Universidade do Minho
Escola de Engenharia

Rita Soares Reis

Plataforma Inteligente de Apoio à Decisão Médica no Transplante de Órgãos

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

Abril de 2019

DECLARAÇÃO

Nome: Rita Soares Reis

Endereço eletrónico: rita.reis.27@hotmail.com

Cartão de Cidadão: 14856800

Título da dissertação: Plataforma Inteligente de Apoio à Decisão Médica no Transplante de Órgãos

Orientador: Professor Doutor António Carlos da Silva Abelha

Ano de Conclusão: 2019

Designação do Mestrado: Mestrado Integrado em Engenharia Biomédica

Área de Especialização: Ramo de Informática Médica

Universidade do Minho, _____/_____/_____

Assinatura:

Aos meus pais.

RESUMO

A alocação adequada de órgãos para transplantação é crítica e crucial. No entanto, o número de órgãos a ser doados não é suficiente dada a quantidade de pacientes em lista de espera. Assim, a determinação do maior número possível de potenciais dadores, de forma eficiente e eficaz torna-se essencial e pode contribuir para melhorar a taxa de sucesso de transplantação de órgãos.

Ao longo dos últimos anos, a utilização de Tecnologias de Informação (TIs) e de ferramentas computacionais em vários setores económicos, incluindo o setor da saúde, cresceu exponencialmente, já que têm potencial para transformar e melhorar a prestação de cuidados de saúde.

Assim, e aliando a necessidade da eficiência na descoberta de potenciais dadores com a emergência das TIs na saúde, surge a necessidade de uma plataforma Web de apoio à decisão clínica. O objetivo desta plataforma é automatizar o processo de descoberta de informação útil e acionável, através da utilização de tecnologias como *Business Intelligence* (BI) e *Data Mining* (DM), ajudando na tomada de decisão clínica diária. Assim, esta é responsável pela recolha, gestão, armazenamento e sinalização de potenciais dadores.

No âmbito deste projeto de dissertação, foi redesenhada e otimizada a plataforma Web Organite, atualmente implementada no Centro Hospitalar do Porto (CHP). Envolveu transformações tanto no *design* da interface do utilizador, como no modo como a informação está organizada na plataforma, de forma a melhorar a experiência do utilizador e a interação com os dados clínicos. Foi ainda desenvolvida uma metodologia, com base em técnicas de *Data Mining,* para construir um modelo preditivo que avalia quais os pacientes que dão entrada no hospital que têm maior probabilidade em ser potenciais dadores de órgãos. O objetivo é tornar mais simples e eficaz o processo de identificação de potenciais dadores, contribuindo positivamente na tomada de decisão do Gabinete de Coordenação de Colheita e Transplantação (GCCT), e impactando na redução da lista de doentes que aguarda um transplante.

ABSTRACT

Proper allocation of organs for transplantation is critical and crucial. However, the number of organs to be donated is not sufficient given the number of patients on the waiting list. Thus, the efficient determination of as many potential donors as possible becomes essential and can contribute to improved organ transplantation success rate.

Over the last few years, the use of Information Technology (IT) and computing tools in different economic sectors, including the health sector, has grown exponentially since they have the potential to transform and improve health care delivery.

Accordingly, and combining the need for efficiency in potential donors discovery with the emergence of IT on healthcare, this dissertation relies on the development of a platform to support clinical decision-making. The goal of this platform is to automate the process of discovering useful and actionable information using technologies such as Business Intelligence (BI) and Data Mining (DM), with the ultimate goal of improving daily clinical decisions. In this way, this platform is responsible for the collection, management, storage and signaling of potential donors.

In the scope of this dissertation project, the Web platform, *Organite*, currently implemented at Centro Hospitalar do Porto (CHP), was redesigned and optimized. It involved transformations both on user interface and on backend tasks, to improve user experience and interaction with the clinical data. Furthermore, a methodology based on Data Mining techniques was developed, with the aim to build a predictive model that evaluates which hospital admitted patients are most likely to be potential organ donors. The goal is to make the process of identifying potential donors easier and more effective, contributing positively to clinical decision-making, and consequently reducing the list of patients awaiting for an organ transplant.

ÍNDICE

Acrónimos 12

1. Introdução 14

1.1 Contextualização e Enquadramento 14

1.2 Motivação 16

1.3 Objetivos 21

1.3.1 Questões de Investigação 22

1.4 Estrutura do Documento 23

2. Estado da Arte 25

2.1 Sistemas de Informação Hospitalar 25

2.1.1 Interoperabilidade 27

2.1.2 Sistemas de Apoio à Decisão Clínica 28

2.2 Business Intelligence e a Informação Clínica 29

2.3 Descoberta de Conhecimento em Base de Dados 32

2.4 Data Mining 34

2.4.1 Taxonomia do Data Mining 34

2.4.2 Algoritmos de Aprendizagem 36

2.4.3 Otimização 44

2.4.4 Balanceamento dos Dados 47

2.4.5 Conjuntos de Treino e de Teste 49

2.4.6 Métricas para Aferição e Avaliação 51

2.5 Aplicações de *Front-end* 54

3. Metodologias de Investigação e Tecnologias 55

3.1 Metodologia de Investigação 55

3.1.1 Design Science Research 55

3.2 Metodologias Técnicas 59

3.2.1 Frameworks de Desenvolvimento Web 59

3.2.2 *CRoss Industry Standard Process para Data Mining* 64

3.3 Metodologia da Prova de Conceito 68

4. Descoberta de conhecimento numa bd de potenciais dadores 70

4.1 Introdução....70
4.2 Definição do Problema e Objetivos da Solução....70
4.3 Exploração dos dados....72
4.3.1 Visualização e Descrição dos Dados....72
4.3.2 Tratamento dos Dados....76
4.4 Modelo de Previsão....78
4.4.1 Escolha de Algoritmos....78
4.4.2 Criação dos Conjuntos de Treino e Teste....80
4.4.3 Criação dos Modelos....80
4.4.4 Avaliação dos Modelos....86
4.4.5 Otimização....89
4.4.6 Reavaliação....90
4.5 Conclusão e Trabalho Futuro....92
5. Plataforma de Apoio à Decisão Clínica no Transplante de Órgãos....96
5.1 Introdução....96
5.2 Definição do Problema e Objetivos da Solução....97
5.3 Desenho e Desenvolvimento....98
5.3.1 Arquitetura BI da Plataforma....98
5.3.2 Arquitetura Web da Plataforma....100
5.3.3 Implementação de um Sistema de Dados Persistente....103
5.3.4 Módulo de Notificações....106
5.3.5 Interface do Utilizador....109
5.4 Conclusão e Trabalho Futuro....114
6. Prova de Conceito....116
7. Conclusão e Trabalho Futuro....119
7.1 Principais Contribuições....120
7.2 Trabalho Futuro....122
8. Bibliografia....124
Anexo I – Designação do Anexo I....132
Anexo II – Designação do Anexo II....134

LISTA DE FIGURAS

Figura 1 – Distribuição nacional de dadores, dador sequencial, dador vivo, dador após morte circulatória e dador após morte cerebral, por tipologia e ano, num período temporal de 4 anos, entre 2014 e 2017 19
Figura 2 – Evolução anual nacional do número de pacientes em lista de espera, num período temporal de 7 anos, entre 2011 e 2017 19
Figura 3 - Arquitetura de um sistema de *Business Intelligence*. 31
Figura 4 – Processo de Descoberta do Conhecimento Base de Dados 34
Figura 5 - Taxonomia de *Data Mining* 36
Figura 6 – Árvore de decisão com testes sobre os atributos X e Y, de modo a classificar instâncias na Classe 1 ou na Classe 2 38
Figura 7 – Algoritmo kNN com k=5 para atribuir a instância x_u a uma das classes (ω_1, ω_2, ω_3) 40
Figura 8 – Representação do hiperplano (w, b) de um problema de classificação binária onde é aplicado o algoritmo SVM. (a) Parâmetros w e b; (b) dois hiperplanos paralelos definem as margens 41
Figura 9 – Arquitetura típica de uma rede neuronal artificial 42
Figura 10 – Curvas *Receiver Operating Characteristic* (ROC) 53
Figura 11 - Representação esquemática da metodologia de investigação DSR 58
Figura 12 – Ciclo de pedido-resposta do cliente-servidor 62
Figura 13 – (a) Metodologia CRISP-DM e (b) metodologia SEMMA 66
Figura 14 – Matriz SWOT 69
Figura 15 – Comunicação entre utilizador, cliente, servidor e base de dados numa tarefa de edição da tabela disponibilizada na página de Diagnósticos Atuais na plataforma Organite. 105
Figura 16 – Arquitetura do sistema de notificações. 107
Figura 17 – Homepage da plataforma Organite 110
Figura 18 – Página de início de sessão do utilizador na plataforma Organite. 110
Figura 19 – *Pop-up* de notificações relativas a eventos neurológicos adversos aquando do início da sessão de um utilizador. 112

Figura 20 – Página de diagnósticos atuais da plataforma Organite. 113
Figura 21 – Protótipo da implementação do modelo de previsão de *Data Mining* na plataforma Organite.....114

LISTA DE TABELAS

Tabela 1 – Dados nacionais, europeus e mundiais, em valor absoluto e por milhão de habitantes (pmh), relativos a dadores falecidos por morte cerebral (DMC) e por morte cardiocirculatória (DPC) e número total de transplantes realizados.....18

Tabela 2 – Matriz de confusão associada a um problema de classificação binária.....51

Tabela 3 – Atributos *Modulo, BD, Sexo, Estado* e *Diag_Recente* da tabela ORGANITE_REPOSITORY, respetiva descrição, variáveis associadas e número de casos (%) de cada variável 73

Tabela 4 – Atributos *Cod_Diag* e *Cod_Especialidade* da tabela ORGANITE_REPOSITORY, respetiva descrição, *top* 5 de variáveis (ID e DESC) com maior número de casos associado e número de casos (%).....73

Tabela 5 – Descrição dos atributos *Idade*, *Episódios_Ant* e *Prognostico* criados; características associadas aos atributos contínuos (*Idade, Episodios*_Ant) e variáveis e número de casos (%) do atributo discreto (*Prognostico*).....74

Tabela 6 – Cálculo dos episódios neurológicos graves antecedentes de um paciente do conjunto de dados de potenciais dadores.....75

Tabela 7 – Discretização dos atributos *Idade* e *Episodios_Ant*, variável associada a cada *bin* e número de casos associados a cada *bin* (%).....77

Tabela 8 – Algoritmos de DM utilizados na fase de modelação, família de algoritmos a que pertencem, algoritmo base e artigos de referência onde são caso de estudo 79

Tabela 9 – Modelos de DM construídos com base em técnicas de balanceamento de dados: *cluster undersampling,* RUS, SMOTE e no conjunto de dados original.....81

Tabela 10 – Representação dos valores relativos ao tamanho do *cluster i das* instâncias maioritárias ($SizeMAi$) e instâncias minoritárias ($SizeMIi$); rácio do número de instâncias MA em relação ao número de instâncias MI, para o *cluster i.*.....82

Tabela 11 – Número de instâncias selecionadas no cluster *i* da classe maioritária ($SSizeMAi$) 83

Tabela 12 – Resultados da aplicação dos algoritmos de DM aos modelos 1, 2, 3 e 4, com a técnica de *k-Fold Cross Validation*, k=10 85

Tabela 13 - Resultados da aplicação dos algoritmos *RBFClassifier* e *NaiveBayes* ao Modelo 1, *RandomForest* e *J48* ao Modelo 2 e *NaiveBayes* e *RBFClassifier* ao Modelo 3, através de *Holdout sampling*.....86
Tabela 14 – Matriz de custo para a aplicação da técnica *Cost-sensitive Classification*..90
Tabela 15 – Resultados da aplicação da técnica *Cost-sensitive Classification* ao Modelo 1 com os algoritmos *RBFClassifier* e *NaiveBayes*, ao Modelo 2 com os algoritmos *RandomForest* e *J48* e ao Modelo 3 com os algoritmos *NaiveBayes* e *RBFClassifier*, através de *Holdout sampling*.....91

ACRÓNIMOS

API – *Application Programming Interface*

AVC – Acidente vascular cerebral

BI – *Business Intelligence*

BSR – *Behavioral Science Research*

CHP – Centro Hospitalar do Porto

CNT – Coordenação Nacional da Transplantação

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

CRUD – *Create, Read, Update, Delete*

CSC – *Cost Sensitive Classification*

DCBD – Descoberta de Conhecimento em Bases de Dados

DM – *Data Mart*

DM – *Data Mining*

DMC – Dador após morte cerebral

DPC – Dador após morte cardiocirculatória

DS – Dador sequencial

DSR – *Design Science Research*

DSRM – *Design Science Research Methodology*

DV – Dador vivo

DW – *Data Warehouse*

EQDM – Direção Europeia para a Qualidade dos Medicamentos e Cuidados de Saúde

ETL – *Extract, Transform, Load*

GCCT- Gabinete de Coordenação de Colheita e Transplantação

HIS – *Health Information Systems*

HTML – *HyperText Markup Language*

HTTP – *Hyper-Text Transfer Protocol*

IDE – *Integrated Development Environment*

IM – Informática Médica

IPST – Instituto Português do Sangue e da Transplantação

JSON – *JavaScript Object Notation*

ORM – *Object Relation Mapping*

RENNDA – Registo Nacional de não Dadores

REST – *REpresentational State Transfer*

SAD – Sistemas de Apoio à Decisão

SADC – Sistemas de Apoio à Decisão Clínica

SGBD - Sistema de Gestão de Bases de Dados

SI – Sistema de Informação

SIH – Sistema de Informação Hospitalar

SNS – Serviço Nacional de Saúde

SPA – *Single Page Application*

SQL – *Structured Query Language*

SU – Serviço de Urgência

5 – Tomografia Computorizada

TI – Tecnologias de Informação

UCI – Unidade de Cuidados Intensivos

UE – União Europeia

URI – Uniform Resource Identifier

URL – Uniform Resource Locator

XML – *eXtensible Markup Language*

1. INTRODUÇÃO

O presente projeto surge no âmbito da dissertação do Mestrado Integrado em Engenharia Biomédica, especialização em Informática médica, e descreve uma plataforma inteligente para o apoio à prática clínica na área de transplantação de órgãos.

O capítulo introdutório encontra-se dividido em quatro secções. Inicia-se através de uma contextualização e enquadramento do projeto (Secção 1.1) e segue para a descrição da motivação que conduziu à sua elaboração (Secção 1.2). Na Secção seguinte (Secção 1.3) são apresentadas as questões de investigação desta dissertação e os objetivos associados. O capítulo conclui (Secção 1.4) com a apresentação da estrutura do documento.

1.1 Contextualização e Enquadramento

O setor da saúde é um dos setores mais críticos da indústria de prestação de serviços, uma vez que é crucial para a vida do ser humano e qualquer erro pode causar resultados irreversíveis. Em particular, e sendo o transplante de órgãos um tratamento vital para a falência crónica de órgãos principais, a alocação adequada de órgãos para transplantação é crítica e crucial. Um dador pode salvar até dez vidas e melhorar significativamente a qualidade de vida de mais algumas dezenas [1].

Em 2015, estimava-se que existiriam em Portugal cerca de 2800 indivíduos em lista de espera para transplantação de rim. O tempo médio de espera varia entre um ano para um transplante de fígado até mais de quatro anos para um transplante renal [2]. Embora a alocação de órgãos seja a única terapia viável para várias doenças terminais, o número de órgãos a ser doados não é suficiente dada a necessidade dos pacientes em lista de espera. Muitas vezes os pacientes aguardam durante um longo período de tempo por um transplante e alguns dos órgãos doados são desperdiçados devido à não deteção de correspondência entre dador e recetor [1].

O tema apresenta um cariz social dado que todos somos potenciais dadores se não estivermos inscritos no RENNDA conforme presente na Lei 12/93, art.º 11º, e também cariz de cidadania dado que o transplante é o único tratamento que requer a

participação da sociedade para ser plenamente desenvolvido, dado que só existe transplante se houver doação [3].

Tanto dadores vivos como dadores falecidos são fontes de órgãos, tecidos e células. Embora existam alguns aspetos comuns a estes dois tipos de dadores, os critérios de seleção podem diferir entre dador vivo e falecido, bem como para a doação de órgãos, tecidos e células. Os dadores falecidos dividem-se em duas categorias: doação após morte cerebral (DMC), doação após morte cardiocirculatória (DPC) [4].

Os pacientes com lesões cerebrais graves são mais propensos a progredir para morte cerebral (MC). A MC é definida como a cessação irreversível das funções de todas as estruturas neurológicas intracranianas, muito embora as funções cardíaca e pulmonar possam estar mantidas artificialmente [4].

Apesar da recuperação da pessoa em situação neurocrítica (NC) ser uma hipótese, a situação mais relevante para este trabalho, prende-se com as alterações irreversíveis que podem ocorrer, constituindo desta forma um potencial dador em morte cerebral (MC). A pessoa com patologia NC apresenta doença do sistema nervoso central (SNC) ou periférico (SNP), primário ou secundário, que requer monitorização e/ou terapêutica intensiva, com potencial risco de vida ou de função do órgão, conforme definido no procedimento de assistência aos doentes neurocríticos. Conforme o guia de boas práticas no processo de doação de órgãos, a existência de um programa orientado especificamente para o tratamento de pacientes neurocríticos melhora a efetividade na transferência de possíveis dadores para as UCIs [3]. Assim, procurou estudar-se o doente em situação NC, potencial dador de órgãos, em contexto de urgência, desde a sua entrada no hospital.

A pertinência do tema prende-se com a notória desproporção entre a emergência de transplantes de órgãos e o número real de transplantes realizados, o que se traduz num grande problema de saúde pública. Neste sentido, o papel de um engenheiro biomédico é fulcral no apoio à tomada de decisão clínica. Mais que os custos inerentes ao transplante, importa despertar para a emergência da sua realização, estando esta diretamente relacionada com a identificação precoce do potencial dador [3].

1.2 Motivação

A transplantação de órgãos é o melhor tratamento para salvar vidas na fase final da falência de órgãos. Existem vários tipos de dadores [4, 5]:

- Dadores falecidos:
    - Dador por morte cerebral (DMC) – dador falecido a quem foi declarada a morte com base em critérios neurológicos, verificando-se a cessação irreversível das funções do tronco cerebral.
    - Dador por morte cardiocirculatória (DPC) – dador falecido que apresenta uma completa e irreversível cessação de toda a função circulatória e respiratória, com a consequente certificação de morte.
- Dadores vivos – pessoa que doou um órgão (rim ou porção de fígado) em vida.
- Dadores sequenciais – recetor de um transplante de órgão (fígado), cujo órgão nativo pode ser considerado para transplantação noutro doente.

Qualquer pessoa, ao falecer, é um potencial dador de órgãos ou tecidos para transplante, desde que, em vida, não se tenha manifestado contra esta possibilidade, nomeadamente através de inscrição no Registo Nacional de Não Dadores (RENNDA). No caso de se tratar de uma pessoa menor de idade ou mentalmente incapaz, é válida a vontade de quem detenha o poder paternal. No entanto, para que possa haver doação de órgãos têm que reunir-se um conjunto de circunstâncias [6]:

- O dador tem que falecer num hospital;
- Depois de se verificar a paragem irreversível das funções cerebrais ou cardiorrespiratórias, o corpo tem que ser mantido artificialmente, desde o momento da morte até ao momento da extração dos órgãos;
- É necessário que se conheça, com exatidão, a causa da morte.

Não são aceites como dadores indivíduos que sejam, na altura da morte, portadores de uma doença infectocontagiosa, de um tumor maligno ou de uma doença com repercussão nos órgãos a transplantar. Também são contraindicações, embora relativas,

para a doação, uma história clínica de hipertensão arterial, de diabetes ou a idade avançada [6].

No que respeita à idade, os dadores mais desejáveis são os que têm entre 15 e 55 anos, mas a idade é valorizada caso a caso, de acordo com o tipo de órgão a utilizar e com o conhecimento da história clínica do dador [6].

Uma vez certificada a morte, e se o cadáver tiver características adequadas à doação (ou seja, se os seus órgãos puderem ser úteis para curar ou melhorar a saúde de outras pessoas), o Gabinete de Coordenação de Colheita e Transplantação (GCCT) tem a obrigação de se informar, por todos os meios ao seu alcance, sobre a vontade expressa em vida por aquele indivíduo em relação à doação. Para este efeito, são consultados o RENNDA e, sobretudo, os familiares próximos do falecido. No caso de não existirem objeções, prosseguir-se-á com o procedimento de colheita [6].

Um dos principais objetivos do GCCT é garantir uma contínua e eficiente cooperação entre as diferentes equipas envolvidas na doação, colheita, preservação, partilha de órgãos e implantação, colheita de tecidos, processamento e armazenamento num banco de tecidos. É, assim, responsabilidade do GCCT identificar o potencial dador, utilizando todas as ferramentas e conhecimento científico disponíveis para expandir, tanto quanto possível, o grupo de dadores [4].

Em Portugal, a atividade encontra-se dividida por 5 GCCT: Centro Hospitalar do Porto, Hospital de S. João, Hospital da Universidade de Coimbra, Centro Hospitalar Lisboa Central e Centro Hospitalar Lisboa Norte. Cada GCCT deve articular-se com as Unidades de Colheita e de Transplantação, bem como com os Coordenadores Hospitalares de Doação e os Centros de Histocompatibilidade, de forma a garantir uma resposta eficaz à referenciação de um potencial dador em qualquer hospital.

No que diz respeito a dados estatísticos, verifica-se que a incidência da doação proveniente de dadores falecidos está desigualmente distribuída nos países da UE [5, 7].

