no CHP e a extração de informação dos mesmos;

garantir que esta esteja disponível no momento de decisão; permitir que os indicadores de infeção sejam facilmente analisados e interpretados pelo utilizador, auxiliando-o na tomada de decisões;

- estudar a aplicabilidade de ferramentas *open-source* de BI na extração e apresentação de informação a partir de dados clínicos;
- implementar o sistema de BI no CHP;
- analisar a importância da plataforma no fluxo de trabalho diário de uma instituição de saúde.

Com este projeto pretende-se ainda estudar a aplicabilidade da tecnologia de *Data Mining* (DM) na realização de previsões clínicas relacionadas com a ocorrência de infeções nosocomiais. Esta tecnologia permite criar modelos de previsão capazes de auxiliar o processo de tomada de decisão, através da descoberta de padrões presentes nos dados. Para esse estudo torna-se necessário: produzir modelos de classificação e avaliar a sua qualidade; aplicar o melhor modelo obtido a dados, para prever o valor da infeção nosocomial; e integrar os resultados na plataforma de BI, de forma a indicar a probabilidade de um doente não pertencer a um grupo de risco passível de contrair uma infeção.

Com estes objetivos tenciona-se responder às seguintes questões:

Questão 1. De que forma a tecnologia de BI se aplica à área da saúde, mais concretamente ao tratamento e à exploração de dados de infeções nosocomiais, de modo a extrair e apresentar as informações presentes nos mesmos?

Questão 2. Qual a contribuição do sistema de BI para o suporte à decisão em ambiente hospitalar e em que medida pode ser utilizado como SADC?

Questão 3. De que forma o sistema desenvolvido beneficia o CHP e a sociedade em geral?

1.3 Estrutura da Dissertação

A presente dissertação encontra-se dividida em sete capítulos.

No capítulo 1 é apresentado o problema a resolver, bem como a motivação, os objetivos do projeto e a organização geral da dissertação.

No capítulo 2 é feita uma contextualização da utilização de TI na área da saúde. Começa-se por expor a importância da implementação de TI nas instituições de saúde e os fatores que contribuem para a sua aceitação por parte dos utilizadores. É também evidenciada a importância do apoio à decisão na área da saúde, bem como o contributo das TI nesse processo.

O capítulo 3 aborda os fundamentos teóricos subjacentes à tecnologia de BI. Inicialmente é explorado o conceito de BI e os principais componentes de um sistema de BI. Em seguida, referem-se os fatores que contribuem para uma implementação bem-sucedida destes sistemas nas instituições de saúde, os principais benefícios da utilização dos mesmos, bem como alguns exemplos de trabalhos relacionados com a utilização de BI na área da saúde.

No capítulo 4 é apresentada a metodologia de investigação considerada no desenvolvimento deste trabalho. Posteriormente, é feita uma breve apresentação das ferramentas computacionais utilizadas para o armazenamento e manipulação dos dados relativos a infeções nosocomiais. São ainda comparadas algumas ferramentas *open-source* de BI, descrevendo-se com mais detalhe a ferramenta escolhida para implementar este projeto.

No capítulo 5 é descrita a solução proposta neste trabalho. Inicialmente é apresentada a metodologia utilizada para a implementação do sistema de BI, seguindo-se a descrição dos indicadores considerados, da arquitetura do sistema criado e dos seus componentes. Referem-se também as considerações e procedimentos realizados durante o desenvolvimento do projeto. Finalmente, são mencionados os principais resultados obtidos com a implementação do sistema de BI e é feita a sua discussão.

No capítulo 6 apresenta-se um estudo sobre a utilização de DM na previsão de infeções nosocomiais. Explica-se a metodologia seguida no desenvolvimento deste estudo e são discutidos os principais resultados obtidos. Referem-se ainda os resultados da integração do melhor modelo obtido na

plataforma de BI.

No capítulo 7 expõem-se as conclusões deste projeto, indicando-se algumas sugestões para futuros trabalhos.

Conclui-se esta dissertação com a apresentação da bibliografia consultada, seguida dos anexos considerados relevantes e complementares ao trabalho. Os anexos incluem resultados obtidos com a implementação do sistema de BI (Anexo A), as publicações científicas realizadas no âmbito desta dissertação (Anexo B) e um glossário com a definição de termos importantes para a compreensão global do trabalho (Anexo C).

Capítulo 2

Tecnologias de Informação na Saúde

Atualmente, a informação e o conhecimento são fatores muito importantes para o sucesso de uma organização, sendo utilizados como uma vantagem competitiva na realização de decisões importantes. A par desta situação encontram-se as instituições de saúde, caracterizadas por um elevado número de departamentos especializados e disciplinas médicas, cujos processos estão extremamente relacionados com informação e conhecimento, requerendo cooperação e coordenação interdisciplinar, pelo que a gestão de informação torna-se preponderante para este tipo de instituições [7].

Segundo Lenz e Reichert [7], muitos estudos têm demonstrado os efeitos positivos da utilização de TI na área da saúde, em particular na prevenção de eventos adversos. Esses eventos são complicações resultantes dos cuidados prestados ao doente e não da doença em si, sendo responsáveis por agravar o estado de saúde do doente, prolongar a sua estadia na unidade de saúde ou até causar incapacidade no momento da alta [8]. A falta de informação e comunicação são os principais fatores que contribuem para a existência desses eventos, mas a utilização de Sistemas de Informação Hospitalar (SIH), através da disponibilização de informação precisa, atempada e útil para o processo de tomada de decisão, possui o potencial de reduzir significativamente a ocorrência dos mesmos [7].

As TI podem ser aplicadas em várias áreas e contextos nas unidades de saúde. Como exemplos disso, podem referir-se diversos SIH, sistemas responsáveis pelo processamento de dados, informação e conhecimento em ambiente hospitalar [9]. Estes são desenvolvidos de modo a permitirem a execução de diversas funções relacionadas com os cuidados prestados, tais como a gestão de doentes e questões financeiras e legais [10]. Destes sistemas destaca-se o Processo Clínico Eletrónico (PCE), um repositório de informação sobre todo o histórico clínico dos doentes numa instituição de saúde.

Com o constante desenvolvimento de novas aplicações para realizar diferentes tarefas de acordo com as necessidades das instituições de saúde, torna-se necessário adotar *standards* para a integração e interoperabilidade dos diferentes sistemas utilizados. Neste sentido, têm surgido diversos *standards* que permitem integrar os diferentes Sistemas de Informação (SI), possibilitando a comunicação e troca de informação entre eles [7, 11].

Em Portugal, os principais SIH utilizados são o Sistema de Gestão de Doentes Hospitalares (SONHO) (focado na gestão de dados administrativos dos doentes), o Sistema de Apoio às Práticas de Enfermagem (SAPE) (para informatizar os registos de enfermagem efetuados) e o Sistema de Apoio ao Médico (SAM) (orientado para a atividade do médico). Por sua vez, no CHP encontra-se também implementado o PCE, um sistema que pode ser acedido em tempo real no momento de prestação de cuidados, auxiliando os profissionais clínicos através do acesso a registos de informação sobre toda a história clínica do doente. Nesta instituição estes sistemas interoperam através da plataforma Agência para a Integração, Difusão e Arquivo de Informação Médica e Clínica (AIDA), um sistema multiagente que garante essencialmente a interoperabilidade e a comunicação entre todos os SI utilizados, permitindo assim integrar, difundir e arquivar grandes volumes de dados provenientes de diferentes fontes [12, 13].

Atualmente, espera-se que a implementação de SIH resulte numa maior qualidade e segurança dos cuidados prestados, contribuindo para que estes se tornem mais orientados para as necessidades dos doentes e, simultaneamente, mais eficientes [9, 14]. Assim, a utilização destes sistemas apresenta-se como uma oportunidade para a melhoria da eficiência, da eficácia, da segurança e

da qualidade dos serviços de saúde prestados, permitindo, também, tornar as atividades económicas mais transparentes e obter informações importantes em tempo real [12, 15, 16]. De facto, a maioria dos estudos sobre a utilização de TI na saúde afirma que existe uma relação significativa entre a dimensão, a produtividade e o bem-estar financeiro de uma instituição de saúde e o nível de TI por esta adotado [15]. Contudo, existe ainda uma discrepância entre o potencial e a utilização real de TI na área da saúde [7].

A utilização de SIH para substituir os registos em papel permite a redução de custos, bem como a recolha, o armazenamento eficiente e fidedigno e a troca de dados, melhorando assim o desempenho e a qualidade dos cuidados prestados pela instituição de saúde [14]. Na prestação de cuidados de saúde é necessário integrar, por exemplo, os dados dos utentes com dados científicos e outras informações complexas relacionadas com a gestão das atividades e recursos da organização, portanto as limitações da utilização de registos em papel, neste processo de integração de informação, são evidentes.

De acordo com Foshay e Kuziemsky [17], numerosos estudos destacam que a implementação de SIH é um processo extremamente complexo. Lluch [14] partilha desta opinião, acrescentando que atualmente os computadores são muito utilizados em ambiente hospitalar porém, nem todos os profissionais de saúde utilizam TI e pouco se sabe sobre as mudanças organizacionais, o tempo e os custos necessários para a implementação bem-sucedida dos SIH. Além disso, a baixa taxa de utilização de TI está essencialmente relacionada com as mudanças organizacionais e com as alterações que os profissionais de saúde têm de incluir no seu trabalho como resultado da implementação destes sistemas [14].

Segundo Chen e Hsiao [10], têm sido realizados estudos que sugerem que a aceitação da tecnologia pode ser uma forma de medir o sucesso dos SI. Assim, a aceitação da tecnologia por parte dos profissionais de saúde é essencial para o sucesso da adoção e implementação de SIH. Deste modo, sendo os principais utilizadores destes sistemas, os médicos desempenham um papel fundamental neste processo de aceitação. Por exemplo, um *design* pouco atrativo de um SIH pode resultar na oposição à sua utilização, numa satisfação reduzida e até colocar a vida dos doentes em risco, especialmente quando

os prestadores de cuidados de saúde consideram que a interface do sistema é difícil de utilizar ou que o sistema é inconveniente para as suas rotinas. Portanto, no desenvolvimento de SIH, é extremamente importante considerar os fatores que afetam a aceitação destes sistemas por parte dos profissionais de saúde, sendo eles a utilidade e a facilidade de utilização do sistema, segundo a opinião desses profissionais [10]. A utilidade é, assim, avaliada através da opinião do utilizador sobre os benefícios do sistema para atingir os objetivos do seu trabalho. Quanto maior for a utilidade do sistema, mais fácil é a sua aceitabilidade e, consequentemente, a integração do mesmo na rotina diária da instituição de saúde. Por outro lado, a facilidade de utilização é a capacidade dos utilizadores se adaptarem ao sistema, sendo que quanto maior esta facilidade mais rapidamente este será aceite.

Além disso, as características mais importantes de um SI são a qualidade do sistema e a qualidade da informação disponibilizada pelo mesmo. O primeiro conceito refere-se às características do SI relacionadas com a sua resposta, a sua fiabilidade e segurança, ao passo que o segundo conceito remete para a integridade, a precisão, a completude e a intemporalidade da informação que este disponibiliza. Estas duas características afetam a satisfação do utilizador, sendo que esta contribui para a aceitação do sistema [10].

2.1 Importância do Apoio à Decisão

No sector da saúde é muito importante tomar decisões rápidas e de qualidade uma vez que estas estão frequentemente relacionadas com a vida humana. Por outro lado, as decisões médicas são tomadas a partir da interpretação de informação específica do doente tendo em consideração o conhecimento médico disponível. Portanto, a tomada de decisão médica tem de integrar as melhores evidências disponíveis com a experiência dos profissionais clínicos e os valores específicos relativos ao estado de saúde do doente [7]. Além disso, o ambiente hospitalar e os processos a ele associados são extremamente complexos, dinâmicos e de natureza multidisciplinar [18]. Por conseguinte, o processo de decisão é sempre muito complexo e requer o acesso a informação de elevada qualidade [7, 17].

Além do mais, um estudo realizado por Foshay e Kuziemsky [17] indica que a falta de meios para o apoio à decisão acarreta implicações negativas significativas numa organização. Afirmam ainda que na ausência de acesso atempado à informação necessária, os responsáveis pela tomada de decisões são obrigados a fazê-lo sem considerar os factos e informação necessários. Para isso, são obrigados a confiar na sua experiência e intuição. Os resultados deste estudo enfatizam que, desta forma, a falta de informação compromete significativamente os processos de tomada de decisão numa organização.

2.2 Tecnologias de Informação e Apoio à Decisão

Atualmente, com a implementação de TI nas instituições de saúde, o volume de dados recolhidos tem aumentado exponencialmente [19]. No entanto, a complexidade destes dados e o seu volume fazem com que estes não possam ser facilmente processados em tempo útil, sem recorrer a métodos automatizados. Desse modo, existe uma necessidade crescente de tratar os dados recolhidos, pois estes possuem muita informação essencial para suportar o processo de tomada de decisão, tanto a nível clínico como a nível administrativo [17, 19]. Assim, a utilização de técnicas de extração de informação torna-se fundamental para as organizações de saúde.

A apresentação dessa informação de forma atrativa, fácil de interpretar e em tempo real pode ajudar a reduzir custos e a melhorar a qualidade, a segurança e a eficiência dos cuidados de saúde prestados, pois permite a tomada de decisões mais racionais e fundamentadas [20]. Estes motivos fazem com que a gestão de informação seja crucial para as instituições de saúde.

Importa também destacar que, nos dias de hoje, as instituições de saúde atuam, frequentemente, sob pressão financeira, pelo que é muito importante possuírem a capacidade de melhorar a eficiência dos seus processos, isto é, aplicar os seus recursos o mais eficientemente possível, enquanto prestam serviços de qualidade [17, 18].

Em ambiente hospitalar as TI podem auxiliar o processo de tomada de decisão clínica de diferentes formas, como por exemplo [7]:

- os sistemas computacionais podem contribuir para melhorar diferentes aspetos da qualidade dos dados, melhorando, deste modo, a informação utilizada na tomada de decisões;
- os sistemas computacionais podem contribuir para melhor monitorizar o estado de saúde de um doente, por exemplo através da apresentação de informações sobre o doente e da criação de mensagens de alerta se algum parâmetro ou combinação de parâmetros se aproximarem de valores considerados perigosos;
- os sistemas computacionais podem ser utilizados para calcular probabilidades de doenças e doses de fármacos através de dados que se encontram registados no sistema, tais como a idade, o sexo ou o peso do doente.

A área da saúde é ainda caracterizada por um ambiente computacional altamente distribuído, onde diferentes sistemas e pessoas necessitam estar em contacto e comunicar, trocando dados e informação necessários à tomada de decisões [13]. Assim, a implementação de Sistemas de Apoio à Decisão (SAD) capazes de auxiliar esse processo é fundamental para as unidades de saúde. A temática dos SAD é uma área de investigação cujo objetivo é desenvolver e estudar SI computorizados capazes de suportar e melhorar o processo de tomada de decisão [21].

A necessidade de tomar boas decisões em ambientes muito competitivos e complexos, como é o caso da área da saúde, de forma rápida e considerando informação de qualidade, pode ser conseguida recorrendo à tecnologia de BI, um campo ligado ao desenvolvimento de SAD. No capítulo 3 são abordados conceitos teóricos subjacentes à tecnologia de BI, bem como a aplicação de sistemas baseados em BI na área da saúde.

Capítulo 3

Business Intelligence

O termo BI foi introduzido por Howard Dresner em 1989 para descrever o conjunto de conceitos e métodos utilizado para melhorar a tomada de decisão numa organização, através da utilização de sistemas computorizados [21, 22].

Um sistema de BI pode ser considerado um SAD. Este engloba todo um conjunto de metodologias e ferramentas, capazes de recolher, integrar, aceder e analisar dados para, a partir destes, obter e apresentar informações sobre as atividades e os processos que decorrem numa organização, promovendo, desse modo, uma tomada de decisão mais fundamentada e, consequentemente, melhores resultados [11, 21, 23]. Por outras palavras, um sistema de BI disponibiliza as tecnologias que permitem transformar os dados de uma organização em informação que, posteriormente, pode ser transformada em conhecimento estratégico e relevante. Esse conhecimento passível de suportar o processo de tomada de decisão permite assim decisões melhores e mais rápidas [24, 25].

A tecnologia de BI melhora o tempo de obtenção e a qualidade dos fatores a ponderar no processo de tomada de decisão [15]. Assim, os sistemas de BI possuem a capacidade de disponibilizar a informação correta e necessária a tempo de auxiliar o processo de tomada de decisão dos seus utilizadores, resultando, por isso, numa vantagem competitiva para a organização que os implementa.

Os principais benefícios da implementação da tecnologia de BI são uma

maior autonomia e flexibilidade dos utilizadores desta tecnologia na análise de informação, uma maior eficiência e suporte na tomada de decisão, bem como maior facilidade e poupança de tempo no acesso e na análise da informação [26].

3.1 Sistemas de *Business Intelligence*

Um sistema de BI (Figura 3.1) deve ser capaz de realizar duas tarefas fundamentais: a integração de grandes volumes de dados provenientes de diferentes fontes heterogéneas e a disponibilização de ferramentas para explorar esses dados e apresentar as informações resultantes dos mesmos [25]. Assim, um sistema de BI integra dados provenientes de diferentes fontes heterogéneas, convertendo-os num formato unificado (1) e carregando-os para um *Data Warehouse* (DW) (2). Posteriormente, os dados armazenados no DW são explorados por ferramentas de BI capazes de extrair e apresentar informações presentes nos mesmos (3). Essas informações são cruciais para auxiliar a tomada de decisões do utilizador do sistema (4).

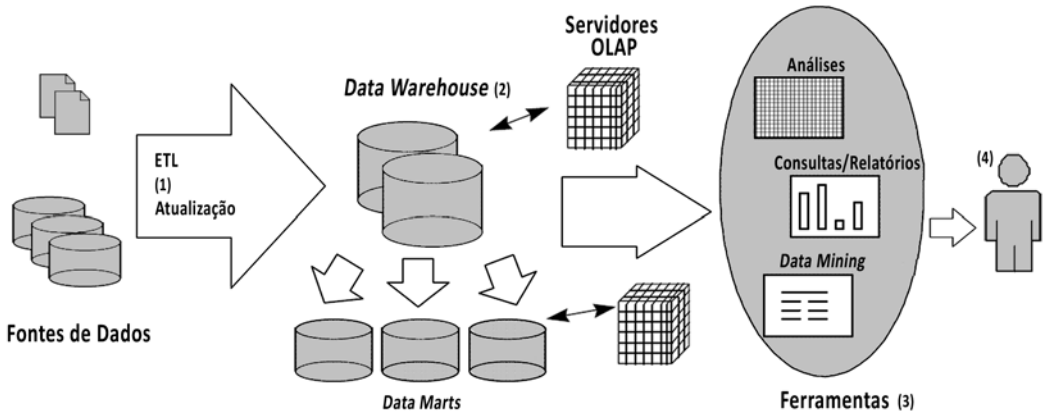

Figura 3.1: Arquitetura típica de um sistema de BI (adaptado de [27]).

A tecnologia de BI integra diferentes fontes de dados e inclui o DW e várias ferramentas de *software* para implementar o processo *Extract Transform Load* (ETL), para análise *On-Line Analytical Processing* (OLAP), para

DM, para consultar informação presente nos dados, para produção de relatórios, entre outras ferramentas [24]. As ferramentas OLAP, as ferramentas de DM, as consultas e os relatórios são diferentes métodos para explorar os dados e extrair informações dos mesmos, através da aplicação de diferentes tecnologias. Estas ferramentas disponibilizam diferentes formas de distribuição e apresentação da informação extraída aos utilizadores, para que esta seja rápida e facilmente compreendida e aplicada na tomada de decisões [28]. Os resultados podem então ser apresentados de diferentes modos, tais como gráficos, tabelas ou *dashboards* [20, 28]. Por sua vez, o DW e o processo ETL constituem a infraestrutura de TI necessária para que as ferramentas anteriormente referidas sejam corretamente implementadas.

Em comparação com os sistemas operacionais, que são utilizados para processar as transações diárias de uma organização, preocupando-se com o processamento rápido e eficiente das mesmas, os sistemas de BI permitem o acesso rápido a informação para análise e criação de relatórios [25].

Nos últimos anos tem-se verificado uma melhoria significativa das ferramentas de BI, não só ao nível da rapidez de recolha de dados, como também ao nível da sofisticação e natureza interativa dos processos de manipulação de dados e extração de informações destes e de apresentação dessas informações ao utilizador [20].

Em seguida, serão apresentadas as principais tecnologias que integram um sistema de BI.

3.1.1 Processo ETL

Num sistema de BI é necessário fazer a recolha de dados provenientes de diferentes fontes, geralmente, não integradas. Normalmente, as diferentes fontes de dados, que necessitam de ser integradas, correspondem a diferentes bases de dados operacionais de vários departamentos da organização. Estas fontes possuem dados com qualidade variável, utilizam representações inconsistentes e diferentes codificações e formatos. Na integração podem assim surgir erros, várias representações para o mesmo valor, valores em falta, entre outros problemas relacionados com a qualidade dos dados [29]. Deste modo,

é muito importante que os problemas de qualidade dos dados sejam detetados e corrigidos antes do carregamento dos mesmos para um DW, pois, como o DW é utilizado para suporte à decisão, é fundamental que os dados que armazena estejam corretos [27]. Assim, a integração é crucial, pois é o passo onde os dados provenientes de diferentes fontes heterogéneas são convertidos num formato adequado ao esquema definido para o DW [11].

Em BI, a tecnologia utilizada para fazer o processo de integração dos dados denomina-se ETL (Figura 3.2). Neste processo, os dados são extraídos das diferentes bases de dados e propagados para a *Data Staging Area* (DSA), onde são transformados num formato unificado e preparados para a sua migração para um DW. Posteriormente, os dados já transformados são carregados para o DW [30].

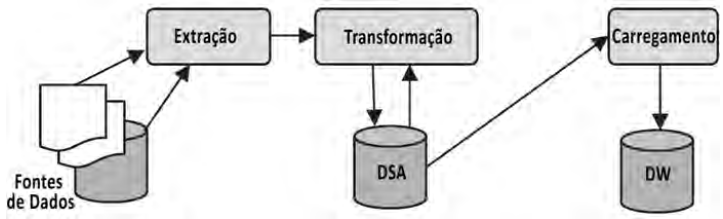

Figura 3.2: Processo ETL (adaptado de [30]).

O processo ETL não é um evento único. De facto, este processo repete-se periodicamente num sistema de BI, pois, à medida que as fontes de dados se alteram, é necessário atualizar o DW. Portanto, os processos ETL devem permitir que o sistema seja facilmente modificado e atualizado [30].

Um processo ETL é composto por três passos consecutivos: *Extração*, *Transformação* e *Carregamento*. O primeiro passo, *Extração*, é responsável pela extração dos dados das diferentes fontes e é constituído por duas fases: a extração inicial e a extração dos dados alterados [30]. A extração inicial corresponde à primeira vez que os dados são extraídos das fontes para serem carregados para o DW, e a extração dos dados alterados corresponde ao momento em que o processo ETL atualiza o DW com os dados que foram modificados e adicionados às fontes desde o momento da última extração.

O segundo passo do processo ETL é a *Transformação*. Esta etapa en-

globa a limpeza, a transformação e a integração dos dados de modo a obter dados coerentes, corretos, completos e inequívocos, num formato adequado à estrutura definida para o DW no qual serão armazenados [30]. Estas alterações não modificam as fontes originais, mas alteram apenas os dados antes de estes serem armazenados no DW [31].

O último passo deste processo, o *Carregamento*, diz respeito ao armazenamento dos dados (extraídos e transformados) no DW, que por sua vez poderá ser acedido pelos utilizadores finais ou por ferramentas de BI [30].

O processo ETL é a fase mais demorada e complexa da construção de um DW [30, 31]. Este processo é crucial para o carregamento eficiente de grandes volumes de dados para o DW e para a descoberta e correção de problemas relacionados com a qualidade dos mesmos, assegurando, desta forma, a qualidade dos dados armazenados no DW [29].

3.1.2 *Data Warehousing*

O componente fundamental de um sistema de BI é o DW, um repositório de dados provenientes de diferentes fontes, que funciona como uma base de dados de grandes dimensões utilizada para armazenar informação relativa às atividades de uma organização [28]. A integração de dados relevantes provenientes de várias fontes numa única localização e formato pode melhorar a velocidade e a eficiência do processo de descoberta de conhecimento, o que contribui para a tomada de decisões melhores, mais rápidas e mais fundamentadas [11].

Inmon [32] define o termo DW como uma coleção de dados orientada por assunto, integrada, não volátil, que varia no tempo e é capaz de suportar o processo de tomada de decisão. Importa clarificar que estas propriedades distinguem os DWs das bases de dados operacionais. Ao contrário das bases de dados operacionais, um DW não é volátil o que significa que os seus dados não são apagados ou alterados, mas vão sendo acrescentados ao repositório à medida que entram no sistema. Deste modo, os DWs variam no tempo, ou seja, cada registo corresponde a um momento específico no tempo e, por isso, permitem o armazenamento de dados históricos. Estes dados possibili-

tam, consequentemente, a disponibilização de informações sobre a evolução das atividades e processos que ocorrem numa organização ao longo de um determinado período de tempo. Além disso, normalmente os DWs possuem maiores dimensões que as bases de dados operacionais e são desenvolvidos especificamente para suportar a tomada de decisão da organização, podendo, por isso, ser vistos como bases de dados de apoio à decisão [30,33].

Os dados presentes num DW estão disponíveis para serem analisados por ferramentas de BI, podendo então ser aplicadas ferramentas para explorar os dados e apresentar a informação extraída dos mesmos, tais como, ferramentas OLAP, ferramentas de DM, ferramentas para consultas e/ou criação de relatórios, e *dashboards* ou outras ferramentas de apresentação de informação.