Segundo a Coordenação Nacional da Transplantação (CNT), unidade orgânica do Instituto Português do Sangue e da Transplantação (IPST), e os valores estatísticos já apresentados, o número de dadores em Portugal tem vindo a aumentar nos últimos anos [5]. Portugal atingiu, em 2017, o segundo lugar na lista mundial de países com mais

órgãos de dadores falecidos, num universo de 50 países, onde estão representados todos os países ocidentais [8]. Na Tabela 1 estão representados dados relativos a transplantes nacionais, europeus e mundiais, em valor absoluto e por milhão de habitantes, relativos ao ano de 2017.

Tabela 1 – Dados nacionais, europeus e mundiais, em valor absoluto e por milhão de habitantes (pmh), relativos a dadores falecidos por morte cerebral (DMC) e por morte cardiocirculatória (DPC) e número total de transplantes realizados [5, 7]

	Dados em valor absoluto e por milhão de habitantes (pmh), 2017		
	Portugal	Europa	Mundo
Dadores falecidos	351 (34,08)	13.098 (18,29)	35.925 (9,67)
Dador após morte cerebral (DMC)	330 (32,04)	11.335 (15,83)	27.911 (7,52)
Dador após morte cardiocirculatória (DPC)	21 (2,04)	1763 (2,46)	8.016 (2,16)
Dadores vivos	89 (8,64)	NA	NA
Total dadores órgãos	440 (42.72)	NA	NA
Total transplantes órgãos	895 (86,89)	43.310 (60,49)	124.682 (33,57)

Analisando os dados da Tabela 1 verifica-se que Portugal supera os valores europeus e mundiais em percentagem por milhão de habitantes, no que diz respeito ao valor total de dadores falecidos, a DMC e ainda ao total de transplantes de órgãos realizados. De notar que este último parâmetro engloba dadores falecidos, vivos e sequenciais.

A distribuição nacional de dadores por tipologia encontra-se representada na Figura 1 para um período temporal de 4 anos, entre 2014 e 2017 [5].

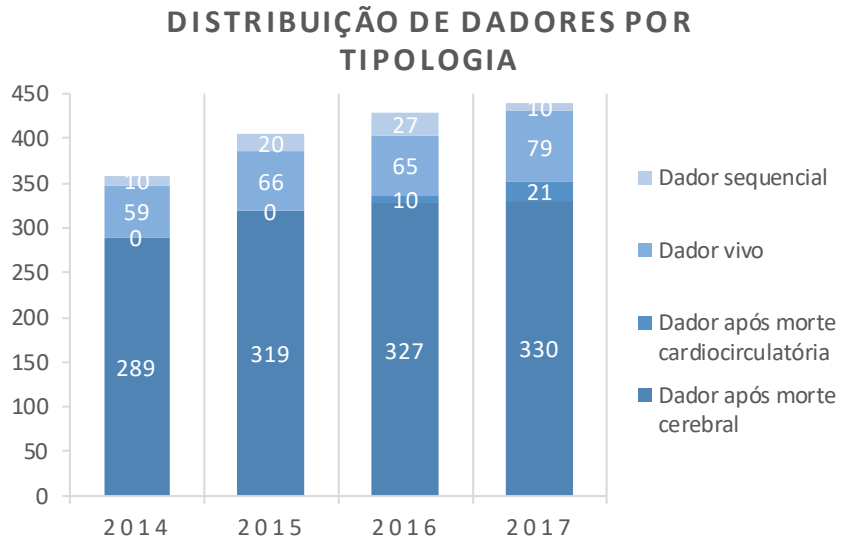


Figura 1 – Distribuição nacional de dadores, dador sequencial, dador vivo, dador após morte circulatória e dador após morte cerebral, por tipologia e ano, num período temporal de 4 anos, entre 2014 e 2017 [5].

Neste período temporal, verifica-se claramente uma predominância de dadores após morte cerebral. Os dadores após morte circulatória surgiram no ano de 2016, tendo-se verificado um aumento de 50% deste número em 2017.

Na Figura 2 está representada a evolução anual do número de pacientes em lista de espera por um transplante, em Portugal, entre 2011 e 2017 [8].

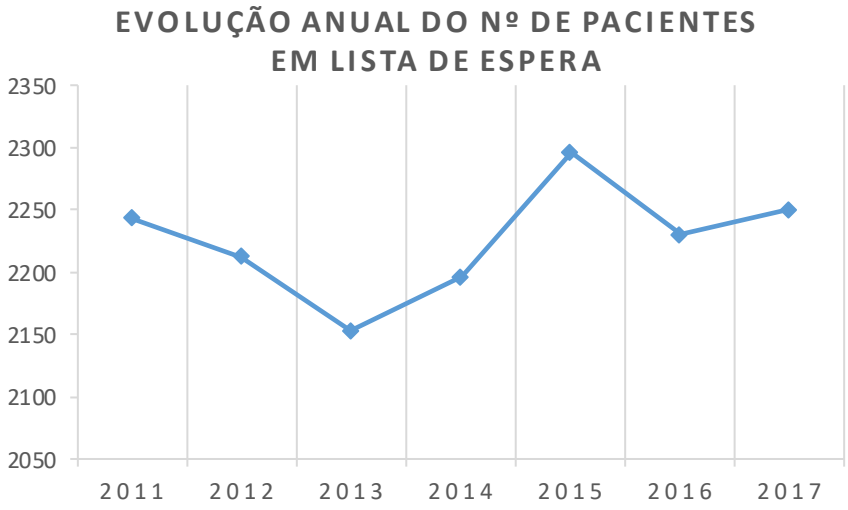


Figura 2 – Evolução anual nacional do número de pacientes em lista de espera, num período temporal de 7 anos, entre 2011 e 2017 [8].

Verifica-se que, de 2011 a 2013 houve um decréscimo do número de doentes em lista de espera por um transplante, havendo uma subida considerável nos dois anos que se seguiram. Em 2016, este número desceu, mas em 2017 voltou a aumentar.

Segundo a Direção Europeia para a Qualidade dos Medicamentos e Cuidados de Saúde (EQDM) do Conselho da Europa, em 2017 mais de 144 000 pacientes faziam parte das listas de espera dos Estados-Membros da União Europeia (UE), representando um aumento de 10% comparativamente a 2013. Ainda relativamente a 2017, 6518 pacientes morreram enquanto aguardavam por um transplante [4]. Morrem, por dia, 18 doentes em lista de espera por não haver órgãos disponíveis. Por hora, são acrescentados às listas de espera europeias cerca de 6 novos pacientes, como relatado no Boletim de Transplantes [9]. A nível nacional e a cada mês, 6 pessoas morrem a aguardar por um transplante. O número de doentes em lista de espera ativa, era de 2250 no final de 2017. O órgão com maior lista de espera é o rim, com 2019 pessoas dependentes de uma doação [8]. Estes números representam a verdadeira dimensão das necessidades dos doentes.

Segundo o guia para a qualidade e segurança dos órgãos para transplantação do conselho da Europa, longos períodos na lista de espera para os órgãos/tecidos podem resultar na deterioração ou morte dos doentes antes da cirurgia. Muito mais pessoas poderiam beneficiar da transplantação de órgãos do que o número das que atualmente recebem transplantes [4].

O fator mais crítico continua a ser a oferta de órgãos para transplantação. A lacuna entre a oferta e a procura de órgãos para transplantação levou à consideração de diferentes estratégias para aumentar a disponibilidade de órgãos, incluindo estratégias de doação em vida e o uso de órgãos provenientes de dadores com critérios de doação alargados ou expandidos e de dadores com riscos não-padronizados [4].

Para qualquer doente que se apresente com um valor menor do que cinco numa Escala de Coma de Glasgow (GSC) ou com uma classificação FOUR (*Full Outline of UnResponsiveness*) menor do que cinco que seja admitido ou permaneça num hospital, a doação de órgãos deve ser considerada como parte do tratamento de fim de vida [2]. É altamente recomendável dar a conhecer cada caso a uma organização de colheita de órgãos e à equipa de coordenação de transplantes para uma avaliação mais pormenorizada [4].

Apenas órgãos recuperados ao abrigo de uma gestão de elevada qualidade no processo de doação são suscetíveis de funcionar satisfatoriamente. A situação do dador e o tempo despendido entre a colheita, a recuperação e a transplantação são parâmetros cruciais com limites estritos. O intervalo de tempo entre a recuperação do órgão e o transplante pode variar entre 4 e 18 horas, dependendo da natureza dos órgãos [4].

A deteção precoce de potenciais dadores é o ponto de partida para a transplantação. Esta é, provavelmente, a parte mais difícil de todo o processo. A única maneira de garantir que não se perdem potenciais dadores é poder identificar e monitorizar possíveis dadores individualmente em hospitais ou áreas geográficas relevantes. A definição de um sistema de avaliação que identifique todas as mortes em hospitais que tenham o potencial de contribuir para a doação de órgãos é crucial.

1.3 Objetivos

Este projeto de dissertação tem como objetivo redesenhar e otimizar a plataforma Web de apoio à decisão clínica, o Organite, atualmente implementada no Centro Hospitalar do Porto (CHP).

Pretende-se transformar o *design* da interface do utilizador e o modo como a informação está organizada na plataforma, de forma a melhorar a experiência do utilizador e a interação com os dados clínicos. Pretende-se ainda desenvolver uma metodologia, com base em técnicas de *Data Mining* (DM), para construir um modelo preditivo que avalie quais os pacientes que dão entrada no hospital que têm maior probabilidade em ser potenciais dadores de órgãos. O objetivo é tornar mais simples e eficaz o processo de identificação de potenciais dadores, contribuindo positivamente na tomada de decisão do Gabinete de Coordenação de Colheita e Transplantação (GCCT), e impactando na redução da lista de doentes que aguarda um transplante.

As questões de investigação propostas para esta dissertação estão descritas na Secção seguinte.

1.3.1 Questões de Investigação

Questão de Investigação nº1: Como se podem proporcionar melhorias ao sistema atual de apoio à decisão no processo de transplantação no GCCT do CHP, de forma a melhorar a prestação de cuidados de saúde?

Objetivos:

- Levantamento dos requisitos do negócio para a melhoria da plataforma de apoio à decisão clínica atual;
- Estudo da arquitetura e funcionalidades da plataforma atual, de modo a identificar fraquezas a transformar em pontos de melhoria que tragam valor aos prestadores de cuidados de saúde;
- Escolha das metodologias e tecnologias envolvidas para o desenho e desenvolvimento de novas funcionalidades;
- Criação e integração de um sistema de notificações de eventos neurológicos adversos;
- Criação e implementação de um sistema de dados clínicos persistente.

Questão de Investigação nº2: Como contribuir no apoio à decisão clínica de forma a maximizar a identificação de potenciais dadores de órgãos e a diminuir a lista de espera por um transplante?

Objetivos:

- Estudo de técnicas de DM de classificação que permitam determinar o prognóstico de um paciente do repositório de potencias dadores;
- Descoberta de padrões úteis na informação do repositório de potenciais dadores;
- Obtenção de um modelo de DM que permite prever se o paciente é um potencial dador;
- Avaliação do modelo de DM através do cálculo de métricas estatísticas sobre os resultados obtidos;
- Otimização do modelo de DM através de técnicas de classificação sensível ao custo.

Questão de Investigação nº3: Qual é a utilidade e usabilidade das ferramentas de apoio à decisão clínica implementadas?

Objetivos:

- Análise crítica e discussão dos resultados para cada caso de estudo;
- Realização de uma prova de conceito às ferramentas para o apoio à decisão à prática clínica.

Questão de Investigação nº4: Quais os pontos de melhoria das ferramentas de apoio à decisão clínica implementadas?

Objetivo:

- Formulação de propostas de alteração e de melhoria ao modelo de *Data Mining* desenvolvido para a previsão de potenciais dadores;
- Formulação de propostas de alteração e de melhoria ao Organite, plataforma de apoio à decisão implementada.

1.4 Estrutura do Documento

A presente dissertação encontra-se dividida em 7 capítulos: Introdução, Estado da Arte, Metodologias de Investigação e Tecnologias, Descoberta de Conhecimento numa BD de Potenciais Dadores, Plataforma de Apoio à Decisão Clínica no Transplante de Órgãos, Prova de Conceito e, por fim, Conclusão e Trabalho Futuro.

O capítulo introdutório apresenta uma contextualização e enquadramento do projeto, assim como a motivação e objetivos definidos.

No capítulo que se segue, Estado da Arte, são apresentados os conceitos teóricos importantes para a realização do trabalho desenvolvido. Para tal, existe uma abordagem genérica aos SIH, seguindo-se a apresentação do conceito de BI. Seguem-se os conceitos da descoberta de conhecimento em bases de dados e *Data Mining*. Este capítulo finaliza com o tema das aplicações *front-end*.

O capítulo das Metodologias de Investigação e Tecnologias apresenta as metodologias adotadas, *Design Science Research* como metodologia de investigação,

CRISP-DM como metodologia de modelação em *Data Mining* e ainda as *frameworks* tecnológicas, como Flask e AngularJS utilizadas no desenvolvimento da solução proposta. Este capítulo termina com a metodologia da Prova de Conceito na defesa da viabilidade e utilidade das ferramentas clínicas desenvolvidas.

Segue-se o capítulo da Descoberta de Conhecimento numa BD de potenciais dadores onde são construídos os modelos de previsão de *Data Mining*. O capítulo inicia com uma breve introdução à qual se segue a definição do problema e objetivos. As restantes secções englobam todos os passos da metodologia de *Data Mining* adotada, CRISP-DM, começando pela exploração dos dados e terminando com a avaliação e otimização dos modelos construídos. O capítulo encerra-se com conclusões e propostas de melhoria para o futuro.

Relativamente ao capítulo da Plataforma de Apoio à Decisão Clínica no Transplante de Órgãos, e depois da introdução e definição de objetivos, é feita uma apresentação da plataforma desenvolvida, em termos de arquitetura, funcionalidades, modo de funcionamento e interface do utilizador. Incluem-se ainda as conclusões e o trabalho proposto para o futuro.

De forma a avaliar as ferramentas desenvolvidas, o capítulo da Prova de Conceito inclui uma análise SWOT, sendo discutidas forças, fraquezas, oportunidades e ameaças.

No capítulo final é feito um balanço do trabalho desenvolvido, sendo apresentadas as principais contribuições e enunciadas perspetivas de trabalho futuro.

De referir ainda que o Apêndice A inclui um glossário com definições úteis e o Apêndice B inclui contributos científicos.

2. ESTADO DA ARTE

A saúde enfrenta vários problemas, incluindo elevadas despesas, qualidade inconsistente e lacunas nos cuidados. Por esta razão, os serviços de saúde representam uma grande parcela dos gastos do governo na maioria dos países. A informatização pode ter um papel importante na redução de custos e numa melhoria dos cuidados de saúde prestados. A importância da recolha e tratamento dos dados clínicos poderá levar a uma redução dos custos com o doente [10].

As Tecnologias de Informação (TI) consistem no conjunto das atividades e soluções computacionais que pretendem produzir, armazenar, transmitir, aceder e utilizar a informação disponível. As Tecnologias de Informação associadas à área médica podem transformar e melhorar radicalmente a prestação de cuidados de saúde, já que tiram verdadeiro partido da informação clínica disponível [11, 12, 13].

No entanto, e apesar de se fazer sentir um crescente aumento na utilização de TI nos últimos anos, o setor da saúde continua a ter uma adoção pouco significativa. Assim, e apesar das evidências relativamente aos benefícios do uso de TI na área médica, ainda se tira pouco proveito das mesmas nas instituições de saúde [11, 14, 15]. A adoção limitada e lenta pode advir do desempenho dececionante das aplicações, da escolha de sistemas que não são interoperáveis ou de fácil utilização ou ainda da reticência em aderir a tecnologias informáticas [14, 16]. Acredita-se que o futuro bem-sucedido das TI no setor da saúde reside no desenho e na implementação de sistemas *user friendly*, centrados nas necessidades reais do dia-a-dia dos profissionais de saúde e que permitam melhorar o acesso e controlo sobre os dados clínicos armazenados [14, 15].

No contexto clínico, destacam-se nas próximas secções deste projeto de dissertação, conceitos, processos, técnicas e ferramentas das TI.

2.1 Sistemas de Informação Hospitalar

Os cuidados de saúde têm como objetivo compreender, prevenir e tratar doenças humanas. Na prática médica, um médico desenvolve, através da educação e experiência, uma compreensão dos processos anatómicos, fisiológicos e patológicos fundamentais, uma compreensão de como aplicar os métodos de diagnóstico, e uma

compreensão dos efeitos das drogas e tratamentos em doentes. Numa situação de tomada de decisão, o médico tem de compreender o estado clínico do doente, o propósito da sua intervenção e as possíveis ações que daí podem advir. Para isso, o médico interpreta um conjunto de dados não estruturados que vão sendo recolhidos ao longo dos episódios clínicos (consulta externa, internamento, intervenção cirúrgica, urgência, hospital, análises clínicas, radiologia). De forma a auxiliar o profissional de saúde durante este processo, surgem os Sistemas de Informação Hospitalar (SIH) que, ao longo dos últimos 30 anos, se têm tornado cada vez mais úteis em ambientes de prestação de cuidados de saúde [17].

Os Sistemas de Informação Hospitalar (SIH) são sistemas computacionais complexos, implementados de modo a gerir processos e atividades administrativas, financeiras e clínicas de um meio hospitalar, e correspondem a uma subcategoria dos Sistemas de Informação (SI) [19]. Por conseguinte, possibilitam a redução de custos hospitalares, melhoram a qualidade da prestação de cuidados de saúde e tornam mais fáceis as tomadas de decisão baseadas em evidências. Para além disso, permitem automatizar atividades de manutenção e gestão de dados clínicos e, deste modo, melhorar a recuperação da informação, reduzir o tempo de resposta e consequentemente tornar mais eficaz a atividade dos profissionais envolvidos [18].

O registo da informação clinica no Centro Hospitalar do Porto (CHP) é assegurada por vários SIH. O Sistema de Gestão de Doentes Hospitalares (SONHO) é um exemplo, implementado em muitos hospitais do Serviço Nacional de Saúde (SNS). O Processo Clínico Eletrónico (PCE) é definido como um repositório eletrónico seguro, organizado e centralizado de informação relativa a pacientes. O Sistema de Urgências é dedicado ao registo, gestão e armazenamento de episódios clínicos que ocorrem nas urgências hospitalares. Mais orientado para a atividade dos profissionais de saúde, surge o SClínico, um sistema que representa, gere e armazena informação clínica e que engloba o SAM (Sistema de Apoio Médico) e o SAPE (Sistema de Apoio à Prática de Enfermagem) [20, 21]. Neste sentido, o SClínico tem funções orientadas para a atividade médica, permitindo efetuar operações como a prescrições de medicamentos, registar e consultar a informação clínica recolhida nas consultas e consultar o histórico clínico do doente e funções orientadas para a atividade de enfermagem, podendo o enfermeiro efetuar ações como registar e consultar os sintomas apresentados pelo doente, registar

e consultar intervenções de enfermagem com base no diagnóstico efetuado e consultar as tabelas de parametrização e codificação da atividade de enfermagem [22].

Nos últimos anos, apesar da implementação de SIH ter aumentado no setor da saúde, ainda é diminuto o número de instituições de saúde que atingiram um nível de maturidade que permita a utilização autónoma destas tecnologias, assim como a partilha e comunicação de informação estruturada entre todas as unidades e profissionais de saúde dos estabelecimentos hospitalares [23, 24, 25]. As principais barreiras para conseguir ultrapassar este desafio são a heterogeneidade das tarefas dos profissionais de saúde, a diversidade das estruturas das organizações, os múltiplos SIH existentes e a dificuldade na adoção dos mesmos em meios hospitalares [19, 25].

Neste contexto, surgem os conceitos descritos nas próximas subsecções, 2.2.1 e 2.2.2. O conceito de interoperabilidade assegura a comunicação e partilha de informação entre sistemas diferentes, e os Sistemas de Apoio a Decisão Clinica (SADC) são ferramentas computacionais cujo potencial na prestação de cuidados de saúde e despoletado pela implementação de SIH interoperáveis.

2.1.1 Interoperabilidade

Atualmente, o contínuo crescimento da informação clinica registada nos SIH tornou o conceito de interoperabilidade nas instituições de saúde numa preocupação na área da Informática Médica (IM) [20, 26]. Desta forma, a interoperabilidade é considerada um requisito nos SIH que interfere na comunicação e cooperação entre sistemas [27]. Genericamente, o conceito de interoperabilidade pode ser definido como a capacidade de um sistema comunicar e partilhar informação com outro sistema. Surge de forma a superar a heterogeneidade e a existência de várias fontes de informação distintas. No setor da saúde, o principal objetivo da interoperabilidade é conectar as aplicações e os dados clínicos de modo a que possam ser compartilhados em toda a organização e distribuídos pelos profissionais de saúde [26, 27]. Assim, advém a necessidade de implementar plataformas dinâmicas, como por exemplo sistemas multiagentes, que permitem o acesso e a partilha de informação entre SI diferentes, de forma simples e eficiente [28].

Agência para Integração, Difusão e Arquivo de Informação Médica (AIDA)

A Agência para Integração, Difusão e Arquivo de Informação Médica (AIDA) consiste numa plataforma baseada no uso de agentes pró-ativos que assegura a interoperabilidade entre os diferentes SIH do CHP e outras entidades complementares como *Radiology Information System* (RIS), *Laboratory Information System* (LIS), *Department Information System (*DIS) e *Administrative Information System* (AIS). É um sistema complexo constituído por subsistemas especializados, definidos como agentes inteligentes, que são responsáveis por tarefas como a comunicação entre sistemas heterogéneos, a gestão e o armazenamento de dados, o envio e a receção de informação, como relatórios clínicos, imagens médicas e prescrições [20, 21].