Um DW é periodicamente atualizado através do processo ETL aplicado às diferentes fontes de dados da organização, sendo que a frequência com que é atualizado depende das necessidades de cada organização [34].

Num sistema de BI os dados também podem ser armazenados, de acordo com o assunto a que se referem, em repositórios de menores dimensões denominados *data marts*, sendo que cada *data mart* possui dados referentes apenas a um determinado departamento ou grupo da organização [28]. As vantagens da utilização de *data marts* em relação à utilização de DWs são a quantidade de dados, que no *data mart* é muito menor, e o facto de este permitir uma análise mais orientada para os objetivos, visto que possui apenas dados referentes a um determinado assunto [28]. Pelos motivos anteriormente apresentados, o desempenho de um *data mart* na execução de consultas poderá ser muito superior ao de um DW.

Em *data warehousing* existem duas abordagens distintas mas também equivalentes para a construção de um DW: o paradigma de Bill Inmon e o paradigma de Ralph Kimball [35].

De acordo com o paradigma de Bill Inmon, a organização possui um DW cujos dados se encontram na 3ª forma normal e que, posteriormente, é utilizado para desenvolver *data marts*, seguindo-se, portanto, uma abordagem *top-down* na construção do sistema de *data warehousing* [35]. Nesta abordagem os dados são extraídos das diferentes bases de dados operacionais, transformados e armazenados numa única localização, o DW. Extrações de

dados deste DW permitem criar bases de dados departamentais de menores dimensões denominadas *data marts*.

Ralph Kimball defende uma abordagem *bottom-up* na construção de um DW, de acordo com a qual os dados são sempre armazenados segundo o modelo dimensional. Os dados provenientes das diferentes fontes são utilizados para alimentar *data marts* individuais que permitem visualizar pequenas porções dos dados organizacionais e, mais tarde, podem ser integrados num DW [35].

Na realidade, a maioria dos sistemas de *data warehousing* encontra-se mais próximo do paradigma de Ralph Kimball, pois normalmente um DW é estruturado ao nível departamental, sob a forma de um *data mart*, evoluindo para DW à medida que novos *data marts* são construídos [35]. Para além disso, normalmente um DW organiza os seus dados segundo um modelo dimensional, o que permite representar mais eficientemente os dados que serão utilizados para apoio à decisão [28]. A maioria dos DWs utiliza ainda a tecnologia de bases de dados relacionais, pois esta oferece uma abordagem eficiente, fiável e robusta para o armazenamento e gestão de grandes volumes de dados [33].

Modelação Dimensional

Segundo Kimball e Ross [36], a modelação dimensional é a técnica preferida para apresentar dados analíticos e armazená-los num DW, pois apresenta os dados de uma forma compreensível para os utilizadores e, simultaneamente, possui um elevado desempenho no processamento de *queries*. Esta técnica é uma atividade crítica para o sucesso de um projeto de desenvolvimento e implementação de um sistema de *data warehousing* [36].

Um modelo dimensional contém os mesmos dados e relacionamentos entre eles do que um modelo normalizado na 3ª forma normal, mas estruturados de forma diferente, num formato especialmente orientado para o suporte à tomada de decisão [36]. Este modelo é constituído por tabelas de factos e de dimensões. As tabelas de factos armazenam as medidas ou factos que correspondem aos objetos em análise nos dados, normalmente valores nu-

méricos [27]. Cada facto está associado a um conjunto de dimensões que o categorizam, isto é, contextualizam-no, dando-lhe significado e tornando-o único [29]. Por outras palavras, as dimensões correspondem às diferentes perspetivas para analisar os factos [34].

As tabelas de dimensões contêm um conjunto de atributos que correspondem a diferentes níveis na dimensão, que estão relacionados através de relações hierárquicas e que são utilizados para agrupar e condicionar factos [29]. Normalmente as tabelas de dimensões contêm menos registos que as tabelas de factos, mas contêm muitos mais atributos [36].

Uma tabela de factos possui os factos e uma ou mais chaves estrangeiras que se encontram relacionadas com as chaves primárias das dimensões, permitindo, deste modo, que os factos se relacionem com as dimensões. Além disso, cada facto encontra-se associado a um nível de detalhe específico, também conhecido com grão, sendo que todos os registos de uma tabela de factos possuem o mesmo grão [36]. O grão de uma tabela de factos diz respeito ao nível de informação mais atómico, sendo que o grão só pode ser agregado, não existindo possibilidade de o desagregar.

Como principais vantagens da utilização do modelo dimensional destacam-se a simplicidade do modelo de dados, a facilidade com que o modelo pode ser alterado, a representação mais eficiente dos dados no DW e os benefícios ao nível do desempenho no acesso aos dados [28,36]. A simplicidade do modelo de dados é um parâmetro crítico porque permite não só que o *software* explore os dados e apresente informações de forma rápida e eficiente, mas também que os utilizadores compreendam facilmente os dados [36].

Existem diferentes configurações para a organização das tabelas de factos e de dimensões num modelo dimensional. A maioria dos DWs utiliza o esquema em estrela (Figura 3.3) para representar os dados segundo o modelo dimensional. Nesta configuração o DW consiste numa única tabela de factos e uma única tabela para cada dimensão, sendo por isso a configuração mais elementar de um esquema dimensional [29].

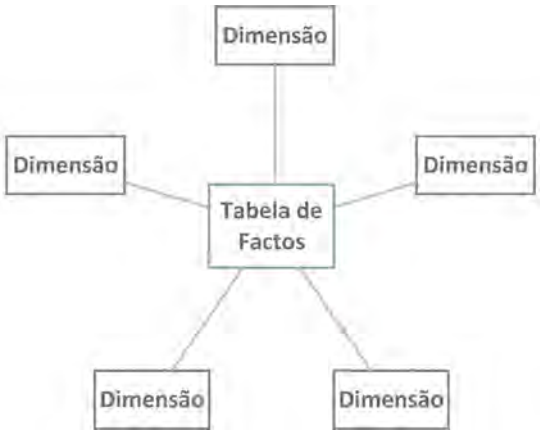

Figura 3.3: Esquema em estrela.

O esquema em estrela não permite representar explicitamente a hierarquia entre os atributos das dimensões. Para essa finalidade utiliza-se um esquema em floco de neve (Figura 3.4), um aperfeiçoamento do esquema em estrela, no qual a hierarquia dimensional é representada através da normalização das dimensões em subdimensões [29]. Este esquema apresenta vantagens ao nível da manutenção das dimensões, mas a estrutura não normalizada do esquema em estrela pode ser mais apropriada para explorar as dimensões quando se pretender conceber um modelo dimensional que proporcione altos níveis de desempenho na execução de *queries* [27].

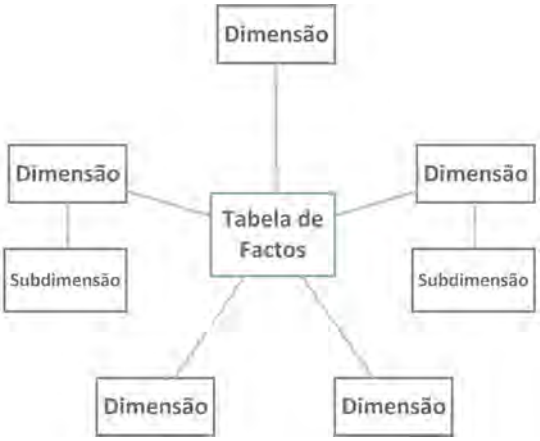

Figura 3.4: Esquema em floco de neve.

Existem ainda esquemas denominados constelações de factos que dizem respeito a estruturas mais complexas, nas quais múltiplas tabelas de factos partilham dimensões [27].

Metodologia de Kimball

A metodologia de Kimball (Figura 3.5) é a metodologia mais conhecida para a implementação de sistemas de *data warehousing* e BI. Esta metodologia indica o fluxo de atividades necessárias para a implementação de um sistema deste tipo, sendo composta por um conjunto de atividades dependentes que ocorrem sequencialmente ou paralelamente.

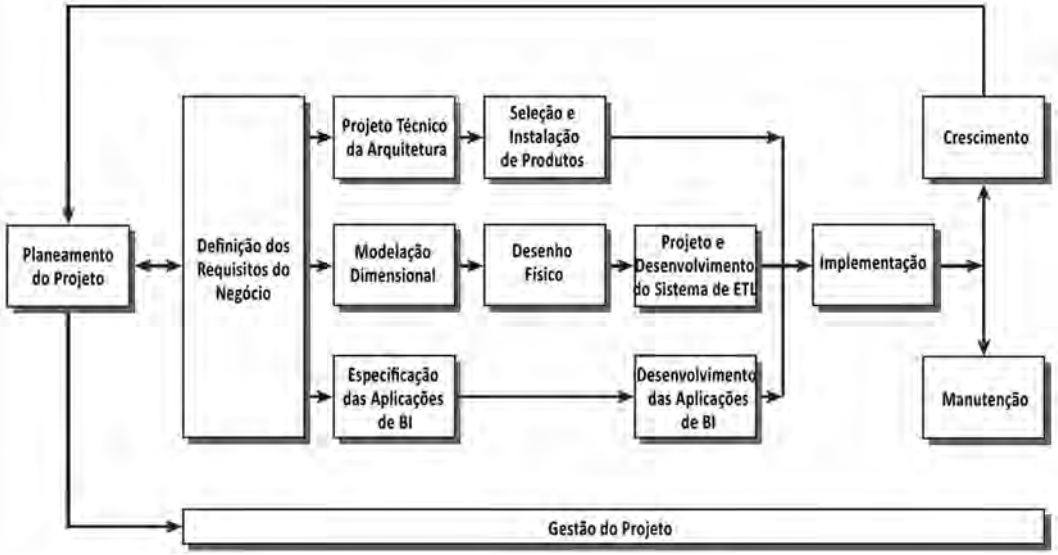

Figura 3.5: Metodologia de Kimball para a implementação de sistemas de *data warehousing* e BI (adaptado de [36]).

Segundo Kimball e Ross [36], um projeto de *data warehousing* e BI inicia-se com *Planeamento do Projeto*, a etapa que define o âmbito do projeto, avalia a capacidade da organização para iniciar o mesmo, faz o planeamento das atividades a realizar, etc. Após esta etapa é iniciada uma outra que será executada durante todo o projeto, a *Gestão do Projeto*, cujo objetivo é gerir todas as atividades. Simultaneamente, começa também a etapa *Definição dos Requisitos do Negócio*, na qual são identificados os requisitos iniciais para o

sistema de *data warehousing* e BI. O sucesso do projeto depende fortemente da compreensão dos requisitos identificados, pois esta é fundamental para a tradução bem-sucedida dos mesmos em especificações de implementação.

Após a *Definição dos Requisitos do Negócio* iniciam-se três linhas de ação distintas que ocorrem paralelamente:

- Linha de ação tecnológica: começa com a atividade *Projeto Técnico da Arquitetura*, etapa na qual é definida a arquitetura do sistema com o intuito de estabelecer os critérios para a seleção das tecnologias/produtos que o irão constituir. Segue-se a *Seleção e Instalação de Produtos* que satisfazem as necessidades da arquitetura do sistema;

- Linha de ação dos dados: inicia-se com a atividade *Modelação Dimensional*, que traduz os requisitos de negócio num modelo dimensional. Segue-se o *Desenho Físico*, isto é, a transformação do modelo dimensional numa estrutura física e, posteriormente, ocorre o *Projeto e Desenvolvimento do Sistema de ETL*;

- Linha de ação de BI: tem início com a *Especificação das Aplicações de BI*, seguindo-se o *Desenvolvimento das Aplicações de BI*. Estas atividades estão relacionadas com o desenho e a implementação de aplicações de BI que serão utilizadas para aceder e explorar os dados do DW.

Por último, realiza-se a *Implementação*, que consiste na integração dos resultados das três linhas de ação. Esta etapa necessita de planeamento para garantir que o sistema funciona corretamente e engloba atividades como a definição da estratégia de formação dos utilizadores do sistema e de suporte aos mesmos. Seguem-se as etapas finais do ciclo, *Manutenção* e *Crescimento*, atividades que permitem, respetivamente, monitorizar o sistema para assegurar que este possui um desempenho ótimo e definir projetos subsequentes que se realizarão no próximo ciclo do projeto.

3.1.3 *On-Line Analytical Processing*

OLAP é uma das tecnologias mais utilizadas para aceder e analisar os dados consolidados num DW ou num *data mart* [31]. Segundo o *OLAP*

Council[1] [37], esta tecnologia permite que os analistas, gestores e executivos adquiram conhecimento presente nos dados através do acesso rápido, consistente e interativo a diferentes perspetivas sobre a informação que foi transformada a partir dos dados brutos. Desse modo, a análise OLAP transforma os dados da organização em informação estratégica. Ao permitir a visualização e a análise da informação segundo diferentes pontos de vista, esta tecnologia disponibiliza um maior poder aos utilizadores ao nível da tomada de decisão.

Os dados a analisar com OLAP encontram-se organizados num cubo OLAP (Figura 3.6), uma estrutura multidimensional formada por factos e dimensões, podendo armazenar valores pré-calculados (tais como contagens ou valores totais), através da agregação dos factos de acordo com a hierarquia das dimensões [36].

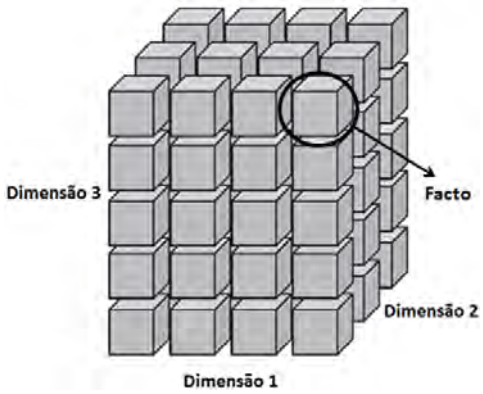

Figura 3.6: Cubo OLAP (adaptado de [36]).

As ferramentas OLAP suportam análises em tempo real, permitindo que o utilizador faça pesquisas mais rápidas e de modo mais estruturado, para gerar mais facilmente gráficos e relatórios [11]. Possibilitam, também, que o utilizador escolha o nível de detalhe a visualizar e as dimensões que pretende examinar [38]. A tecnologia OLAP admite assim a realização de operações como [27, 29, 34]:

- *Drill-down*: operação que permite diminuir o nível de agregação dos

[1] O *OLAP Council* é uma organização cujo objetivo é educar sobre a tecnologia OLAP e promover normas para esta tecnologia.

factos de modo a expor um maior detalhe. Esta pode ser realizada através da exploração da hierarquia de uma determinada dimensão.

- *Roll-up*: operação que possibilita o aumento do nível de agregação dos factos de forma a expor um menor detalhe. Esta pode ser realizada explorando a hierarquia dos atributos de uma determinada dimensão.
- *Slice and dice*: operação que cria um cubo mais específico, através da redução da dimensionalidade do mesmo. *Slice* realiza uma seleção numa única dimensão, ao passo que *dice* efetua uma seleção em duas ou mais dimensões.
- *Pivot* ou *rotate*: operação que altera a perspetiva de análise dos factos através da rotação do eixo de visualização do cubo.

As figuras 3.7, 3.8 e 3.9 ilustram as operações anteriormente descritas apresentando um exemplo sobre a contabilização do total de vendas de um determinado produto a um determinado cliente, num certo momento no tempo. Neste caso, o *drill-down* é efetuado na dimensão associada ao tempo, permitindo assim a evolução da análise do número total de vendas do nível mensal para o nível diário (Figura 3.7). A operação *roll-up* é responsável pelo processo inverso (Figura 3.7).

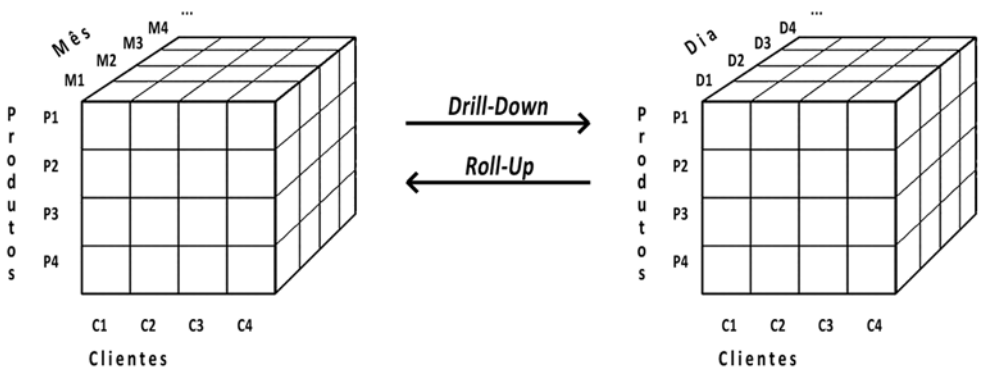

Figura 3.7: Exemplo das operações *drill-down* e *roll-up*.

Neste exemplo, a operação *slice and dice* (Figura 3.8) realiza-se através da seleção do mês *M1* (*slice*) ou com a seleção dos meses *M1* e *M2* e, simultaneamente, dos clientes *C1* e *C2* (*dice*).

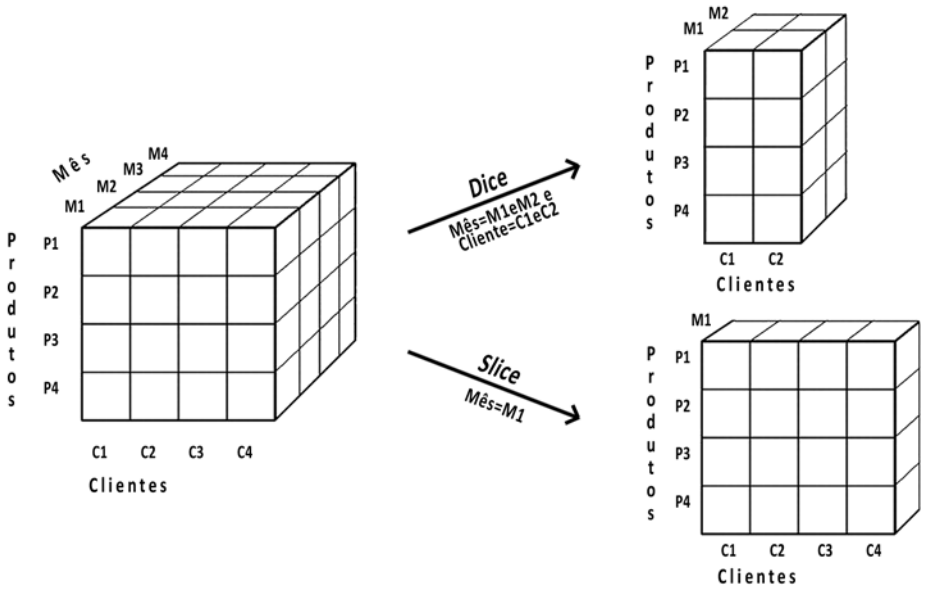

Figura 3.8: Exemplo da operação *slice and dice*.

A operação *pivot* realiza-se através da rotação dos eixos (Figura 3.9).

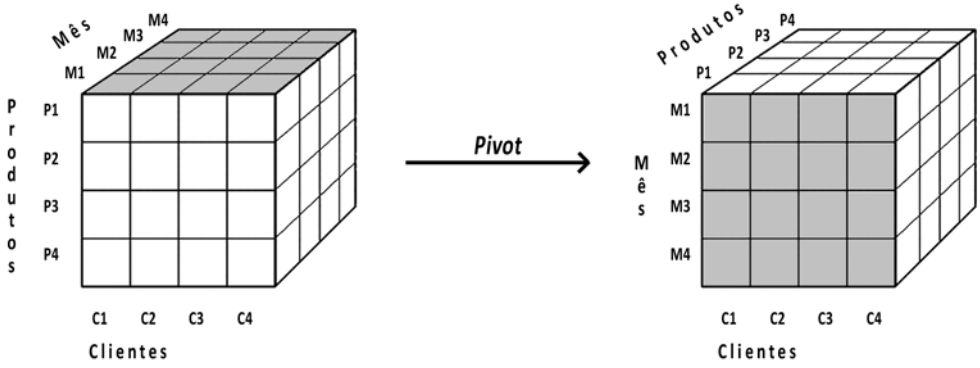

Figura 3.9: Exemplo da operação *pivot*.

Os servidores OLAP armazenam os dados nos cubos e atuam como intermediários no acesso aos mesmos por parte das ferramentas OLAP. Dependendo da tecnologia utilizada para realizar OLAP, estes servidores podem ser classificados como [27, 29, 35]:

- *Multidimensional On-Line Analytical Processing* (MOLAP)
  Estes servidores suportam a análise OLAP através do armazenamento

direto dos dados em *arrays* multidimensionais, conhecidos por cubos OLAP. O processamento, isto é, a pré-computação e armazenamento de dados, e as consultas são implementados diretamente nestes servidores. Como todo o processamento é efetuado no momento de construção do cubo, estes sistemas apresentam como principal vantagem o reduzido tempo de resposta a consultas e têm como aspeto negativo as limitações na quantidade de dados que podem ser manipulados, pois não é possível incluir grandes volumes de dados no cubo.

- *Relational On-Line Analytical Processing* (ROLAP)
  Estes servidores atuam como intermediários entre um DW implementado numa base de dados relacional e as ferramentas analíticas utilizadas para analisar os dados. Neste caso, as *queries* são colocadas diretamente ao DW, no momento da consulta. Esta arquitetura permite a manipulação de grandes volumes de dados, pois possui a escalabilidade de um sistema relacional. O seu desempenho, porém, pode ser lento, pois o tempo de consulta depende do volume de dados presentes na base de dados relacional e, além disso, a consulta à base de dados relacional é limitada pelas funcionalidades da linguagem de consulta *Structured Query Language* (SQL).

- *Hibrid On-Line Analytical Processing* (HOLAP)
  Estes servidores combinam características dos sistemas ROLAP e MOLAP, beneficiando do elevada velocidade no processamento dos sistemas MOLAP e da grande escalabilidade dos sistemas ROLAP.

3.1.4 *Data Mining*

O termo DM refere-se ao processo responsável por encontrar padrões e tendências, anteriormente desconhecidos, em grandes volumes de dados complexos [38,39]. Com esta tecnologia, o auxílio da tomada de decisão efetua-se através da descoberta desses padrões e tendências que de forma manual seriam muito difíceis ou até mesmo impossíveis de encontrar [39].

As ferramentas de DM identificam esses padrões válidos, compreensíveis,

potencialmente úteis e novos utilizando métodos automáticos [39]. Uma ferramenta de DM permite fazer uma exploração profunda e detalhada dos dados, bem como construir modelos de previsão capazes de ajudar a responder a questões específicas. Esta análise vai para além da disponibilizada pelas ferramentas OLAP [29]. Se por um lado, a tecnologia OLAP requer interação humana para a descoberta de relacionamentos entre os dados, a tecnologia de DM, por sua vez, é capaz de encontrar muitas dessas relações automaticamente.

De acordo com Fayyad *et al.* [40], o processo de DM é apenas uma etapa do processo de Descoberta de Conhecimento em Bases de Dados (DCBD) (Figura 3.10), sendo que este último se refere à descoberta de conhecimento útil a partir dos dados, enquanto o primeiro se refere apenas à aplicação de algoritmos capazes de extrair padrões dos dados. As restantes quatro etapas tradicionais do processo DCBD, como a seleção dos dados, o seu pré-processamento, a sua transformação, a interpretação e avaliação dos resultados da etapa de DM, são essenciais para assegurar a extração de conhecimento útil. Deste modo, na etapa de seleção dos dados (1) são escolhidos os mais úteis para resolver o problema em análise. Posteriormente, ocorre a etapa de pré-processamento (2), responsável pela realização de procedimentos de limpeza para obter dados consistentes. Por sua vez, a etapa de transformação (3) refere-se à manipulação dos dados de modo a torná-los mais adequados aos algoritmos de DM a aplicar seguidamente (4). A última etapa (5) consiste na interpretação e avaliação dos resultados obtidos com a etapa de DM, bem como na sua aplicação na tomada de decisão [40, 41]. Nesta última etapa pode ser necessário regressar a qualquer uma das anteriores [40]. A qualidade dos resultados de DM depende da qualidade dos dados nos quais os resultados se baseiam [39]. Assim, a qualidade dos dados utilizados interfere diretamente na qualidade do conhecimento obtido, sendo que este é um fator preponderante na tomada de boas decisões.

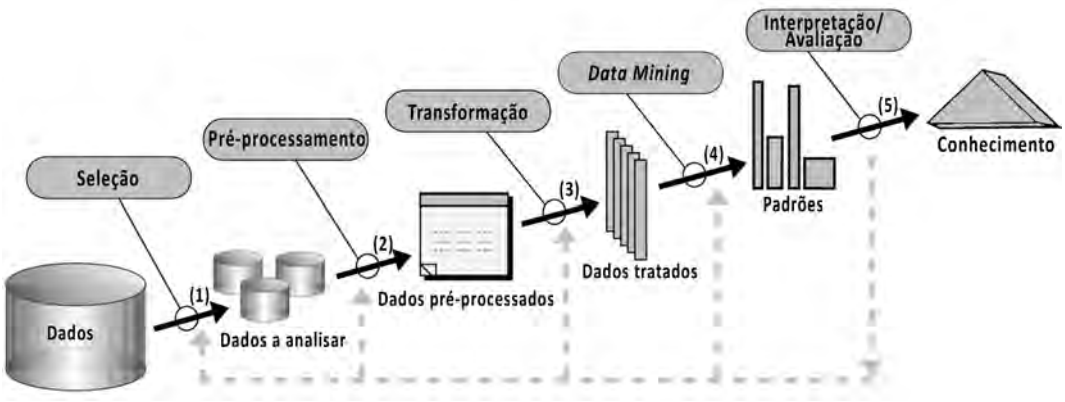

Figura 3.10: Processo de DCBD (adaptado de [40]).