Assim, e a partir da AIDA, é possível integrar, disseminar e arquivar grandes conjuntos de dados de fontes distintas, proporcionando-se fácil acesso a informação registada, facilitando a investigação médica e a aplicação e desenvolvimento de ferramentas computacionais como Sistemas de Apoio à Decisão Clínica (SADC). O objetivo principal é otimizar os serviços prestados numa instituição de saúde [26, 27, 28]

2.1.2 Sistemas de Apoio à Decisão Clínica

Os Sistemas de Apoio à Decisão (SAD) são a área da disciplina de SI que está focada em sistemas que suportam e melhoram a tomada de decisões de gestão. Recorrem a dados e modelos para resolverem problemas não estruturados ou semiestruturados e juntam a capacidade humana de decidir com a informação computorizada, contribuindo assim para um aumento da qualidade da tomada de decisão [29].

Um sistema com a capacidade de analisar os dados clínicos, ligá-los, integrá-los com o conhecimento de um especialista de domínio, e procurar o conhecimento necessário, se possível, noutros sistemas conectados é de enorme utilidade. Este tipo de sistema é um Sistema de Apoio à Decisão Clínica (SADC). Os SADC são SI projetados para ajudar o profissional de saúde na tomada de decisão relativamente à situação clínica dos doentes [10].

Os SADC estão a desempenhar um papel cada vez mais importante na prestação de serviços de saúde, fornecendo recomendações específicas para um determinado paciente através da utilização do conjunto de dados individuais do paciente, orientação

clínica computadorizada e estatísticas populacionais. A integração de SADC em plataformas de registos eletrónicos de saúde mostrou não só enriquecer a qualidade do atendimento clínico, mas também melhorar os resultados clínicos dos pacientes [30].

O reconhecimento da importância dos SADC deve-se ao crescimento inexorável da informação clínica armazenada, ao desafio de oferecer serviços médicos personalizados para cada paciente e também à introdução de requisitos obrigatórios relativamente à implementação deste tipo de ferramentas computacionais no setor da saúde [31, 32].

Conceptualmente, os SADC permitem não só a recuperação de informação relevante, como também a partilha e comunicação da informação no contexto clínico, disponibilizando assim informações específicas e recomendações. Assim, os SADC podem criar e enviar alertas ou recomendações para os pacientes e/ou profissionais da instituição de saúde, organizar e apresentar a informação em *dashboards*, diagramas, documentos e relatórios, de forma a facilitar, acelerar e melhorar o processo de tomada de decisão [31, 32, 33].

O objetivo principal dos SADC é assistir, e não substituir, o profissional de saúde no seu trabalho do dia-a-dia, consistindo numa ferramenta que permite reduzir a incidência de erros médicos, melhorar a qualidade dos serviços prestados aos utentes e reduzir os custos e desperdícios desnecessários associados [31, 32].

2.2 Business Intelligence e a Informação Clínica

Business Intelligence (BI) pode ser definido como sendo um *umbrella term* que combina arquiteturas, bases de dados, ferramentas analíticas, aplicações e metodologias [34, 35].

Com o crescimento acentuado nas duas últimas décadas do número de produtos e serviços de BI oferecidos, bem como na adoção destes por parte das organizações, a área tem vindo a ser identificada como essencial para a melhoria da quantidade e da qualidade da informação disponível para a tomada de decisão nas organizações [34, 36]. Através destes, os profissionais de saúde têm mais facilidade em tomar decisões, de uma forma menos intuitiva e mais fundamentada na informação.

Assim, e segundo Santos e Ramos [37], poder-se-á dizer que os sistemas de BI combinam a recolha de dados operacionais, permitem o seu posterior armazenamento

em repositórios adequados, que por sua vez vão permitir a gestão de conhecimento através de diferentes ferramentas de análise, exploração e apresentação da informação, dita essencial, para a tomada de decisão. De um modo geral, um sistema de BI tem como objetivo [37]:

- Analisar dados passados ou atuais;
- Prever fenómenos e tendências;
- Analisar e comparar dados do passado com novos dados de forma a perceber o que mudou;
- Permitir o acesso *ad-hoc* a dados para responder a questões que não se encontram predefinidas;
- Analisar a organização de modo a obter um conhecimento mais profundo das suas atividades.

No setor da saúde, são geradas grandes quantidades de dados devido aos requisitos obrigatórios progressivamente mais exigentes relacionados com a prestação de cuidados de saúde, impulsionando, assim, a necessidade de uma melhor manutenção e gestão dos registos clínicos [38, 39]. Este vasto conjunto de dados, denominado de *big data*, tem uma enorme potencialidade se for trabalhado de forma a mostrar informação útil e estruturada [35, 39]. Assim, o conceito de BI tem visibilidade por parte dos profissionais de saúde por trazer benefícios relacionados com o acesso a registos clínicos organizados, contribuindo na tomada de melhores decisões num menor intervalo de tempo [40]. Por conseguinte, a implementação de um sistema deste tipo pode contribuir de forma eficiente e precisa no desenvolvimento de uma organização de saúde, criando o conhecimento necessário para ações futuras, de modo a minimizar falhas do passado e maximizar o seu desempenho.

Ferramentas que recorrem às tecnologias de BI incluem a aplicação de uma série de processos incluindo o *Extract*, *Transform* e *Load* (ETL) que se encarrega da extração, limpeza, normalização e carregamento dos dados, a construção de *Data Warehouses* (DW) para estruturação dos dados e ainda a visualização, análise e interpretação da informação representada [35, 36, 40].

Através da revisão de literatura efetuada foi possível identificar várias arquiteturas de sistemas de BI diferentes entre si, de acordo com cada autor e o contexto onde estas estavam a ser aplicadas [36, 41, 42]. Na Figura 3 é apresentada uma arquitetura geral de um sistema de BI incorporando os conceitos apresentados pelos diversos autores, em particular pela arquitetura apresentada por Chaudhuri, Dayal, e Narasayya [36], mostrando de forma simples cada elemento que integra este tipo de sistemas.

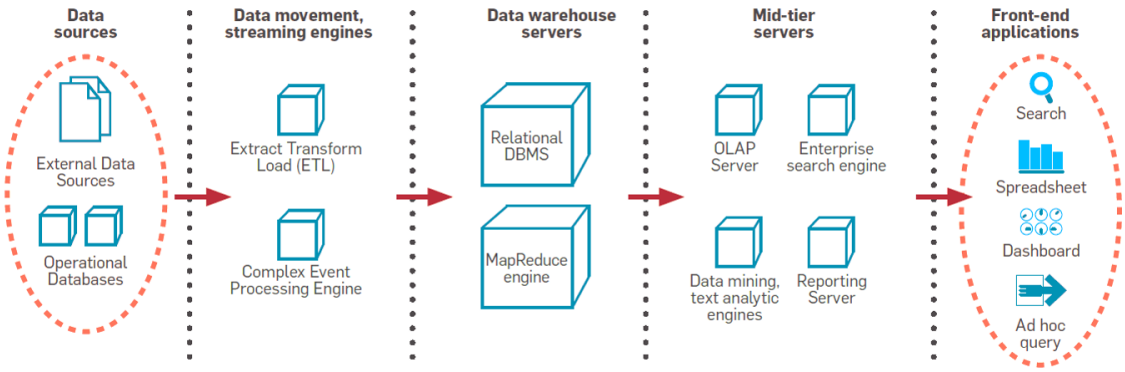


Figura 3 - Arquitetura de um sistema de *Business Intelligence* [36].

A definição de uma arquitetura em camadas facilita a identificação das fases de implementação e tecnologias necessárias para concretizar sistemas de BI. Esta arquitetura é constituída por cinco camadas, correspondendo cada uma destas a um determinado ambiente.

No ambiente de fontes de dados encontram-se todas as origens dos dados que vão suportar o sistema. Estas podem ser internas à organização, como os sistemas bases de dados operacionais, ou externas, como ficheiros Excel. Os sistemas de bases de dados operacionais, conhecidos por *On-Line Transaction Processing* (OLTP), são sistemas concebidos para registar todas as operações do dia-a-dia de uma organização, através das operações de inserção, modificação e eliminação de informação na base de dados num determinado período de tempo.

Na camada seguinte encontra-se o ambiente de movimentação de dados. Aqui realiza-se todo o processo de ETL, que inclui um conjunto de ferramentas especializadas de extração, transformação e carregamento, que permitem tratar a complexidade encontrada nos dados, tratando da sua homogeneidade, a sua posterior limpeza e o

respetivo carregamento para o *Data Warehouse*. Por outras palavras, é o processo de recuperação e transformação dos dados dos sistemas OLTP para o seu posterior armazenamento no DW [43].

O ambiente de DW integra, tal como o nome indica, o DW e os diversos *Data Marts* da organização. Segundo Han e Kamber [41], um DW é considerado um repositório de dados consistentes, através do qual se constitui um modelo de dados de suporte à decisão, armazenando informação relevante para a tomada de decisão estratégica da organização. São uma cópia de registos informacionais de uma transação, estruturados para que sobre eles se possam efetuar interrogações e análises. Os *Data Marts* são repositórios de dados multidimensionais, mais pequenos do que os DW, que reúnem um conjunto de tabelas dimensionais de suporte a um determinando processo de negócio [44].

A penúltima camada representa o ambiente de servidores *mid-tier*. Aqui será possível trabalhar os dados, acedendo ao *Data Warehouse* ou aos *Data Marts*, com recurso a várias técnicas, como OLAP e *Data Mining*, de forma a gerar informação relevante para a tomada de decisão [36].

A última camada representa o ambiente de análise de negócio. O objetivo é permitir o acesso e a manipulação da informação, utilizando ferramentas como *reports* e *dashboards*, de modo a identificar tendências e padrões e a retirar informação útil e acionável [36].

2.3 Descoberta de Conhecimento em Base de Dados

O método tradicional de transformar dados em conhecimento depende de análise e interpretação manual da informação apresentada. Na área da saúde, é comum que os especialistas analisem periodicamente as tendências atuais e as alterações nos dados clínicos, disponibilizando relatórios com uma análise detalhada na organização de saúde. Este relatório torna-se a base para futuras tomadas de decisão e planeamento na gestão de cuidados médicos. Esta análise manual é lenta, cara e altamente subjetiva. Na verdade, com o crescimento exponencial do volume de dados, este tipo de análise de dados manual torna-se impraticável [45].

O KDD é o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e, em última instância, compreensíveis nos dados. A expressão 'não trivial', de acordo com Fayyad et al. [45], pretende transmitir a ideia de que alguma procura ou inferência está envolvida; isto é, não é um cálculo direto de quantidades predefinidas, como calcular o valor médio de um conjunto de números. Os padrões descobertos devem ser válidos em novos dados com algum grau de certeza, assim como potencialmente úteis e que geram benefício para os utilizadores. Finalmente, os padrões devem ser compreensíveis, se não imediatamente, depois de algum pós-processamento.

O KDD é composto por vários passos, que envolvem a preparação dos dados, a procura de padrões, a avaliação e reavaliação do conhecimento, todos estes passos repetidos em múltiplas iterações [45]. As cinco fases principais encontram-se descritas abaixo e representadas na Figura 4.

- Seleção: seleção do conjunto de dados a utilizar no processo de descoberta de conhecimento;
- Pré-processamento: inclui a limpeza e o processamento dos dados, de modo a torná-los consistentes e fiáveis;
- Transformação: os dados são modificados, de acordo com a variável alvo selecionada;
- *Data Mining*: definição dos objetivos que se pretendem atingir e o tipo de resultado que se pretende obter. De acordo com o resultado desejado, é definido o tipo de metodologia a seguir e identificado o tipo de técnica a utilizar. Posteriormente, é aplicada a técnica de DM selecionada ao conjunto de dados, de modo a originar a obtenção de padrões;
- Interpretação/Avaliação: consiste na interpretação e avaliação dos padrões obtidos. A validade dos resultados obtidos pode ser avaliada aplicando esses padrões a novos conjuntos de dados.

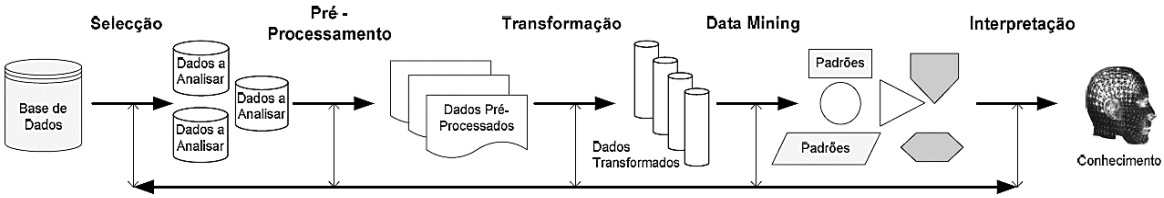


Figura 4 – Processo de Descoberta do Conhecimento Base de Dados (adaptado de Fayyad et al. [45]).

2.4 Data Mining

Data Mining (DM) é a extração de informação implícita dos dados, previamente desconhecida, e potencialmente útil [46]. A ideia subjacente ao processo de DM é criar um modelo que permita analisar e extrair padrões dos dados de forma automática, transformando-os numa estrutura compreensível e utilizável nos passos seguintes do processo de DCBD. Estes padrões são utilizados na deteção de dependências entre os dados, para explicação, compreensão, predição ou classificação de novos dados [46]. Este termo é muitas vezes usado como sinónimo, ou simplesmente como um passo essencial no processo de KDD [41, 46]. Assim, pode afirmar-se que DM é o conjunto de técnicas e estratégias usadas na extração de padrões discerníveis, relações e tendências nos dados.

Como tecnologia integrante de um sistema de BI, o DM pode ser aplicado sobre diversas áreas, como no comércio/retalho, na saúde, na banca, na indústria agrícola e na indústria imobiliária.

2.4.1 Taxonomia do Data Mining

O processo de DM é composto por várias fases entre as quais a modelação. É nesta fase que são aplicados algoritmos ao conjunto de dados e se identificam padrões que ajudam na classificação. Os algoritmos aprendem o padrão dos dados, criam uma função para aplicar aos novos dados, tentando aproximar-se o mais possível da classificação correta dos mesmos. Os algoritmos podem ser classificados de acordo com a forma como é realizada a aprendizagem, supervisionada e não supervisionada [47, 48]:

- Aprendizagem supervisionada: ocorre a partir de conhecimento prévio da classificação dos casos no conjunto de treino do algoritmo, classificação essa que posteriormente é utilizada no processo de aprendizagem. É uma abordagem direta, que é seguida quando o utilizador pretende testar a validade de uma hipótese formulada por si, através dos dados, de modo a comprová-la ou a negá-la;

- Aprendizagem não supervisionada: o algoritmo infere correlações ou agrupa os dados de acordo com algumas características comuns, desconhecendo a classificação de cada um dos casos. É uma abordagem indireta, onde a ênfase está no sistema, que descobre automaticamente a informação relevante nos dados.

As funcionalidades de DM são utilizadas para especificar tipos de padrões que podem ser encontrados através de tarefas que se dividem em duas categorias:

- Descrição: orientada à interpretação dos dados, permitindo aumentar o conhecimento acerca dos dados analisados e a relação entre eles, através da descrição dos mesmos, do reconhecimento de padrões regulares e da definição de regras e critérios que podem ser facilmente compreendidos [41, 49]. Por exemplo, uma organização de saúde pode utilizar técnicas de DM para agrupar os pacientes por faixa etária e perceber em qual existe maior incidência de determinada doença, de forma a monitorizar os pacientes mais críticos.

- Previsão: orientada à identificação de modelos de comportamento, capazes de prever os valores de uma ou mais variáveis a partir de valores já conhecidos de outras variáveis, relacionadas com a amostra, utilizando para isso atributos que se encontram na base de dados [41, 49]. Por exemplo, uma organização de saúde pode utilizar este tipo de técnica de DM para prever as camas das urgências que poderão estar ocupadas num determinado período do dia.

A Figura 5, adaptada de Maimon e Rokach [49], apresenta de forma simples uma taxonomia de DM, em termos de abordagens, técnicas e objetivos.

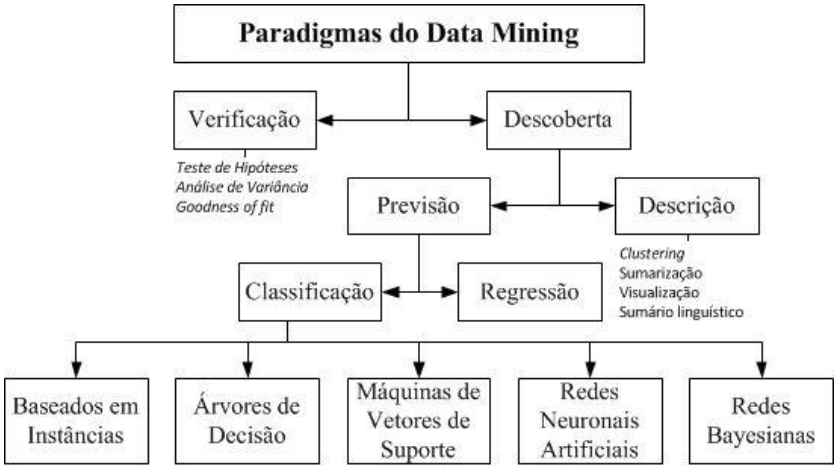


Figura 5 - Taxonomia de *Data Mining* (adaptada de Maimon e Rokach [49]).

Os modelos de Classificação são baseados na aprendizagem supervisionada uma vez que permitem descobrir relações entre os atributos de entrada (variáveis independentes) e os atributos de saída (variáveis dependentes). Envolvem a criação de um modelo ou função capaz de distinguir classes de dados, de forma a que, quando aplicado a novos dados, seja possível a previsão da classe a que esses novos dados pertencem [50, 51]. As técnicas de Classificação mais comuns em DM são as Árvores de Decisão, as *Support Vector Machines* (SVM), as Redes *Bayesianas* e as Redes Neuronais Artificiais [49, 52].

Os modelos de Regressão consistem na descoberta de uma função capaz de representar, de forma aproximada, a distribuição dos dados. Com base nos atributos disponíveis, o objetivo é prever, aproximadamente, o valor da variável alvo para cada observação [53]. A função de regressão mais conhecida é a Regressão Linear.

2.4.2 Algoritmos de Aprendizagem

São considerados vários tipos de aprendizagem: *Inductive Learning* (IL), aprendizagem baseada em instâncias (*Instance-Based Learning* (IBL)) e aprendizagem probabilística (*Statistical Learning* (SL)). A caracterização de cada uma tem diferenças na literatura, pelo que surgem diferenças entre autores. No entanto, verificam-se características comuns entre os algoritmos agrupados em cada um destes conjuntos.

Neste processo é importante ter a perceção da capacidade de aprendizagem do algoritmo. O efeito de *overfitting* acontece quando o modelo gerado se adapta bem ao

conjunto de treino disponível, no entanto apresenta resultados insatisfatórios quando aplicado a um conjunto de teste. Por outro lado, o *underfitting* é um problema que surge quando o algoritmo generaliza a aprendizagem, apresentando fraca capacidade para classificar corretamente novos casos, tratando-os todos como pertencentes a uma determinada classe. Assim, um bom algoritmo é aquele que não se especializa nem generaliza em demasia, estabelecendo uma relação de compromisso [54].

Os algoritmos que se incluem na aprendizagem IL são as árvores de decisão e as regras de indução (*Rule-Based Learning* (RBL)). Neste caso, os algoritmos tentam construir um qualquer conjunto de regras que permitam classificar um novo caso a partir dos atributos que constituem o conjunto de dados.

Árvores de Decisão (AD)

As árvores de decisão (AD) são dos algoritmos de DM mais relevantes [49]. Baseiam-se na classificação de uma variável, construindo para o efeito uma árvore que tem por base a partição recursiva dos casos pertencentes ao conjunto de treino, e cuja classificação e valores dos atributos são conhecidos [55]. Cada partição é designada por nó. A árvore é construída até se atingir um nó folha onde, de acordo com o ramo da árvore e o caminho percorrido, se classificam as amostras que venham a fazer parte desse nó. De acordo com o número de casos conhecidos que façam parte desse nó, pode estipular-se a probabilidade de um determinado caso pertencer a uma dada classe. O desafio é a escolha da melhor divisão em cada nó por forma a criar uma árvore que classifique corretamente e de forma eficiente uma nova amostra. O processo envolve a escolha em cada nó da árvore, do atributo mais significativo, ou seja, que separa o maior número de casos de classes diferentes. O processo repete-se até atingir o critério de homogeneidade ou até que sejam satisfeitos critérios de paragem impostos à árvore [46]. Na Figura 6 está representada uma árvore de decisão simples para representar graficamente o processo de classificação através deste algoritmo.

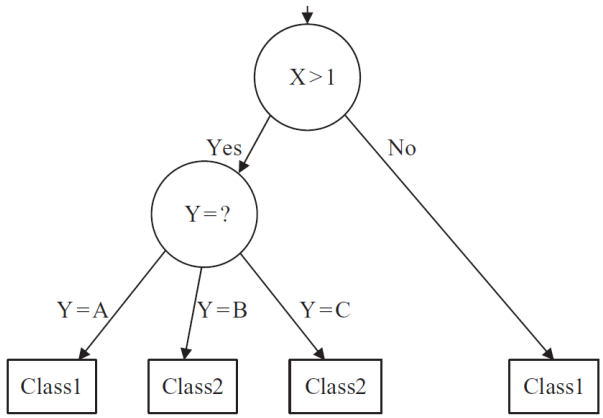


Figura 6 – Árvore de decisão com testes sobre os atributos X e Y, de modo a classificar instâncias na Classe 1 ou na Classe 2 [47].

Uma desvantagem das AD é a especialização na deteção de casos similares aos presentes no conjunto de treino. Por outro lado, têm a vantagem de não serem "caixas negras", ou seja, é possível analisar como o modelo é construído e explicá-lo através de regras. Esta vantagem faz com que sejam amplamente utilizadas na área da saúde [56].

Algumas implementações das árvores de decisão que se destacam incluem o ID3, C4.5, C5 de Quinlan [57] e o CART (*Classification and Regression Trees*) de Breiman et al. [58].

Regras de Indução

As regras de indução analisam o conjunto de treino e tentam encontrar um conjunto de regras que permitam classificar com o maior grau de certeza os novos casos. As regras podem decompor-se sob a forma de $\{Condição\} \rightarrow Y$, em que Y ocorre quando determinada condição é satisfeita. Assim, é possível classificar uma nova ocorrência que satisfaça uma condição previamente conhecida com um determinado grau de confiança, essencial na construção das regras.

Algumas implementações conhecidas deste tipo de aprendizagem são os algoritmos *Apriori*, *Eclat* e *1-Rule* [59, 60].

Relativamente à aprendizagem IBL, os algoritmos mais conhecidos são o *K-Nearest Neighbors* (KNN), as *Support Vector Machines* (SVM) e as *Artificial Neural Networks* (ANN). Esta família de algoritmos tem em conta as instâncias do conjunto de dados, e não um atributo ou conjunto de atributos em especial. A classificação de novas instâncias depende do grau de verosimilhança entre os casos conhecidos e os novos casos. Fazendo uma analogia com os casos onde a aprendizagem é feita tentando inferir regras que permitem classificar novas amostras, os algoritmos desta família guardam as instâncias pertencentes ao conjunto de treino e classificam as novas instâncias comparando-as com as anteriores. Tem como principal desvantagem o espaço necessário de armazenamento das instâncias que compõem o conjunto de treino e servem de base para comparação com as novas instâncias [46].

K-Nearest Neighbors (kNN)

O algoritmo kNN, ou algoritmo dos vizinhos mais próximos, é um método baseado em distâncias sendo a distância Euclidiana a mais usual. A classificação de uma nova instância é feita considerando as k instâncias mais próximos, ou seja, a nova instância é classificada com base nas instâncias do conjunto de treino mais próximas a ela. Em problemas de classificação, considera-se a categoria mais frequente entre os k-vizinhos considerados [47].

Este algoritmo é um '*lazy learner*', ou seja, os modelos não são construídos explicitamente como no caso de *Support Vector Machines* (SVM), por exemplo. A fase de treino consiste simplesmente em determinar o valor de k. Assim, o kNN simplesmente memoriza todas as instâncias do conjunto de treino para posteriormente comparar com as instâncias do conjunto de teste. É por esta razão que este algoritmo está incluído na aprendizagem baseada em instâncias [47].

Na Figura 7 é apresentado um exemplo prático de aplicação do algoritmo kNN, onde três classes (ω_1, ω_2, ω_3) estão representadas por conjuntos de treino. O objetivo é encontrar a classe para a instância x_u pertencente ao conjunto de teste. Neste caso, utilizou-se a distância Euclidiana e um valor de k=5 vizinhos mais próximos [47].

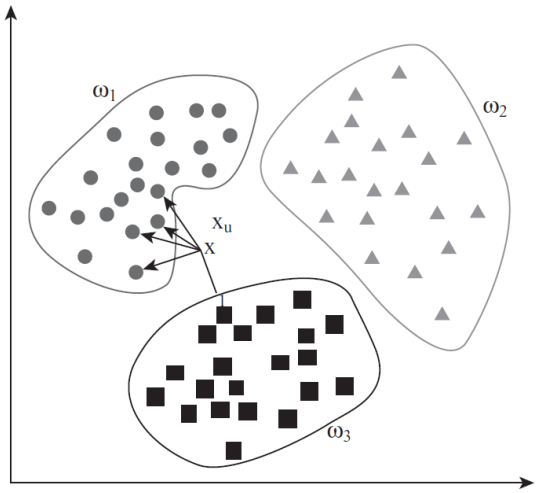


Figura 7 – Algoritmo kNN com k=5 para atribuir a instância x_u a uma das classes (ω_1, ω_2, ω_3) [47].

Dos cinco vizinhos mais próximos, quarto pertencem à classe ω_1 e um pertence à classe ω_3. Assim, a instância x_u é assignada à classe ω_1 pois é a classe predominante da vizinhança.

Support Vector Machines (SVM)

O SVM é um algoritmo supervisionado, inicialmente proposto por Cortes e Vapnick [60], e hoje em dia amplamente utilizado em diversas áreas, entre as quais a bioinformática e as ciências médicas. Considerando o exemplo mais básico de classificação de um caso em uma de duas classes, o objetivo consiste em mapear os casos para um espaço multidimensional, conhecido por espaço das características, e encontrar um hiperplano que separa os casos de cada uma das classes [60]. Dada a possibilidade de existência de vários hiperplanos que separam os casos das duas classes, o objetivo é encontrar o melhor. Assim, o hiperplano deve ter uma margem máxima entre os casos pertencentes a cada uma das classes para minimizar o erro de generalização [61].

Considere-se um problema em que é necessário separar instâncias em duas classes binárias, $y \in \{-1,1\}$. O hiperplano é definido como $< w, x > + b = 0$. A notação $< w, x >$ denota o produto escalar entre os vetores w e x. O conjunto de vetores é dito como otimamente separado pelo hiperplano se for separado sem erro e a distância dos

vetores mais próximos do hiperplano for máxima [47]. A ilustração desta separação de forma gráfica é dada na Figura 8 (a).

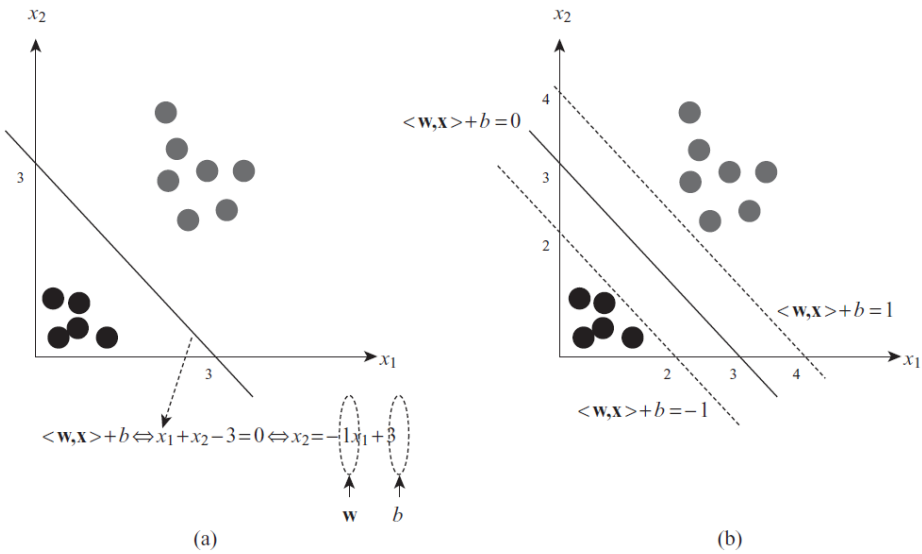


Figura 8 – Representação do hiperplano (w, b) de um problema de classificação binária onde é aplicado o algoritmo SVM. (a) Parâmetros w e b; (b) dois hiperplanos paralelos definem as margens [47].

De forma a que o hiperplano separe de forma correta as duas classes, este tem que respeitar duas condições:

$$< w, x^i > +b > 0 \; para \; todo \; y^i = 1$$
$$< w, x^i > +b < 0 \; para \; todo \; y^i = -1$$

O que é equivalente a dizer que as instâncias têm de permanecer no sítio correto do espaço das características, tendo em conta a sua classe [47].

Na Figura 8 (b) estão representadas duas linhas a tracejado, que representam quando a função $< w, x^i > +b$ é igual a -1 (abaixo e à esquerda do hiperplano) e quando a função $< w, x^i > +b$ é igual a 1 (acima e à direita do hiperplano). De forma a maximizar as margens do hiperplano, as linhas paralelas a tracejado devem estar equidistantes do mesmo e maximizar a distância entre elas, enquanto satisfazem a condição de que as instâncias permanecem no espaço das características da classe que lhes foi associada [41, 47].

Artifical Neural Networks (ANN)

O termo rede neuronal aplicado à classificação de dados teve a sua origem pela analogia que existe entre a rede cerebral, composta por neurónios interligados, e uma rede de dados com elementos de processamento, também interligados, em que ambos respondem a estímulos (no caso do cérebro) ou a dados introduzidos (no caso da classificação de dados). Tal como as SVM, também as redes neuronais fazem parte da família de algoritmos IBL, dado que a aprendizagem é feita instância a instância, adaptando os pesos de cada um dos nós que constituem as redes neuronais [47].

No contexto computacional, os neurónios biológicos são substituídos por nós e as ligações entre neurónios, sinapses, são substituídas por conexões [62]. A cada conexão é atribuído um peso aleatório no início da aprendizagem e que é ajustado durante a fase de treino quando surge uma nova instância. A cada nó cabe o cálculo dos pesos das conexões que entram nesse nó e que ativam a saída caso seja ultrapassado um determinado valor [47].

A rede neuronal é usualmente constituída por três camadas, uma com as entradas no sistema, outra com as saídas e uma camada intermédia, também designada por camada oculta. Na Figura 9 está representada a arquitetura de uma rede neuronal simples, onde os círculos representam os nós da rede e as setas entre círculos representam as conexões entre os nós.

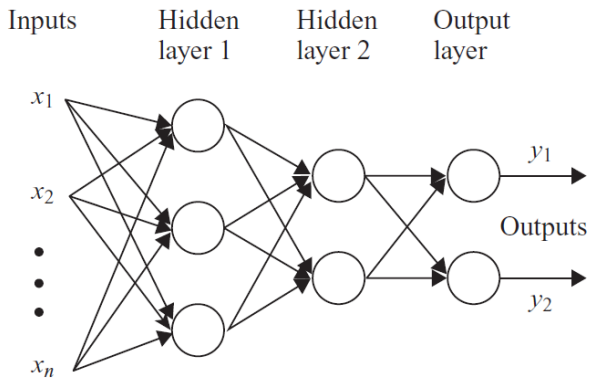


Figura 9 – Arquitetura típica de uma rede neuronal artificial [47].

Este algoritmo é, de entre os métodos conhecidos, um dos mais eficazes, de que é exemplo o algoritmo de *backpropagation*, que aprende os pesos de uma rede multicamada, dada uma rede com conjunto fixo de unidades e interligações [41].

Relativamente à aprendizagem probabilística (SL), esta é baseada no estudo probabilístico do conjunto de dados, sendo que em vez de inferir relações entre os atributos ou instâncias, é feito um estudo estatístico com base nos atributos do conjunto de dados, determinando a probabilidade de uma nova instância pertencer a uma classe. Os algoritmos baseados no teorema de *Bayes* são uma das famílias mais representativas deste tipo de aprendizagem, sendo o *Naive Bayes* (NB) um dos algoritmos mais simples e mais utilizado no processo de DM. O algoritmo *Logistic Regression* é outro algoritmo de DM que tira partido do ramo da estatística.

Naive Bayes (NB)

O NB é um algoritmo probabilístico baseado no teorema de *Bayes*, representado sob a expressão

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

onde A e B são dois acontecimentos e $P(A|B)$ é a probabilidade de A dado B [47].

Tem como pressuposto de que todos os atributos são independentes entre si face ao atributo classificador [63]. Esta assunção torna este algoritmo eficiente, especialmente nos casos cujos atributos não são fortemente relacionados.

É um algoritmo supervisionado, que simplifica a aprendizagem e que na prática compete frequentemente com algoritmos mais sofisticados. Calcula a probabilidade explícita da hipótese, permitindo determinar a incerteza associada ao modelo, sendo robusto no que diz respeito ao ruído nos dados de entrada [64].

Logistic Regression (LR)

A regressão linear é utilizada para modelar funções com valores contínuos. No entanto, modelos generalizados de regressão linear representam a base teórica de que a regressão linear pode ser aplicada a modelos com variáveis categóricas. *Logistic Regression* (LR) é um modelo generalizado de regressão linear que calcula a probabilidade de determinado evento ocorrer como uma função linear de uma série de variáveis preditivas. Em vez de prever o valor da variável dependente, o LR estima a

probabilidade desta variável pertencer à classe 1 ou à classe 0. Este método é utilizado unicamente em problemas de classificação binária. Demostra ser simples e robusto na aplicação a problemas de DM de previsão binária [47].

2.4.3 Otimização

Durante a Secção anterior, descreveram-se diferentes tipos de aprendizagem, aprendizagem indutiva (IL), aprendizagem baseada em instâncias (IBL) e aprendizagem probabilística (SL), que classificam novas instâncias de acordo com um modelo singular construído para o efeito. Embora este possa produzir bons resultados, o objetivo é procurar obter o melhor desempenho possível. Para que isso seja possível, o processo de modelação engloba uma componente de otimização.

Ensemble Learning

A metodologia *ensemble learning* tem como objetivo combinar os resultados de vários modelos preditivos, construídos a partir de conjuntos de treino. Mesmo que isoladamente estes modelos tenham uma prestação fraca, em conjunto podem formar um modelo com melhor desempenho. Assim, e através da combinação de múltiplos e independentes '*decision makers*', as decisões corretas resultantes da modelação são reforçadas, reduzindo-se a taxa de erro [47]. Existem duas técnicas que aplicam o *ensemble learning: boosting* e *bagging*.

A técnica *boosting* consiste na criação de diversos modelos de forma iterativa, e na atribuição de um peso a cada uma das instâncias do conjunto de dados. O peso de cada instância é tido em conta na probabilidade de determinada instância constar no conjunto de dados seguinte. Assim, na primeira iteração todas as instâncias têm igual peso, sendo que o peso de uma instância vai variando de acordo com a correta ou incorreta classificação em cada iteração. Se a classificação for correta, o seu peso diminui, caso contrário, aumenta, de forma a aumentar a probabilidade dessa amostra ser incluída na iteração seguinte. Cada iteração seguinte terá como foco os casos mal classificados nas iterações anteriores. No final, é dado um peso aos algoritmos usados em cada iteração

de acordo com a sua precisão [41, 46]. Uma das implementações com maior sucesso e representativas desta técnica é o *AdaBoost*.

A técnica *bagging*, nome derivado de *bootstrap aggregation*, distingue-se por gerar *n* modelos usando o mesmo algoritmo mas com distribuições de dados diferentes. Cada modelo é construído com uma fração do conjunto de treino, usando a técnica de *bootstrap* que consiste na escolha aleatória de amostras com reposição. A classificação final do modelo dá-se por maioria de votos. Por exemplo, se o *bagging* for composto por três modelos simples e dois deles classificarem uma instância como positiva e outra como negativa, a resposta do modelo final indicará que essa instância é positiva. Uma implementação de sucesso desta técnica é o *Random Forest* [62].

Random Forest

Random Forest é resultante da combinação de um grande número de modelos baseados em árvores de decisão. A obtenção destes modelos é feita através da utilização da técnica *bagging*. Em cada nó da árvore é escolhido um pequeno conjunto de atributos aleatoriamente a partir dos quais se escolhe a melhor partição dos dados desse nó. Esta técnica permite obter um conjunto aleatório de árvores, cada uma especializada na deteção de determinadas especificidades do conjunto de dados [62].

Embora este algoritmo seja comparável com o *Adaboost* em termos de precisão, é mais robusto no que diz respeito à deteção de erros e *outliers*. Para além disso, e como considera poucos atributos em cada partição, é eficiente em grandes conjuntos de dados [62].

Cost-sensitive Classification

Nos problemas de classificação mede-se o desempenho dos classificadores através da taxa de erro. A taxa de erro é a proporção dos erros encontrados entre todas as instâncias e é indicativa da performance global do classificador. Grande parte dos algoritmos de classificação assume que todos os erros têm o mesmo custo e são, normalmente, desenhados para minimizar o número de erros. Nestes casos, a taxa de erro é equivalente à atribuição do mesmo custo para todos os erros de classificação, ou seja, por exemplo,

no caso de uma classificação binária, os falsos positivos teriam o mesmo custo que os falsos negativos. No entanto, em muitas situações do dia-a-dia, cada erro poderá ter um diferente custo associado. Assim, é importante ter em conta os diferentes custos associados às decisões, ou seja, às classificações obtidas.

Na área da medicina, por exemplo, considere-se um diagnóstico de um paciente com cancro como a classe positiva e um diagnóstico de paciente saudável como a classe negativa. Se porventura um paciente com cancro for incorretamente classificado como saudável, ou seja, for classificado incorretamente como pertencente à classe negativa, sendo um Falso Negativo (FN), então é muito mais grave e tem um custo superior ao erro inverso, de um Falso Positivo (FP). O paciente pode perder a vida devido ao diagnóstico incorreto e consequente atraso na iniciação do tratamento [65].

Assim, a classificação sensível ao custo (*Cost-sensitive Classification*) é uma técnica que permite diferenciar os custos associados a cada classe a prever num modelo de DM. É uma das áreas de pesquisa mais ativas e importantes em *machine learning* e desempenha um papel importante em aplicações de DM reais [66].

Esta técnica aplica-se aos problemas e modelos de DM, através de meta-classificadores. Os meta-classificadores podem ser aplicados sobre qualquer classificador já construído e fornecer previsões alteradas, no sentido de minimizar os custos de má classificação. O *CostSensitiveClassifier*, um meta-classificador desenvolvido por Witten e Frank [46], permite duas abordagens para tornar um classificador sensível aos custos de má classificação: manipulação dos dados de treino ou manipulação dos resultados.

Na primeira abordagem, considera o custo total atribuído a cada classe para atribuir diferentes pesos às instâncias de treino, ou seja, faz *reweighting*. Na segunda abordagem, ajusta o modelo para que a classe prevista seja a que tem menores custos esperados de má classificação, e não a mais frequente (*relabeling*). Como para uma determinada instância, os algoritmos de classificação calculam internamente uma probabilidade para cada classe, torna-se possível alterar o resultado da classificação, através do ajuste da matriz de custo (FN e FP no caso da classificação binária).

2.4.4 Balanceamento dos Dados

O desequilíbrio entre classes numa classificação binária pode ser encontrado em muitos problemas de classificação reais. Na tentativa de prever eventos como a intrusão numa rede privada, fraude bancária e o diagnóstico clínico de um paciente, o objetivo passa por identificar as instâncias da classe minoritária, isto é, a classe sub-representada no conjunto de dados. Como os modelos preditivos tendem a ignorar a classe minoritária, constroem modelos precisos, mas que não respondem ao problema, ao privilegiarem a classe com maior número de instâncias [67]. Existem várias técnicas para balanceamento das classes de um conjunto de dados, sendo que todas têm como base o *undersampling* ou o *oversampling*.

A técnica de *undersampling* consiste no balanceamento através da redução das amostras da classe maioritária. O método mais simples denomina-se por *Random undersampling* (RUS) e tem como objetivo remover, de forma aleatória, amostras da classe maioritária até perfazer o número de amostras da classe minoritária. Tem como desvantagem a possibilidade de remover casos relevantes da classe maioritária, casos estes que poderiam ter impacto na classificação de novos casos [68].

Por sua vez, a técnica de *oversampling* consiste na replicação aleatória das amostras da classe minoritária até perfazer o número de amostras da classe maioritária. Tem como principal desvantagem a hipótese de se verificar *overfitting* do algoritmo, onde este tende a especializar-se na classificação dos casos replicados, diminuindo a precisão na classificação de novos dados [68].

De forma a ultrapassar as dificuldades encontradas com ambas as técnicas, surgiram métodos inteligentes para retirar amostras do conjunto de dados e também para sintetizar novas amostras.

Cluster undersampling

Segundo Show-Jane e Yue-Shi [69], podem existir grupos de dados com características distintas (*clusters*) no conjunto de dados disponível. O *undersampling* baseado em *clustering* tem como objetivo selecionar um número representativo de instâncias da classe maioritária de cada *cluster*, considerando o rácio entre o número de amostras da classe maioritária e o número de amostras da classe minoritária.

O primeiro passo da metodologia é dividir o conjunto de dados de treino em *K clusters* com número de instâncias idêntico. Assume-se que o número de amostras no conjunto de dados tem o valor N, incluindo MA instâncias da classe maioritária e MI instâncias da classe minoritária. O tamanho de MA é representado por $Size_{MA}$ e o de MI por $Size_{MI}$, sendo $Size_{MA}$ muito superior a $Size_{MI}$ no conjunto de dados. O número de instâncias do conjunto maioritário e o número de instâncias do conjunto minoritário no *cluster i* ($1 \leq i \leq K$) é igual a $Size^i_{MA}$ e $Size^i_{MI}$ respetivamente. Assim, o rácio do número de MA para o número de MI no *cluster i* será de $Size^i_{MA}/Size^i_{MI}$. Supõe-se que o rácio de $Size_{MA}$ para $Size_{MI}$ no conjunto de treino é m:1, sendo $m \geq 1$ [69]. O número de instâncias a selecionar no *cluster i* será de acordo com a expressão

$$SSize^i_{MA} = (m \times Size_{MI}) \times \frac{Size^i_{MA} \Big/ Size^i_{MI}}{\sum_{i=1}^{K} Size^i_{MA} \Big/ Size^i_{MI}}$$

em que $(m \times Size_{MI})$ representa o número total de instâncias MA do conjunto de treino final e $\sum_{i=1}^{K} Size^i_{MA} \Big/ Size^i_{MI}$ representa o rácio total do número de instâncias MA em relação ao número de instâncias MI em todos os *clusters* [69].

Assim, a expressão determina as instâncias da classe maioritária a selecionar em cada *cluster* construído. O valor de $SSize^i_{MA}$ é tanto maior quanto mais instâncias MA e menos MI o *cluster i* possuir [69].