Segundo Fayyad *et al.* [40] os objetivos do processo de DCBD podem ser a verificação de uma hipótese, a previsão de variáveis futuras ou a descrição dos dados através da descoberta de padrões nestes. Estes objetivos podem ser alcançados através da aplicação de diferentes métodos de DM, tais como classificação de dados, regressão linear, segmentação de dados em conjuntos, associação entre variáveis, entre outros [28, 40]. Por sua vez, cada um destes métodos pode ser implementado utilizando diferentes técnicas ou algoritmos, tais como árvores de decisão, raciocínio baseado em casos, algoritmos genéticos, redes neuronais artificiais, algoritmos de regressão, etc [29, 40, 42]. Os algoritmos a aplicar variam de acordo com o objetivo e o método de DM pretendidos, sendo que não existe uma técnica de DM universal e ideal, uma vez que cada uma é adequada para determinado tipo de problemas [40].

Em suma, a abordagem do processo de DM consiste em: selecionar um subconjunto de dados do DW ou do *data mart*, um *dataset*; efetuar análises complexas nos dados selecionados através da aplicação de algoritmos; identificar aspetos estatísticos importantes [29].

Existem várias metodologias que permitem a implementação do processo de DCBD, sendo que a mais utilizada é a metodologia *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) (Figura 3.11).

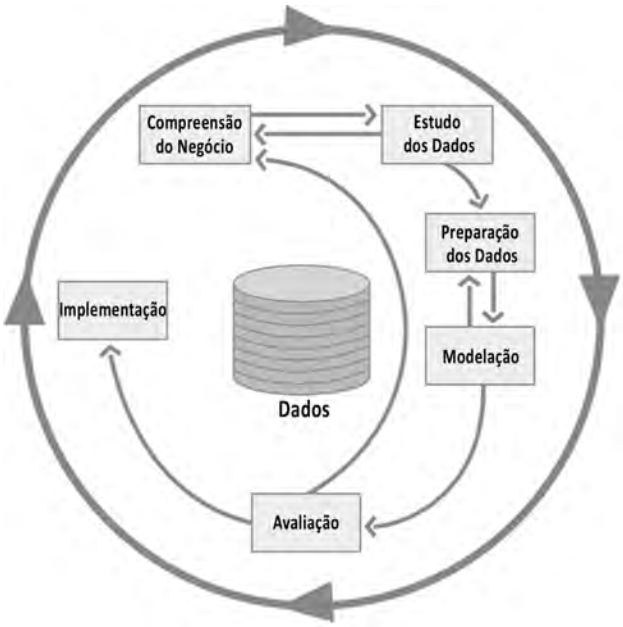

Figura 3.11: Metodologia CRISP-DM (adaptado de [43]).

A metodologia CRISP-DM consiste num ciclo composto pelas seguintes etapas [19, 41, 43]:

- *Compreensão do Negócio:* compreensão dos objetivos e necessidades do projeto, segundo uma perspetiva de negócio; conversão desse conhecimento num problema de DM e num plano preliminar para o projeto;
- *Estudo dos Dados:* recolha dos dados com interesse para o estudo, avaliação da qualidade desses dados e identificação de eventuais problemas nos mesmos;
- *Preparação dos Dados:* todas as atividades necessárias para criar o *dataset* final, incluindo operações como seleção, transformação e limpeza dos dados, podendo ser necessário repetir esta etapa várias vezes até os dados apresentarem qualidade;
- *Modelação:* seleção e aplicação de técnicas de modelação, de acordo com os objetivos de DM definidos, sendo que no final desta etapa pode ser necessário regressar à etapa anterior com o intuito de adequar os dados às técnicas de DM;

- *Avaliação:* os modelos gerados são avaliados a fim de verificar se cumprem ou não os objetivos de negócio definidos na primeira etapa do ciclo e, mediante os resultados dessa avaliação, escolhe-se a etapa a realizar em seguida (*Compreensão do Negócio* ou *Implementação*);

- *Implementação:* realização do relatório final e aplicação dos resultados obtidos com o estudo. O conhecimento extraído deve ser organizado e disponibilizado ao utilizador para que este possa beneficiar dele.

A sequência destas etapas não é rígida, uma vez que frequentemente é necessário regressar a fases anteriores, pois o resultado de cada etapa determina as ações que é necessário implementar na etapa seguinte. Além disso, o processo de DM é cíclico, não terminando com a última fase descrita, visto que o conhecimento obtido durante o processo CRISP-DM pode dar origem a novas questões de negócio, frequentemente mais focadas, que impliquem o recomeço do ciclo [43].

3.1.5 Consultas e Relatórios

Para obter informações existentes nos dados, a tecnologia de BI permite a consulta de dados presentes no DW ou em *data marts*, através da utilização de uma linguagem de consulta, como por exemplo SQL. Existem muitas ferramentas de consulta e de criação de relatórios que disponibilizam uma interface gráfica que disfarça os detalhes técnicos relacionados com o acesso aos dados e extração da informação, utilizando, por exemplo, menus e botões para especificar os parâmetros de seleção. Estas ferramentas podem ser aplicadas para realizar um conjunto de *queries* pré-definidas ou *queries ad hoc* (*queries* específicas, criadas para responder às necessidades de um contexto de tomada de decisão particular), permitindo gerar relatórios com os resultados obtidos [28]. Por sua vez, as ferramentas para a produção de relatórios criam eficientemente relatórios e, normalmente, apresentam informação quantitativa sob a forma de números, de gráficos ou de tabelas [38].

3.2 *Business Intelligence* na Saúde

Foshay e Kuziemsky [17] afirmam que a implementação bem-sucedida de um sistema de BI requer "elevada qualidade dos dados, o acesso apropriado dos utilizadores e uma integração efetiva com outros sistemas". Para a obtenção de bons resultados com a implementação de ferramentas de BI e para que o sistema tenha sucesso, é necessário que o utilizador possa confiar nos resultados deste e aplicá-los na sua tomada de decisão. Deste modo, é necessário que o sistema seja capaz de apresentar informações de elevada qualidade, extraídas a partir de dados de qualidade, relevantes para a análise que se pretende realizar. Na área da saúde é, por isso, fundamental convencer os profissionais de saúde de que os sistemas de BI efetivamente apresentam informações credíveis e podem auxiliá-los no seu trabalho diário. Portanto, é extremamente importante considerar e tratar problemas relacionados com a qualidade dos dados, pois este processo contribui para uma maior qualidade da informação apresentada pelo sistema [17]. Assim, os dados com qualidade possuem um elevado valor para a organização [23]. De facto, quanto maior a sua qualidade, mais utilidade apresentam para serem utilizados no suporte à decisão.

Na área da saúde, os dados recolhidos apresentam frequentemente problemas como valores em falta, inconsistências, falta de um formato comum e de um vocabulário padrão, e provêm de diferentes fontes heterogéneas, de grandes dimensões e complexas [39]. Por outro lado, os dados recolhidos não são apenas quantitativos. De facto, são maioritariamente qualitativos, isto é, dados não estruturados e à base de texto, pois os processos associados aos cuidados de saúde ainda não estão suficientemente padronizados [19]. Portanto, é necessário detetar estes problemas de qualidade dos dados e corrigi-los antes dos mesmos serem carregados para o DW, recorrendo-se, para isso, ao processo ETL anteriormente descrito (Secção 3.1.1).

3.2.1 Vantagens da sua Implementação

A implementação de sistemas baseados na tecnologia de BI apresenta-se como um método eficiente e adequado para integrar e explorar os dados

clínicos recolhidos pelas instituições de saúde, pois, com a aplicação desta tecnologia, os dados não são apenas recolhidos e armazenados em bases de dados, mas são também utilizados para suporte à decisão. Estes sistemas tratam os dados, analisam-nos e extraem informações dos mesmos, sendo que essas informações podem ser muito relevantes para a identificação, análise e monitorização das atividades e processos que decorrem na organização. Deste modo, torna-se possível a descoberta de problemas e oportunidades de melhoria. De acordo com Mettler e Vimarlund [15], a utilização de sistemas de BI na área da saúde disponibiliza novas formas de trabalhar, permitindo medir resultados em tempo real e integrar informação e organizações.

A implementação de BI nas instituições de saúde pode ajudar assim na melhoria da qualidade e da segurança dos cuidados de saúde prestados, na melhoria da eficiência e do desempenho financeiro da organização, na implementação de práticas baseadas em evidências, bem como na utilização eficiente de recursos. Estas melhorias podem ocorrer, pois esta tecnologia auxilia os gestores e os profissionais de saúde a tomar melhores decisões, através da análise de dados que fornecem informações relevantes acerca das atividades e processos que decorrem na unidade de saúde [11,17]. Além disso, estas unidades estão associadas a um ambiente muito complexo e em constante mudança, pelo que a utilização de ferramentas de BI é fundamental para auxiliar o processo de tomada de decisão.

Tal como já foi referido, o processo de tomada de decisão na área da saúde é muito complexo e requer informação de elevada qualidade. Segundo Popovič *et al.* [25], a implementação de sistemas de BI pode contribuir para melhorar a qualidade da informação utilizada pela organização através do acesso mais rápido à mesma, maior facilidade em consultar e em explorar dados, e melhoria da consistência dos dados como resultado dos processos de integração dos mesmos, como por exemplo a sua limpeza, realizados antes do seu armazenamento no DW. De acordo com Foshay e Kuziemsky [17], a qualidade da informação apresentada pelo sistema de BI é assim uma consequência de uma implementação bem-sucedida do mesmo.

Na opinião destes autores, apesar dos potenciais benefícios da implementação de sistemas de BI nas instituições de saúde, existe ainda pouco conhe-

cimento sobre os fatores que realmente contribuem para a implementação bem-sucedida destes sistemas. A maioria dos estudos sobre a aplicação de BI na área da saúde foca-se apenas nas ferramentas que podem ser utilizadas para implementar estes sistemas ou nos resultados que se pretende obter com os mesmos, isto é, a melhoria do processo de tomada de decisão [17].

3.2.2 Trabalhos Relacionados

Atualmente a necessidade de extrair informação dos dados tem vindo a aumentar, sendo que estes provêm dos diferentes SIH utilizados pela instituição de saúde. Na maioria das unidades de saúde, os dados são armazenados em sistemas distintos e, por vezes, é necessário correlacionar os dados relevantes de um sistema com outros de outros sistemas. Geralmente, estes sistemas estão pouco integrados, o que implica a existência de diferentes formatos de dados e de diferentes mecanismos de acesso aos mesmos para cada um deles [20]. Esta complexidade dificulta a obtenção de informação presente nos dados, existindo, por isso, interesse em desenvolver aplicações que simplifiquem o acesso aos dados clínicos e a extração de informação destes, para que a mesma possa ser fácil e rapidamente aplicada na tomada de decisão. As ferramentas de BI são capazes de trabalhar eficientemente com estes dados [23]. Por estes motivos, a introdução deste tipo de tecnologia no ambiente hospitalar tem adquirido muita popularidade.

A tecnologia de BI tem sido utilizada para desenvolver portais de pesquisa de informação. Horvath *et al.* [44] desenvolveram um portal de pesquisa que utiliza ferramentas de BI para gerar relatórios com os resultados das pesquisas. Recorrendo a este portal, os investigadores e os especialistas responsáveis pela melhoria da qualidade da organização podem obter informação clínica presente no DW, sem terem de conhecer e compreender a linguagem estruturada utilizada pelas *queries* ou o modelo da base de dados subjacente ao sistema. O portal possui uma interface gráfica que permite que a criação de *queries* seja um processo simples e que estas sejam executadas eficientemente. Posteriormente, a informação filtrada através das *queries* SQL colocadas ao sistema é apresentada sob a forma de relatórios com gráficos e

tabelas. A principal vantagem da utilização de uma aplicação como o portal anteriormente descrito é este permitir que os seus utilizadores possam aceder diretamente aos dados armazenados, atribuindo-lhes, portanto, uma maior autonomia na extração de informação a partir dos mesmos.

Segundo Bonney [24], muitos SIH, como o PCE, possuem grande quantidade de informação de elevada relevância para o processo de tomada de decisão clínica. Este autor afirma que a aplicação de BI ao conteúdo do PCE é imprescindível para assegurar que a informação existente nos dados clínicos é eficientemente aproveitada, pois esta tecnologia permite extrair informação relevante e de qualidade a partir dos dados armazenados. Essa informação pode ser utilizada pelos profissionais de saúde para auxiliar a sua tomada de decisão em tempo real, possibilitando uma tomada de decisão baseada em evidências e contribuindo, consequentemente, para a melhoria dos resultados obtidos. Além disso, a aplicação de BI ao PCE permite explorar grandes volumes de dados que podem ser utilizados para estudos epidemiológicos e outras investigações médicas de carácter estatístico [24]. Apesar de existirem muitos desafios à implementação da tecnologia BI no PCE na prática clínica, as plataformas de BI são cada vez mais utilizadas e sofisticadas [24].

As ferramentas de BI também têm sido aplicadas à área da Imagiologia com o intuito de gerar *Key Performance Indicators*(KPIs) para avaliar o desempenho de departamentos de Imagiologia [11, 45]. Os administradores e os gestores de um departamento destes necessitam de dados provenientes de diferentes sistemas para o apoio à decisão e análise de tendências [11]. Esta necessidade não é apenas para avaliar a eficiência e a performance financeira do departamento, mas também para monitorizar a qualidade e a segurança dos serviços prestados, e obter um conhecimento mais detalhado de todos os fatores que estão envolvidos num determinado processo ou atividade.

De modo a avaliar o desempenho de uma unidade de saúde, nomeadamente um departamento de Imagiologia, é necessário definir medidas objetivas denominadas KPIs [11, 20]. No entanto, como os dados se encontram armazenados em diferentes sistemas, esta tarefa não é facilmente concretizável [11]. A tecnologia de BI pode ser utilizada para fazer a integração e a consolidação de dados em KPIs, que são essenciais para apoiar a tomada de

decisões [22]. Assim, estes indicadores são gerados pelas ferramentas de BI a partir dos dados da organização, e podem ser utilizados como uma vantagem competitiva porque permitem medir o desempenho dos processos que ocorrem na organização. As análises realizadas pelas ferramentas de BI podem ser efetuadas sistemática e regularmente, permitindo, neste caso, fazer a monitorização dos KPIs ao longo do tempo, ou podem ser *ad hoc*, isto é, relacionadas com um contexto de tomada de decisão específico.

Os responsáveis pela tomada de decisão em vários níveis organizacionais aplicam o conhecimento retirado da análise dos KPIs nas suas decisões. Esse conhecimento pode ser utilizado para ajustar ou alterar o comportamento atual da organização e para auxiliá-la a atingir os seus objetivos [11]. Esta operação de monitorização do funcionamento das unidades de saúde é muito importante porque permite uma melhoria contínua do desempenho das mesmas, contribuindo para melhorar a qualidade e a segurança dos cuidados prestados. É importante salientar que a qualidade dos indicadores gerados depende muito da qualidade dos dados nos quais se baseiam.

Prevedello *et al.* [11] utilizaram ferramentas *open-source* para construir um DW para a análise de KPIs, de modo a possibilitar a integração de dados provenientes de todos os SI do departamento de Imagiologia e a visualização de informação presente nos mesmos. O método proposto demonstrou ser particularmente útil quando aplicado a casos em que os dados estão constantemente a ser gerados e os relatórios precisam de ser regularmente criados com base em dados atualizados.

Nagy *et al.* [45] apresentaram um sistema automatizado para a extração, processamento e visualização de KPIs utilizados na identificação de problemas e oportunidades para melhorar a performance de um departamento de Imagiologia. Tal como no caso anteriormente apresentado, os dados necessários são extraídos dos diferentes SI utilizados e armazenados num DW. Os indicadores são gerados num ambiente *web* dinâmico, a partir dos dados desse DW. Desse modo, é possível facilitar a visualização e a análise dos KPIs. Os resultados obtidos ao longo de 24 meses após a implementação deste sistema de BI sugerem que este permitiu obter informações significativas para melhorar a eficácia da gestão do departamento.

Capítulo 4

Materiais e Métodos

4.1 Metodologia de Investigação

Este trabalho iniciou-se com uma revisão bibliográfica acerca de várias áreas de investigação: TI na área da saúde, apoio à decisão na saúde, sistemas de BI e aplicação da tecnologia de BI à área da saúde. Foi também realizada uma pesquisa sobre ferramentas *open-source* de BI.

Além disso, no desenvolvimento do projeto optou-se por uma metodologia de investigação *action research*. Esta metodologia caracteriza-se pela pesquisa orientada à resolução progressiva de um problema. Esta é assim um processo cíclico que inclui: a identificação de um problema; o planeamento e realização de ações para o resolver; e a reflexão sobre os efeitos dessas ações. Assim, o investigador vai aprendendo, aplicando os conhecimentos que obtém na resolução dos problemas. Este processo ocorre repetidamente, até que os problemas estejam resolvidos [46].

No caso do trabalho apresentado nesta dissertação, o problema identificado foi a necessidade de facilitar o tratamento dos dados relativos a infeções nosocomiais e a sua utilização para gerar informações relevantes sobre as mesmas. Posteriormente, foi planeado e desenvolvido um sistema baseado na tecnologia de BI para simplificar essas mesmas ações. Este sistema foi testado e melhorado ao longo do trabalho, até permitir verificar que, de facto, a sua implementação pode beneficiar o estudo deste tipo de infeções.

4.2 Armazenamento e Manipulação de Dados

4.2.1 Base de Dados *Oracle*

Para armazenar e manipular todos os dados necessários para a realização deste trabalho optou-se por uma base de dados *Oracle*. O sistema de gestão de bases de dados *Oracle* é um dos sistemas de gestão de base de dados mais utilizados atualmente, é altamente escalável e robusto e disponibiliza confiabilidade e segurança dos dados. Este possui um bom desempenho mesmo na presença de grandes quantidades de dados. Além disso, no CHP também são utilizadas bases de dados *Oracle*.

Para a manipulação dos dados armazenados na base de dados *Oracle* recorreu-se à ferramenta *Oracle SQL Developer*. Esta ferramenta é uma plataforma de desenvolvimento e administração de bases de dados gráfica e gratuita que simplifica a criação e a manutenção de bases de dados. Esta permite, entre outras funcionalidades: a execução de *queries* e *scripts* em SQL; o desenvolvimento e *debugging* de aplicações *Procedural Language/Structured Query Language* (PL/SQL); a manipulação e a exportação de dados [47].

4.2.2 Modelação Dimensional

Oracle SQL Developer Data Modeler, ou simplesmente *Data Modeler*, é uma ferramenta de modelação e *design* de bases de dados que é uma extensão da ferramenta *Oracle SQL Developer* e disponibiliza um ambiente para a captura, modelação, gestão e exploração de metadados [47]. Neste trabalho, utilizou-se esta ferramenta para desenhar o modelo dimensional a implementar no DW, ou seja, para definir as características das tabelas de factos e de dimensões, bem como os relacionamentos entre as mesmas.

4.2.3 *Data Mining*

Para a criação dos modelos de classificação de DM e aplicação dos mesmos a dados recorreu-se à ferramenta *Oracle Data Miner*. Esta ferramenta é uma extensão da ferramenta *Oracle SQL Developer* que permite trabalhar

diretamente com os dados da base de dados, explorá-los graficamente, construir e avaliar modelos de DM, bem como aplicar os modelos criados a novos dados [48].

4.3 Ferramentas *Open-source* de *Business Intelligence*

No processo de implementação de um sistema de BI é sempre necessário selecionar as ferramentas a utilizar de acordo com as necessidades do sistema e os resultados que se pretende obter. Atualmente, existem no mercado várias ferramentas de BI que podem ser utilizadas em instituições de saúde para assistir os profissionais de saúde na sua tomada de decisões. Para isso, estas ferramentas realizam o processamento de dados e a apresentação de informações relevantes extraídas destes. Porém, as instituições de saúde atuam sob pressão financeira pelo que a implementação de novo *software* deve ter isso em consideração, de modo a não sobrecarregar financeiramente estas instituições. Torna-se então importante encontrar soluções que possibilitem diminuir os custos de desenvolvimento, como é o caso do *software* *open-source*. Este tipo de *software* permite ainda que o seu código-fonte seja modificado consoante as necessidades [49, 50].

Neste trabalho é necessário utilizar uma ferramenta de BI para realizar a extração de informações dos dados do DW e criar uma plataforma *web* onde estas informações possam ser apresentadas em *dashboards* interativos. Pretende-se ainda utilizar uma ferramenta sem qualquer custo. Com o intuito de selecionar a ferramenta *open-source* de BI mais adequada a este projeto compararam-se algumas ferramentas, que serão brevemente apresentadas em seguida.

SpagoBI[1]

SpagoBI é uma ferramenta *open-source* de BI desenvolvida pela *SpagoWorld*. Ao contrário do que acontece com outras ferramentas de BI, esta

[1] http://www.spagoworld.org/xwiki/bin/view/SpagoBI/

ferramenta possui apenas uma versão totalmente gratuita. Este *software* oferece várias funcionalidades analíticas, destacando-se ferramentas para: a produção e a exportação de relatórios; análises OLAP; a criação de gráficos; a criação de *dashboards* interativos; a produção de relatórios *ad hoc*; DM; ETL; etc [51, 52]. Todas estas funcionalidades fazem de *SpagoBI* uma ferramenta completa, robusta e com um elevado potencial sem, no entanto, implicar qualquer custo para o utilizador [52].

Pentaho Community Edition[2]

Pentaho Community Edition é uma ferramenta *open-source* de BI criada pela *Pentaho Corporation*. Este *software* é uma das ferramentas de BI mais poderosas e é totalmente gratuito. *Pentaho Community Edition* permite a criação de relatórios e *dashboards*, a aplicação de técnicas de DM e de OLAP, a exportação de dados e oferece uma gama completa de ferramentas cada vez mais avançadas, que auxiliam os utilizadores a visualizar e a analisar dados [53]. Existe também uma versão paga desta ferramenta que contém mais funcionalidades (*Pentaho Enterprise Edition*) [52, 53].

Jaspersoft Community Edition[3]

Jaspersoft Community Edition é uma ferramenta *open-source* de BI da *TIBCO Software*. Esta é constituída por diversos produtos individuais que permitem, entre outras funcionalidades: a criação e a publicação de relatórios sofisticados com, por exemplo, gráficos, tabelas e imagens; o acesso a dados provenientes de qualquer fonte de dados; a implementação de ETL; a realização de OLAP [54]. Esta ferramenta é robusta, confiável e é também bastante completa, porém não possui uma funcionalidade que permita fazer DM nem possui um componente para a criação de *dashboards* [53]. Tal como acontece com a ferramenta *Pentaho*, existe igualmente uma versão paga e mais completa (*Jaspersoft Enterprise Edition*) [52, 53].

[2] http://community.pentaho.com
[3] http://community.jaspersoft.com

4.3.1 Comparação das Ferramentas

Todas as ferramentas de BI anteriormente apresentadas são muito semelhantes em termos das funcionalidades que disponibilizam. As ferramentas *Pentaho Community Edition* e *SpagoBI* são as mais completas. De facto, a ferramenta *SpagoBI*, apesar de estar disponível apenas numa versão gratuita, é mais completa do que algumas versões *Enterprise* de outras ferramentas [52].

Após uma análise, instalação e experiências realizadas com algumas destas ferramentas, verificou-se que a ferramenta *SpagoBI*, apesar de ser muito completa e totalmente gratuita, é muito difícil de instalar. Constatou-se também que a ferramenta *Pentaho Community Edition* não é tão completa como a anterior, porém é fácil de instalar e de utilizar e permite realizar com sucesso muitas das tarefas de BI que atualmente uma organização necessita. A ferramenta *Jaspersoft Community Edition* é também completa, mas, ao contrário das anteriores, não permite a criação de *dashboards*, um critério importante para selecionar a ferramenta de BI a utilizar neste projeto. Assim, após esta comparação e análise, optou-se por selecionar a ferramenta *Pentaho Community Edition*, pois as suas funcionalidades são capazes de satisfazer todas as necessidades deste projeto.

4.3.2 *Pentaho Community Edition*

A ferramenta *Pentaho Community Edition* é constituída por vários módulos, dos quais se destacam: um servidor *web* que é o componente principal (*Business Analytics Platform*), um módulo para a implementação de ETL (*Kettle* ou *Pentaho Data Integration*), uma ferramenta para a criação de relatórios (*Pentaho Reporting*) e uma ferramenta de DM (*Weka*).

Uma das grandes vantagens da ferramenta *Pentaho Community Edition*, pelo facto de ser *open-source*, é a possibilidade de estender as suas funcionalidades, nomeadamente através do *download* e instalação de novos *plug-ins* e componentes no módulo *Business Analytics Platform*. O utilizador também pode contribuir com novos *plug-ins* para a comunidade de utilizadores desta ferramenta.

Neste trabalho recorreu-se apenas ao módulo *Business Analytics Platform* e, em seguida, apresentam-se os componentes deste módulo e *plugin-ins* que foram utilizados.

Community Dashboard Editor

Community Dashboard Editor (CDE) é um *plug-in* integrado na *Business Analytics Platform*, que foi desenvolvido para simplificar a criação, a edição e a apresentação de *dashboards*. É uma ferramenta muito poderosa e completa que combina uma interface gráfica com fontes de dados e componentes personalizados [55]. CDE permite criar rápida e facilmente *dashboards* complexos, dinâmicos e visualmente apelativos, que simplificam a análise das informações que apresentam. Com esta ferramenta é possível criar o *layout* do *dashboard* através da combinação de recursos como linhas, colunas, espaços e elementos *HyperText Markup Language* (HTML) como texto ou imagens. Os diferentes componentes que constituem o *dashboard*, tais como tabelas, gráficos, parâmetros ou caixas de texto, também podem ser facilmente criados e personalizados e é ainda possível definir as diferentes fontes de dados que serão utilizadas pelos componentes do *dashboard*, tais como ficheiros *eXtended Markup Language* (XML) ou *queries* SQL [55].

Mondrian

Para implementar e operar em cubos OLAP, *Mondrian* é um servidor OLAP *open-source* desenvolvido em Java e integrado na *Business Analytics Platform*. Este servidor encontra-se implementado segundo uma arquitetura ROLAP, podendo, por isso, lidar com grandes volumes de dados armazenados em bases de dados relacionais [56]. Este processa ainda *queries Multidimensional Expressions* (MDX), uma linguagem de consulta para bases de dados OLAP [56, 57]. Em suma, este servidor executa *queries* MDX definidas por ferramentas OLAP através da leitura de dados provenientes de bases de dados relacionais, apresentando os resultados dessas consultas num formato multidimensional. Essas *queries* são realizadas sobre cubos OLAP cuja estrutura foi previamente definida. A estrutura desses cubos pode ser

definida, por exemplo, utilizando o componente *Data Source Model Editor* da *Business Analytics Platform*. Este componente possibilita a configuração do modelo dimensional de um cubo através da escolha de um conjunto de dimensões e factos, bem como a especificação da forma como os atributos dessas dimensões se relacionam através da criação de relações hierárquicas [58].

Plug-in OpenI

O *OpenI* é um *plug-in* que pode ser instalado no *Pentaho Community Edition*, possibilitando a criação relatórios OLAP [59]. Este *plug-in* apresenta uma interface muito simples e fácil de utilizar, permitindo explorar detalhadamente os dados presentes em cubos OLAP, através da escolha dos factos a visualizar e das dimensões a considerar na análise. Este *software* possibilita também a realização de operações como *slice and dice*, *pivot*, *drill-down* e *roll-up* segundo as diferentes dimensões do cubo OLAP, possibilitando ao utilizador: visualizar os resultados sob a forma de gráficos ou tabelas dinâmicas; exportar os resultados para PDF ou folha de cálculo do Excel; criar as suas próprias *queries* MDX sobre o cubo OLAP ou simplesmente utilizar o mecanismo *drag-and-drop* de factos e dimensões para criar essas *queries*. Este *software* pode ainda comunicar com servidores *Mondrian* [59]. Neste trabalho, optou-se por utilizar esta ferramenta por ser facilmente integrável no *Pentaho* e por realizar OLAP de um modo simples e bastante intuitivo.

Capítulo 5

Sistema de *Business Intelligence* para o Estudo de Infeção Nosocomial

O objetivo primordial deste trabalho consiste no desenvolvimento de uma plataforma de BI que permita o estudo da incidência de infeção nosocomial entre doentes internados nas Unidades de Medicina do CHP. A plataforma apresenta um conjunto de indicadores clínicos relevantes para o estudo da incidência de infeção nosocomial. Estes são informações obtidas a partir dos dados do CHP, capazes de auxiliar neste estudo através da identificação de fatores de risco e de parâmetros importantes para caracterizar a incidência de infeção nosocomial nas Unidades de Medicina.

A motivação para o desenvolvimento da plataforma advém da necessidade de facilitar a extração de informações importantes dos dados relativos a infeções nosocomiais, bem como a sua interpretação. Deste modo, pretende-se auxiliar o trabalho dos profissionais de saúde do CHP responsáveis pelo estudo destas infeções e pela realização de ações neste âmbito, nomeadamente os profissionais da CCI desta instituição (incumbidos do planeamento e promoção de ações relacionadas com a deteção, a prevenção e o controlo de infeções). Através da plataforma, estes podem monitorizar e compreender melhor as infeções nosocomiais. Assim, são capazes de tomar decisões mais

fundamentadas, bem como definir medidas de controlo e prevenção de infeção específicas e mais orientadas para as necessidades reais das Unidades de Medicina.

Com a plataforma de BI o processo de obtenção de informações relevantes é automatizado e otimizado, permitindo, desse modo, que a informação esteja sempre disponível no momento de decisão. Esta plataforma dá também utilidade ao grande volume de dados recolhidos no CHP, permitindo a apresentação de informação presente nestes e, consequentemente, a criação de conhecimento útil a partir desta.

Assim, a plataforma pode beneficiar o estudo da infeção nosocomial na medida em que permite:

- maior apoio na tomada de decisões, através da organização de informação dispersa e da disponibilização de informações relevantes;

- analisar e monitorizar a incidência de infeções nosocomiais, tornando possível a identificação de processos e atividades com grande impacto na ocorrência destas infeções;

- definir e implementar medidas de controlo e prevenção de infeção específicas e adequadas à realidade da unidade de saúde, bem como avaliar os efeitos dessas medidas na diminuição da taxa de infeção;

- análises de informações mais rápidas e simples, bem como maior autonomia dos utilizadores nas mesmas.

A plataforma de BI faz parte de um sistema implementado através da aplicação de métodos e ferramentas de BI, capaz de extrair e tratar dados referentes à ocorrência de infeções nosocomiais, gerar um conjunto de indicadores clínicos relacionados com essas infeções e apresentá-los na plataforma. Tal como anteriormente mencionado, um dos objetivos deste trabalho é a implementação deste sistema no CHP.

5.1 Aplicação da Metodologia de Kimball ao Desenvolvimento do Projeto

O sistema foi desenvolvido tendo em consideração a metodologia de Kimball para a implementação de sistemas de *data warehousing* e BI descrita na secção 3.1.2. Tendo em conta cada fase desta metodologia, procedeu-se à aplicação dos conceitos teóricos anteriormente apresentados ao caso prático referente a este trabalho.

Assim, numa primeira fase (*Planeamento do Projeto*) definiu-se o âmbito do projeto como sendo o estudo da incidência de infeção nosocomial nas Unidades de Medicina do CHP através da implementação da plataforma de BI. Nesta etapa também se definiram e planearam as atividades a executar ao longo de todo o projeto. Apresentou-se ainda a motivação para o desenvolvimento da plataforma como sendo a necessidade de facilitar o tratamento de dados referentes a infeções nosocomiais, de modo a gerar informações relevantes, capazes de auxiliar o estudo da incidência dessas infeções.

Ao longo de todo o projeto foi feita uma monitorização do mesmo (*Gestão do Projeto*) com o intuito de identificar eventuais problemas ao longo da implementação do sistema.

Realizou-se também um levantamento e análise dos indicadores de infeção nosocomial a apresentar com o sistema de BI (*Definição dos Requisitos do Negócio*). Estes indicadores serão descritos detalhadamente na secção 5.2.

Posteriormente, estabeleceu-se a arquitetura do sistema de acordo com os requisitos do projeto (*Projeto Técnico da Arquitetura*). A arquitetura do sistema será apresentada na secção 5.3.

Após a definição da arquitetura do sistema, realizou-se uma análise do *software* necessário para implementar o projeto, selecionou-se o *software* mais adequado e procedeu-se à instalação do mesmo (*Seleção e Instalação de Produtos*). Tal como referido anteriormente, neste trabalho utilizou-se: uma base de dados *Oracle* para realizar todas as tarefas de armazenamento e manipulação de dados (Secção 4.2.1); a ferramenta *Oracle SQL Developer* para facilitar a realização dessas tarefas (Secção 4.2.1); a ferramenta *Oracle Data Modeler* para ajudar no desenho físico do modelo dimensional do DW (Sec-

ção 4.2.2); e a ferramenta *Pentaho Community Edition* como ferramenta de BI para fazer a extração e apresentação de informações (Secção 4.3.2).

No que concerne ao DW, definiu-se o seu modelo dimensional através da identificação dos factos e dimensões necessários para gerar os indicadores (*Modelação Dimensional*). Posteriormente, criou-se o modelo dimensional na base de dados (*Desenho Físico*) e foram ainda criados procedimentos em PL/SQL para implementar o processo ETL (*Projeto e Desenvolvimento do Sistema de ETL*). Todas estas atividades serão detalhadamente apresentadas na secção 5.3.2.

Em relação à plataforma de BI, realizou-se, em primeiro lugar, um levantamento das necessidades que esta tem de satisfazer e das funcionalidades que deverá disponibilizar (*Especificação das Aplicações de BI*). Em seguida, tendo em consideração essa informação, procedeu-se ao desenvolvimento da plataforma com a ferramenta *Pentaho Community Edition* (*Desenvolvimento das Aplicações de BI*). A plataforma será descrita na secção 5.3.3.

Após estas atividades, procedeu-se à integração e teste de todos os componentes do sistema, de forma a validar todas as funcionalidades do mesmo (*Implementação*).

Por último, o sistema foi criado tendo em consideração a eventual expansão ou modificação do mesmo, de modo a que este esteja sempre adequado à realidade do CHP e às suas necessidades, havendo também a possibilidade do sistema ser aplicado noutros contextos ou unidades hospitalares (*Manutenção* e *Crescimento*). Estas ações podem ser realizadas, por exemplo, através da alteração do DW, através da definição de novos indicadores e/ou através da alteração da plataforma de BI.

Ao longo das próximas secções deste capítulo apresentam-se e explicam-se os componentes do sistema de BI implementado segundo a metodologia descrita, assim como as considerações e as atividades realizadas durante o seu desenvolvimento. Por fim, apresentam-se e discutem-se os principais resultados obtidos com este sistema.

5.2 Indicadores de Infeção Nosocomial

Os indicadores utilizados neste trabalho são parâmetros que possibilitam a caracterização e o estudo da incidência de infeção nosocomial, bem como a análise da relação entre esta e certos fatores considerados de risco. Estes indicadores permitem sumariar informações importantes presentes nos dados clínicos e são os utilizados no CHP para estudar a incidência de infeção nosocomial em doentes internados nas Unidades de Medicina.

Em seguida, serão apresentados os indicadores considerados neste projeto e será feita uma breve descrição dos mesmos.

5.2.1 População Estudada

Com o intuito de caracterizar a população estudada foi considerado um conjunto de indicadores que analisa informações gerais referentes aos formulários de infeção nosocomial preenchidos no CHP em situações de internamento. Este conjunto de indicadores inclui os seguintes parâmetros:

- lotação: número médio de camas disponíveis;
- dias de internamento: número médio de dias de internamento, calculado considerando a data de internamento e a data de alta;
- doentes saídos: número de doentes que tiveram alta;
- total de registos disponíveis: número de formulários de infeção nosocomial iniciados;
- percentagem de registos: percentagem de formulários de infeção nosocomial totalmente preenchidos.

Pretende-se contabilizar o total de cada um destes parâmetros para os três serviços em análise, Medicina A, Medicina B e Medicina C, bem como o seu valor total, isto é, a soma dos valores obtidos por serviço.

5.2.2 Fatores de Risco Intrínseco por Serviço

Este conjunto de indicadores avalia a relação entre a presença de certos fatores de risco intrínseco com relevância para a análise e a presença de infeção nosocomial. Esta análise realiza-se apenas para os serviços Medicina A, Medicina B e Medicina C e pretende-se obter os seguintes indicadores para cada um dos serviços mencionados:

- número total de doentes com o fator de risco intrínseco em análise;
- número total de doentes que possuem o fator de risco intrínseco em análise e, simultaneamente, apresentam infeção nosocomial;
- percentagem de doentes, com o fator de risco intrínseco em análise, que possuem infeção nosocomial ($\frac{\text{nº de doentes c/ fator de risco e infeção}}{\text{nº de doentes c/ fator de risco}} \times 100$).

Nesta análise são considerados relevantes os seguintes fatores de risco intrínseco: alcoolismo, diabetes, doença hepática cronica, transplantes, corticoides, imunossupressores, desnutrição, *Human Immunodeficiency Virus* (HIV), traqueostomia, coma, doença pulmonar obstrutiva crónica, perturbação da deglutição e imunodeficiência. Estes fatores de risco intrínseco foram identificados por especialistas nesta área e são os considerados no CHP para estudar a incidência de infeção nosocomial em doentes internados nas Unidades de Medicina.

5.2.3 Fatores de Risco Extrínseco por Serviço

Este conjunto de indicadores tem como objetivo estudar a influência de fatores de risco extrínseco, isto é, dispositivos invasivos, na ocorrência de infeções nosocomiais. Esta análise efetua-se apenas ao nível dos serviços Medicina A, Medicina B e Medicina C e, para cada um dos serviços mencionados, pretende-se averiguar:

- o número total de doentes com o dispositivo invasivo em análise;
- o número total de doentes que utilizam o dispositivo invasivo em análise e, simultaneamente, apresentam infeção nosocomial;

- a percentagem de doentes, com o dispositivo invasivo em análise, que possuem infeção nosocomial ($\frac{\text{nº de doentes c/ dispositivo invasivo e infeção}}{\text{nº de doentes c/ dispositivo invasivo}} \times 100$).

Os dispositivos invasivos considerados fatores de risco extrínseco são o cateter urinário, o cateter central e o cateter periférico como formas de cateterismo, bem como a entubação naso-gástrica e a entubação naso-traqueal como formas de entubação. Estes fatores de risco extrínseco foram identificados por especialistas nesta área, sendo os considerados no CHP para estudar a incidência de infeção nosocomial em doentes internados nas Unidades de Medicina.

5.2.4 Infeções por Tipo e Serviço

Este conjunto de indicadores pretende caracterizar a infeção nosocomial ao nível dos tipos de infeção verificados. Pretende-se obter os seguintes indicadores:

- número de infeções nosocomiais por tipo de infeção e serviço clínico;
- número total de infeções nosocomiais registadas em cada serviço;
- número total de infeções associadas à presença de uma infeção nosocomial por serviço, isto é, a soma por serviço dos valores obtidos para cada tipo de infeção;
- valor global de cada um dos parâmetros anteriormente mencionados independentemente do serviço clínico;
- percentagem de infeções nosocomiais registada em cada um dos serviços clínicos em análise.

Nesta análise consideram-se apenas os serviços Medicina A, Medicina B e Medicina C e os seguintes tipos de infeção: infeção do trato urinário, sépsis, infeção respiratória e outras infeções. Estes tipos de infeção, frequentemente associados a infeções nosocomiais, foram identificados por especialistas nesta área e são os considerados no CHP para analisar a incidência de infeção nosocomial em doentes internados nas Unidades de Medicina.

5.3 Arquitetura do Sistema

O sistema de BI para o estudo de infeção nosocomial organiza-se segundo uma arquitetura composta por três níveis (Figura 5.1), baseada na arquitetura típica de sistemas de BI apresentada na figura 3.1 (Secção 3.1).

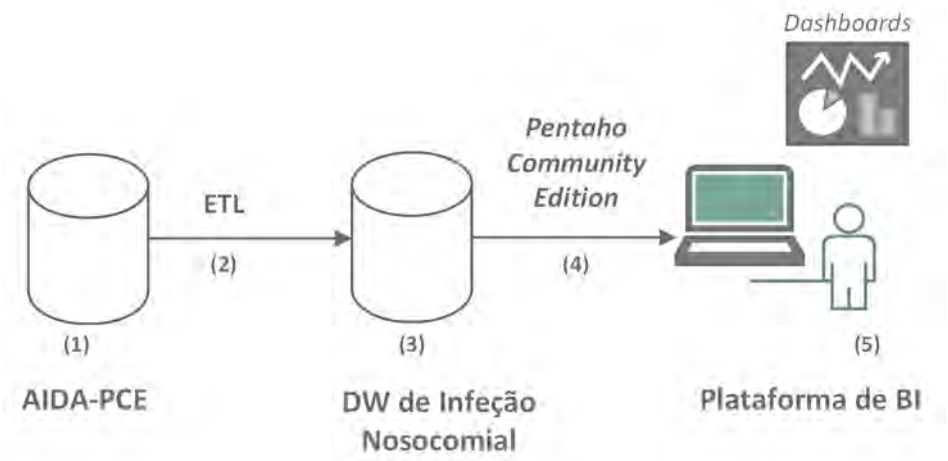

Figura 5.1: Arquitetura do sistema de BI para o estudo da incidência de infeção nosocomial.

O primeiro nível deste sistema diz respeito às fontes de dados, neste caso, o módulo de PCE da AIDA (AIDA-PCE) (1) que contém os dados relevantes para a análise.

No segundo nível encontra-se o DW construido para o estudo de infeção nosocomial, composto por dois *data marts* que são povoados a partir das fontes de dados do primeiro nível, através de processos ETL (2). Estes últimos permitem extrair os dados da AIDA-PCE, transformá-los de acordo com as necessidades do sistema e carregá-los para o DW (3). Neste DW encontram-se todos os dados necessários para gerar os indicadores relevantes para o estudo.

Por fim, o terceiro nível do sistema diz respeito à plataforma de BI para apresentar os indicadores de infeção nosocomial. Esta plataforma *web* disponibiliza as ferramentas de BI que permitem a interação entre o utilizador e os dados armazenados no DW (4). A plataforma é constituída por *dashbo-*

ards (5) que apresentam os indicadores em gráficos e em tabelas dinâmicas que permitem a análise OLAP dos indicadores. Existe ainda um *dashboard* que expõe a probabilidade de um doente adquirir uma infeção nosocomial, com base na tecnologia de DM. Este *dashboard* e o trabalho relativo ao seu desenvolvimento serão descritos no capítulo 6.

Convém destacar que, pelo facto de ser uma aplicação *web*, a plataforma pode ser consultada em qualquer local dentro do hospital e por qualquer dispositivo, desde que o utilizador tenha acesso à rede e privilégios de acesso.

5.3.1 Caracterização dos Dados

Neste estudo são considerados os dados presentes nos registos de infeção nosocomial armazenados no CHP durante todo o ano de 2013, pelo que a análise refere-se exclusivamente a esse ano. Além disso, tal como já foi mencionado, a análise efetua-se unicamente ao nível das Unidades de Medicina, ou seja, são considerados apenas dados referentes aos serviços clínicos Medicina A, Medicina B e Medicina C. De notar que, atualmente, o sistema contém apenas dados referentes ao ano de 2013, mas no futuro, através da atualização dos *data marts*, este poderá incluir dados atuais, de modo a permitir obter informações em tempo real.

Os registos de infeção nosocomial do CHP são formulários preenchidos e registados eletronicamente pelos médicos no momento da alta. Com estes formulários é registado um conjunto de informações importantes para a compreensão da infeção nosocomial, tais como, informação relativa ao período de internamento e diagnóstico efetuado e características intrínsecas do doente. É ainda registada informação relativa à ocorrência ou não de infeção nosocomial, dispositivos invasivos utilizados durante o período de internamento, tratamentos realizados, antibióticos administrados, etc.

Estes formulários são caracterizados por uma interface amigável para o profissional de saúde (composta por *combo boxes*, *check boxes*, entre outros elementos de interface gráfica), restringindo desta forma os valores que este pode preencher nos diferentes campos apresentados. Deste modo, é possível obter uniformidade nos dados recolhidos, o que permite o armazenamento

de dados corretos, capazes de oferecer informações úteis e importantes para estudos estatísticos e comparativos [3]. Com o preenchimento de cada formulário gera-se um ficheiro XML contendo todos os valores registados, sendo que todos os ficheiros gerados são armazenados na AIDA-PCE juntamente com alguns dados relevantes associados aos mesmos (nomeadamente informações pessoais do doente, o serviço clínico onde o registo é efetuado, a data de criação do registo, a versão do formulário associado ao ficheiro XML).

Durante o período de tempo em análise e para os serviços clínicos referidos foram preenchidos 2118 formulários de infeção nosocomial. Desses 2118 registos apenas 1669 possuem o campo do formulário referente à presença ou ausência de infeção preenchido. Por sua vez, desses 1669 registos apenas 173 indicam que efetivamente ocorreu uma infeção.

5.3.2 *Data Warehouse*

Neste trabalho optou-se pelo paradigma de Ralph Kimball para a construção do DW, pelo que o DW foi construído segundo uma abordagem *bottom-up*. Por conseguinte, começou-se pela construção dos *data marts* necessários para implementar o sistema.

O DW é constituído por dois *data marts*, um para armazenar informações relativas aos indicadores relacionados com a caracterização da população estudada e outro para armazenar informações relativas aos restantes conjuntos de indicadores: Fatores de Risco Intrínseco por Serviço, Fatores de Risco Extrínseco por Serviço e Infeções por Tipo e Serviço.

Os *data marts* foram implementados segundo a técnica de modelação dimensional descrita na secção 3.1.2 e cada um deles tem em consideração as especificidades de cada grupo de indicadores que representa. Estes *data marts* foram cuidadosamente projetados e implementados, de modo que os dados necessários para a obtenção dos indicadores estejam corretamente armazenados e permitam obter as informações necessárias para o estudo da incidência de infeção nosocomial.

Modelo Dimensional

De modo a modelar adequadamente o problema em análise, o modelo dimensional dos *data marts* foi definido tendo em consideração os indicadores a gerar com o sistema. Após a definição do modelo dimensional, procedeu-se à criação e ao povoamento das diferentes tabelas de factos e de dimensões na base de dados.

O *data mart População Estudada* (Figura 5.2) foi criado para armazenar os dados necessários para gerar o conjunto de indicadores que permite caracterizar a população estudada e é constituído por três dimensões (*Data*, *Especialidade* e *Infeção Nosocomial*) e uma tabela de factos (*População Estudada*). Este *data mart* possui um esquema em estrela.

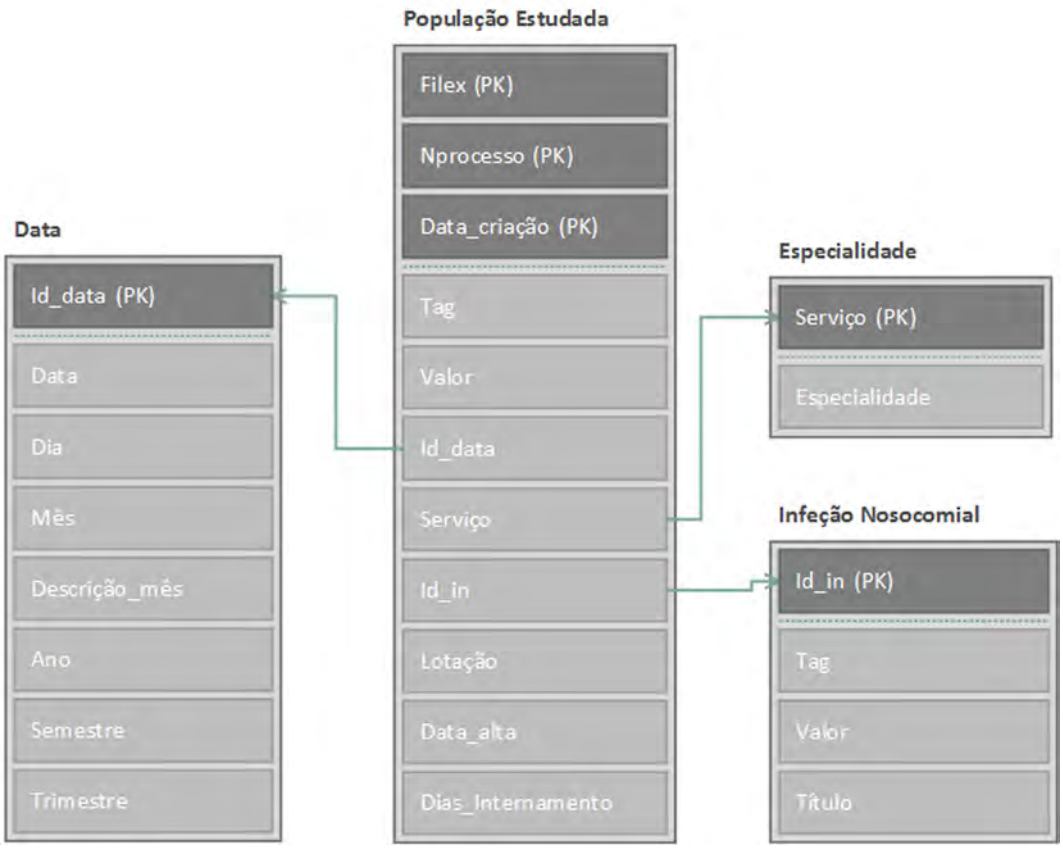

Figura 5.2: Modelo dimensional do *data mart População Estudada*.

A tabela de factos *População Estudada* (Tabela 5.1) contém os factos

necessários para gerar os indicadores relativos à caracterização da população em estudo. Os atributos *filex*, *nprocesso* e *data_criação* constituem a chave primária desta tabela de factos e permitem identificar univocamente cada registo presente na mesma. Esta tabela possui ainda os factos *lotação*, *data_alta* e *dias_internamento*.

Esta tabela de factos relaciona-se com as dimensões *Data*, *Especialidade* e *Infeção Nosocomial* através das chaves estrangeiras *serviço*, *id_data* e *id_in*, respetivamente.

Tabela 5.1: Estrutura da tabela de factos *População Estudada*.

Atributo	**Tipo**	**Chave**	**Descrição**
Filex	Varchar2 (100 BYTE)	Primária	Código da versão do formulário
Nprocesso	Number (15, 0)	Primária	Número do processo hospitalar do doente no CHP
Data_criação	Date	Primária	Data de criação do registo
Id_data	Integer	Estrangeira	Código associado à data de criação do registo
Serviço	Varchar2 (50 BYTE)	Estrangeira	Código da especialidade no CHP
Id_in	Integer	Estrangeira	Código associado ao valor de infeção nosocomial registado
Lotação	Integer		Lotação total média para o serviço
Data_alta	Date		Data de alta do doente
Dias_internamento	Integer		Número de dias de internamento do doente

O *data mart Infeção Nosocomial* (Figura 5.3), referente à caracterização da incidência de infeção nosocomial é constituído por seis dimensões (*Data*, *Especialidade*, *Infeção Nosocomial*, *Fatores de Risco*, *Fatores de Risco Intrínseco* e *Fatores de Risco Extrínseco*) e uma tabela de factos (*Factos de Infeção Nosocomial*). Este *data mart* possui um esquema em estrela.