Depois de determinar o número de instâncias MA a selecionar em cada *cluster* construído, estas são escolhidas de forma aleatória dos K *clusters*. Acrescentam-se todas as instâncias MI, ficando construído o conjunto de treino final. O rácio de $Size_{MA}$ em relação ao $Size_{MI}$ é representado por m:1 neste novo conjunto [69].

SMOTE

Chawla et al. [70] propõem uma abordagem de *oversampling* através da criação de exemplos sintéticos, ao invés da replicação de amostras existentes, dando ênfase ao 'espaço das características' e não ao 'espaço dos dados'.

Assim, o método SMOTE, *Synthetic Minority Over-sampling Technique*, permite aumentar o número de amostras do conjunto minoritário, gerando amostras sintetizadas a partir de um determinado número de amostras vizinhas de cada uma das amostras [68, 70]. As amostras vizinhas consideradas para a sintetização da nova amostra também pertencem ao conjunto minoritário. A nova amostra é gerada tendo em conta a diferença entre a amostra e as amostras vizinhas, isto é, dado por

$$S_i = X_i + \gamma(R_i - X_i)$$

tal que S representa o conjunto das novas amostras, R o conjunto das amostras vizinhas e X o conjunto de amostras da classe minoritária. γ define uma variável aleatória no intervalo [0,1], sendo independente das outras variáveis [70]. O conjunto de treino resultante é composto por todas as amostras do conjunto minoritário, as amostras sintetizadas e as amostras escolhidas aleatoriamente do conjunto maioritário.

As amostras vizinhas são proporcionalmente utilizadas, dependendo na quantidade necessária de amostras a sintetizar. Utilizando a técnica SMOTE, mais regiões associadas à classe minoritária são aprendidas, permitindo aos algoritmos classificar de forma mais eficaz dados novos pertencentes a esta classe [68].

2.4.5 Conjuntos de Treino e de Teste

Um dos passos do processo de DM é a realização de testes para validação dos resultados, por aplicação dos algoritmos no conjunto de dados. O objetivo é que os modelos construídos possam ser aplicados a novos casos, sendo desejável que não se especializem unicamente na classificação dos casos conhecidos, mas consigam também classificar, de forma eficiente, instâncias desconhecidas [46].

Normalmente são utilizados dois conjuntos de dados. Um primeiro para uso na aprendizagem, designado por conjunto de treino, e um segundo para casos de teste, designado por conjunto de teste. Visto que em muitos casos os dados disponíveis são escassos, é necessário encontrar uma abordagem que permita simular situações que se aproximem o mais possível com a realidade. Duas das técnicas mais utilizadas são: *Holdout sampling* e *k-Fold Cross Validation*.

Holdout sampling

O *Holdout sampling* é uma técnica simples e geralmente aceite nos processos de DM. Consiste na divisão do conjunto de dados em dois subconjuntos, um para treino e outro para teste. Estes devem ser disjuntos de forma a avaliar o seu desempenho em casos potencialmente distintos daqueles sob os quais se efetua o treino. Esta disjunção é garantida pela escolha aleatória dos casos que fazem parte de cada um dos subconjuntos, sendo a percentagem que deve constar em cada um dos conjuntos parametrizável, tendo em conta a sua dimensão [71].

Relativamente a esta percentagem, quanto maior for o conjunto de treino, melhor será o classificador, até um determinado tamanho, a partir do qual o classificador piora. Quanto maior for o conjunto de testes, maior será a estimativa de erro [46]. É importante que seja mantida a representatividade dos dados em ambos os conjuntos, ou seja, verificar-se uma relação de compromisso entre uma melhor aprendizagem, menor estimativa de erro e maior precisão. Na prática, uma distribuição geralmente aceite e amplamente utilizada, é de $^1/_3$ dos dados para teste e os restantes $^2/_3$ para treino [72].

Note-se que a preservação da proporcionalidade entre as classes no conjunto de treino e de teste é importante para que a aprendizagem não vicie o resultado da classificação. No entanto, não é possível assegurar que os conjuntos sejam efetivamente representativos. O problema é minimizado repetindo o processo diversas vezes de forma aleatória, obtendo vários pares de conjuntos de treino e teste, aos quais são aplicados os algoritmos de aprendizagem. O erro global é obtido pela média ponderada dos erros dos conjuntos de treino e teste. Quantos mais testes forem realizados, menor será o erro global [46].

k-Fold Cross Validation

Um dos problemas da técnica de *Holdout sampling* é a não garantia de representatividade do conjunto de dados. Assim, e aproveitando a noção de repetição do processo para obter representatividade dos dados, foi desenvolvida uma técnica conhecida por *k-Fold Cross Validation*. Uma das implementações mais conhecidas e geralmente aceite é o *10-Fold Cross Validation* [46].

Esta técnica consiste na divisão do conjunto de dados em dez partes iguais, mantendo a proporcionalidade entre os casos de cada uma das classes. São feitas dez iterações, sendo que em cada iteração será usada uma parte diferente do conjunto de dados para teste. Em cada uma das iterações, o conjunto de treino é constituído por 90% dos casos e o conjunto de teste é composto pelos restantes 10%. Todo o conjunto é testado, pelo que é necessário a realização de dez aprendizagens e correspondentes testes. A aplicação desta técnica pode não ser suficiente para garantir uma boa estimativa de erro, pelo que a repetição desta técnica por dez vezes, é um procedimento habitual [46].

2.4.6 Métricas para Aferição e Avaliação

A escolha de um bom método de avaliação é essencial para a obtenção do modelo mais adequado ao problema identificado. São usadas várias métricas para avaliação e comparação dos modelos: matriz confusão, sensibilidade, especificidade, curva *Receiver Operating Characteristic* (ROC) e *Area under the ROC Curve* (AUC).

A matriz de confusão facilita a visualização do número de classificações corretas e do número de classificações previstas para cada classe, de um determinado conjunto de dados, segundo o classificador em análise. Torna-se uma ferramenta útil para analisar a qualidade do classificador no reconhecimento de instâncias de diferentes classes [41]. Quando um conjunto de dados tem apenas duas classes, é muitas vezes considerada uma como positiva e a outra como negativa. As entradas da matriz de confusão são referidas como Verdadeiros Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) e Verdadeiros Negativos (VN) [62].

A Tabela 2 apresenta a matriz de confusão para um problema de classificação binária.

Tabela 2 – Matriz de confusão associada a um problema de classificação binária

		Classe a prever	
		Positiva	Negativa
Classe real	Positiva	VP	FN
	Negativa	FP	VN

As instâncias VP referem-se ao número de exemplos da classe positiva corretamente classificados e as instâncias FN representam o número de exemplos da classe positiva incorretamente classificados como sendo pertencentes à classe negativa. Da mesma foram, as instâncias VN são relativas a exemplos da classe negativa corretamente classificados e as instâncias FP exemplos da classe negativa incorretamente classificados na categoria positiva.

Com base nas entradas da matriz de confusão, o número total de classificações corretas feitas pelo classificador é a soma de TP e TN e o número total de classificações incorretas corresponde à soma de FP e FN.

Para além disso, é ainda possível o cálculo de diferentes métricas associadas aos valores da matriz: exatidão, sensibilidade e especificidade [52].

- Exatidão: corresponde à percentagem total de concordância entre as classificações corretamente efetuadas e todas as classificações realizadas. Mede-se através da proporção entre todos os resultados corretos ($VP + VN$) e todos os casos possíveis de serem obtidos ($VP + VN + FP + FN$).

$$Exatidão = \frac{VP + VN}{VP + VN + FP + FN}$$

- Sensibilidade: é a capacidade de detetar corretamente a ocorrência de casos positivos. Consiste no resultado da razão entre os valores VP e todos os valores correspondentes a positivos ($VP + FN$)

$$Sensibilidade = \frac{VP}{VP + FN}$$

- Especificidade: corresponde à capacidade de identificar corretamente num modelo a não ocorrência de casos positivos. É medida através da razão entre os valores corretamente identificados como negativos (VN) e todos os valores correspondentes a negativos (VN + FP)

$$Especificidade = \frac{VN}{VN + FP}$$

De acordo com Zweig et al. [73], as curvas ROC são uma ferramenta fundamental na avaliação dos modelos por permitirem uma visualização rápida e intuitiva dos resultados. É um método utilizado frequentemente em técnicas de DM em casos clínicos para comparação eficiente de modelos, consistindo na representação gráfica das métricas de sensibilidade e especificidade. Assim, é possível identificar visualmente o desempenho de um modelo e ao mesmo tempo dar indicações sobre como otimizá-lo considerando os objetivos propostos.

A Figura 10 ilustra três tipos de curvas ROC. O desempenho de cada modelo é medido por proximidade da classificação ótima, ou seja, quando todos os casos são classificados corretamente, as curvas de resposta estão representadas junto ao canto superior esquerdo. A curva 1 distingue os casos de cada uma das classes pelo que se aproxima da classificação perfeita, contrariamente à curva 3 onde os casos estão sobrepostos e como tal a classificação é aleatória. A curva 2 ilustra a situação mais habitual onde a maioria dos casos são distinguíveis e existe apenas uma faixa onde estes se sobrepõem. É a redução desta faixa que otimiza um classificador. O resultado final trata-se de uma relação de compromisso entre as duas vertentes, sensibilidade e especificidade [74, 75].

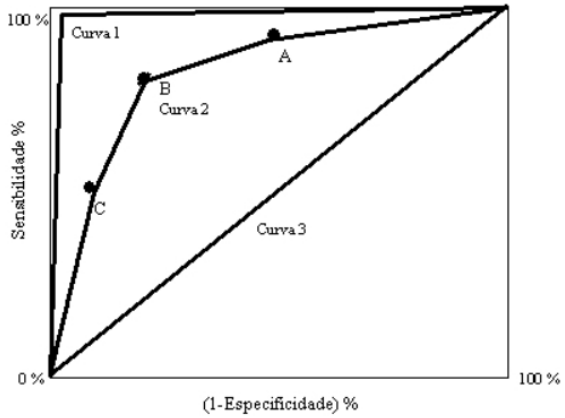


Figura 10 – Curvas *Receiver Operating Characteristic* (ROC).

Já que quanto mais a curva se aproxima do canto superior esquerdo melhor é a qualidade do teste, quanto maior for o valor da área, isto é, quanto mais próximo estiver da unidade, maior é a capacidade para discriminar as classes.

A AUC, *Area Under ROC Curve*, é obtida a partir da análise da curva ROC, e usada como medida de qualidade do desempenho de um modelo [75, 76]. A área abaixo da curva

ROC é um dos índices de precisão mais utilizados para avaliar a qualidade da curva. Num caso real, e para um dado paciente que sobreviveu e outro que faleceu, ambos escolhidos ao acaso, a área abaixo da curva, é uma medida que permite aferir qual a probabilidade do paciente que faleceu obter um resultado verdadeiro positivo e do paciente que sobreviveu obter um resultado verdadeiro negativo.

2.5 Aplicações de *Front-end*

Um sistema de BI deve permitir ao utilizador a interação e o entendimento dos dados, permitindo a sua manipulação, monitorização e compreensão, a fim, por exemplo, de fornecer informação adequada para a tomada de decisão [41]. Uma das questões mais importantes e principais determinantes do seu sucesso, a partir da perspetiva do utilizador final, é a interface. Fornecer aplicações com uma interface *user-friendly* que oferecem a capacidade de criação de relatórios e análises e que agregam num único painel (acessível e de forma imediata) a informação considerada relevante para suportar o processo de tomada de decisão, torna-se num fator crítico de sucesso e uma vantagem no desenvolvimento de um sistema deste tipo.

As tecnologias de *Data Warehousing* e *Data Mining*, são utilizadas para trabalhar e fornecer os dados com intuito de proporcionar às organizações de saúde o acesso à informação. Por sua vez, as aplicações de *front-end* disponibilizam a interface do utilizador, permitindo o acesso aos dados de uma forma simples, sem grande complexidade e mais atrativa. Nestas, encontram-se portais para pesquisa, visualização de KPIs através de painéis visuais, interação com os dados apresentados e possibilidade de manipulação dos mesmos, análise/geração de modelos de DM [36]. A apresentação da informação nestas aplicações ocorre de diversas formas como *dashboards*, tabelas, gráficos e ferramentas interativas de análise multidimensional.

Assim, torna-se indispensável a disponibilização da informação presente num DW e eventualmente resultante da aplicação de um algoritmo de DM, na interface do utilizador em tempo quase real ("*near real-time*"). O objetivo é permitir que a informação seja acionável e ajude no processo de tomada de decisões diárias numa organização.

3. METODOLOGIAS DE INVESTIGAÇÃO E TECNOLOGIAS

A área das TI exige uma investigação de metodologias e tecnologias que estejam disponíveis, sejam válidas e viáveis para o desenvolvimento do produto final pretendido. As várias escolhas realizadas no âmbito deste projeto de dissertação estão relacionadas com questões de vantagem de uma metodologia ou tecnologia sobre outra semelhante e também com questões limitativas que levaram à exclusão de opções.

A metodologia de investigação no qual este projeto assenta é designada por *Design Science Research* (DSR). Uma metodologia robusta, útil e rigorosa utilizada na avaliação de soluções de TI.

No que diz respeito às fases de conceção do projeto, foi analisado um conjunto de metodologias, tecnologias e ferramentas e selecionado desse conjunto as que melhor se adequavam ao desenvolvimento das soluções previamente projetadas.

De forma a testar a viabilidade e utilidade das ferramentas de apoio à decisão clínica que pertencem ao escopo deste projeto, realizou-se uma prova de conceito.

Nas próximas secções deste capítulo estão descritas cada uma das metodologias e tecnologias enquadradas no desenvolvimento deste projeto de dissertação.

3.1 Metodologia de Investigação

3.1.1 Design Science Research

Existem várias discussões na comunidade científica sobre a metodologia de investigação indicada para o desenvolvimento de estudos na área de SI. Segundo Hevner et al. [77], a investigação em SI ocorre no encontro entre pessoas, organizações e tecnologias; desta forma, dois paradigmas distintos e complementares são necessários para adquirir as informações necessárias para melhorar os SI: (1) *behavioral-science* e (2) *design-science*. *Behavioral-science* aborda a investigação através do desenvolvimento e justificação de teorias que explicam ou preveem fenómenos relacionados à necessidade de negócio identificada. *Design-science* aborda a investigação através da construção e avaliação de artefactos projetados para responder à necessidade identificada.

Assim, a ideia chave é a existência de um ciclo de investigação complementar entre *behavioral-science* e *design-science*. A verdade (teoria justificada) e a utilidade (artefactos eficazes) são dois lados da mesma moeda. Assim, o rigor da investigação realizada deve ser tão valorizável como a relevância prática do resultado obtido [77].

Estudos conduzidos pela *Design Science Research Methodology* (DSRM) envolvem a criação de artefactos e a sua incorporação em ambientes físicos, psicológicos, económicos, sociais e virtuais. No contexto da saúde, se estes artefactos forem portadores de um bom design, trouxerem inovação, utilidade e sustentabilidade, vão criar valor e reduzir a resistência por parte dos profissionais de saúde à implementação de novas tecnologias. No contexto concreto deste projeto, cada artefacto construído tem como objetivo auxiliar nos desafios clínicos diários da equipa do CHP, proporcionando soluções apropriadas para os problemas identificados, e incitando, ainda, novo conhecimento, tanto para a organização como a nível científico. Assim, este projeto de dissertação foi conduzido pela metodologia de investigação *Design Science Research* (DSR).

Na área das TI, o principal objetivo da utilização da metodologia de investigação DSR encontra-se relacionado com o desenvolvimento e a avaliação de "artefactos" que permitem o processamento de informação organizacional e o desencadeamento de ações face a um problema [78]. No contexto de resolução de problemas de negócio reais, é fundamental a melhoria da relevância e utilidade do artefacto, sendo que este tem de corresponder a uma solução tecnológica viável que vá de encontro às necessidades identificadas, e a sua utilidade, qualidade e eficácia devem ser rigorosamente demonstradas através de métodos de avaliação bem executados [77].

Hevner et al. [77] fornecem regras práticas para a aplicação da metodologia de DSR na disciplina de TI, na forma de diretrizes que descrevem características de estudos bem estruturados e construídos. A mais importante destas diretrizes é a construção de um "artefacto criado para resolver um problema". O artefacto deve ser relevante para a solução de um "problema importante até então não resolvido" e características como "utilidade, qualidade e eficácia" devem ser rigorosamente avaliadas. A pesquisa deve representar uma contribuição verificável e o rigor deve ser aplicado tanto no desenvolvimento do artefacto como posteriormente na sua avaliação. O desenvolvimento do artefacto deve envolver um processo de pesquisa que se baseie em

teorias existentes para encontrar uma solução para o problema identificado. Finalmente, o estudo deve ser efetivamente comunicado às audiências apropriadas.

Segue-se uma descrição mais detalhada de cada uma das atividades presentes na metodologia DSR.

A atividade 1, identificação do problema e motivação, tal como o próprio nome indica, define o problema de investigação e justifica o valor de uma possível solução. Como a definição do problema será usada para desenvolver um artefacto que pode efetivamente fornecer uma solução, pode ser útil detalhar o problema conceptualmente para que a solução possa capturar a sua complexidade. Justificar o valor de uma solução motiva o investigador a atingir a solução assim como auxilia o entendimento da mesma pelo público da investigação. Os recursos necessários para esta atividade incluem o conhecimento do estado do problema e a importância da sua solução [79, 80].

Na atividade 2, definição dos objetivos, infere-se os objetivos de uma solução a partir da definição do problema e conhecimento do que é possível e viável. Os objetivos podem ser quantitativos, como por exemplo uma solução melhor do que a solução atual, ou qualitativos, como uma descrição de como um novo artefacto deve sustentar soluções para problemas até então não abordados. Os recursos necessários para esta atividade incluem o conhecimento do estado dos problemas e soluções atuais, se estas últimas existirem, assim como a sua eficácia [79].

Relativamente à atividade 3, *design* e desenvolvimento, esta tem como objetivo a criação do(s) artefacto(s). Tais artefactos podem ser construções, modelos, métodos ou instanciações. Conceptualmente, um artefacto de DSR pode ser qualquer objeto projetado no qual uma contribuição de investigação é incorporada no design. Esta atividade inclui a determinação da funcionalidade desejada do artefacto, a sua arquitetura e a sua criação propriamente dita. Os recursos necessários para transformar objetivos em design e desenvolvimento incluem o conhecimento da teoria que pode ser usada para suportar uma solução [79, 77].

A demonstração do uso do artefacto para resolver as instâncias do problema é a atividade 4. Pode envolver experimentação, simulação, caso de estudo, prova de conceito ou outra atividade apropriada. Os recursos necessários para esta atividade incluem o conhecimento efetivo de como usar o artefacto para resolver o problema [79].

A atividade 5 é caracterizada pela avaliação, onde se observa e perceciona o nível a que o artefacto justifica e sustenta a solução para o problema. Esta atividade envolve comparar os objetivos da solução aos resultados reais observados na utilização do artefacto na demonstração. Requer conhecimento de métricas relevantes e técnicas de análise. Dependendo da natureza do problema identificado e do artefacto construído, a avaliação pode assumir várias formas. Pode incluir uma comparação da funcionalidade do artefacto com os objetivos definidos na atividade 2, medidas quantitativas de desempenho, resultados de inquéritos de satisfação, feedback do cliente ou simulações. Pode ainda incluir medidas quantificáveis de desempenho do sistema, como o tempo de resposta e a disponibilidade. No final desta atividade, os investigadores podem decidir se é necessário repetir a atividade 3 para melhorar a eficácia do artefacto ou continuar para a atividade seguinte, deixando possíveis melhorias para projetos subsequentes [79, 77].

O último passo da metodologia é a comunicação, sendo esta a atividade 6. O objetivo é a comunicação do problema e da sua importância, do artefacto, da sua utilidade e caráter inovador, do rigor do seu design e da sua eficácia para investigadores e outros públicos relevantes [79, 77]. Estas atividades encontram-se representadas no esquema ilustrado na Figura 11.

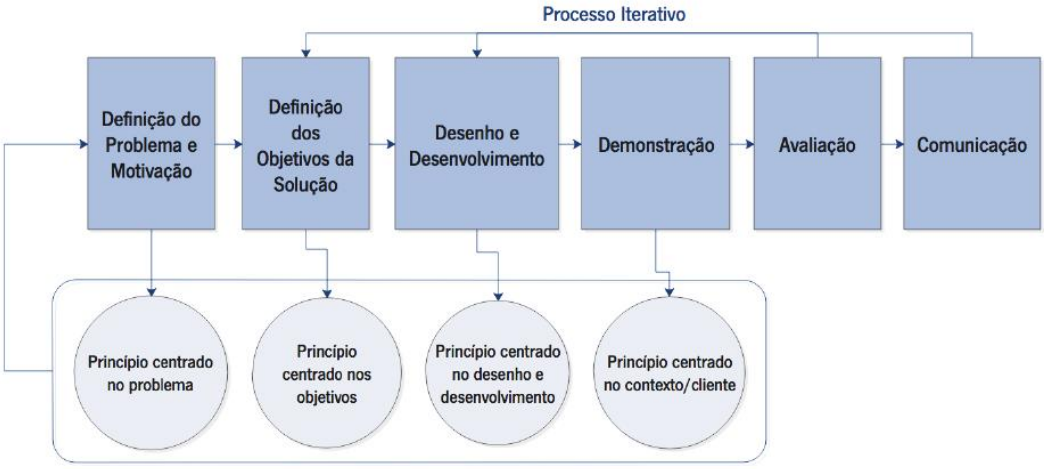


Figura 11 - Representação esquemática da metodologia de investigação DSR (adaptado de Peffers et al. [79]).