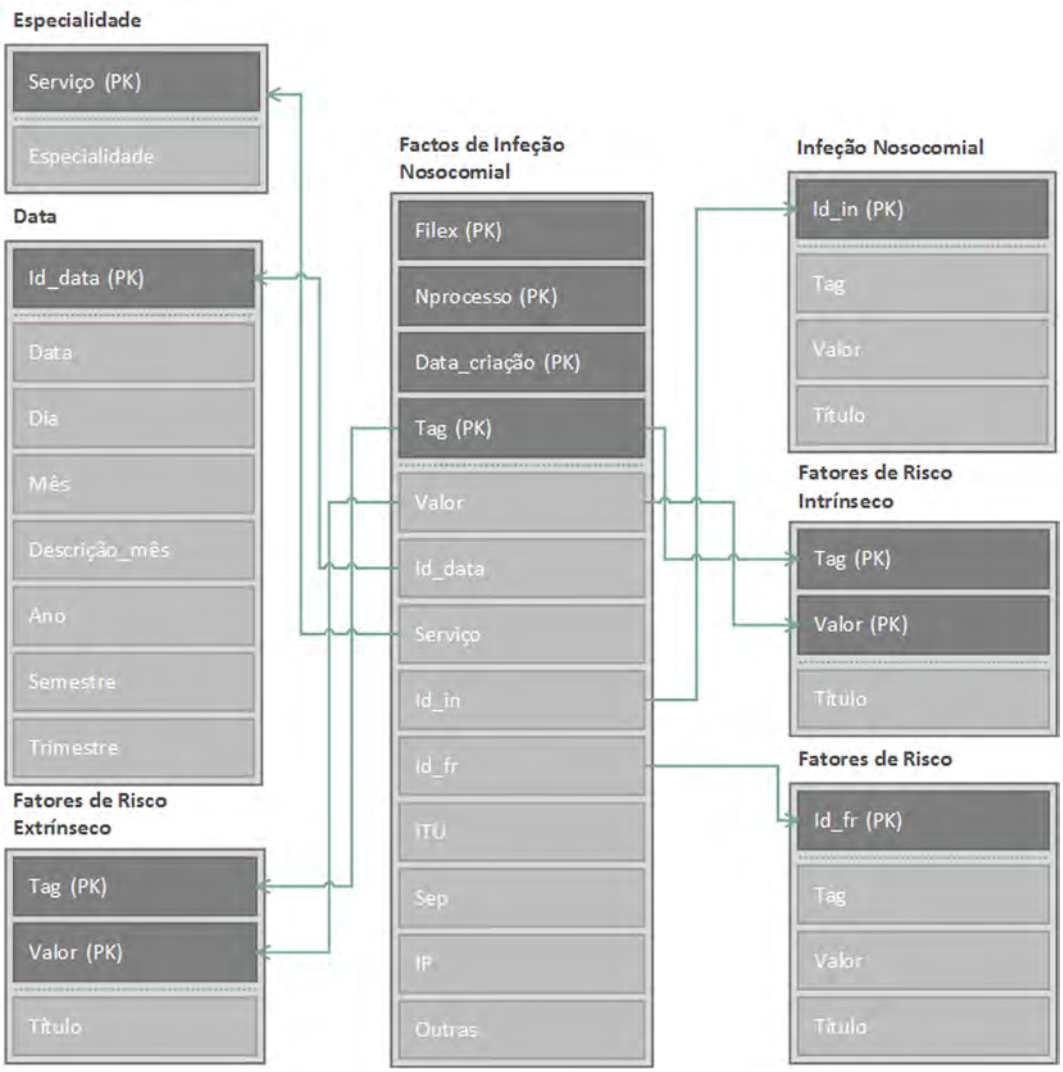

Figura 5.3: Modelo dimensional do *data mart Infeção Nosocomial*.

A tabela de factos *Factos de Infeção Nosocomial* (Tabela 5.2) contém os factos necessários para gerar os indicadores relacionados com os fatores de risco intrínseco, fatores de risco extrínseco e tipos de infeção. Os atributos *filex*, *nprocesso*, *data_ criação* e *tag* constituem a chave primária desta tabela e permitem identificar univocamente cada um dos seus registos. Esta tabela possui ainda os factos *itu*, *sep*, *ip* e *outras* que indicam, respetivamente, o valor associado à infeção do trato urinário, à sépsis, à infeção respiratória e a outras infeções, informando se o doente tem ou não a infeção em questão.

Esta tabela de factos relaciona-se com as dimensões *Data*, *Especialidade*, *Infeção Nosocomial*, *Fatores de Risco*, *Fatores de Risco Intrínseco* e *Fatores de Risco Extrínseco* através das chaves estrangeiras *serviço*, *id_data*, *id_in*, *id_fr*, *tag* e *valor*.

Tabela 5.2: Estrutura da tabela de factos *Factos de Infeção Nosocomial*.

Atributo	Tipo	Chave	Descrição
Filex	Varchar2 (100 BYTE)	Primária	Código da versão do formulário
Nprocesso	Number (15, 0)	Primária	Número do processo hospitalar do doente no CHP
Data_criação	Date	Primária	Data de criação do registo
Tag	Varchar2 (50 BYTE)	Primária Estrangeira	Código de cada campo do formulário no ficheiro XML
Valor	Varchar2 (50 BYTE)	Estrangeira	Valor associado ao campo do formulário
Id_data	Integer	Estrangeira	Código associado à data de criação do registo
Serviço	Varchar2 (50 BYTE)	Estrangeira	Código da especialidade no CHP
Id_in	Integer	Estrangeira	Código associado ao valor de infeção nosocomial registado
Id_fr	Integer	Estrangeira	Código associado ao valor dos fatores de risco
ITU	Varchar2 (50 BYTE)		Valor da infeção do trato urinário
Sépsis	Varchar2 (50 BYTE)		Valor da sépsis
IP	Varchar2 (50 BYTE)		Valor da infeção respiratória
Outras	Varchar2 (50 BYTE)		Valor associado a outras infeções

A dimensão *Data* (Tabela 5.3) é utilizada nos factos com necessidade de análise por data, por exemplo nas situações em que é necessário agregar os dados por data (ano, semestre, trimestre, mês ou dia). Esta dimensão é

comum aos dois *data marts* e contém um registo para cada dia do ano em análise, neste caso, para cada dia de 2013.

Tabela 5.3: Estrutura da tabela de dimensão *Data*.

Atributo	Tipo	Chave	Descrição
Id_data	Integer	Primária	Chave única que identifica cada registo
Data	Date		Data completa
Dia	Integer		Dia do mês
Mês	Integer		Número correspondente ao mês
Descrição_mês	Varchar2 (50 BYTE)		Nome correspondente ao mês
Ano	Integer		Número correspondente ao ano
Semestre	Varchar2 (50 BYTE)		Número correspondente ao semestre
Trimestre	Varchar2 (50 BYTE)		Número correspondente ao trimestre

A dimensão *Especialidade* (Tabela 5.4) é utilizada nos factos com necessidade de análise por serviço, é comum aos dois *data marts* e contém apenas informação referente aos três serviços em análise: Medicina A, Medicina B e Medicina C.

Tabela 5.4: Estrutura da tabela de dimensão *Especialidade*.

Atributo	Tipo	Chave	Descrição
Serviço	Varchar2 (50 BYTE)	Primária	Código da especialidade no CHP
Especialidade	Varchar2 (100 BYTE)		Designação da especialidade

A dimensão *Infeção Nosocomial* (Tabela 5.5) é utilizada nos factos com necessidade de agrupar os dados pelo valor associado à infeção nosocomial,

isto é, nas situações em que é necessário distinguir os doentes que tiveram infeção nosocomial dos doentes que não tiveram infeção. Esta dimensão é comum aos dois *data marts* e contém apenas informação referente aos três valores possíveis para a ocorrência de uma infeção nosocomial: presente, ausente ou desconhecido.

Tabela 5.5: Estrutura da tabela de dimensão *Infeção Nosocomial*.

Atributo	Tipo	Chave	Descrição
Id_in	Integer	Primária	Código da infeção nosocomial
Tag	Varchar2 (50 BYTE)		Código do campo de infeção nosocomial no ficheiro XML
Valor	Varchar2 (50 BYTE)		Valor do campo de infeção nosocomial
Título	Varchar2 (50 BYTE)		Nome do campo de infeção nosocomial no ficheiro XML

A dimensão *Fatores de Risco* (Tabela 5.6) é utilizada nos factos com necessidade de impor condições aos dados pela presença ou ausência de fatores de risco. Esta dimensão pertence apenas ao *data mart Infeção Nosocomial* e contém apenas informação referente aos três valores possíveis associados aos fatores de risco: com fatores de risco, sem fatores de risco ou desconhecido.

Tabela 5.6: Estrutura da tabela de dimensão *Fatores de Risco*.

Atributo	Tipo	Chave	Descrição
Id_fr	Integer	Primária	Código do fator de risco
Tag	Varchar2 (50 BYTE)		Código do campo dos fatores de risco no ficheiro XML
Valor	Varchar2 (50 BYTE)		Valor do campo dos fatores de risco
Título	Varchar2 (50 BYTE)		Nome do campo dos fatores de risco no ficheiro XML

A dimensão *Fatores de Risco Intrínseco* (Tabela 5.7) é utilizada nos factos

com necessidade de agrupar os dados de acordo com o valor associado a cada fator de risco intrínseco. Esta dimensão pertence somente ao *data mart Infeção Nosocomial* e contém informação referente aos três valores possíveis associados a cada fator de risco intrínseco: presente, ausente ou desconhecido.

Tabela 5.7: Estrutura da tabela de dimensão *Fatores de Risco Intrínseco*.

Atributo	Tipo	Chave	Descrição
Tag	Varchar2 (50 BYTE)	Primária	Código do fator de risco intrínseco no ficheiro XML
Valor	Varchar2 (50 BYTE)	Primária	Valor do fator de risco intrínseco
Título	Varchar2 (50 BYTE)		Nome do fator de risco intrínseco no ficheiro XML

A dimensão *Fatores de Risco Extrínseco* (Tabela 5.8) é utilizada nos factos em que é necessário verificar se cada fator de risco extrínseco está presente. Esta dimensão pertence apenas ao *data mart Infeção Nosocomial* e contém informação referente aos três valores possíveis associados a cada fator de risco extrínseco: presente, ausente ou desconhecido.

Tabela 5.8: Estrutura da tabela de dimensão *Fatores de Risco Extrínseco*.

Atributo	Tipo	Chave	Descrição
Tag	Varchar2 (50 BYTE)	Primária	Código do fator de risco extrínseco no ficheiro XML
Valor	Varchar2 (50 BYTE)	Primária	Valor do fator de risco extrínseco
Título	Varchar2 (50 BYTE)		Nome do fator de risco extrínseco no ficheiro XML

ETL

Antes de serem carregados para os *data marts*, os dados relevantes para este trabalho foram extraídos da AIDA-PCE, analisados, manipulados e sujeitos a operações de limpeza, de modo a adequá-los aos modelos dimensionais

definidos para os *data marts*. Uma vez que os indicadores apresentados pela plataforma são extraídos dos dados presentes nos *data marts*, é muito importante que estes dados possuam elevada qualidade. Deste modo, tornou-se necessário dedicar bastante tempo ao tratamento de dados.

Neste trabalho foram detetados valores armazenados sob a forma de XML que foi necessário extrair e transformar. Foi também necessário eliminar caracteres indesejados, substituir valores não preenchidos por "desconhecido"e excluir registos não relevantes para a análise, bem como registos repetidos. Por fim, povoaram-se os *data marts* com os dados transformados. Todas estas operações de ETL foram automatizadas com a execução de procedimentos em PL/SQL.

5.3.3 Plataforma de *Business Intelligence*

A plataforma de BI para indicadores de infeção nosocomial é uma aplicação *web* que, recorrendo a uma ferramenta de BI, faz análise OLAP e consultas de dados para gerar indicadores, e apresenta essas informações relevantes em gráficos e em tabelas.

Para criar a plataforma recorreu-se ao componente CDE da ferramenta *Pentaho Community Edition* (Secção 4.3.2), pois este permite desenvolver muito facilmente *dashboards* interativos e apelativos para o utilizador, facilitando a interpretação das informações que os mesmos apresentam. Este possibilita também a exportação dos resultados e formatação de todos os componentes dos *dashboards*, dando, por isso, uma grande autonomia no desenvolvimento de *dashboards*. A plataforma que apresenta os indicadores de infeção nosocomial é composta por um *dashboard* inicial e três *dashboards* que apresentam informações mais detalhadas. Esta possui todos os indicadores de infeção nosocomial anteriormente mencionados e permite a navegação entre as diferentes páginas que a constituem.

No *dashboard* inicial o utilizador pode selecionar o ano que pretende analisar e visualizar um conjunto de gráficos que resumem cada conjunto de indicadores, com valores referentes apenas ao ano escolhido. Este *dashboard* expõe uma visão geral sobre a incidência de infeção nosocomial no ano sele-

cionado, sendo os indicadores que mais rapidamente permitem tirar ilações dos dados apresentados em seis gráficos: lotação média e duração média do internamento; percentagem de registos por serviço; percentagem de infeções nosocomiais por serviço; percentagem de infeções por tipo e por serviço; percentagem de doentes com o fator de risco intrínseco e infeção nosocomial por serviço; percentagem de doentes com o fator de risco extrínseco e infeção nosocomial por serviço. Estes gráficos são criados através de *queries* SQL executadas sobre os *data marts* e os seus dados podem ser exportados para folha de cálculo do Excel. Através de botões presentes nesta página é possível aceder aos restantes *dashboards* que constituem a plataforma. Na figura 5.4 apresenta-se um excerto do *dashboard* inicial, onde é possível observar as funcionalidades anteriormente descritas.

Figura 5.4: Excerto do *dashboard* inicial.

Os restantes três *dashboards* que compõem a plataforma contêm informação associada a cada conjunto de indicadores, detalhada por data e serviço, independentemente do ano. O *dashboard Caracterização da População Estudada* apresenta informação detalhada sobre o conjunto de indicadores que caracterizam a população em estudo. No *dashboard Caracterização da Infeção Nosocomial* é apresentada informação detalhada alusiva ao conjunto de indicadores Infeções por Tipo e Serviço. Por fim, o *dashboard Fatores de Risco e Infeção Nosocomial* expõe informação detalhada relativa aos conjuntos de indicadores Fatores de Risco Intrínseco por Serviço e Fatores de Risco Extrínseco por Serviço.

Estes *dashboards* são constituídos por tabelas dinâmicas com resultados da análise OLAP que podem ser exploradas em tempo real e cada um deles, por meio de um botão, permite regressar ao *dashboard* inicial. A informação apresentada nestas tabelas pode ser configurada pelo utilizador através da alteração do que este pretende visualizar nas linhas ou nas colunas das mesmas, ou ainda através da aplicação de filtros aos dados. As tabelas dinâmicas podem também ser impressas ou exportadas para PDF ou folha de cálculo do Excel.

Como exemplo, na figura 5.5 encontra-se um excerto do *dashboard Fatores de Risco e Infeção Nosocomial*, podendo ser observadas as funcionalidades anteriormente mencionadas.

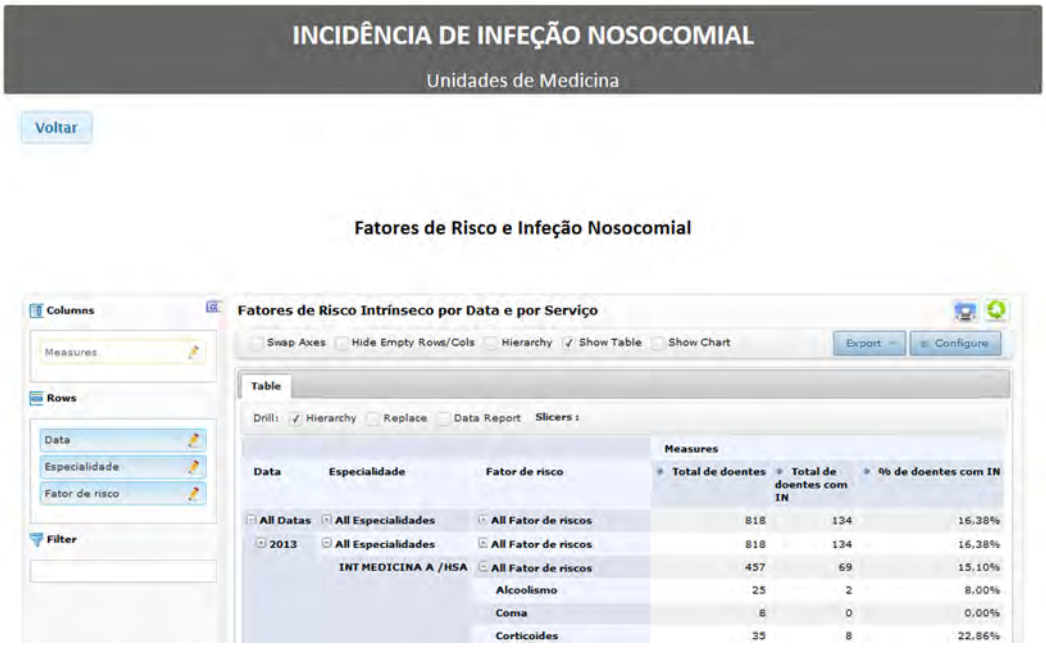

Figura 5.5: Excerto do *dashboard* *Fatores de Risco e Infeção Nosocomial*.

Análise OLAP

A implementação de um DW segundo a técnica de modelação dimensional permite a realização de OLAP. Esta possui inúmeras vantagens ao nível da análise de informação porque efetua operações que possibilitam uma exploração profunda da mesma, tais como *drill-down*, *slice and dice*, *pivot*, entre outras. As ferramentas OLAP permitem ainda a apresentação dos resultados em gráficos ou tabelas, bem como a análise rápida, interativa e em tempo real dos mesmos, segundo as diferentes dimensões do modelo de dados. Deste modo, considerou-se relevante a utilização de uma ferramenta deste tipo para a realização de uma análise mais detalhada dos indicadores de infeção nosocomial. Assim, foi necessário criar cubos OLAP com os dados dos *data marts* que, posteriormente, foram explorados com uma ferramenta OLAP.

Neste trabalho utilizou-se o servidor OLAP *Mondrian* (Secção 4.3.2), integrado na ferramenta *Pentaho Community Edition*, para implementar os cubos OLAP e operar sobre eles. Foram criados quatro cubos, sendo que cada um deles representa um dos conjuntos de indicadores definidos na sec-

ção 5.2 e contém apenas as medidas e os factos necessários para representar cada um desses conjuntos. Estes cubos consistem numa seleção de factos e dimensões dos *data marts*, a fim de obter um conjunto de dados mais específico, e definem a hierarquia entre atributos pertencentes a uma determinada dimensão.

Como exemplo, a figura 5.6 apresenta a estrutura de um cubo criado para representar o conjunto de indicadores referentes à caracterização da população em estudo. Este contém os factos que se pretende analisar com esses indicadores, bem como as diferentes dimensões necessárias para os analisar e os respetivos atributos. A ordem dos atributos na dimensão define a sua estrutura hierárquica, possibilitando a realização de operações como *drill-down* e *roll-up*. Neste trabalho, estas operações podem ser efetuadas, por exemplo, ao nível da dimensão *Data*, visto que esta possui atributos que podem ser utilizados para agregar os factos segundo diferentes níveis de detalhe (dia, mês, trimestre, semestre ou ano).

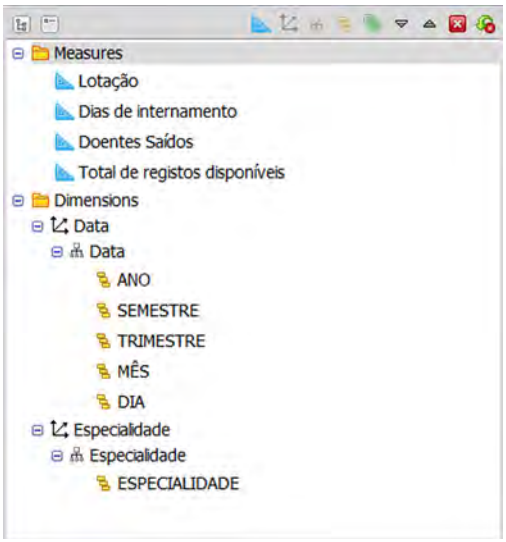

Figura 5.6: Estrutura do cubo OLAP criado para o conjunto de indicadores População Estudada.

Posteriormente, os cubos foram explorados com o *plug-in OpenI* (Secção 4.3.2), um *software* que realiza OLAP sobre os cubos, permitindo a escolha das dimensões e dos factos a visualizar. Com este *software* procedeu-se à

elaboração de tabelas dinâmicas através da criação de *queries* MDX, sendo que as tabelas criadas permitem: apresentar os indicadores definidos nos cubos OLAP; a realização de operações OLAP, de acordo com as hierarquias definidas para os atributos das dimensões; a sua exploração em tempo real. Deste modo, torna-se possível explorar detalhadamente os indicadores de infeção nosocomial.

5.4 Apresentação e Discussão dos Resultados

Para a análise dos resultados obtidos com este trabalho escolheram-se apenas alguns dos indicadores expostos pela plataforma de BI. Os resultados não apresentados nesta secção encontram-se em anexo (Anexo A).

Analisando em primeiro lugar os dados de infeção nosocomial de 2013 ao nível da caracterização da população estudada neste trabalho (Figura 5.7), verifica-se que, por exemplo, o serviço Medicina A teve, nesse ano, a lotação média mais elevada, tendo sido esse valor 49 camas.

		Measures				
Data	Especialidade	Lotação	Dias de internamento	Doentes Saídos	Total de registos disponíveis	% Registos
All Datas	All Especialidades	41,54	14,07	1669	2118	78,80%
2013	All Especialidades	41,54	14,07	1669	2118	78,80%
	INT MEDICINA A /HSA	49,00	14,30	1018	1318	77,24%
	INT MEDICINA B /HSA	32,78	15,43	278	366	75,96%
	INT MEDICINA C /HSA	26,26	12,23	373	434	85,94%

Figura 5.7: Excerto do *dashboard* *Caracterização da População Estudada*: indicadores que caracterizam a população estudada, por serviço e ano.

Ainda no período de tempo em análise, verifica-se que Medicina B registou o maior número médio de dias de internamento, tendo sido esse valor 15.43 dias. O menor valor médio de dias de internamento foi 12.23 dias e foi registado em Medicina C.

Medicina A registou também o maior número de doentes saídos (altas) e o maior número de registos disponíveis (formulários de infeção nosocomial totalmente finalizados), sendo que a diferença entre estes valores neste serviço

e nos outros serviços estudados é bastante significativa e poderá, em parte, ser justificada pela maior capacidade deste serviço.

Observa-se ainda que o serviço Medicina C, apesar de possuir uma lotação média menor do que Medicina B, registou um maior número de casos, quer ao nível de doentes saídos, quer ao nível do total de registos disponíveis. Esta observação poderá estar relacionada com a duração do internamento, que em Medicina C é significativamente inferior.

Medicina C registou a maior percentagem de formulários de infeção nosocomial preenchidos, tendo verificado uma percentagem de registos de 85.94%. A percentagem de registos no geral foi de 78.80% o que significa que em 100 casos de internamento nas Unidades de Medicina apenas 78.8 formulários de infeção nosocomial foram corretamente terminados.

Na figura 5.8 apresenta-se um exemplo de *drill-down* realizado sobre a tabela dinâmica anteriormente apresentada na figura 5.7. Neste caso o *drill--down* é efetuado ao nível da dimensão *Data*, permitindo a visualização do conjunto de indicadores que carateriza a população em estudo segundo qualquer atributo pertencente a esta dimensão.

		Measures				
Data	Especialidade	Lotação	Dias de internamento	Doentes Saídos	Total de registos disponíveis	% Registos
All Datas	All Especialidades	41,54	14,07	1669	2118	78,80%
2013	All Especialidades	41,54	14,07	1669	2118	78,80%
1º Semestre	All Especialidades	41,79	13,92	1030	1272	80,97%
1º Trimestre	All Especialidades	42,06	13,82	623	797	78,17%
Janeiro	All Especialidades	41,90	13,74	200	241	82,99%
	INT MEDICINA A /HSA	49,00	13,28	121	137	88,32%
	INT MEDICINA B /HSA	36,00	15,02	35	59	59,32%
	INT MEDICINA C /HSA	28,00	13,47	44	45	97,78%
Fevereiro	All Especialidades	42,32	13,27	218	311	70,10%

Figura 5.8: Exemplo de *drill-down* nos indicadores que caracterizam a população estudada, efetuado ao nível da dimensão *Data*.

Neste exemplo, é possível observar os valores mensais dos diferentes indicadores, mais concretamente, os valores referentes ao mês de janeiro de 2013.

Relativamente aos resultados relacionados com os indicadores que permitem caracterizar as infeções nosocomiais ao nível do tipo de infeção e da taxa de infeção nosocomial verificada, observa-se que, no ano de 2013, a percentagem de infeções nosocomiais foi idêntica em todos os serviços em análise. Este valor foi inferior no serviço Medicina A (9.43%) e superior em Medicina B (12.95%) (Figura 5.9). Convém assinalar que Medicina A foi o serviço com mais formulários de infeção nosocomial registados e Medicina B foi o serviço com menos formulários preenchidos.