Apesar das etapas desta metodologia de investigação estarem estruturadas numa ordem nominalmente sequencial, a investigação em questão pode exigir não prosseguir numa ordem sequencial, ou seja, da atividade 1 até à atividade 6. Se a investigação resultar da observação do problema ou de pesquisas futuras sugeridas num trabalho de um projeto anterior, estamos perante uma abordagem centrada no problema. A investigação inicia então pela atividade 1. Por outro lado, numa solução centrada num objetivo, a sequência começa com a atividade 2, podendo ser desencadeada por uma necessidade identificada pelo negócio que pode ser colmatada pelo desenvolvimento de um artefacto. Uma abordagem centrada no design e no desenvolvimento começa com a atividade 3. É resultante da existência de um artefacto ainda não formalmente definido como uma solução para o domínio explícito do problema no qual será usado. Finalmente, uma solução iniciada pelo cliente, que pode ser baseada na observação de uma solução prática que funciona, começa na atividade 4, resultando numa solução de DSR se os investigadores retrocederem para aplicar rigor ao processo [79].

3.2 Metodologias Técnicas

3.2.1 Frameworks de Desenvolvimento Web

Uma arquitetura REST (*Representational State Transfer*), é uma arquitetura web cliente-servidor para aplicações de rede, que particiona tarefas pelos servidores e pelos que solicitam determinado conteúdo ou serviço, os clientes. Estes dois comunicam através de uma rede de computadores em máquinas separadas usando pedidos HTTP, sendo estes iniciados pelo cliente, que aguardam uma resposta do servidor. O protocolo HTTP é um protocolo de comunicação entre SI, que permite a transferência de dados entre redes de computadores. Baseia-se no conceito de pedido e resposta. O cliente inicia o pedido a ser atendido pelo Web server e este retorna dados como resposta no formato especificado [81].

As principais características de um sistema REST são definidas como [82, 83]:

- Cliente-servidor – há uma separação entre o servidor que fornece o serviço e o cliente que o utiliza;
- Sem estado – cada solicitação do cliente para o servidor contém todas as informações necessárias para entender a solicitação e não pode aproveitar qualquer contexto armazenado no servidor, o que significa que o estado da sessão é mantido exclusivamente no cliente;
- Interface uniforme – o método de comunicação entre um cliente e um servidor é uniforme;
- Sistema em camadas – a comunicação entre um cliente e um servidor deve ser padronizada de tal forma que permita que os intermediários respondam às solicitações em vez do servidor final, sem que o cliente tenha que fazer nada diferente;
- *Cacheable* – o servidor deve indicar ao cliente se as solicitações podem ser armazenadas em cache ou não.

Podem ainda ser identificadas como propriedades de um sistema REST, a visibilidade, a confiabilidade e a escalabilidade. Como um único pedido por parte do cliente contém toda a informação necessária para a resposta por parte do servidor, este último tem visibilidade para determinar os requisitos do pedido completo. A recuperação de falhas parciais é mais simples num sistema com este tipo de arquitetura, havendo maior fiabilidade. Quanto à propriedade de escalabilidade, esta é possível já que o servidor não armazena o estado da sessão entre os pedidos do cliente, libertando assim recursos rapidamente. Assim, a implementação é simples, pois o servidor não precisa de fazer uma gestão da utilização de recursos nas várias solicitações [82].

Os *Web services* APIs que seguem uma arquitetura REST são designados por RESTful web *Application Programming Interfaces*, ou simplesmente por RESTful APIs. Nestes sistemas, os recursos são representados por URIs. Os clientes enviam pedidos para estes URIs usando os métodos definidos pelo protocolo HTTP e, como resultado, uma ação é executada [82, 83]. RESTful APIs baseadas no protocolo HTTP usam os seus métodos *GET* para ler, *POST* para criar, *PUT* para atualizar e *DELETE* para remover - CRUD (*Create, Read, Update, Delete*) [84].

A arquitetura REST é flexível, não requerendo um formato específico para os dados fornecidos com os pedidos e respostas. Assim, o formato das mensagens do sistema pode ser escolhido de acordo com as necessidades específicas do sistema. Os formatos mais comuns são JSON, XML e texto puro, mas em teoria qualquer formato pode ser usado [84].

No contexto deste projeto, os *Web services* enviam pedidos às bases de dados, na forma de *queries*, e o cliente recebe as respostas no formato JSON. Este formato é um tipo de estrutura de dados que deriva da sintaxe de *JavaScript*.

Os componentes de uma aplicação Web podem ser divididos em duas partes: *back-end* e *front-end*. Estes dois distinguem a separação entre a camada de acesso a dados - *back-end* - e a camada de apresentação – *front-end*.

Front-end - AngularJS

As ferramentas de apoio à decisão clínica foram desenvolvidas com o auxílio da *framework* AngularJS. Esta consiste numa ferramenta *JavaScript open-source*, que auxilia na construção de SPAs.

Single Page Applications (SPA) são aplicações compostas por componentes individuais que podem ser substituídos ou atualizados de forma independente, para que a página inteira não precise de ser recarregada em cada ação do utilizador. O objetivo é tornar o carregamento da página mais rápido comparativamente ao ciclo tradicional de solicitação-resposta. Na Figura 12 está representado um esquema de um pedido do cliente ao servidor e a forma como este pedido é tratado com o ciclo tradicional de solicitação/resposta e com uma SPA [85].

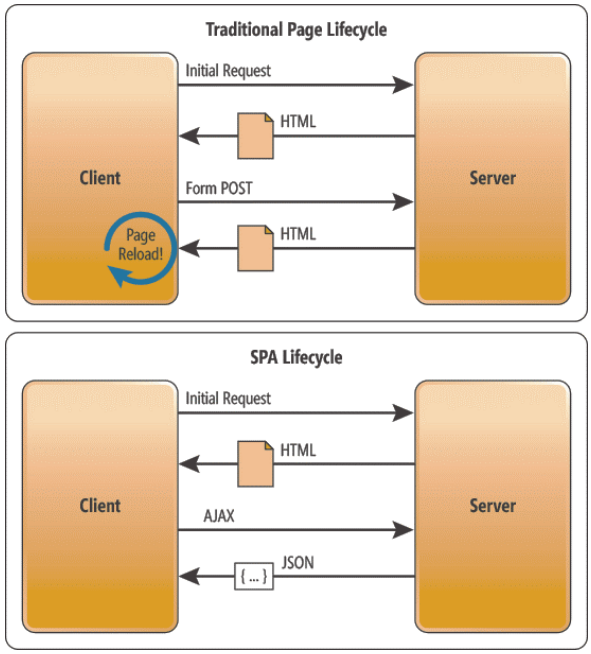


Figura 12 – Ciclo de pedido-resposta do cliente-servidor (adaptado de Jadhav et al. [85]).

A escolha do AngularJS, *framework* SPA, deve-se à sua produtividade no desenvolvimento de aplicações *Web*, bem como devido à sua compatibilidade com o ambiente de implementação das aplicações no CHP, isto é, as máquinas de produção.

As aplicações *Web* desenvolvidas recorrendo à *framework* AngularJS baseiam-se na arquitetura MVC (*Model-View-Controller*, isto é, Modelo-Vista-Controlador), que consiste num padrão arquitetural usado para desenvolver aplicações *Web* [86]. O desenvolvimento de cada aplicação *Web* em AngularJS inicia-se pela criação de um controlador, responsável por controlar o modelo através da vista que é atualizada cada vez que é executada uma modificação [87]. As três componentes da arquitetura MVC podem ser resumidamente descritas da seguinte forma [86, 87]:

- Modelo: responsável pela gestão dos dados da aplicação. Responde ao pedido da vista e às instruções do controlador para se atualizar;
- Vista: consiste na apresentação dos dados num formato específico, desencadeada pela decisão do controlador para apresentar os dados;

- Controlador: responde à entrada do utilizador, procede à sua validação e, de seguida, executa operações sobre os objetos dos modelos de dados que modificam o seu estado atual.

Uma aplicação em AngularJS tem várias vantagens associadas. Exige menos linhas de código para concluir uma tarefa comparativamente com outras *frameworks*. O código é simples, limpo, já que a lógica é transferida para componentes reutilizáveis e fora da vista construída. A maior parte do código desenvolvido é focado na lógica de negócio e consequentemente na funcionalidade principal da aplicação. Oferece uma técnica designada por ligação de dados bidirecional que permite que os dados sejam atualizados sempre que existir uma alteração nas vistas. É uma *Meta-framework* de *Single Page Application* (SPA), onde os controladores comunicam com o servidor sendo responsáveis por controlar o comportamento do SPA [85, 88].

Bootstrap

O *Bootstrap* foi projetado para tornar o *front-end* de uma aplicação mais rápido e mais fácil, tornando as páginas Web dinâmicas e flexíveis. Oferece um grande conjunto de ferramentas para personalizar o modo como o conteúdo é exibido de forma a oferecer uma melhor experiência ao utilizador [89].

A sua capacidade de adaptação a diferentes dispositivos faz do *Bootstrap* uma *framework* única, já que evita a necessidade de criar vários ambientes para diferentes dispositivos, como *smartphones, tablets* ou PCs. Assim, os utilizadores têm a mesma experiência de navegação, quer seja em dispositivos móveis ou *desktop*.

D3.js

D3.js (*Data-Driven Documents*) é uma biblioteca *JavaScript* usada para manipular documentos com base em dados. Com o *D3.js*, é possível associar dados arbitrários a um DOM (*Document Object Model*) e gerar tabelas HTML ou criar gráficos SVG (*Scalable Vector Graphics*) interativos [90].

O nvd3.js é uma coleção de componentes gráficos reutilizáveis baseados no D3.js, simplificando a sua implementação sem perder o seu potencial. A integração do nvd3.js

é simples, intuitiva e eficiente ao personalizar as visualizações de relatórios e dados, especialmente na forma de painéis [91].

Back-end - Flask

Escolher uma *framework* para desenvolvimento web *back-end* pode ser uma tarefa complicada devido ao número de tecnologias existentes disponíveis. Uma *framework* é utilizada para ajudar o *developer* a criar aplicações fiáveis, escaláveis e de fácil manutenção. No caso deste projeto, a escolha recaiu sobre *Flask*.

O *Flask* é uma micro *framework*, minimalista, criada em 2010, com uma estrutura extensível, baseada em *Jinja2* e *Werkzeug* [92]. Foi construído com um núcleo robusto que inclui as funcionalidades básicas utilizadas por todos as aplicações Web, sendo extensível para responder às necessidades do *developer.* Não há suporte em Flask para aceder a bases de dados, validar formulários da web, autenticar utilizadores ou outras tarefas de alto nível. Estas e outras tarefas que a maioria das aplicações Web precisa, estão disponíveis por meio de extensões que se integram com os pacotes principais. Assim, o *developer* consegue fazer uma escolha das extensões que melhor se adaptam ao projeto em questão. Esta é uma característica vantajosa já que, numa *framework* de maior escala, a maioria das opções foi previamente decidida e é difícil ou mesmo impossível de ser alterada [93].

Numa arquitetura MVC, o Flask cobre V e C, ou seja, *View* e *Controller*. O Flask não fornece uma camada de modelo integrada pronta para utilização. Esta camada é adaptada às necessidades do *developer* e eventualmente do utilizador, através de extensões já criadas e de uma camada adaptada à solução a construir [92].

3.2.2 *CRoss Industry Standard Process para Data Mining*

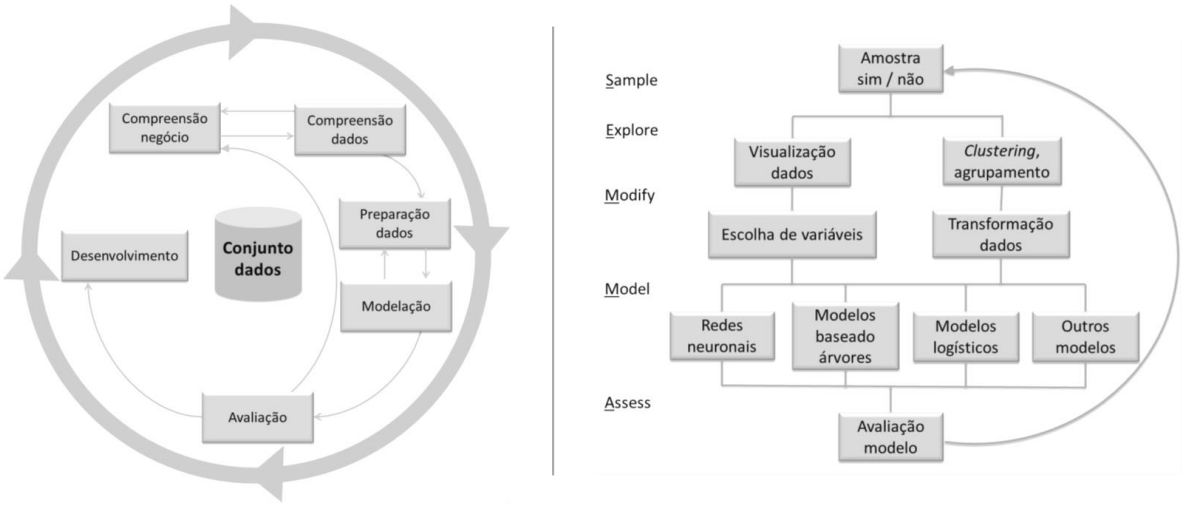
O termo KDD, em português DCBD, foi formalizado em 1989 como uma referência ao conceito mais amplo de procura de conhecimento em dados, e é um processo que envolve a identificação e o reconhecimento de padrões numa base de dados de uma forma automática. Assim, podemos definir KDD como o processo de descoberta de novas correlações, padrões e tendências, por meio de análise a um conjunto de dados

de elevada dimensão e complexidade. O processo KDD está dividido em várias fases, descritas na Secção 2.4, sendo que as de maior relevância incluem pré-processamento, o processo de DM e pós-processamento [45].

No final do século XX, várias entidades seguiram as suas próprias estratégias e métodos, com produção de resultados distintos. Surgiu assim a necessidade de definir uma metodologia que servisse de referência para o desenvolvimento de projetos de KDD. Atualmente, destacam-se duas das metodologias propostas: SEMMA (*Sample, Explore, Modify, Model and Assess*) e CRISP-DM *(Cross Industry Standard Process for Data Mining)* [37].

A SEMMA desenvolvida pelo SAS Inc, usualmente interpretada como uma metodologia refere-se a um processo central de DM. Esta metodologia, representada na Figura 13 (b) inicia com uma amostra estatisticamente representativa dos dados, avançando para a exploração da mesma com a procura de relações e tendências não antecipadas, através de técnicas estatísticas e de visualização. Segue-se a modificação dos dados através da criação, seleção e transformação das variáveis. A modelação usa ferramentas analíticas para procurar uma combinação dos dados que preveja de maneira fiável o resultado. O último passo é a avaliação do modelo pela exatidão do mesmo e utilidade dos resultados obtidos [37].

A metodologia CRISP-DM [52], concebida em 1996 por um consórcio de empresas, foi motivada pelo mercado de DM e pela necessidade de um processo padronizado, que desse respostas fidedignas e fáceis de gerir. Além disso, foi baseada em tentativas anteriores para definir metodologias de descoberta de conhecimento [94]. A última versão baseia-se em princípios académicos e em aplicações práticas de DM. Apresenta-se em 6 fases, ilustradas na Figura 13 (a): (i) compreender os objetivos e condições necessárias do projeto; (ii) estudo dos dados; (iii) preparação dos dados; (iv) modelação; (v) avaliação e (vi) implementação [94].



(a) (b)

Figura 13 – (a) Metodologia CRISP-DM e (b) metodologia SEMMA (adaptado de [52] e [37])

As metodologias SEMMA e CRISP-DM estruturam ambas o processo de KDD em fases que se encontram relacionadas entre si, convertendo o processo de descoberta de conhecimento num processo iterativo.

A metodologia SEMMA encontra-se mais direcionada às características do desenvolvimento das técnicas e do processo, enquanto que a CRISP-DM mantém uma perspetiva mais ampla em relação aos objetivos do projeto. Esta diferença verifica-se logo na primeira fase, isto é, a SEMMA preocupa-se com a recolha de uma amostra de dados, enquanto a CRISP-DM realiza uma análise do problema para o transformar num problema técnico. Assim, pode-se dizer que a última é mais dirigida para uma conceção de um projeto real. Outra diferença significativa é a relação com ferramentas comerciais. A metodologia SEMMA está muito ligada a produtos SAS, enquanto que a metodologia CRISP-DM foi desenhada como uma metodologia neutra em relação à ferramenta que é utilizada, sendo que a sua distribuição é livre e gratuita [52,37].

Assim, a metodologia escolhida para a descoberta de conhecimento numa base de dados de potenciais dadores, foi a CRISP-DM, por ser completa e neutra, se adequar ao projeto e por ser a que de maior expressão entre as metodologias existentes.

O ciclo de vida do projeto de DM desta dissertação segue cada uma das fases abaixo descritas da metodologia CRISP-DM [52, 94, 95]:

- Compreensão do negócio: a fase inicial de um projeto de DM concentra-se em compreender os objetivos e os requisitos do projeto, de uma perspetiva do negócio, definindo-se um plano de implementação preliminar;
- Compreensão dos dados: envolve a recolha e exploração dos dados. Nesta fase são observados os dados de forma mais pormenorizada e é determinado o nível de abordagem para o problema de DM formulado. São também identificados possíveis problemas de qualidade, podendo optar-se por remover ou adicionar dados;
- Preparação de dados: nesta fase estão contempladas todas as tarefas envolvidas na criação da fonte de informação estruturada que será usada para a construção dos modelos. Tarefas de preparação de dados são suscetíveis de serem realizadas várias vezes, sem uma ordem específica. Estas incluem a construção da tabela de cenários, a seleção dos atributos, a limpeza e a transformação dos dados. Além disso, nesta fase, é possível a criação de novos atributos obtidos a partir de atributos existentes. Esta etapa é muito importante para a melhoria significativa da qualidade do conhecimento extraído através do DM;
- Modelação: nesta fase são aplicadas várias técnicas de modelação aos dados (algoritmos de árvores de decisão, regras de associação, regressão linear, redes neuronais, entre outros). É possível que durante esta etapa seja realizada a calibração dos parâmetros para otimização, pelo que é comum o retorno à fase de preparação dos dados.
- Avaliação: corresponde à seleção do modelo que parece ter maior qualidade de uma perspetiva de análise dos dados. A avaliação pode ser realizada recorrendo a medidas estatísticas. É necessário durante esta fase verificar se o modelo está de acordo com os objetivos do negócio;
- Implementação: consiste na utilização do modelo de DM em casos semelhantes aos utilizados para a construção do modelo. A implementação pode envolver scoring (aplicação de modelos em novos dados), a extração de detalhes do modelo (por exemplo, as regras de uma árvore de decisão), ou a integração de modelos de DM em aplicações, infraestruturas de DW, ferramentas de *reporting*.

Após a implementação dos modelos e execução dos mesmos em novos dados é obtido novo conhecimento útil.

Note-se que, apesar das 6 fases ilustradas na Figura 13 a) estarem integradas num processo iterativo, é possível o retrocesso a fases anteriores para incluir novos dados ou alterar decisões. Para além disso, o fluxo de um processo num projeto de DM não termina quando uma determinada solução é implementada. Os resultados desencadeiam novas questões de negócios que, por sua vez, podem ser usadas para desenvolver modelos com maior exatidão e precisão [37].

3.3 Metodologia da Prova de Conceito

A metodologia de investigação da prova de conceito consiste num modelo prático que tem como objetivo provar ou validar se um conceito ou teoria é bem-sucedido, viável e suscetível de ser explorado de maneira útil. A realização de uma prova de conceito é frequentemente apontada como sendo um dos passos mais importantes no processo de desenho, desenvolvimento e implementação de um protótipo de uma solução na área das TI, já que estabelece se determinada solução satisfaz a sua finalidade, isto é, cumpre os requisitos definidos para a qual foi inicialmente projetada. Por outro lado, permite também a identificação de falhas ou erros na solução desenvolvida [96].

Assim, uma prova de conceito permite demonstrar, na prática, os conceitos, as metodologias e as tecnologias envolvidas na elaboração de determinado projeto, de forma a validar a solução proposta através da prova da sua viabilidade e utilidade.

A análise SWOT (*Strengths, Weaknesses, Opportunities and Threats*) é uma ferramenta utilizada para estruturar um planeamento estratégico, promovendo uma análise dos pontos fortes e fracos (fatores internos), assim como das oportunidades e ameaças (fatores externos), apresentando-os numa matriz de forma a facilitar a visualização das características de determinada solução [97, 98].

No contexto da aplicação da análise SWOT a ferramentas de TI, cada uma das características pode ser resumida e esquematicamente representada (Figura 14) como [97]:

- *Strengths*: estão relacionadas com as vantagens, ou pontos fortes, que uma ferramenta apresenta em relação a concorrentes;
- *Weaknesses*: correspondem aos pontos fracos que interferem ou prejudicam de algum modo uma ferramenta;
- *Opportunities*: são fatores externos que influenciam positivamente uma ferramenta;
- *Threats*: são fatores externos que influenciam negativamente uma ferramenta.

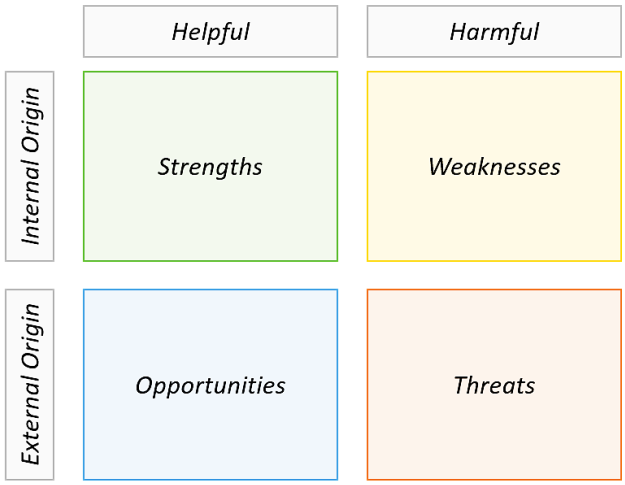


Figura 14 – Matriz SWOT (adaptada de [97]).