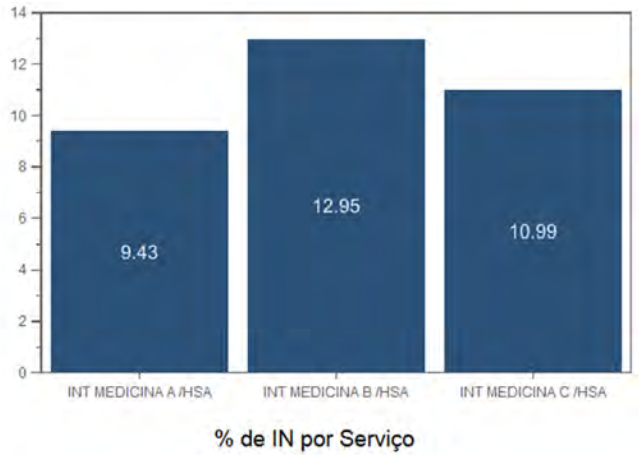

Figura 5.9: Excerto do *dashboard* inicial: percentagem de infeções nosocomiais, por serviço em 2013.

Em relação às percentagens de infeções nosocomiais por serviço, constata-se ainda que poderá existir uma relação entre a duração do internamento e a ocorrência deste tipo de infeções, pois Medicina B foi o serviço estudado com maior percentagem de infeções nosocomiais e obteve também a maior duração média do internamento. Este resultado poderá justificar-se pelo facto dos doentes que permanecem mais tempo hospitalizados possuírem uma exposição ao ambiente hospitalar mais prolongada e, portanto, um maior risco de contrair uma infeção.

No que concerne aos tipos de infeções associados a infeções nosocomiais (Figura 5.10), verifica-se que a maior parte das infeções nosocomiais estiveram associadas a infeções do trato urinário, sendo que a percentagem destas

infeções oscilou entre 38.71% (Medicina A) e 41.67% (Medicina C).

Constata-se ainda que, de um modo geral, os tipos de infeções nosocomiais mais frequentes foram a infeção do trato urinário e a infeção respiratória. Estes resultados poderão estar relacionados com os procedimentos clínicos efetuados nestes serviços, bem como com os dispositivos invasivos utilizados.

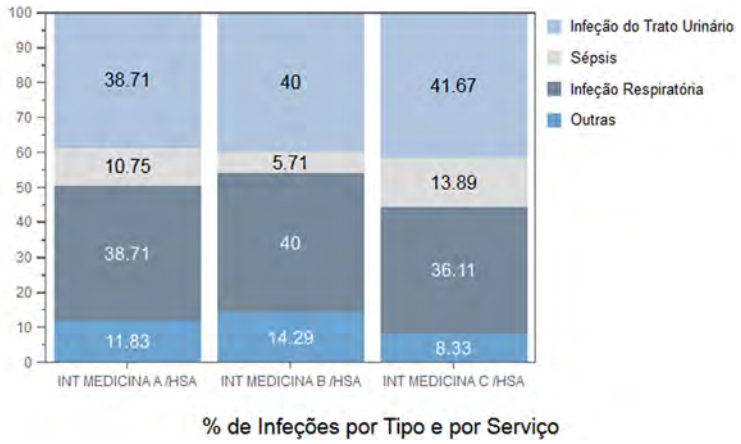

Figura 5.10: Excerto do *dashboard* inicial: percentagem de infeções, por tipo e serviço em 2013.

Relativamente aos resultados do conjunto de indicadores que tem como objetivo estudar a influência de fatores de risco extrínseco na ocorrência de infeções nosocomiais (Figura 5.11) verifica-se que, em todos os serviços estudados, o cateter periférico foi o dispositivo invasivo mais utilizado nos doentes e a entubação naso-traqueal foi o menos frequente.

Nos casos em que se verificaram infeções nosocomiais, observa-se também que o cateter periférico foi o dispositivo invasivo mais vezes utilizado e a entubação naso-traqueal foi o que apresentou menos ocorrências de infeções.

			Measures		
Data	Especialidade	Dispositivo invasivo	Total de doentes	Total de doentes com IN	% de doentes com IN
All Datas	All Especialidades	All Dispositivo invasivos	2118	333	15,72%
2013	All Especialidades	All Dispositivo invasivos	2118	333	15,72%
	INT MEDICINA A /HSA	All Dispositivo invasivos	1276	187	14,66%
		Cateter Central	39	8	20,51%
		Cateter Periférico	789	86	10,90%
		Cateter Urinário	352	67	19,03%
		Ent. Naso-Gástrica	85	22	25,88%
		Ent. Naso-Traqueal	11	4	36,36%
	INT MEDICINA B /HSA	All Dispositivo invasivos	363	60	16,53%
		Cateter Central	11	3	27,27%
		Cateter Periférico	213	26	12,21%
		Cateter Urinário	111	26	23,42%
		Ent. Naso-Gástrica	26	4	15,38%
		Ent. Naso-Traqueal	2	1	50,00%
	INT MEDICINA C /HSA	All Dispositivo invasivos	479	86	17,95%
		Cateter Central	17	7	41,18%
		Cateter Periférico	288	35	12,15%
		Cateter Urinário	126	31	24,60%
		Ent. Naso-Gástrica	42	11	26,19%
		Ent. Naso-Traqueal	6	2	33,33%

Figura 5.11: Excerto do *dashboard Fatores de Risco e Infeção Nosocomial*: indicadores que relacionam fatores de risco extrínseco com a presença de infeção nosocomial, por serviço e ano.

No que diz respeito à percentagem de doentes, com o fator de risco, que apresentam infeção nosocomial, constata-se que, em 2013, as percentagens obtidas foram relativamente elevadas sendo que a percentagem mais elevada foi 50% e é referente à utilização de entubação naso-traqueal no serviço Medicina B. A percentagem mais baixa foi 10.90% e é referente à utilização de cateter periférico em Medicina A. Apesar da utilização de cateter periférico estar associada ao maior número de doentes com infeção nosocomial em todos os serviços, foi também o dispositivo invasivo mais utilizado pelo que a percentagem de infeções na presença deste dispositivo foram as menores para todos os serviços em análise. Com a entubação naso-traqueal ocorre o oposto. Por conseguinte, na prestação de cuidados de saúde deve ser dada atenção à relação existente entre a utilização de dispositivos invasivos, em especial a entubação naso-traqueal, e a ocorrência de infeção nosocomial.

Os resultados anteriormente discutidos permitem verificar que os indicadores apresentados pelo sistema são extremamente importantes, uma vez que possibilitam o estudo da incidência de infeção nosocomial nas Unidades de Medicina do CHP. Estes auxiliam, por exemplo na identificação de: tipos de infeções mais frequentemente associados a infeções nosocomiais; serviços clínicos que possuem maior taxa deste tipo de infeção; dispositivos invasivos utilizados que poderão ter maior influência na ocorrência destas infeções; e fatores de risco subjacentes ao estado de saúde do doente que poderão ter maior influência na sua ocorrência. Portanto, o sistema permite caracterizar as infeções nosocomiais e monitorizá-las. Deste modo, o planeamento, a implementação e a avaliação das medidas utilizadas para prevenir e controlar estas infeções deverão considerar a informação disponibilizada por estes indicadores para obter melhores resultados. Através do conhecimento obtido com a análise dos indicadores, o sistema poderá, por conseguinte, auxiliar na prevenção e diminuição da taxa de incidência de infeção nosocomial nas Unidades de Medicina do CHP. Assim, poderá contribuir para a melhoria dos cuidados prestados e, consequentemente, do bem-estar e segurança dos doentes, bem como para a redução dos custos associados à ocorrência destas infeções.

Por outro lado, o sistema desenvolvido permite que o utilizador analise informações obtidas a partir de dados referentes a infeções nosocomiais de forma rápida e simples. Assim, a plataforma de BI pode beneficiar o CHP na medida em que, por exemplo: aumenta a flexibilidade e a autonomia dos profissionais de saúde responsáveis pelo o estudo de infeção na análise de informações; permite dar utilidade ao grande volume de dados clínicos armazenados; contribui para suportar a tomada de decisões clínicas; possibilita o estudo e monitorização de infeções nosocomiais.

Tal como já foi mencionado, atualmente a plataforma apresenta apenas indicadores relativos ao ano de 2013, mas no futuro poderá abranger outros períodos de tempo. Deste modo, será possível acompanhar a evolução da incidência de infeção nosocomial a longo prazo, bem como verificar os efeitos que as medidas de combate à infeção produzem na diminuição da taxa de ocorrência desta nas Unidades de Medicina do CHP.

O sistema implementado revela-se, assim, extremamente útil ao permitir que os dados clínicos sejam utilizados para extração de informações com métodos automatizados, em vez de serem simplesmente armazenados. No entanto, a implementação de um sistema de BI é um processo que requer muitas transformações nos dados e um planeamento cuidado da arquitetura e requisitos do sistema de acordo com os resultados que se pretendem obter no final. Verifica-se que a criação de *data marts* facilita o processo de consulta de dados e permite a realização de OLAP, disponibilizando dados de qualidade num formato que facilita o seu acesso e a aplicação de ferramentas analíticas. Utilizando um modelo dimensional como o que foi descrito anteriormente, os dados clínicos podem ser facilmente explorados por ferramentas de BI, de forma a obter os indicadores desejados. Constata-se ainda que processo ETL é crucial para o sucesso de um projeto de *data warehousing* e BI, uma vez que garante a qualidade e a consistência dos dados armazenados nos *data marts* e, consequentemente, a qualidade das informações extraídas destes.

Com este trabalho, verifica-se que ferramentas *open-source* de BI, tais como a ferramenta *Pentaho Community Edition*, podem ser utilizadas em SADC, mais concretamente no estudo de incidência de infeção nosocomial, sem, neste caso, resultarem em custos adicionais para as instituições de saúde. Estas ferramentas são úteis para manipular dados desta área e, através dos resultados que apresentam, permitem obter informações muito úteis para auxiliar na tomada de decisão dos profissionais de saúde que dependem das mesmas para realizar as tarefas associadas ao seu trabalho. Além disso, estas ferramentas possibilitam a apresentação das informações em *dashboards* apelativos e interativos, constituídos por gráficos, tabelas e botões, facilitando assim a interpretação das mesmas. Por sua vez, as ferramentas OLAP, como o *plug-in OpenI*, permitem a exploração de informações presentes nos dados do DW em tempo real, através de operações como *drill-down*, *pivot* e *slice and dice*. Desta forma, estas ferramentas facilitam a análise dos resultados por parte do utilizador final e permitem que este adeque os mesmos às suas necessidades analíticas (análises *ad hoc*). A integração de uma ferramenta OLAP no sistema revela-se então extremamente útil na exploração dos indicadores, possibilitando uma análise detalhada, rápida e interativa dos mesmos.

Por outro lado, a utilização de uma plataforma *web* para apresentar os resultados facilita a acessibilidade por parte dos seus utilizadores, permitindo que estes acedam à aplicação a partir de qualquer equipamento dentro do hospital e a qualquer momento, desde que estejam ligados à rede e tenham privilégios de acesso. Deste modo, assegura-se uma elevada disponibilidade e acessibilidade dos resultados e facilita-se a sua apresentação a todos os utilizadores finais. Por conseguinte, garante-se que os indicadores estão sempre disponíveis para auxiliar as decisões dos profissionais de saúde responsáveis pelo estudo de infeções nosocomiais.

Capítulo 6

Data Mining para Previsão de Infeções Nosocomiais

A tecnologia de DM pode ser aplicada à área da saúde para construir modelos capazes de realizar previsões em ambientes reais, utilizando, para isso, dados reais. No caso das infeções nosocomiais é importante saber quando estas poderão ocorrer. Este estudo pode ser efetuado através da aplicação da tecnologia de DM para prever a probabilidade de ocorrência de infeção na presença de certas variáveis. Desta forma, com este estudo de DM, pretende-se desenvolver modelos de DM capazes de fazer a classificação de um doente como pertencente a um grupo de risco passível de contrair uma infeção nosocomial, de acordo com os fatores de risco que descrevem a sua condição clínica. Estas previsões clínicas permitem que os profissionais de saúde associados ao estudo de infeções e à realização de ações nesse âmbito compreendam melhor a incidência de infeção e possam planear e implementar atempadamente as medidas preventivas adequadas.

Este módulo de DM para previsão de infeções nosocomiais faz parte do sistema de BI desenvolvido no âmbito desta dissertação (Capítulo 5). O melhor modelo obtido com este estudo de DM pode ser incluído num SADC, podendo ser utilizado para fazer a classificação de novos doentes. Desta forma, é capaz de auxiliar na identificação dos doentes mais propensos ao desenvolvimento de infeções nosocomiais. Por conseguinte, o melhor modelo

encontra-se integrado na plataforma de BI apresentada na secção 5.3.3.

6.1 Descrição do Estudo segundo CRISP-DM

Para fazer a previsão de variáveis recorrendo a DM é necessário seguir um conjunto de etapas que permitem o tratamento dos dados a utilizar, a aplicação de técnicas de DM para gerar os modelos de previsão e, posteriormente, a análise dos resultados obtidos. Neste estudo optou-se pela metodologia CRISP-DM (Secção 3.1.4) para implementar o processo de DCBD, uma vez que esta é a metodologia mais frequentemente utilizada neste tipo de projeto. Todas as operações de manipulação e armazenamento de dados necessárias para realizar o estudo foram realizadas numa base de dados *Oracle*, através da ferramenta *Oracle SQL Developer* (Secção 4.2.1).

6.1.1 Compreensão do Negócio

Tal como já foi referido, o principal objetivo da aplicação de DM aos dados relativos a infeções nosocomiais é prever a ocorrência destas infeções na presença de certos fatores de risco. Por outras palavras, pretende-se induzir modelos para prever o valor da variável associada à infeção nosocomial, mediante a condição clínica do doente.

O objetivo de DM deste estudo é, portanto, a previsão de infeções nosocomiais através da categorização de doentes e, para isso, é necessário recorrer a técnicas de DM para classificação de variáveis. A classificação consiste na previsão de uma variável alvo, que possui diferentes classes e permite mapear elementos de um conjunto de dados nessas classes predefinidas [39, 40].

O problema a resolver foi traduzido num problema de DM e formulado como "Qual a probabilidade de um doente não pertencer a um grupo de risco de ocorrência de infeção nosocomial, quando fatores de risco intrínseco ou fatores de risco extrínseco estão presentes na sua condição clínica?". Posteriormente, procedeu-se à seleção, à análise e ao pré-processamento dos dados capazes de conter uma relação entre as variáveis a estudar e a variável de infeção nosocomial. Depois, procedeu-se à indução de modelos de DM a

partir desses dados.

6.1.2 Estudo dos Dados

Tendo em consideração a questão formulada para o problema, selecionaram-se os dados capazes de conter uma relação entre as variáveis a estudar e a variável de infeção nosocomial. Neste estudo foi considerada uma amostra de dados constituída por formulários de infeção nosocomial do CHP registados na AIDA-PCE entre 30 de setembro de 2013 e 31 de dezembro de 2013. Para além disso, a análise inclui apenas as Unidades de Medicina, ou seja, os serviços clínicos Medicina A, Medicina B e Medicina C. Durante este período de 93 dias foram preenchidos 391 formulários de infeção nosocomial, dos quais 33 correspondem a casos em que esta ocorreu.

A qualidade das possíveis variáveis a utilizar no processo de DM foi analisada. Nem todos os atributos presentes nos formulários de infeção nosocomial registados são relevantes ou possuem qualidade suficiente para serem utilizados como variáveis no processo de DM. Deste modo, foi efetuada uma seleção cuidada de atributos a fim de escolher apenas os mais representativos para o estudo. Selecionaram-se os seguintes atributos:

- **Infeção Nosocomial**: variável alvo que dita o resultado do processo de diagnóstico simulado pelas técnicas de DM e possui dois valores possíveis ("Sim"ou "Não");

- **Idade, Sexo, Especialidade Clínica e Dias de Internamento**: variáveis que caracterizam o doente e o seu internamento;

- **Fatores de Risco**: variável que representa a presença ou a ausência de algum fator de risco intrínseco, como por exemplo, alcoolismo, diabetes, coma, HIV, ou qualquer outro fator anteriormente apresentado na secção 5.2.2;

- **Cateter Urinário, Cateter Periférico, Cateter Central, Entubação Naso-gástrica e Entubação Naso-traqueal**: variáveis que

representam a presença ou ausência de diferentes dispositivos invasivos (fatores de risco extrínseco) durante o período de internamento do doente.

As variáveis selecionadas modelam o problema em estudo, permitindo a previsão da variável alvo.

6.1.3 Preparação dos Dados

Após a seleção dos dados e das variáveis a utilizar na indução de modelos de DM, procedeu-se ao pré-processamento dos dados, uma etapa que permite a construção de um *dataset* com todos os casos de interesse, ao qual as técnicas de DM serão aplicadas. Esta etapa reduz o espaço de procura, pois elimina todos os valores nulos e ruído presentes nos dados, assim como colunas ou linhas sem interesse, deixando o *dataset* em análise apenas com os registos que têm relevância para o estudo. Desta forma, no final desta etapa o *dataset* ficou reduzido a 283 registos, dos quais apenas 26 correspondem a situações em que ocorreu infeção nosocomial.

Durante esta etapa, procedeu-se também à agregação de dados e de variáveis em classes de modo a modelar o problema mais corretamente, criando-se as seguintes classes:

- **Classe de Idade**: agregação da idade dos doentes em intervalos de idades, correspondentes a diferentes faixas etárias;
- **Entubação**: agregação de todos os dispositivos invasivos relacionados com entubação numa única classe (entubação naso-gástrica e entubação naso-traqueal);
- **Cateterismo**: agregação de todos os dispositivos invasivos relacionados com cateterismo numa única classe (cateter urinário, cateter periférico e cateter central).

Nesta etapa, foi ainda aplicado *oversampling* aos dados para replicar dados correspondentes aos casos em que a infeção nosocomial esteve presente. Esta técnica consiste na replicação dos dados da classe minoritária,

aumentando o peso da mesma. Este processo é necessário, uma vez que os classificadores tendem a produzir maiores erros na presença de classes minoritárias [60]. Neste caso recorreu-se a este procedimento pois a diferença entre registos referentes à ocorrência de infeção e registos referentes a casos em que não houve infeção era muito elevada, pelo que o significado da classe de doentes com infeção nosocomial podia perder-se devido à sua reduzida ocorrência na população em estudo. Desta forma, foi possível obter um número de registos com infeção aproximadamente igual ao número de registos sem infeção, que era bastante superior. Assim, a aplicação de *oversampling* permitiu obter um *dataset* com 517 registos.

Com esta etapa da metodologia CRISP-DM foram criados três *datasets* distintos: um com os dados sem replicação, outro com os dados replicados e outro com os dados replicados e a variável da idade agrupada em classes.

6.1.4 Modelação

Tal como referido na secção 3.1.4, existem várias técnicas ou algoritmos que podem ser utilizados em DM. Para este estudo são necessárias técnicas de classificação para obter os modelos de previsão, tendo sido utilizadas as técnicas *Support Vector Machines*, Árvores de Decisão e *Naïve Bayes*. Estas três técnicas de DM foram utilizadas para induzir automaticamente os modelos de classificação, com recurso à ferramenta *Oracle Data Miner* (Secção 4.2.3). A seleção das técnicas foi realizada considerando a sua eficiência e a facilidade de interpretação dos modelos que geram.

Support Vector Machines são algoritmos poderosos que se baseiam na teoria da aprendizagem estatística [42, 61]. Estes algoritmos encontram os vetores ou planos de decisão que melhor separam conjuntos de objetos pertencentes a diferentes classes. Estes possuem uma elevada capacidade de adaptação do modelo a novos dados e podem modelar problemas complexos [61].

As Árvores de Decisão baseiam-se em probabilidades condicionais. Segundo este algoritmo, o *dataset* é recursivamente dividido em subcategorias discretas, de modo a maximizar a distância entre as diferentes classes indi-

viduais. Deste modo, a árvore encontra-se estruturada numa sequência de questões, sendo que a resposta a essas questões descreve uma trajetória ao longo da árvore que culmina num valor que dita a previsão para a categoria [42, 61]. Este algoritmo é rápido e produz bons resultados e modelos interpretáveis [61].

A técnica *Naïve Bayes*, tal como as Árvores de Decisão, baseia-se em probabilidades condicionais. Este algoritmo realiza as previsões utilizando o Teorema de Bayes que calcula a probabilidade de ocorrência de um evento dada a probabilidade de um outro evento que já ocorreu, e tem como vantagem ser rápido e altamente escalável na construção do modelo [61].

Tendo em conta as variáveis citadas anteriormente e as possíveis combinações entre elas, consideraram-se vários cenários na construção dos modelos de DM. Os seguintes quatro cenários, combinações distintas entre diferentes variáveis e a variável alvo, foram considerados:

- **Ausência de Fatores de Risco (*Cenário 1*)**: todas as variáveis foram utilizadas no modelo, exceto a variável Fatores de Risco, isto é, foram consideradas as variáveis Idade ou Classe de Idade, Sexo, Especialidade Clínica, Dias de Internamento, Entubação e Cateterismo e a variável alvo Infeção Nosocomial;

- **Ausência de Entubação (*Cenário 2*)**: todas as variáveis foram utilizadas no modelo, exceto a variável Entubação, ou seja, foram consideradas as variáveis Idade ou Classe de Idade, Sexo, Especialidade Clínica, Dias de Internamento, Fatores de Risco e Cateterismo e a variável alvo Infeção Nosocomial;

- **Ausência de Cateterismo (*Cenário 3*)**: todas as variáveis foram utilizadas no modelo, exceto a variável Cateterismo, isto é, foram consideradas as variáveis Idade ou Classe de Idade, Sexo, Especialidade Clínica, Dias de Internamento, Fatores de Risco e Entubação e a variável alvo Infeção Nosocomial;

- **Todas as variáveis (*Cenário 4*)**: todas as variáveis foram utilizadas no modelo, ou seja, consideraram-se as variáveis Idade ou Classe de

Idade, Sexo, Especialidade Clínica, Dias de Internamento, Fatores de Risco, Entubação e Cateterismo e a variável alvo Infeção Nosocomial.

Estes cenários foram modelados segundo três abordagens distintas que correspondem aos três *datasets* de interesse criados na etapa anterior: abordagem que considera os dados sem replicação (*Abordagem A*), abordagem que considera o *dataset* composto por dados replicados (*Abordagem B*) e abordagem que considera o *dataset* com dados replicados e as idades agrupadas em classes (*Abordagem C*).

As técnicas de DM foram então aplicadas a todas as combinações de cenários e abordagens (situações), de modo a criar novo conhecimento e obter o melhor modelo para modelar o problema a resolver. Portanto, os quatro cenários foram modelados pelas três técnicas de classificação, pelos três *datasets* e por uma variável alvo, o que significa que, por fim, foram obtidos 36 modelos de DM.

Os modelos gerados analisam a relação entre as diferentes variáveis consideradas na sua construção, bem como o impacto destas no valor da variável alvo. Estes modelos podem ser representados pela seguinte expressão:

$$M_n \equiv < A_f, S_i, TDM_y > .$$

Segundo esta expressão, o Modelo de Previsão n (M_n) pertence à Abordagem de DM f (A_f) e é composto pela Situação i (S_i) e pela Técnica de DM y (TDM_y), onde para a *Abordagem A*

$$A_f \in \{Classificação\}$$

$$TDM_y \in \{Support\ Vector\ Machines, Árvores\ de\ Decisão, Naïve\ Bayes\}$$

$$S_i \in \{Cenário\ 1\ e\ Abordagem\ A,\ ....,\ Cenário\ 4\ e\ Abordagem\ A\},$$

para a *Abordagem B*

$$A_f \in \{Classificação\}$$

$$TDM_y \in \{Support\ Vector\ Machines, Árvores\ de\ Decisão, Naïve\ Bayes\}$$

$$S_i \in \{Cenário\ 1\ e\ Abordagem\ B,\ ....,\ Cenário\ 4\ e\ Abordagem\ B\}$$

e para a *Abordagem C*

$$A_f \in \{Classificação\}$$

$$TDM_y \in \{Support\ Vector\ Machines, Árvores\ de\ Decisão, Naïve\ Bayes\}$$

$$S_i \in \{Cenário\ 1\ e\ Abordagem\ C,\ ....,\ Cenário\ 4\ e\ Abordagem\ C\}.$$

Na construção dos modelos procedeu-se à separação dos dados em dados de treino e dados de teste, sendo que 70% dos dados foram utilizados para treino e os restantes 30% para teste.

6.1.5 Avaliação

Como resultado do processo de DM obtém-se um conjunto de modelos que devem ser avaliados para que seja possível aferir a sua qualidade e, deste modo, escolher o que permita obter melhores resultados e esteja de acordo com os objetivos iniciais do estudo. A técnica mais utilizada para fazer a avaliação dos modelos é a Matriz de Confusão, uma matriz que mede a correção das previsões efetuadas por um modelo [61].

No caso de uma variável alvo de duas classes existem quatro resultados distintos que podem ser contabilizados numa Matriz de Confusão (Tabela 6.1): Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) ou Falso Negativo (FN). Um resultado VP corresponde a um caso positivo corretamente classificado e um resultado FP refere-se a um caso incorretamente classificado como positivo. Por sua vez, um resultado VN diz respeito a um caso negativo corretamente classificado e um resultado FN corresponde a um caso erradamente classificado como negativo. Através destes resultados é possível aplicar um conjunto de métricas estatísticas para avaliar a qualidade dos modelos, sendo as métricas mais utilizadas a sensibilidade, a especificidade e a acuidade (Tabela 6.1).

Tabela 6.1: Matriz de confusão e expressões que definem a sensibilidade, a especificidade e a acuidade.

	Resultado Positivo	**Resultado Negativo**	**Sensibilidade**	**Especificidade**	**Acuidade**
Valor Positivo	VP	FN	$\frac{VP}{VP+FN}$	$\frac{VN}{VN+FP}$	$\frac{VP+VN}{VP+FP+VN+FN}$
Valor Negativo	FP	VN			

A sensibilidade é a capacidade do modelo detetar a ocorrência de um evento quando presente. Esta métrica é a proporção entre o número de resultados VP e o número total de valores positivos (VP+FN) [62].

Por sua vez, a especificidade diz respeito à capacidade do modelo classificar corretamente a não ocorrência de um evento. O seu valor corresponde ao rácio entre o número de resultados VN e todos os valores negativos (VN+FP) [62].