A análise SWOT consegue, assim, maximizar as oportunidades de uma solução através dos seus pontos fortes, ao mesmo tempo que minimiza as ameaças externas e os pontos fracos da mesma [98].

4. DESCOBERTA DE CONHECIMENTO NUMA BD DE POTENCIAIS DADORES

4.1 Introdução

A deteção de potenciais dadores de órgãos numa instituição de saúde é uma tarefa complexa. Os critérios de identificação são vários, as bases de dados existentes são múltiplas, o volume de dados existente é massivo. Enquanto por exemplo os potenciais dadores mais jovens são mais facilmente referenciados pelas equipas de urgência, os mais velhos, vítimas de AVC, são colocados em medidas de conforto e não encarados como possíveis dadores. O GCCT no CHP tem como principal função a deteção de pacientes cujos órgãos possam ser utilizados para doação. A grande maioria destes pacientes têm eventos neurológicos devastadores e podem eventualmente atingir o estado de morte cerebral. Assim, a monitorização diária destes pacientes é crucial. A acumulação de funções da equipa do GCCT faz com que seja difícil a visita diária sistematizada aos SU e UCI para identificar todos os potenciais dadores.

Nas próximas secções encontra-se descrito a modelação de um modelo preditivo de *Data Mining* que tem como objetivo ajudar na tomada de decisão clínica, tornando a monitorização e análise de pacientes mais focada e efetiva.

4.2 Definição do Problema e Objetivos da Solução

Atualmente, no CHP, existe um repositório de pacientes sinalizados como possíveis dadores. Como este repositório tem um elevado volume de dados, a equipa do GCCT pode não conseguir detetar um potencial dador, falhar no processo de transplantação dos seus órgãos e o paciente recetor não sobreviver por falta de órgãos disponíveis. Assim, compreende-se a urgência de uma monitorização diária eficiente destes pacientes.

A utilização de modelos de *Data Mining* para a descoberta de conhecimento de uma forma automatizada é, então, fulcral para o dia-a-dia de uma equipa como a do GCCT. De forma a encontrar potenciais dadores de órgãos no repositório de dados, a pergunta

que se levanta para o modelo de DM responder é: "*Um paciente do repositório de potenciais dadores vai sobreviver ou falecer?*".

Um modelo de DM cujo objetivo é a previsão de uma variável, pode ter necessidade de avaliação e interpretação do resultado distintos, consoante o conjunto de dados e o objetivo. Por exemplo, quanto temos classes desequilibradas, a métrica exatidão pode não ser apropriada para avaliar certos problemas de DM. Considerando um caso real de classificação de pixéis em imagens de mamografia, o conjunto de dados pode conter 98% de pixéis normais e 2% de pixéis anómalos. Um simples exercício de DM de previsão sobre este conjunto de dados poderia retornar valores de exatidão de 98%. O mesmo aconteceria num exercício de uma companhia aérea para identificar possíveis passageiros terroristas a embarcar num voo. Os dados neste domínio são maioritariamente relativos a passageiros normais, sendo uma minoria os casos de terroristas identificados. Neste tipo de problemas de classificação, a métrica exatidão retorna valores enganadores, já que não avalia de forma apropriada a performance de um modelo de DM [99].

O mesmo acontece com o problema identificado neste projeto de dissertação. O número de casos de pacientes falecidos no repositório de potenciais dadores é consideravelmente menor do que o número de casos de pacientes que sobreviveram. Neste sentido, e para avaliar corretamente o modelo de DM construído, deve ser dada ênfase à classificação correta dos dados relativos a pacientes falecidos. Em termos estatísticos, a métrica a maximizar denomina-se sensibilidade, tendo esta como objetivo encontrar todas as instâncias relevantes num conjunto de dados. Esta métrica calcula-se através da divisão entre casos verdadeiros positivos (VP) e a soma destes mesmos casos com casos Falsos Negativos (FN) [73, 100]. Neste caso concreto, os casos VP referem-se a pacientes falecidos classificados corretamente pelo modelo, enquanto os casos FN se referem a pacientes classificados incorretamente como sobreviventes.

Com a maximização da métrica de sensibilidade, a especificidade do modelo diminui. A especificidade é a capacidade do modelo classificar corretamente os pacientes saudáveis, ou seja, casos verdadeiros negativos (VN) [73, 100].

Se a relevância do primeiro parâmetro parece indiscutível dado que a finalidade do modelo é detetar um potencial dador, o segundo parâmetro também é importante já que diminui o conjunto de pacientes que a equipa do GCCT tem de monitorizar. Sendo

que é difícil obter sensibilidade e especificidade máximas, a relação de compromisso que se pretende é a aproximação da métrica de sensibilidade do valor ideal, tornando igual ou próximo de zero o valor de casos FN, e a redução dos casos Falsos Positivos (FP), de forma a atingir maior especificidade.

4.3 Exploração dos dados

Neste capítulo são descritos os procedimentos usados para a compreensão e exploração dos dados, inseridos na etapa 2 do CRISP-DM:

1. Visualização e descrição dos dados
2. Tratamento de dados
3. Balanceamento do conjunto de dados

4.3.1 Visualização e Descrição dos Dados

Neste projeto de dissertação foi utilizado o WEKA *(Waikato Environment for Knowledge Analysis*), versão 3.8, uma ferramenta desenvolvida utilizando linguagem JAVA, na Universidade de Waikato, na Nova Zelândia. O WEKA inclui uma coleção de algoritmos de *machine learning* e ferramentas de processamento de dados. Foi desenhada de forma a ser simples e flexível na aplicação de vários métodos a novos conjuntos de dados [46]. As suas funcionalidades, às quais acrescem os pacotes de *plugins* disponíveis para instalação, fazem deste ambiente uma boa ferramenta para efetuar os passos necessários num processo de DCBD.

Os dados são disponibilizados a partir da tabela ORGANITE_REPOSITORY, cuja estrutura é descrita na Secção 5.3.1 e que contém informação relativa a pacientes potenciais dadores e cujos atributos são descritos na tabela 3.1. A amostra compreende o período entre 1 de janeiro de 2013 e 31 de outubro de 2017, correspondendo a 29.410 registos e a 13 atributos. Estes registos compreendem 11.398 pacientes únicos.

Na Tabela 3 estão presentes as variáveis dos atributos discretos, assim como a percentagem de ocorrência de cada uma delas no conjunto de dados. Relativamente aos atributos *Cod_Diagnosis* e *Cod_Especialidade*, como possuem 366 e 68 variáveis distintas, respetivamente, estão apresentadas na Tabela 4 as cinco variáveis com um maior número de casos associado.

Tabela 3 – Atributos *Modulo, BD, Sexo, Estado* e *Diag_Recente* da tabela ORGANITE_REPOSITORY, respetiva descrição, variáveis associadas e número de casos (%) de cada variável

Atributo	Descrição	Variável	Casos (%)
MODULO	Módulo clínico do episódio	URG	51,16
		INT	41,18
		BLO	6,38
		CON	1,00
		HDI	0,29
BD	Base de dados de onde provém a informação	SGD	71,24
		ALERT	25,42
		PCE	3,34
SEXO	Género do paciente	Masculino	52,55
		Feminino	47,45
ESTADO	Condição clínica do paciente	Alta	81,00
		Falecido	16,96
		Internado	0,87
		Entrada	0,54
		Desconhecido	0,46
		Urgências	0,17
DIAG_RECENTE	Identifica o episódio mais recente e os episódios antigos dos pacientes	Recente	58,36
		Antigo	41,64

Tabela 4 – Atributos *Cod_Diag* e *Cod_Especialidade* da tabela ORGANITE_REPOSITORY, respetiva descrição, *top* 5 de variáveis (ID e DESC) com maior número de casos associado e número de casos (%)

Atributo	Descrição	Variável (ID)	Variável (DESC)	Casos (%)
COD_DIAGNOSIS	Código ICD-9-CM do diagnóstico registado. Para um episódio podem ser associados diferentes códigos de diagnóstico, cada um deles representando uma nova linha	43401	Trombose Cerebral, Com Enfarte Cerebral;	28,77
		431	Hemorragia Intracerebral;	20,31
		4340	Trombose Cerebral;	18,60
		4321	Hemorragia Subdural;	16,27
		852	Hemorragia Subaracnoídea, Subdural Ou Extradural, Pós-traumática.	16,06
COD_ESPECIALIDADE	Código da especialidade onde o paciente foi visto e/ou para onde foi reencaminhado	1	Urgência Geral	64,42
		31701	INT T.C.E. /HSA	11,20
		31700	INT NEUROCIRURGIA/HSA	9,30
		31800	INT NEUROLOGIA/HSA	8,45
		6	NEUROCIRURGIA	6,63

Pela análise dos valores da Tabela 3 verifica-se desde logo um desequilíbrio entre o número de incidências em que o estado dos pacientes é "Alta" e o número de incidências dos restantes estados. É importante referir que o estado dos pacientes vai sendo alterado ao longo do tempo através de uma instrução SQL de *INSERT OVERWRITE* à tabela. Assim, não é possível analisar o histórico de estados do paciente, mas sim o estado do episódio mais recente do paciente. Os estados "Alta" e "Falecido" preenchem 97,96% dos dados, já que são estados de diagnóstico finais.

Verifica-se ainda que apenas cerca de 17% dos casos se referem a pacientes falecidos. Trata-se de um conjunto de dados desequilibrado, onde existem mais casos negativos do que positivos. Esta situação pode ter impacto na modelação dos dados e será abordada na Secção 4.4.3.

Como o conjunto de atributos disponibilizado é reduzido, foram facilmente selecionados os atributos com relevância para a construção do modelo de previsão. No mesmo seguimento, foram criados novos atributos que traduzem informação não relevante da BD para informação útil e da qual se pode inferir conhecimento. Na Tabela 5 estão representados os atributos criados.

Tabela 5 – Descrição dos atributos *Idade*, *Episódios_Ant* e *Prognostico* criados; características associadas aos atributos contínuos (*Idade, Episodios*_Ant) e variáveis e número de casos (%) do atributo discreto (*Prognostico*)

Atributo	**Descrição**	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**
IDADE	Idade do paciente à data do diagnóstico	18	89	66,62	16,27
EPISODIOS_ANT	Contagem do número de episódios neurológicos graves antecedentes do paciente	0	12	0,72	1,14
PROGNOSTICO	Prognóstico do diagnóstico do paciente por data do diagnóstico distinta	**Variável**		**Casos (%)**	
		Sobrevive (0)		92,03	
		Falece (1)		7,97	

No cálculo do atributo *idade*, consideraram-se pacientes com idades superiores a 18 anos e inferiores a 90 anos. A limitação deveu-se à escolha de uma análise somente em

pacientes adultos. Para além disso, e embora possa não haver nenhuma idade máxima definitiva para doação, como a presença de comorbidades é maior com o aumento da idade, desencadeando uma situação de menor aceitabilidade [4], limitou-se também o conjunto de pacientes mais velhos.

O atributo *Episodios_Ant* foi calculado com o objetivo de perceber se um paciente tem maior probabilidade de ser potencial dador se tiver histórico de episódios neurológicos graves associados. Assim, e utilizando os atributos *Num Sequencial e Dt_Diag*, contabilizou-se o número de episódios antecedentes por paciente e data de diagnóstico distinta.

Na Tabela 6, encontra-se representado um exemplo real relativo ao paciente com o N*um_Sequencial '*303123'. O primeiro registo de um evento neurológico grave data do dia 2 de dezembro de 2014, sendo o valor do atributo *Episodios_Ant* igual a zero. No início do ano seguinte, o paciente teve dois registos associados à mesma data, com dois códigos de diagnóstico diferentes. Os episódios antecedentes referentes a estes dois registos têm ambos o valor 1 associado. No ano de 2016 a variável volta a incrementar três vezes, por data de diagnóstico distinta. Assim, e à data do episódio mais recente, o paciente apresenta quatro episódios antecedentes de eventos neurológicos graves, valores que podem ser úteis aquando da construção do modelo de DM e aplicação do mesmo a novos casos.

Tabela 6 – Cálculo dos episódios neurológicos graves antecedentes de um paciente do conjunto de dados de potenciais dadores

NUM_SEQUENCIAL	***COD_DIAGNOSIS***	***DTA_DIAG***	***EPISODIOS_ANT***
303123	431	02/12/2014	0
303123	2252	26/01/2015	1
303123	431	26/01/2015	1
303123	431	07/03/2016	2
303123	2252	06/05/2016	3
303123	431	06/05/2016	3
303123	2252	27/05/2016	4
303123	431	27/05/2016	4

Relativamente ao atributo *prognostico*, este será a variável alvo do modelo de DM de previsão a construir. A base deste atributo provém da coluna *estado*, com 6 variáveis:

'Alta', 'Falecido', 'Internado', 'Desconhecido', 'Urgências'. Como anteriormente referido, o conjunto de dados não dá visibilidade do estado real dos pacientes à data do diagnóstico, mas sim do estado mais recente do paciente. O conhecimento que se pode inferir a partir dos dados desta coluna é a sobrevivência ou não sobrevivência dos pacientes. Assim, transformaram-se as variáveis 'Alta', 'Internado', 'Desconhecido', 'Urgências' em 'Sobrevive' e 'Falecido' em 'Falece'.

Como o objetivo deste estudo é a determinação dos pacientes com maior probabilidade de se tornarem potenciais dadores de órgãos, ou seja, de não sobreviver a eventos neurológicos devastadores, o *prognostico* é a variável alvo do modelo de previsão. Neste sentido, os casos positivos do modelo referem-se a pacientes falecidos e os casos negativos a pacientes que sobreviveram.

4.3.2 Tratamento dos Dados

Dados imprecisos, incompletos e inconsistentes são propriedades comuns das bases de dados do mundo real dos dias de hoje, devido ao seu tamanho (geralmente vários gigabytes), à utilização de fontes de dados heterogéneas, à inserção errada de dados tanto por erro humano como computacional, a limitações da própria tecnologia, como tamanho de *buffer* limitado para coordenar a transferência de dados a sincronizar. Dados que não são considerados fiáveis podem causar modelos de DM imprecisos [41].

Para resolver este problema e garantir qualidade e consistência dos dados, utilizou-se como técnica de pré-processamento a limpeza dos dados, tendo-se analisado o conjunto de dados de forma a encontrar erros, omissão de dados e perceber a integridade dos mesmos.

As linhas com múltiplas colunas não preenchidas, com erros de escrita e caracteres não reconhecidos, foram eliminadas. De forma a não perder informação relevante, os códigos de diagnóstico mal definidos foram corrigidos para os valores corretos por comparação com semelhantes, e através da análise de uma coluna de anotações dos profissionais de saúde. As linhas com códigos de diagnóstico nulos ou com múltiplos códigos de diagnóstico atribuídos, foram eliminadas.

As linhas cujos atributos de interesse se repetiam também foram eliminadas. Estas linhas advinham da existência da mesma informação proveniente de bases de dados distintas (coluna BD). Para além disso, e quando um paciente é reencaminhado de uma

especialidade para outra (diferentes valores para a coluna *Cod_Especialidade*), a informação dos atributos de interesse pode também repetir-se.

No passo seguinte, deu-se relevância a outra técnica de pré-processamento, a transformação de dados. O objetivo da transformação é melhorar a precisão e a eficiência dos algoritmos de DM [41]. Foi feita a discretização dos valores contínuos em intervalos de valores dos atributos *idade* e *Episodios_Ant*, recorrendo à ferramenta *Weka*, ao filtro *Discretize* disponibilizado no conjunto de filtros *unsupervised* da aplicação *Explorer.* O atributo idade foi dividido em 4 *bins* e o atributo *Episodios_Ant* foi dividido em 3 *bins,* conforme valores da Tabela 7. Prosseguiu-se para a transformação dos valores "Masculino" e "Feminino" do atributo *sexo,* primeiramente de forma manual para valores numéricos, 0 e 1, respetivamente, e depois para valores binários na ferramenta *Weka*, com a utilização do filtro *NumericToBinary*, do conjunto de filtros *unsupervised.*

Tabela 7 – Discretização dos atributos *Idade* e *Episodios_Ant*, variável associada a cada *bin* e número de casos associados a cada *bin* (%)

Atributo	Descrição	Variável	Casos (%)
IDADE	Discretização valores contínuos relativos à idade dos pacientes	[18-49]	15,87
		[50-66]	26,56
		[67-78]	29,73
		[79-90]	27,84
EPISODIOS_ANT	Discretização valores contínuos relativos aos episódios antecedentes dos pacientes	0	58,34
		1	23,70
		>1	17,96

Para alem destas ações, e utilizando a ferramenta *Weka*, foi utilizado o filtro *unsupervised NumericToNominal* no atributo *prognostico* e também no atributo *Cod_Diagnosis* para que estes sejam reconhecidos pela ferramenta não como valores contínuos, mas sim como valores discretos.

No fim desta fase, o conjunto de dados compreendeu um total de 24.481 registos.

4.4 Modelo de Previsão

De acordo com a metodologia CRISP-DM, descrita na Secção 3.2.2, os passos que se seguem são a modelação, a avaliação dos modelos e a implementação dos mesmos a novos dados. O objetivo é desenvolver modelos de DM, através da utilização de algoritmos de diferentes famílias, de diferentes técnicas de balanceamento de dados e ainda de abordagens de otimização de forma a escolher o modelo com melhor desempenho e que melhor responde ao problema em questão. Assim, foi criado um processo de modelação, avaliação e implementação composto pelas etapas:

1. Escolha de algoritmos
2. Criação dos conjuntos de treino e teste
3. Criação dos modelos
4. Avaliação dos modelos
5. Otimização
6. Reavaliação

4.4.1 Escolha de Algoritmos

Os esquemas de aprendizagem são variados pelo que é necessário escolher os algoritmos e técnicas adequadas, de modo a criar um modelo capaz de classificar uma amostra com base no conhecimento adquirido na observação, e encontrar padrões em dados existentes.

Para a realização desta tarefa, a compreensão dos diversos algoritmos e técnicas (descritos no subcapítulo 2.4.2) é essencial de modo a proceder à sua correta parametrização e aplicação. Não existe um processo instituído para a escolha dos algoritmos. A escolha dos algoritmos para o problema descrito na Secção teve por base três aspetos: algoritmos utilizados em problemas de medicina semelhantes; algoritmos usualmente utilizados em comparações de desempenho nos problemas de DM; algoritmos representativos de diferentes famílias de algoritmos.

Foi dada ênfase ao primeiro aspeto dado que o uso de um algoritmo em casos de domínio semelhante dá uma boa noção da adequação de um determinado algoritmo para este problema. Serviram de base os trabalhos desenvolvidos por Asil Oztekin et al. [56], Ali Dag et al. [101], Hamed Zolbanin et al. [102], Hsueh-Yi Lu et al. [103], Kyung Yoo

et al. [104] e Monika Gandhi et al. [105], cujo âmbito se enquadra na aplicação de técnicas de DM na área médica. Também foi tido em conta o trabalho de Wu et al. [106], onde foram escolhidos algoritmos usualmente utilizados em comparações de desempenho e alguns representativos de diferentes famílias. Os algoritmos que mais se destacam estão representados na Tabela 8, assim como a família a que pertencem. Num estudo recente relacionado com a sobrevivência de pacientes, as árvores de decisão, as redes neuronais artificiais e as *support vector machines* (SVM), obtiveram melhores resultados de previsão. As redes neuronais artificiais têm sido um algoritmo de DM frequentemente utilizado em medicina dada a sua boa performance na previsão de variáveis [56]. Por outro lado, as árvores de decisão têm a vantagem de não serem "caixas negras", ou seja, é possível analisar como o modelo é construído e explicá-lo através de regras. Esta vantagem faz com que sejam utilizadas na área da saúde [56]. As *support vector machines* (SVM) e a *logistic regression* são normalmente utilizadas em aplicações biomédicas, como na deteção de cancro e em estudos de análise de sobrevivência após transplantação [101].

Tabela 8 – Algoritmos de DM utilizados na fase de modelação, família de algoritmos a que pertencem, algoritmo base e artigos de referência onde são caso de estudo

Algoritmo	**Família**	**Algoritmo base**	**Artigos**
RBFClassifier	IBL	ANN	[102]; [103]
J48	DT	C4.5	[56]; [103]
REPTree	DT	CART	[101]; [104]
Logistic	SL	Logistic Regression	[56]; [101]; [102]
SMO	IBL	SVM	[101]
NaiveBayes	SL	Naive Bayes	[103]
BayesNet	SL	Naive Bayes	[103]
RandomForest	DT	Meta-algoritmo: *bagging*	[102]; [104]

4.4.2 Criação dos Conjuntos de Treino e Teste

A avaliação de um modelo implica a validação deste através de um conjunto aleatório de novas amostras. A separação em conjuntos de treino e teste é uma parte relevante para a avaliação de um modelo de DM. Neste sentido, o primeiro deve ser representativo do segundo e permitir a sua validação [46, 72].

O conjunto de testes é obtido através da adoção de uma estratégia como o *Holdout Sampling* descrito na Secção 2.4.5 O *Holdout* é uma boa opção dada a sua simplicidade, aceitação generalizada e pela sua aplicabilidade ao problema em causa. É necessária a parametrização do rácio do conjunto de dados para treino e teste. Optou-se pela divisão genericamente usada nos processos de DM, em que $^2/_3$ dos dados são utilizados no conjunto de treino e $^1/_3$ no conjunto de teste [46, 72]. A representatividade dos dados é preservada através da escolha aleatória dos casos que fazem parte de cada subconjunto e manutenção da proporcionalidade entre casos positivos e negativos, no conjunto de treino e teste.