Por fim, a acuidade é a concordância entre os valores detetados corretamente e os valores reais. Esta medida é a proporção entre todos os resultados medidos corretamente (VP+VN) e todos os casos obtidos (VP+FP+VN+FN) [62].

A qualidade dos resultados obtidos com os modelos de DM gerados neste trabalho foi avaliada com estas medidas estatísticas. Neste caso, pretende-se fazer a previsão da não ocorrência da infeção nosocomial e, por isso, considerou-se a sensibilidade como a capacidade de prever a não ocorrência de infeção nosocomial e a especificidade como a capacidade de prever a ocorrência de infeção.

Alguns dos modelos gerados permitiram obter os quatro melhores resultados gerais para cada uma das técnicas de DM utilizadas (Tabela 6.2). Os melhores modelos foram selecionados com base no valor de sensibilidade, pois é importante identificar todas as não ocorrências de infeção nosocomial. Sabendo as não ocorrências de infeção nosocomial é possível considerar todas as outras previsões como grupos de risco passíveis de contrair uma infeção.

Tabela 6.2: Quatro melhores modelos para cada técnica de DM utilizada (adaptado de [63]).

Support Vector Machines			
	Especificidade	**Sensibilidade**	**Acuidade**
Cenário 1 e Abordagem B	0.763	0.919	0.838
Cenário 2 e Abordagem B	0.741	0.942	0.831
Cenário 1 e Abordagem C	0.731	0.793	0.766
Cenário 3 e Abordagem C	0.675	0.845	0.754
Árvores de Decisão			
	Especificidade	**Sensibilidade**	**Acuidade**
Cenário 1 e Abordagem C	0.855	0.798	0.818
Cenário 4 e Abordagem C	0.673	0.982	0.786
Cenário 4 e Abordagem B	0.673	0.982	0.786
Cenário 2 e Abordagem B	0.670	1	0.786
Naïve Bayes			
	Especificidade	**Sensibilidade**	**Acuidade**
Cenário 1 e Abordagem B	0.733	0.941	0.825
Cenário 2 e Abordagem B	0.733	0.941	0.825
Cenário 4 e Abordagem B	0.733	0.941	0.825
Cenário 4 e Abordagem C	0.733	0.941	0.825

6.1.6 Implementação

Após a avaliação dos modelos, é possível identificar a combinação de variáveis que mais provavelmente poderá causar uma infeção nosocomial e, deste modo, justificar as medidas de prevenção e controlo de infeção a implementar. Com os resultados da avaliação dos modelos, é também possível identificar os que apresentam qualidade e aplicá-los a novos dados. O conhecimento resultante desse processo pode ser utilizado para prever a ocorrência de infeções nosocomiais na presença dos fatores dos risco estudados. Por conseguinte, o melhor modelo obtido foi integrado na plataforma de BI apresentada na secção 5.3.3 para implementar um SADC capaz de fazer a previsão da não ocorrência de infeção em novos doentes, através do modelo gerado a partir de dados antigos. Esta parte da plataforma será apresentada na secção 6.3.

6.2 Discussão dos Resultados dos Modelos

Utilizando as técnicas de DM escolhidas foi possível obter resultados aceitáveis para cada modelo, sendo que os quatro melhores modelos obtidos para cada técnica utilizada se encontram na tabela 6.2.

É importante salientar que, neste caso, os modelos com percentagens de sensibilidade elevadas são capazes de prever corretamente a não ocorrência da variável alvo e, de facto, o comportamento ideal para os modelos de classificação seria que os valores de sensibilidade das previsões fossem, de um modo geral, superiores a 90%. Neste estudo, obtiveram-se valores de sensibilidade superiores a 91.90%, pelo que, de um modo geral, os resultados são aceitáveis e os modelos são capazes de prever, com bastante certeza, a não ocorrência de uma infeção.

A melhor combinação de cenário e abordagem para todas as técnicas de DM utilizadas foi a situação *Cenário 2 e Abordagem B*, pois esta possui o valor de sensibilidade mais elevado para todas as técnicas. Esta situação também possui valores de acuidade elevados para todas as técnicas.

O melhor valor de sensibilidade foi 100% e foi obtido para a situação *Cenário 2 e Abordagem B* com a técnica Árvores de Decisão. Deste modo,

conclui-se que, para este modelo, a não ocorrência de infeção pode ser prevista com muita certeza. Portanto, de todas as combinações de situações e técnicas, este é o modelo que melhor prevê a não ocorrência de infeção nosocomial. Tendo em consideração que o valor de sensibilidade do modelo é 100%, espera-se que todos os casos, exceto os classificados como não infetados, correspondam a grupos de risco suscetíveis de contrair infeção. Neste caso obtiveram-se valores de acuidade e especificidade de 78.60 % e 67%, respetivamente.

A tabela 6.3 apresenta, para cada um dos algoritmos aplicados à situação *Cenário 2 e Abordagem B*, o número de casos que o modelo classificou incorreta e corretamente, assim como a percentagem de casos corretamente classificados.

Tabela 6.3: Número de casos incorreta e corretamente classificados e percentagem de casos corretamente classificados para a situação *Cenário 2 e Abordagem B*, em cada um dos algoritmos de DM aplicados.

	Incorreto	**Correto**	**% de Corretos**
Support Vector Machines	26	128	83.12
Árvores de Decisão	33	121	78.57
Naïve Bayes	27	127	82.47

De acordo com os resultados apresentados na tabela 6.3, o algoritmo mais eficiente aplicado à situação *Cenário 2 e Abordagem B* foi o algoritmo *Support Vector Machines*, pois permitiu obter a maior percentagem de respostas corretas (83.12%). Apesar desta situação possuir o maior valor de sensibilidade quando modelada com Árvores de Decisão, possui também a menor especificidade, pelo que, nesse caso, o seu valor de acuidade não é muito elevado.

De um modo geral, os valores de especificidade são aceitáveis, sendo que o menor foi 67% e o maior valor foi 85.50%. Portanto, os modelos são capazes de detetar a ocorrência de uma infeção nosocomial, ainda que essa classificação possa introduzir algum erro. O melhor valor de especificidade (85.50%) foi obtido para a situação *Cenário 1 e Abordagem C* com a técnica de DM

Árvores de Decisão. Neste caso, um valor de acuidade de 81.80% e um valor de sensibilidade de 79.80% foram obtidos.

A tabela 6.4 apresenta, para cada um dos algoritmos aplicados à situação *Cenário 1 e Abordagem C*, o número de casos que o modelo classificou incorreta e corretamente, assim como a percentagem de casos corretamente classificados.

Tabela 6.4: Número de casos incorreta e corretamente classificados e percentagem de casos corretamente classificados para a situação *Cenário 1 e Abordagem C*, em cada um dos algoritmos de DM utilizados (adaptado de [63]).

	Incorreto	**Correto**	**% de Corretos**
Support Vector Machines	36	118	76.62
Árvores de Decisão	28	126	81.82
Naïve Bayes	30	124	80.52

De acordo com os resultados apresentados na tabela 6.4, o algoritmo mais eficiente aplicado à situação *Cenário 1 e Abordagem C* foi o algoritmo Árvores de Decisão, pois permitiu obter a maior percentagem de respostas corretas (81.82%).

O valor mais baixo de acuidade foi 75.40% e o mais elevado foi 83.80% o que significa que, de um modo geral, existe concordância entre os valores corretamente detetados e o valor real.

Verificou-se ainda que, de um modo geral, os valores de sensibilidade são superiores aos valores de especificidade, o que significa que os modelos obtidos são melhores a prever a não ocorrência de uma infeção nosocomial do que a prever os casos em que a infeção está presente.

Este estudo demonstrou que, através da utilização de técnicas de classificação e de dados antigos de doentes internados nas Unidades de Medicina do CHP, é possível obter modelos de DM de classificação, capazes de prever se um doente internado nestes serviços adquirirá ou não uma infeção nosocomial. Os modelos obtidos são satisfatórios e podem auxiliar a tomada de decisão dos profissionais de saúde responsáveis pelo estudo de infeções

nosocomiais, permitindo a aplicação de medidas preventivas adequadas, necessárias para assegurar o bem-estar e segurança dos doentes em risco.

Sendo as infeções nosocomiais um problema de extrema relevância para todas as instituições de saúde, convém registar que este estudo de DM pode ser adaptado para gerar modelos de previsão destas infeções relativos a outras instituições.

6.3 *Dashboard Previsão de Infeções Nosocomiais*

Para integrar o módulo de DM na plataforma de BI, selecionou-se o modelo correspondente à situação *Cenário 2 da Abordagem B* para prever a probabilidade de não ocorrência de infeção nosocomial, pois este foi o melhor modelo obtido para estes casos.

Ainda utilizando a ferramenta *Oracle Data Miner* (Secção 4.2.3), foi efetuada a previsão da variável da infeção nosocomial em dados registados com os formulários de infeção nosocomial durante todo ano de 2013. O resultado da previsão foi uma tabela com o valor previsto, a probabilidade da previsão ser "Não", bem como o valor de uma série de atributos que caracterizam o doente e que foram utilizados na construção do modelo de classificação utilizado (Especialidade Clínica, Idade, Sexo, Dias de Internamento, Fatores de Risco, Cateterismo e Entubação).

Com a ferramenta CDE (Secção 4.3.2) criou-se um *dashboard* para apresentar os resultados obtidos com as previsões de DM. Através de uma *combo box*, este *dashboard* permite que o utilizador escolha o episódio a prever. Um episódio corresponde a um código atribuído ao doente em cada situação de internamento.

O resultado da previsão efetuada é apresentado num gráfico (Figura 6.1). Este expõe a probabilidade de não ocorrência de uma infeção para o episódio selecionado, através de um ponteiro que indica o valor previsto para essa probabilidade. Se o resultado apresentado for 100, significa que há uma grande probabilidade do doente associado ao episódio em questão não contrair uma

infeção nosocomial. Pelo contrário, se este valor for inferior a 100 significa que o doente pertence a um grupo de risco passível de contrair uma infeção nosocomial, sendo que quanto mais próximo de zero o ponteiro se encontrar maior o risco do doente contrair uma infeção.

Figura 6.1: Excerto do *dashboard* *Previsão de Infeções Nosocomiais*: gráfico da probabilidade de não ocorrência de infeção, prevista para o episódio selecionado.

No caso do exemplo da figura 6.1, verifica-se que o doente associado ao episódio selecionado (episódio 12033354), possui cerca de 80% de probabilidade de não adquirir uma infeção nosocomial.

Neste *dashboard* é ainda apresentada uma tabela com algumas características do doente associado ao episódio em análise (Figura 6.2).

Figura 6.2: Excerto do *dashboard* *Previsão de Infeções Nosocomiais*: características do doente associado ao episódio selecionado.

No exemplo apresentado na figura 6.2 verifica-se que o episódio selecionado (12033354) está ainda associado a um doente de 76 anos, do sexo feminino e que esteve internado 56 dias em Medicina A. Esta doente apresentou fatores de risco, foi submetida a cateterismo, mas não foi sujeita a entubação.

Este *dashboard* auxilia os seus utilizadores a identificar os doentes que pertencem a grupos de risco passíveis de contrair uma infeção nosocomial. Deste modo, estes podem constatar que doentes devem ser mais acompanhados e as aplicar medidas preventivas adequadas. Por conseguinte, torna-se possível reduzir os custos associados à ocorrência de infeções, bem como assegurar o bem-estar e segurança desses doentes.

É importante mencionar que, neste trabalho, o melhor modelo de DM obtido foi utilizado apenas em dados de 2013. No entanto, verificou-se que os resultados apresentados pelo *dashboard Previsão de Infeções Nosocomiais* poderão ser muito úteis em ambiente hospitalar e, desse modo, o modelo deverá ser aplicado a dados atuais com o intuito de fazer a previsão da ocorrência de infeções em tempo real. Este modelo poderá também ser aplicado a dados relativos a infeções nosocomiais verificadas noutras especialidades clínicas do CHP.

Capítulo 7

Conclusões

Conclui-se que a compreensão e a monitorização das atividades e dos processos que decorrem em ambiente hospitalar são tarefas fundamentais para a descoberta de problemas e oportunidades de melhoria, sendo que a análise de dados clínicos permite realizar estas tarefas. A tecnologia de BI apresenta-se como um método automatizado eficiente e adequado para integrar e explorar o grande volume de dados clínicos recolhidos pelas instituições de saúde. Assim, a aplicação de BI à área da saúde é fundamental para as instituições de saúde e para o auxílio do processo de tomada de decisão.

7.1 Contributos

Neste trabalho implementou-se um sistema para o estudo da incidência de infeção nosocomial nas Unidades de Medicina do CHP. O sistema foi desenvolvido através da aplicação de conceitos e ferramentas de BI, sendo constituído por dois *data marts* e uma plataforma de BI. Os dados são extraídos das fontes de dados e adequadamente armazenados nos *data marts*. Posteriormente, a ferramenta *Pentaho Community Edition* extrai indicadores dos dados dos *data marts*, sendo que estes parâmetros sumariam informações importantes presentes nos dados relativos a infeções nosocomiais. Estes indicadores permitem ainda a caracterização e o estudo da incidência destas infeções, bem como a análise da relação entre as mesmas e certos fatores

considerados de risco. Através da plataforma, o sistema apresenta estes indicadores, permitindo que o utilizador faça uma análise interativa e em tempo real dos mesmos. As informações apresentadas pelo sistema possuem elevada qualidade porque se baseiam em dados clínicos, cuidadosamente extraídos das fontes e transformados.

A solução proposta neste trabalho demonstrou ser um método automatizado, eficiente para tratar e explorar os dados de infeções nosocomiais. Permitiu ainda estudar a incidência destas infeções no CHP, uma vez que a plataforma de BI apresenta indicadores relevantes para esse estudo, tais como a percentagem de infeções nosocomiais por serviço e ano ou a percentagem de infeções por tipo de infeção, serviço e ano. Através da análise destes indicadores, verificou-se que, por exemplo, a percentagem de infeções nosocomiais nas Unidades de Medicina em 2013 variou entre 9.43% e 12.95% e as infeções do trato urinário e as infeções respiratórias foram as infeções mais frequentemente associadas a infeções nosocomiais.

Tal como inicialmente proposto, a plataforma de BI desenvolvida permite que os profissionais de saúde responsáveis pelo estudo de infeções tenham uma maior autonomia e flexibilidade na análise dos dados e que sejam capazes de analisar e interpretar rápida e facilmente as informações extraídas destes. Portanto, a plataforma facilita o trabalho destes profissionais, sendo que estes podem utilizá-la para monitorizar as infeções nosocomiais, identificar os fatores de risco que contribuem fortemente para a ocorrência destas infeções, bem como planear medidas de controlo e prevenção de infeção específicas e mais orientadas para as necessidades reais de cada serviço clínico. Assim, para prevenir e diminuir eficientemente a incidência de infeção nosocomial, a implementação de medidas de combate à infeção deverá considerar as informações apresentadas pela plataforma.

Conclui-se que ferramentas *open-source* de BI, como o *Pentaho Community Edition* e o *plug-in OpenI*, são uma grande ajuda na exploração de dados clínicos, porque permitem a criação de novo conhecimento em tempo real, sem que a sua utilização resulte em custos adicionais para a organização.

A solução proposta para a criação e apresentação de indicadores de infeção nosocomial pode ser aplicada a outros dados da área da saúde ou utilizada

para gerar outros indicadores relacionados com esta infeção, pois este sistema foi concebido tendo em consideração a eventual necessidade da sua expansão. Atualmente, a plataforma disponibiliza apenas informações referentes ao ano de 2013, mas no futuro o sistema poderá conter dados de outros períodos de tempo, sendo que o principal objetivo será ter dados em tempo real. Desta forma, será possível acompanhar a evolução da infeção nosocomial a longo prazo e avaliar os efeitos das medidas de combate à infeção implementadas. Esta solução é ainda válida para dados de outras instituições de saúde.

Através da utilização de dados reais de infeção nosocomial do CHP, o módulo de DM demonstrou que é possível obter modelos de DM para classificação, capazes de prever se um doente que apresente certos fatores de risco contrairá ou não uma infeção nosocomial. Neste estudo de DM foi seguida a metodologia CRISP-DM e foram aplicadas três técnicas de classificação distintas: *Support Vector Machines*, Árvores de Decisão e *Naïve Bayes*. Tal como referido em [63], conclui-se que a utilização de técnicas de classificação e de dados clínicos reais possibilita a previsão de infeções nosocomiais, tendo-se obtido modelos capazes de prever bastante bem a não ocorrência destas infeções (sensibilidades superiores a 91.90%). A integração do melhor modelo de DM obtido na plataforma de BI é uma mais-valia, uma vez que permite a sua utilização para classificar doentes. Assim, a criação de modelos de previsão da ocorrência infeção nosocomial pode ser vista como um grande contributo para o desenvolvimento de SADC para esta área.

Este trabalho permitiu estudar a aplicabilidade da tecnologia de BI à área da saúde e demonstrou a utilidade e importância dos sistemas de BI no tratamento e na análise de dados clínicos, mais concretamente no estudo de infeção nosocomial. É fundamental que estas infeções sejam prevenidas e controladas, pois o seu diagnóstico e tratamento implica custos adicionais para as unidades de saúde, e podem colocar em causa a segurança e o bem-estar dos doentes ou profissionais de saúde. Desse modo, o tratamento de dados e extração de informações destes é um método eficaz para caracterizar as infeções e identificar atividades e fatores de risco que contribuem para a sua ocorrência. Em resposta à Questão 1, verificou-se que a tecnologia de BI pode ser aplicada ao estudo de infeção nosocomial com grande sucesso e utili-

dade. Conclui-se que atualmente a sua implementação, apesar de não ser um processo simples, constitui-se preponderante para as instituições de saúde, uma vez que resulta em informações de qualidade, atempadas e estratégicas que podem ser utilizadas para tomar decisões mais rápidas e fundamentadas, melhorando, assim, o fluxo de trabalho diário da instituição.

Como resposta à Questão 2, um sistema de BI, composto por *dashboards* que apresentam indicadores adequados e úteis e uma componente de DM capaz de prever a ocorrência de infeção, pode disponibilizar informações muito úteis e relevantes para apoiar a tomada de decisão dos seus utilizadores, possibilitando a monitorização, a análise e a previsão de infeções nosocomiais. Deste modo, conclui-se que o sistema desenvolvido é capaz atuar como SADC, podendo auxiliar os profissionais de saúde responsáveis pelo estudo de infeções nosocomiais nas suas decisões.

Em resposta à Questão 3 considera-se que, de um modo geral, o trabalho apresentado nesta dissertação pode ser visto como uma mais-valia não só para o CHP, como também para a sociedade em geral. Isto acontece porque, tal como concluído em [63], o sistema de BI desenvolvido é capaz de auxiliar na prevenção e diminuição da incidência de infeções nosocomiais em instituições de saúde, diminuindo, dessa forma, o risco de complicações para os doentes e melhorando o seu bem-estar e segurança.

É importante mencionar que o trabalho desenvolvido nesta dissertação deu origem a três publicações. A primeira, um capítulo de livro (Anexo B.1), foca a importância da utilização da tecnologia de BI na área da saúde para o auxílio do processo de tomada de decisão, através da utilização de técnicas de DM para gerar modelos de classificação e da apresentação de um conjunto de indicadores de infeção nosocomial. A segunda, um capítulo de livro (Anexo B.2), demonstra a utilidade da incorporação da tecnologia de DM em sistemas de BI para prever a ocorrência de infeções nosocomiais. Por último, um artigo (Anexo B.3) que discute a importância da plataforma de BI na apresentação de indicadores capazes de auxiliar o estudo da incidência de infeção nosocomial.

7.2 Trabalho Futuro

Apesar da plataforma implementada possuir todas as funcionalidades inicialmente propostas, considera-se que existem aspetos que podem ser melhorados. Neste sentido, pensa-se que seria interessante apresentar os indicadores de infeção nosocomial com outras ferramentas de BI, nomeadamente outras ferramentas OLAP e outras ferramentas para criar *dashboards*, a fim de comparar as funcionalidades e o desempenho das mesmas, bem como concluir qual a mais adequada à realidade deste trabalho.

De modo a explorar mais profundamente as potencialidades do sistema de BI, seria interessante estender o estudo a outros anos, serviços clínicos e/ou até outros indicadores com interesse para a análise da incidência de infeção nosocomial. Isto pode ser feito, por exemplo, através da criação de novos *data marts*, alteração dos *data marts* criados e integração de novos *dashboards* na plataforma. A integração de dados atuais nos *data marts* permitirá ainda disponibilizar informação em tempo real, o que auxilia o processo de tomada de decisão no momento.

Em relação ao estudo de DM para previsão de infeções nosocomiais, considera-se relevante a repetição do estudo para outros dados e outras técnicas de DM, assim como a incorporação de outras variáveis nos modelos de previsão. O melhor modelo de DM obtido deverá também ser aplicado a dados atuais para prever casos futuros de ocorrência de infeções nosocomiais, permitindo, dessa forma, a obtenção de resultados em tempo real.

Seria interessante automatizar mais o processo ETL e agendar a sua execução periódica, de forma que os dados presentes nos *data marts* estejam sempre atualizados e permitam a obtenção de informações atuais. Para além disso, o acesso à plataforma de BI deverá ser disponibilizado aos seus utilizadores para que estes possam usufruir das funcionalidades da mesma. A plataforma criada poderá ainda ser integrada na AIDA-BI, um módulo de BI da plataforma AIDA que é utilizado para extrair e apresentar informações de interesse para o CHP.

Por último, a avaliação da usabilidade e da funcionalidade da plataforma seria uma mais-valia, pois permitiria encontrar aspetos a melhorar, quer ao

nível do desempenho do sistema, quer ao nível das funcionalidades da plataforma que os utilizadores considerem importante disponibilizar.

Bibliografia

[1] E. S. Berner and T. J. La Lande, "Overview of clinical decision support systems," in *Clinical Decision Support Systems: Theory and Practice*, 2nd ed., ser. Health Informatics Series, E. S. Berner, Ed. New York, NY, USA: Springer, 2007, pp. 3–22.

[2] K. Inweregbu, J. Dave, and A. Pittard, "Nosocomial infection," *Continuing Education in Anaesthesia, Critical Care and Pain*, vol. 5, no. 1, pp. 14–17, 2005.

[3] H. Rigor, J. Machado, A. Abelha, J. Neves, and C. Alberto, "A web-based system to reduce the nosocomial infection impact in healthcare units," in *Proceedings of the WEBIST 2008 - International Conference on Web Information Systems*, Funchal, Portugal, 2008, pp. 264–268.

[4] Clean Care is Safer Care Team, "Report on the burden of endemic health care-associated infection worldwide: A systematic review of the literature," World Health Organization, Geneva, Switzerland, Tech. Rep., 2011, [Online]. Available: http://apps.who.int/iris/bitstream/10665/80135/1/9789241501507_eng.pdf?ua=1. [Accessed on June 15, 2014].

[5] N. N. Damani, *Manual of Infection Control Procedures*, 2nd ed. New York, NY, USA: Greenwich Medical Media, 2003.

[6] Hospital do Futuro. (2014) Infecção hospitalar: Um problema do mundo, um problema de todos. [Online]. Available: http://www.hospitaldofuturo.com/profiles/blogs/1967198:BlogPost:42140. [Accessed on June 15, 2014].

[7] R. Lenz and M. Reichert, "IT support for healthcare processes - Premises, challenges, perspectives," *Data & Knowledge Engineering*, vol. 61, no. 1, pp. 39–58, 2007.

[8] M. Garrouste-Orgeas, F. Philippart, C. Bruel, A. Max, N. Lau, and B. Misset, "Overview of medical errors and adverse events," *Annals of Intensive Care*, vol. 2, no. 1, pp. 2–9, 2012.

[9] R. Haux, "Health information systems – Past, present, future," *International Journal of Medical Informatics*, vol. 75, no. 3-4, pp. 268–281, 2006.

[10] R.-F. Chen and J.-L. Hsiao, "An investigation on physicians' acceptance of hospital information systems: A case study," *International Journal of Medical Informatics*, vol. 81, no. 12, pp. 810–820, 2012.

[11] L. M. Prevedello, K. P. Andriole, R. Hanson, P. Kelly, and R. Khorasani, "Business intelligence tools for radiology: Creating a prototype using open-source tools," *Journal of Digital Imaging*, vol. 23, no. 2, pp. 133–141, 2010.

[12] L. Cardoso, F. Marins, C. Quintas, F. Portela, M. Santos, A. Abelha, and J. Machado, "Interoperability in healthcare," in *Cloud Computing Applications for Quality Health Care Delivery*, A. Moumtzoglou and A. Kastania, Eds. Hershey, PA, USA: IGI Global Book, 2014, pp. 78–101.

[13] J. Duarte, M. Miranda, A. Abelha, M. Santos, J. Machado, J. Neves, C. Alberto, M. Salazar, C. Quintas, A. Ferreira, and J. ao Neves, "Agent-based group decision support in medicine," in *Proceedings of the 2009 International Conference on Artificial Intelligence, ICAI 2009*, Las Vegas, NV, USA, 2009, pp. 115–121.

[14] M. Lluch, "Healthcare professionals' organisational barriers to health information technologies — A literature review," *International Journal of Medical Informatics*, vol. 80, no. 12, pp. 849–862, 2011.

[15] T. Mettler and V. Vimarlund, "Understanding business intelligence in the context of health care," *Health Informatics Journal*, vol. 15, no. 3, pp. 254–264, 2009.

[16] L. Cardoso, F. Marins, F. Portela, M. Santos, A. Abelha, and J. Machado, "The next generation of interoperability agents in healthcare," *International Journal of Environmental Research and Public Health*, vol. 11, no. 5, pp. 5349–5371, 2014.

[17] N. Foshay and C. Kuziemsky, "Towards an implementation framework for business intelligence in healthcare," *International Journal of Information Management*, vol. 34, no. 1, pp. 20–27, 2014.

[18] Á. Rebuge and D. R. Ferreira, "Business process analysis in healthcare environments: A methodology based on process mining," *Information Systems*, vol. 37, no. 2, p. 99–116, 2012.

[19] M. Spruit, R. Vroon, and R. Batenburg, "Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands," *Computers in Human Behavior*, vol. 30, no. 0, pp. 698 – 707, 2013.

[20] L. M. Prevedello, K. P. Andriole, and R. Khorasani, "Business intelligence tools and performance improvement in your practice," *Journal of the American College of Radiology*, vol. 5, no. 12, pp. 1210–1211, 2008.

[21] D. J. Power, "Understanding data-driven decision support systems," *Information Systems Management*, vol. 25, no. 2, pp. 149–154, 2008.

[22] M. Ghazanfari, M. Jafari, and S. Rouhani, "A tool to evaluate the business intelligence of enterprise systems," *Scientia Iranica*, vol. 18, no. 6, pp. 1579–1590, 2011.

[23] J. Glaser and J. Stone, "Effective use of business intelligence," *Healthcare Financial Management*, vol. 62, no. 2, pp. 68–72, 2008.

[24] W. Bonney, "Applicability of business intelligence in electronic health record," *Procedia - Social and Behavioral Sciences*, vol. 73, no. 0, pp. 257 – 262, 2013.

[25] A. Popovič, R. Hackney, P. S. Coelho, and J. Jaklič, "Towards business intelligence systems success: Effects of maturity and culture on analytical decision making," *Decision Support Systems*, vol. 54, no. 1, pp. 729 – 739, 2012.

[26] B. Hočevar and J. Jaklič, "Assessing benefits of business intelligence systems - A case study," *Management: Journal of Contemporary Management Issues*, vol. 15, no. 1, pp. 87–119, 2010.

[27] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *SIGMOD Record*, vol. 26, no. 1, pp. 65–74, 1997.

[28] D. Loshin, *Business Intelligence: The Savvy Manager's Guide*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.

[29] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Communications of the ACM*, vol. 54, no. 8, pp. 88–98, 2011.

[30] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. E. Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University – Computer and Information Sciences*, vol. 23, no. 2, pp. 91–104, 2011.

[31] J. Ferreira, M. Miranda, A. Abelha, and J. Machado, "O processo ETL em sistemas data warehouse," in *INForum 2010 - II Simpósio de Informática*, Braga, Portugal, 2010, p. 757 – 765.

[32] W. H. Inmon, *Building the Data Warehouse*, 3rd ed. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[33] S. Chaudhuri, U. Dayal, and V. Ganti, "Database technology for decision support systems," *Computer*, vol. 34, no. 12, pp. 48–55, 2001.

[34] T. Thalhammer, M. Schrefl, and M. Mohania, "Active data warehouses: Complementing OLAP with analysis rules," *Data & Knowledge Engineering*, vol. 39, no. 3, pp. 241–269, 2001.

[35] 1Keydata. (2014) Data warehousing concepts. [Online]. Available: http://www.1keydata.com/datawarehousing/concepts.html. [Accessed on June 31, 2014].

[36] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN, USA: John Wiley & Sons, Inc., 2013.

[37] OLAP Council. (1997) OLAP and OLAP server definitions - OLAP: On-line analytical processing. [Online]. Available: http://www.olapcouncil.org/research/glossaryly.htm. [Accessed on July 31, 2014].

[38] H. Baars and H.-G. Kemper, "Management support with structured and unstructured data - An integrated business intelligence framework," *Information Systems Management*, vol. 25, no. 2, pp. 132–148, 2008.

[39] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management*, vol. 19, no. 2, pp. 64–72, 2005.

[40] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[41] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," in *Proceedings of the IADIS European Conference on Data Mining 2008*, Amesterdam, Netherlands, 2008, pp. 182–185.

[42] V. Paramasivam, T. S. Yee, S. K. Dhillon, and A. S. Sidhu, "A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 3, pp. 139–145, 2014.

[43] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," Tech. Rep., 2000, [Online]. Available: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf. [Accessed on June 4, 2014].

[44] M. M. Horvath, S. Winfield, S. Evans, S. Slopek, H. Shang, and J. Ferranti, "The DEDUCE guided query tool: Providing simplified access to clinical data for research and quality improvement," *Journal of Biomedical Informatics*, vol. 44, no. 2, pp. 266–276, 2011.

[45] P. G. Nagy, M. F. Warnock, M. Daly, C. Toland, C. D. Meenan, and R. S. Mezrich, "Informatics in radiology: Automated web-based graphical dashboard for radiology operational business intelligence," *RadioGraphics*, vol. 29, no. 7, pp. 1897–1906, 2009.

[46] R. L. Baskerville, "Investigating information systems with action research," *Communications of the Association for Information Systems*, vol. 2, no. 3, 1999.

[47] Oracle. (2014) Oracle SQL developer documentation. [Online]. Available: http://docs.oracle.com/cd/E39885_01/index.htm. [Accessed on June 4, 2014].

[48] Oracle. Oracle data miner. [Online]. Available: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/dataminerworkflow-168677.html. [Accessed on June 4, 2014].

[49] B. Janamanchi, E. Katsamakas, W. Raghupathi, and W. Gao, "The state and profile of open source software projects in health and medical informatics," *International Journal of Medical Informatics*, vol. 78, no. 7, pp. 457–472, 2009.

[50] J. Marsan and G. Paré, "Antecedents of open source software adoption in health care organizations: A qualitative survey of experts in Canada," *International Journal of Medical Informatics*, vol. 82, no. 8, pp. 731–741, 2013.

[51] Engeneering Group. (2014) Spagobi: the 100% open source, complete and flexible business intelligence suite. [Online]. Available: http://www.spagoworld.org/spw-resources/Presentations/SpagoBI-ENG-May2014.pdf. [Accessed on Sept. 2, 2014].

[52] M. Tereso and J. Bernardino, "Open source business intelligence tools for SMEs," in *2011 6th Iberian Conference on Information Systems and Technologies (CISTI)*, Chaves, Portugal, 2011, pp. 1–4.

[53] M. Golfarelli, "Open source bi platforms: A functional and architectural comparison," in *Data Warehousing and Knowledge Discovery*, ser. Lecture Notes in Computer Science, T. Pedersen, M. Mohania, and A. Tjoa, Eds. Springer, 2009, vol. 5691, pp. 287–297.

[54] Jaspersoft. (2014) Jaspersoft community. [Online]. Available: https://community.jaspersoft.com/. [Accessed on Sept. 16, 2014].

[55] Webdetails. CDE: Community dashboard editor. [Online]. Available: http://www.webdetails.pt/ctools/cde.html. [Accessed on Sept. 16, 2014].

[56] Pentaho. (2006) Mondrian documentation. [Online]. Available: http://mondrian.pentaho.com/documentation/olap.php. [Accessed on Sept. 15, 2014].

[57] C. Thomsen and T. Pedersen, "A survey of open source tools for business intelligence," in *Data Warehousing and Knowledge Discovery*, ser. Lecture Notes in Computer Science, A. Tjoa and J. Trujillo, Eds. Springer, 2005, vol. 3589, pp. 74–84.

[58] Pentaho. (2014) Tour the data source model editor. [Online]. Available: https://help.pentaho.com/Documentation/5.1/0L0/0A0/070. [Accessed on Sept. 1, 2014].

[59] OpenI. (2012) OpenI 3.0.1 is here (pentaho plugin for olap data visualization). [Online]. Available: http://openi.org/2012/openi-3-0-1-is-here-pentaho-plugin-for-olap-data-visualization/. [Accessed on July 31, 2014].

[60] S. Barua, M. M. Islam, and K. Murase, "ProWSyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, J. Pei, V. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Springer, 2013, vol. 7819, pp. 317–328.

[61] Oracle. (2008) Oracle data mining concepts. [Online]. Available: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm. [Accessed on June 4, 2014].

[62] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning - Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, vol. 30, no. 2-3, pp. 271–274, 1998.

[63] E. Silva, A. Alpuim, L. Cardoso, F. Marins, C. Quintas, C. F. Portela, M. F. Santos, J. Machado, and A. Abelha, "Business intelligence and nosocomial infection decision making," in *Integration of Data Mining in Business Intelligence Systems*, A. Azevedo and M. F. Santos, Eds. Hershey, PA, USA: IGI Global Book, 2014, pp. 196–218.

Anexo A

Resultados da Plataforma de *Business Intelligence*

A.1 *Dashboard* Inicial

Figura A.1: Excerto do *dashboard* inicial (Parte 1/2).

Figura A.2: Excerto do *dashboard* inicial (Parte 2/2).

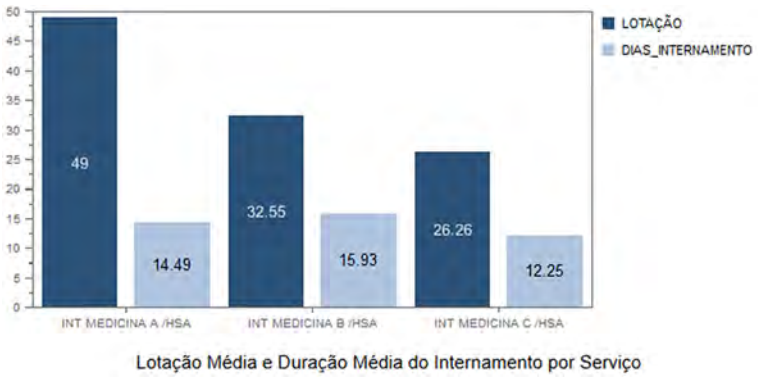

Figura A.3: Excerto do *dashboard* inicial: lotação média e duração média do internamento, por serviço em 2013.

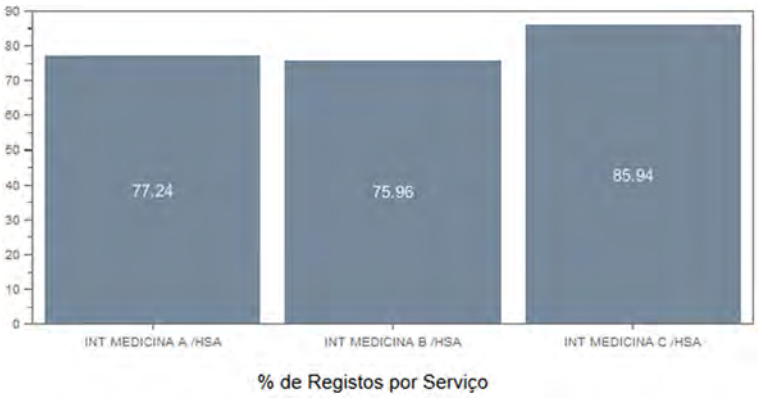

Figura A.4: Excerto do *dashboard* inicial: percentagem de registos de infeção nosocomial, por serviço em 2013.

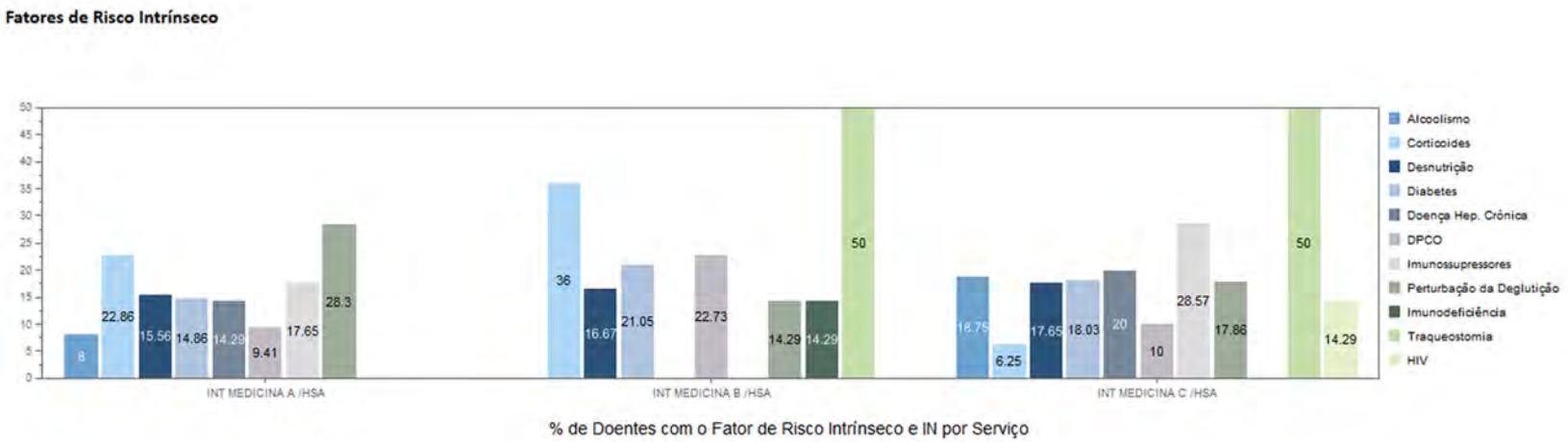

Figura A.5: Excerto do *dashboard* inicial: percentagem de doentes com o fator de risco intrínseco e infeção, por serviço e fator de risco intrínseco em 2013.

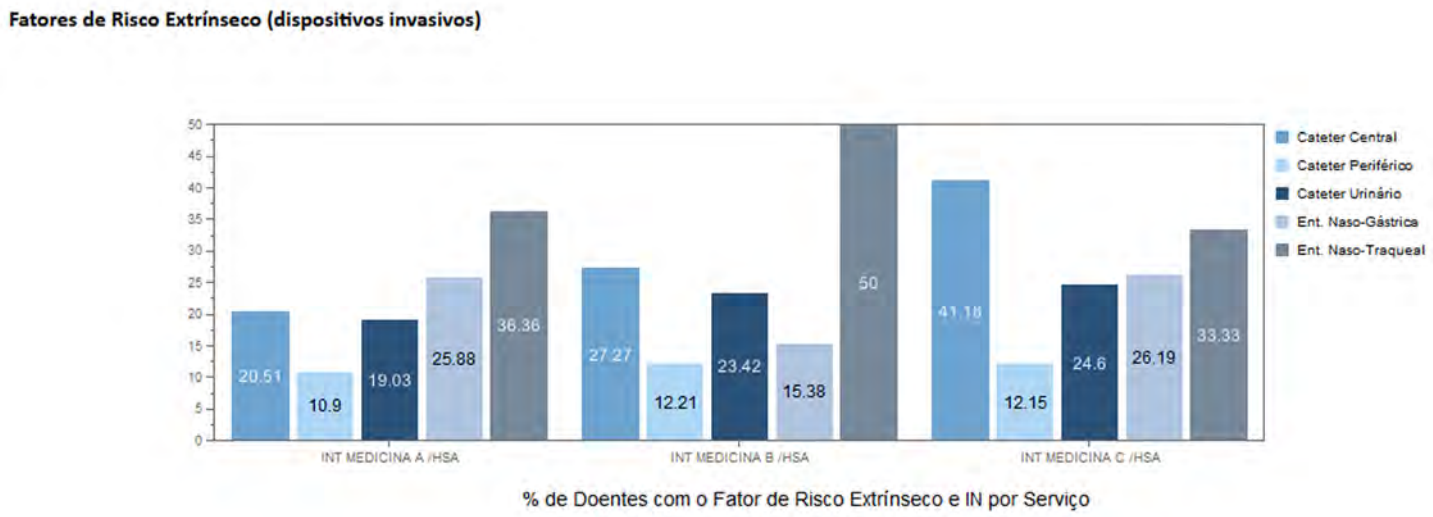

Figura A.6: Excerto do *dashboard* inicial: percentagem de doentes com o fator de risco extrínseco e infeção, por serviço e fator de risco extrínseco em 2013.

A.2 *Dashboard Fatores de Risco e Infeção Nosocomial*

Fatores de Risco Intrínseco por Data e por Serviço

Swap Axes | Hide Empty Rows/Cols | Hierarchy | ✓ Show Table | Show Chart | Export | Configure

Table

Drill: ✓ Hierarchy | Replace | Data Report | Slicers :

Data	Especialidade	Fator de risco	Measures: Total de doentes	Total de doentes com IN	% de doentes com IN
All Datas	All Especialidades	All Fator de riscos	818	134	16,38%
2013	All Especialidades	All Fator de riscos	818	134	16,38%
	INT MEDICINA A /HSA	All Fator de riscos	457	69	15,10%
		Alcoolismo	25	2	8,00%
		Coma	8	0	0,00%
		Corticoides	35	8	22,86%
		DPCO	85	8	9,41%
		Desnutrição	45	7	15,56%
		Diabetes	148	22	14,86%
		Doença Hep. Crónica	28	4	14,29%
		HIV	3	0	0,00%
		Imunodeficiência	7	0	0,00%
		Imunossupressores	17	3	17,65%
		Perturbação da Deglutição	53	15	28,30%
		Transplante	1	0	0,00%
		Traqueostomia	2	0	0,00%
	INT MEDICINA B /HSA	All Fator de riscos	149	30	20,13%
		Alcoolismo	6	0	0,00%
		Coma	0	0	0,00%
		Corticoides	25	9	36,00%
		DPCO	22	5	22,73%
		Desnutrição	18	3	16,67%
		Diabetes	38	8	21,05%
		Doença Hep. Crónica	5	0	0,00%
		HIV	0	0	0,00%
		Imunodeficiência	7	1	14,29%
		Imunossupressores	5	0	0,00%
		Perturbação da Deglutição	21	3	14,29%
		Transplante	0	0	0,00%
		Traqueostomia	2	1	50,00%

Figura A.7: Excerto do *dashboard Fatores de Risco e Infeção Nosocomial*: indicadores que relacionam fatores de risco intrínseco com a presença de infeção nosocomial, por serviço e ano (Parte 1/2).

INT MEDICINA C /HSA	All Fator de riscos	212	35	16,51%
	Alcoolismo	16	3	18,75%
	Coma	1	0	0,00%
	Corticoides	16	1	6,25%
	DPCO	30	3	10,00%
	Desnutrição	34	6	17,65%
	Diabetes	61	11	18,03%
	Doença Hep. Crónica	10	2	20,00%
	HIV	7	1	14,29%
	Imunodeficiência	0	0	0,00%
	Imunossupressores	7	2	28,57%
	Perturbação da Deglutição	28	5	17,86%
	Transplante	0	0	0,00%
	Traqueostomia	2	1	50,00%

Figura A.8: Excerto do *dashboard Fatores de Risco e Infeção Nosocomial*: indicadores que relacionam fatores de risco intrínseco com a presença de infeção nosocomial, por serviço e ano (Parte 2/2).

A.3 *Dashboard Caracterização da Infeção Nosocomial*

Figura A.9: Excerto do *dashboard Caracterização da Infeção Nosocomial*: indicadores que caracterizam a infeção nosocomial, por tipo de infeção, serviço e ano.

A.4 *Dashboard Previsão de Infeções Nosocomiais*

Figura A.10: *dashboard Previsão de Infeções Nosocomiais.*

Anexo B

Publicações

B.1 *Business Intelligence and Nosocomial Infection Decision Making*

Autores: Eva Silva, Ana Alpuim, Luciana Cardoso, Fernando Marins, César Quintas, Carlos Filipe Portela, Manuel Filipe Santos, José Machado e António Abelha

Livro: *Integration of Data Mining in Business Intelligence Systems*, Ana Azevedo e Manuel Filipe Santos (ed)

Editora: IGI Global Book

Ano: 2014

Estado: Publicado

Abstract: *The implementation of Business Intelligence tools in healthcare organizations helps the managers and the healthcare professionals in their decision making process through data manipulation and data analysis. The main goal of this chapter is to evaluate the applicability of the Business Intelligence tools and concepts to healthcare and their performance as a Clinical Decision Support System, analyzing the evolution of nosocomial infection in the* Centro Hospitalar do Porto*, by defining a set of indicators that can help*

the nosocomial infection management and inducing Data Mining models to predict the occurrence of nosocomial infections by (sensitivity of 91%). The knowledge obtained with the analysis of the indicators and the knowledge obtained with the nosocomial infection prediction can be applied by healthcare professionals in their decision making. Through the analysis of the data collected, Business Intelligence tools help overcome the problems associated with the complexity, heterogeneity and distributiveness present in the healthcare environment.

Keywords: *Business Intelligence, Clinical Decision Support System, Data Mining, Data Warehouse, Electronic Health Record, ETL, Health Information System, Nosocomial Infection.*

B.2 *Nosocomial Infection Prediction using Data Mining Technologies*

Autores: Eva Silva, Luciana Cardoso, António Abelha e José Machado

Livro: *Applying Business Intelligence to Clinical and Healthcare Organizations*, José Machado e António Abelha (ed)

Editora: IGI Global Book

Ano: 2014

Estado: Submetido

Abstract: *The existence of nosocomial infection prevision systems in healthcare environments can contribute to improve the quality of the healthcare institution and also to reduce the costs with the treatment of the patients that acquire these infections. The analysis of the information available allows to efficiently prevent these infections and to build knowledge that can help to identify their eventual occurrence. This chapter presents a Business Intelligence (BI) platform responsible for predicting the occurrence of nosocomial*

infections in Centro Hospitalar do Porto *(CHP), a hospital center in the north of Portugal. It presents the results of the application of predictive models to real clinical data. Good models, induced by the Data Mining (DM) classification techniques Support Vector Machines, Decision Trees and Naïve Bayes, were achieved (sensitivities higher than 91.90%). Therefore, with this system that be able to predict these infections may allow the prevention and, consequently, the reduction of nosocomial infection incidence. The platform presents important information, supporting healthcare professionals in their decisions, namely in planning infection prevention measures. So, the system acts as a Clinical Decision Support System (CDSS) capable of reducing nosocomial infections and the associated costs, improving the healthcare and, increasing patients' safety and well-being.*

Keywords: *Business Intelligence, Clinical Decision Support System, CRISP-DM, Data Mining, Knowledge Discovery in Databases, Nosocomial Infection, Open Source.*

B.3 *Business Intelligence Platform for Nosocomial Infection Incidence*

Autores: Eva Silva, Luciana Cardoso, Fernando Marins, António Abelha e José Machado

Revista: *Journal of Convergence Information Technology*

Ano: 2014

Estado: Submetido

Abstract: *Nosocomial infection prevention is essential for patients' safety and well-being. It can be efficiently performed through the analysis of the information available. With this analysis it is possible to build knowledge that helps to identify the risk factors and the activities related to the nosocomial infection occurrence and it also allows characterizing the infection.This*

paper presents a Business Intelligence (BI) system built to allow the study of nosocomial infection incidence in the Medicine Units of Centro Hospitalar do Porto *(CHP), a hospital centre in the north of Portugal. This BI platform is responsible for presenting nosocomial infection indicators. This platform enables to query important information and to analyze it, supporting healthcare professionals in their decisions. The knowledge obtained with this analysis allows preventing, monitoring and reducing nosocomial infections. So, the system acts as a Clinical Decision Support System (CDSS) capable of increasing patient's safety and well-being. The platform developed shows that, for example, in 2013 the rate of nosocomial infection in CHP Medicine Units varied between 9.43% and 12.95% and the respiratory and the urinary tract infections were the most frequent nosocomial infections. This work and the platform developed demonstrate that BI technology can be applied to healthcare with great utility and success.*

Keywords: *Business Intelligence, Clinical Decision Support System, Data Warehousing, Nosocomial Infection, On-Line Analytical Processing.*

Anexo C

Glossário

cateterismo

Introdução de uma sonda ao longo de um canal do organismo (por exemplo veias ou uretra) para retirar o seu conteúdo, inserir substâncias, determinar pressões sanguíneas, etc. 53, 80, 91

corticoides

Substâncias sintéticas cuja ação é semelhante à das hormonas segregadas pelo córtex das glândulas suprarrenais. Possuem um efeito anti-inflamatório e imunossupressor. 52

dashboard

Apresentação visual de informações importantes para a gestão de uma organização, tais como os valores dos KPIs. Um *dashboard* pretende eliminar a necessidade do seu utilizador realizar *queries* manualmente para obter as informações, bem como facilitar a tomada de decisão através da disponibilização, em tempo real, de informações relevantes e fáceis de analisar. 17, 20, 41–44, 54, 55, 64–67, 69, 71–73, 75, 90–92, 96, 97, 105–112

data warehousing

Todo o processo de *design*, desenvolvimento e manutenção de um sistema de DW. 20, 21, 24, 25, 49, 75

entubação

Introdução de um tubo num canal ou cavidade do organismo, geralmente através da boca ou do nariz. Normalmente o tubo é introduzido até à traqueia com o objetivo de assegurar a circulação do ar, ou até ao estômago com o intuito de alimentar o doente ou aspirar o conteúdo gástrico. 53, 80, 91

imunossupressores

Agentes que diminuem ou suprimem a resposta imunológica do doente. 52

Key Performance Indicators

Conjunto de medidas quantificáveis que refletem os objetivos de uma organização e a auxiliam a medir o seu desempenho e progresso na execução de determinadas atividades, de acordo com as suas prioridades e objetivos. Estes parâmetros permitem, deste modo, a avaliação dos fatores que são cruciais para o sucesso da organização. xix, 37, 38, 117

open-source

Software cujo código fonte está disponível para uso e/ou modificação, podendo ser livremente utilizado e partilhado (na forma modificada ou não) por qualquer utilizador. 6, 7, 38, 39, 41–44, 75, 94

oversampling

Técnica que, através da replicação dos dados, aumenta o peso da classe de dados minoritária, de modo que o modelo de classificação tenha um bom desempenho. 80, 81

sépsis

Resposta inflamatória generalizada do organismo, geralmente causada pela presença de agentes infeciosos na corrente sanguínea. 53, 59

traqueostomia

Intervenção cirúrgica que consiste na incisão da traqueia, de forma a permitir a introdução de uma cânula para a passagem de ar. 52