4.4.3 Criação dos Modelos

Normalmente, em bases de dados médicas verifica-se um desequilíbrio entre amostras de casos positivos e negativos no conjunto de dados. Esse desequilíbrio pode não permitir uma correta classificação dos casos minoritários. O mesmo acontece com o problema identificado neste projeto de dissertação. O número de casos de pacientes falecidos no repositório de potenciais dadores é consideravelmente menor do que o número de casos de pacientes que sobreviveram. Uma alternativa a este problema é adotar técnicas que permitem balancear o conjunto de dados.

Na Tabela 9 estão representados os diferentes modelos testados, com diferentes técnicas de balanceamento do conjunto de dados associadas. O objetivo é perceber qual a técnica de balanceamento de dados que aplicada ao conjunto de dados da tabela ORGANITE_REPOSITORY apresenta melhores resultados e, consequentemente, origina um melhor modelo de DM.

As técnicas abordadas incluem *undersampling* e *oversampling*, descritas na Secção 2.4.4 sendo que na primeira incluem-se as técnicas utilizadas no modelo 1 e 3, *cluster undersampling* e RUS, e na segunda inclui-se a técnica utilizada no modelo 2, SMOTE.

Tabela 9 – Modelos de DM construídos com base em técnicas de balanceamento de dados: *cluster undersampling,* RUS, SMOTE e no conjunto de dados original

Modelo	Técnica	Ratio	Descrição
1	Cluster Undersampling	1:1	Aplicação da técnica *cluster undersampling* ao conjunto de dados original. nº total de instâncias da classe maioritária: 1319 nº total de instâncias da classe minoritária: 1319
2	SMOTE	~1:1	Aplicação da técnica SMOTE ao conjunto de dados original. nº total de instâncias da classe maioritária: 3751 nº total de instâncias da classe minoritária: 3706
3	RUS	1:1	Aplicação da técnica RUS ao conjunto de dados original. nº total de instâncias da classe maioritária: 1319 nº total de instâncias da classe minoritária: 1319
4	Nenhuma	~11:1	Conjunto de dados original. nº total de instâncias da classe maioritária: 22.530 nº total de instâncias da classe minoritária: 1951

Balanceamento dos dados

A estratégia da técnica de *undersampling* é manter o conjunto minoritário e retirar amostras do conjunto maioritário até que o conjunto final tenha o mesmo número de amostras dos dois conjuntos.

Assim, e para a construção do modelo 1, foi utilizada a técnica de *undersampling* baseada em *clustering*, adaptada de Show-Jane e Yue-Shi [69]. Os autores defendem que podem existir grupos de dados com características distintas (*clusters*) no conjunto de dados disponível. Assim, se um determinado *cluster* possuir maior número de instâncias da classe maioritária e menor da classe minoritária, vai comportar-se de

acordo com a classe maioritária. Por outro lado, se um *cluster* possuir maior número de instâncias da classe minoritária e menor da classe maioritária, não possui características suficientes relativas à classe maioritária e, portanto, comporta-se mais de acordo com a classe minoritária. Assim, a metodologia apresentada seleciona um número representativo de instâncias da classe maioritária de cada *cluster*, considerando o rácio entre o número de amostras da classe maioritária e o número de amostras da classe minoritária.

O número de amostras no conjunto de dados tem o valor 16.321, incluindo 15.002 instâncias da classe maioritária (MA) e 1319 instâncias da classe minoritária (MI). Assumiu-se um rácio 1:1 (m=1) entre o número de instâncias MA e o número de instâncias MI do conjunto de treino. Este conjunto foi dividido em cinco *clusters* com número de instâncias idêntico (entre 3.264 e 3.265).

Na Tabela 10 estão representados os valores relativos ao tamanho do *cluster i* no que diz respeito a instâncias maioritárias ($Size_{MA}^{i}$) e instâncias minoritárias ($Size_{MI}^{i}$). São ainda apresentados os valores referentes ao rácio do número de instâncias MA em relação ao número de instâncias MI, para o *cluster i.*

Tabela 10 – Representação dos valores relativos ao tamanho do *cluster i das* instâncias maioritárias ($Size_{MA}^{i}$) e instâncias minoritárias ($Size_{MI}^{i}$); rácio do número de instâncias MA em relação ao número de instâncias MI, para o *cluster i.*

Cluster ID	$\boldsymbol{Size_{MA}^{i}}$	$\boldsymbol{Size_{MI}^{i}}$	$\boldsymbol{Size_{MA}^{i}}/\boldsymbol{Size_{MI}^{i}}$
1	3.001	263	11,4
2	2.972	292	10,2
3	2.996	268	11,2
4	3.023	241	12,5
5	3.010	255	11,8
TOTAL	15.002	1319	$\sum_{i=1}^{K} Size_{MA}^{i}/Size_{MI}^{i} = 57{,}1$

O número de instâncias da classe maioritária selecionadas em cada cluster foi determinado de acordo com a expressão (1), presente na Secção 2.4.4 e está representado na Tabela 11.

Tabela 11 – Número de instâncias selecionadas no cluster *i* da classe maioritária ($SSize_{MA}^{i}$)

Cluster ID	Nº de instâncias selecionadas da classe maioritária ($SSize_{MA}^{i}$)
1	264
2	235
3	258
4	290
5	273

Assim, a expressão determina as instâncias da classe maioritária a selecionar em cada *cluster* construído. O valor de $SSize_{MA}^{i}$ é tanto maior quanto mais instâncias MA e menos MI o *cluster i* possuir.

Depois de determinar o número de instâncias MA a selecionar em cada *cluster* construído, estas são escolhidas de forma aleatória dos K *clusters*. Acrescentam-se todas as instâncias MI, ficando construído o conjunto de treino final. O rácio de $Size_{MA}$ em relação ao $Size_{MI}$ é representado por m:1 neste novo conjunto.

No que diz respeito ao modelo 2, foi utilizada uma técnica de *oversampling* que consiste no aumento das amostras do conjunto minoritário até que este iguale as amostras do conjunto maioritário. O SMOTE, técnica detalhada na Secção 2.4.4, gera amostras sintetizadas da classe minoritária a partir de um determinado número de amostras vizinhas de cada uma das amostras.

O volume de amostras do conjunto maioritário é cerca de onze vezes superior ao do conjunto minoritário, o que implica a criação de um elevado número de instâncias de forma a gerar um rácio de 1:1 entre conjunto maioritário e minoritário. Como a quantidade de amostras do conjunto minoritário é muito reduzida, os k vizinhos mais próximos utilizados na sintetização das amostras podem incluir ruído que prejudicam os resultados da aplicação do modelo criado a dados novos. O modelo especializa-se em casos não reais e, consequentemente, diminui a precisão na classificação das amostras que não tem conhecimento prévio, ocorrendo o problema de *overfitting*.

De forma a contornar este problema, criaram-se quatro *clusters* a partir do total de amostras maioritárias no conjunto de treino disponível. O passo seguinte englobou a

junção de todas as amostras do conjunto minoritário a cada *cluster* previamente criado. O SMOTE foi aplicado a cada um dos clusters, na Secção de pré-processamento do conjunto de dados da ferramenta *Weka*, com um número de k-*nearest neighbours* igual a 5 e de forma a gerar um rácio de 1:1.

Este processo de balanceamento seguiu-se de uma aplicação experimental dos algoritmos descritos na Tabela 8 da Secção 4.2.2 em cada um dos *clusters* construídos. O *cluster* com melhores resultados é composto pelas amostras da classe maioritária melhor representativas do conjunto de dados total, eliminando assim possível ruído e *outliers*. Desta forma, o modelo 2 é constituído pelo *cluster* balanceado pela técnica de SMOTE que obteve melhores resultados na análise experimental.

O modelo 3 englobou a técnica de *undersampling* RUS, detalhada na Secção 2.4.4 e aplicada através da ferramenta *Weka*. A ferramenta efetua o *undersampling* de modo aleatório até atingir uma percentagem definida previamente no parâmetro do filtro. Esta foi definida com o valor de 50% para que as duas classes fossem representadas pelo menos número de instâncias, ou seja, com um rácio de 1:1.

Relativamente ao modelo 4, nenhuma técnica de balanceamento foi aplicada, correspondendo este ao conjunto de treino original.

Aplicação dos algoritmos

Os algoritmos de DM previamente selecionados e representados na Secção 4.4.1, foram aplicados ao conjunto de treino para a construção de um modelo que faça a previsão do prognóstico de um paciente: sobrevivente ou falecido.

Para ser possível uma correta validação dos resultados, fizeram-se dois blocos de testes. Primeiramente, e numa fase mais experimental, usando a técnica *k-Fold Cross Validation* e numa segunda fase usando a técnica *Holdout Sampling*.

Os resultados da primeira fase de testes estão representados na Tabela 12. Incluem a aplicação das métricas de avaliação exatidão, sensibilidade, especificidade e AUC, descritas na Secção 2.4.6, para cada modelo da Tabela 9 com a técnica de *k-Fold Cross Validation*, onde k=10.

O objetivo desta fase, e a razão de utilização desta técnica, é garantir que os resultados da aplicação de determinado algoritmo são representativos do conjunto de dados. Pretende-se, assim, validar que os resultados são independentes da distribuição e convergem para um mesmo valor, excluindo a hipótese de existir um conjunto ou balanceamento mais favorável, cujo resultado não seja reproduzível em situações semelhantes. O objetivo é evitar problemas como o *overfitting* e o *underfitting* e ter uma visão de como o modelo se comporta num conjunto de dados desconhecido.

Tabela 12 – Resultados da aplicação dos algoritmos de DM aos modelos 1, 2, 3 e 4, com a técnica de *k-Fold Cross Validation*, k=10

10-fold Cross Validation Modelo	Métrica Avaliação \ Algoritmo	BayesNet	J48	Logistic	NaiveBayes	Random Forest	RBF Classifier	REPTree	SMO
1	Exatidão	**0,690**	0,681	0,687	**0,690**	0,671	**0,690**	0,675	0,659
	Sensibilidade	0,700	0,669	0,688	0,701	0,679	**0,705**	0,659	0,649
	Especificidade	0,680	**0,693**	0,685	0,679	0,663	0,675	0,691	0,669
	AUC	0,748	0,721	0,748	0,749	0,711	**0,753**	0,717	0,659
2	Exatidão	0,701	**0,722**	0,707	0,701	0,719	0,709	0,718	0,686
	Sensibilidade	0,718	**0,732**	0,715	0,719	**0,732**	0,719	0,717	0,651
	Especificidade	0,684	0,712	0,700	0,684	0,706	0,700	0,720	**0,721**
	AUC	0,777	0,767	0,780	0,777	**0,785**	0,782	0,772	0,686
3	Exatidão	0,687	0,673	0,688	0,688	0,665	**0,694**	0,661	0,670
	Sensibilidade	0,707	0,694	0,698	**0,710**	0,685	0,707	0,670	0,680
	Especificidade	0,666	0,652	0,679	0,666	0,645	**0,681**	0,652	0,660
	AUC	0,757	0,712	0,754	**0,758**	0,715	0,756	0,713	0,670
4	Exatidão	0,919	0,920	0,921	0,920	0,917	**0,922**	0,921	0,921
	Sensibilidade	0,038	0,000	0,048	0,035	**0,066**	0,026	0,045	0,046
	Especificidade	0,996	1,000	0,996	0,996	0,991	**0,999**	0,997	0,997
	AUC	0,719	0,500	**0,758**	0,719	0,719	0,754	0,720	0,522

Numa segunda fase, e utilizando os algoritmos com melhor performance para cada modelo, aplicou-se a técnica de *Holdout Sampling*, utilizando os modelos 1,2 e 3 treinados na fase anterior sobre o conjunto de treino no conjunto de teste. O conjunto de teste é composto por $^1/_3$ do conjunto de dados original, conforme definido na Secção 4.4.2 O modelo 4 não foi incluído nesta segunda fase dado servir como um modelo de controlo para a fase experimental.

Os resultados representados na Tabela 13 incluem a matriz de confusão e a aplicação das métricas de avaliação exatidão, sensibilidade, especificidade e AUC para cada modelo da Tabela 9, com a técnica *Holdout Sampling*.

Tabela 13 - Resultados da aplicação dos algoritmos *RBFClassifier* e *NaiveBayes* ao Modelo 1, *RandomForest* e *J48* ao Modelo 2 e *NaiveBayes* e *RBFClassifier* ao Modelo 3, através de *Holdout sampling*

		Modelo 1		**Modelo 2**		**Modelo 3**	
Algoritmo / Métrica Avaliação		**RBF Classifier**	**NaiveBayes**	**Random Forest**	**J48**	**NaiveBayes**	**RBF Classifier**
Matriz de Confusão	VP	469	455	500	465	458	458
	VN	5069	5093	5314	5387	4964	5070
	FP	2459	2435	2214	2141	2564	2458
	FN	163	177	132	167	174	174
Exatidão		0,679	0,680	0,713	0,717	0,664	0,677
Sensibilidade		0,742	0,720	0,791	0,736	0,725	0,725
Especificidade		0,673	0,677	0,706	0,716	0,659	0,673
AUC		0,774	0,768	0,825	0,779	0,761	0,77

4.4.4 Avaliação dos Modelos

Na fase de avaliação é importante analisar os resultados dos modelos construídos, através de indicadores de avaliação, e ainda os resultados obtidos aquando da aplicação destes a um novo conjunto de dados. O conjunto de teste, previamente definido na Secção 4.4.2, é independente do conjunto de treino utilizado, servindo como exemplo de aplicação dos modelos a um conjunto desconhecido.

Iniciando a análise pelos valores da Tabela 12, verificou-se desde logo a discrepância de valores entre o modelo 4 e os restantes modelos. Relativamente à métrica exatidão, esta representa o número de amostras que o modelo classifica corretamente, incluindo amostras positivas (pacientes falecidos) e negativas (pacientes que sobreviveram). No modelo 4, e para todos os algoritmos, o valor desta métrica foi cerca de 25 a 30% superior aos modelos 1, 2 e 3. Por outro lado, os valores da métrica sensibilidade para o modelo 4 foram inferiores a 0,1%. Esta métrica representa a capacidade do modelo classificar corretamente os casos verdadeiros positivos, ou seja, os casos de pacientes

falecidos e, portanto, potenciais dadores. Como o conjunto de treino não apresenta balanceamento entre classe maioritária e minoritária, sendo o número de amostras da classe maioritária cerca de 11 vezes superior ao número de amostras da classe minoritária, o modelo especializou-se de forma quase total na classe com maior número de amostras. Assim, e embora os valores de exatidão sejam elevados, estamos perante um mau modelo que não cumpre o objetivo proposto, a previsão de potenciais dadores. Entende-se também a razão para os valores de especificidade serem elevados para este modelo, já que a sensibilidade e a especificidade são métricas de avaliação que se relacionam de forma inversa.

No que diz respeito aos restantes modelos, o comportamento dos algoritmos quando aplicados a cada um deles não foi uniforme.

No modelo 1, onde foi aplicada a técnica de balanceamento de *cluster undersampling*, o algoritmo que apresentou melhores resultados para três métricas, exatidão, sensibilidade e AUC, foi o *RBFClassifier*, baseado em *artificial neural networks*. Sendo que a métrica de estudo mais relevante é a sensibilidade, o algoritmo que se seguiu com bons resultados foi o *NaiveBayes*, da família de aprendizagem probabilística.

No modelo 2, onde se utilizou o SMOTE para balanceamento das classes, dois dos algoritmos com melhores resultados foram árvores de decisão: *J48* e *RandomForest*. Em termos de exatidão, o J48 apresentou o resultado mais elevado, enquanto em termos da métrica AUC, o *RandomForest* obteve um resultado superior. O valor de sensibilidade obtido foi o mesmo para os dois algoritmos neste modelo.

Relativamente ao modelo 3, a técnica de *undersampling* RUS foi aplicada ao conjunto de treino, e os algoritmos com melhor comportamento foram o *RBFClassifier* e o *NaiveBayes*. Exatidão e especificidade apresentaram valores superiores para o *RBFClassifier*, enquanto as métricas sensibilidade e AUC foram superadas pelo *NaiveBayes*.

Numa análise geral, verificou-se claramente a necessidade de balanceamento do conjunto de treino para a construção de um modelo de previsão fiável, que não se especialize somente numa das classes e obtenha bons resultados quando aplicado a um conjunto de dados desconhecido.

No que diz respeito à comparação entre técnicas de balanceamento, o *oversampling* através do SMOTE superou as duas técnicas de *undersampling* utilizadas. Entre o RUS e

o *cluster undersampling,* ambas técnicas onde são removidas amostras do conjunto maioritário, o RUS obteve melhores resultados. Em teoria, a técnica que envolve a criação de *clusters* e a seleção das amostras da classe maioritária relevantes para a criação do conjunto de treino final, deveria obter melhor resultados do que a seleção destas amostras de forma aleatória. No entanto, e tal como é sugerido na literatura [107], a abordagem de *cluster undersampling* é apropriada para conjuntos de dados onde há confiança que as amostras das classes positiva e negativa definem corretamente as propriedades de uma classe positiva e negativa, respetivamente. No entanto, e em particular em conjuntos de dados clínicos, não há garantia que classe positiva e negativa reflitam as características reais de determinado registo.

Como a métrica AUC é obtida a partir da análise da curva ROC, e usada como medida de qualidade do desempenho de um modelo, foi possível inferir os algoritmos que geraram os melhores modelos para o caso de estudo. Estes variaram conforme a técnica de balanceamento utilizada. As árvores de decisão (*J48* e *RandomForest*) obtiveram os valores de AUC mais elevados e, portanto, um melhor comportamento quando foi aplicado o SMOTE ao conjunto de treino. As redes neuronais artificiais (*RBFClassifier*) e a aprendizagem probabilística (*NaiveBayes*) obtiveram melhores resultados quando foram aplicadas técnicas de *undersampling*.

De forma a perceber o comportamento dos modelos construídos num conjunto de dados desconhecido e real, aplicaram-se os algoritmos com melhores resultados de cada modelo no conjunto de teste. Este é composto por 612 amostras da classe positiva e 7528 amostras da classe negativa.

Analisando os resultados apresentados na Tabela 13, e relativamente aos dois testes efetuados no conjunto de teste com o modelo 1, os resultados foram idênticos. No entanto, aquando da aplicação do modelo com o algoritmo *RBFClassifier*, o valor de AUC foi superior, traduzindo-se num modelo melhor construído. De facto, este acertou em 77% de casos verdadeiros positivos e em 67% de casos verdadeiros negativos, enquanto o *NaiveBayes* acertou em 74% dos primeiros e 68% dos segundos. Estes resultados estão de acordo com os da análise exploratória, já que o modelo 1 com este algoritmo tinha apresentado melhores resultados.

Nos testes com o modelo 2, e analisando a matriz de confusão originada pela ferramenta *Weka*, verificou-se que a utilização do algoritmo *RandomForest* acertou na

previsão de 500 amostras relativas a pacientes falecidos, correspondendo aproximadamente a 82% do total de amostras da classe positiva. O valor de AUC foi significativamente mais elevado do que em qualquer um dos outros testes efetuados aquando da aplicação deste algoritmo. A métrica sensibilidade obteve também um valor superior aos dos restantes modelos. Sendo que um dos objetivos principais do problema em questão é a maximização da previsão de potenciais dadores, o modelo 2 com o algoritmo *RandomForest* é o que melhor cumpre o objetivo.

No que diz respeito ao modelo 3, os resultados foram diferentes dos da fase experimental. O algoritmo com um valor inferior de AUC, o *RBFClassifier* comportou-se melhor quando foi aplicado ao conjunto de teste. Embora o valor da métrica sensibilidade fosse igual ao do algoritmo *NaiveBayes*, o valor da métrica especificidade foi mais elevado, o que levou a que o modelo acertasse mais casos relativamente a pacientes sobreviventes. Este também é um dos objetivos propostos inicialmente para este problema. Maximizar a métrica sensibilidade afetando o mínimo possível o valor da especificidade.

4.4.5 Otimização

A fase de avaliação dos modelos de DM construídos serve de base na orientação de uma possível otimização dos mesmos. Revendo a definição do problema, este centra-se na descoberta de potenciais dadores num repositório de dados que inclui pacientes que tiveram episódios neurológicos devastadores. Estatisticamente, e como já foi descrito na Secção 1.2, os pacientes com lesões cerebrais graves são mais propensos a progredir para morte cerebral e a tornarem-se potenciais dadores de órgãos.

Neste sentido, deve ser dada ênfase à classificação correta dos dados relativos a pacientes falecidos, sendo a métrica a maximizar a de sensibilidade. Note-se que com a maximização da métrica de sensibilidade, a especificidade do modelo diminui, sendo importante encontrar uma relação de compromisso entre as duas.

Analisando de forma geral os diferentes modelos gerados, os valores relativos às métricas de sensibilidade e especificidade foram sempre relativamente idênticos. Poderá então haver possibilidade de subir o valor de sensibilidade, embora em detrimento da especificidade, de forma a atingir o principal objetivo proposto.

Cost-sensitive Classification (CSC), descrita na Secção 2.4.3, é uma técnica que permite diferenciar os custos associados a cada classe a prever num modelo de DM. Em problemas como o desta dissertação, em que o custo de perder um potencial dador de órgãos é superior a não identificar um paciente que sobreviveu, faz sentido criar uma matriz de custo para que os resultados finais sejam influenciados pelos *inputs* desta matriz.

A ferramenta *Weka* apresenta um algoritmo que faz parte dos classificadores *meta* designado por *CostSensitiveClassifier* que permite aplicar a qualquer algoritmo a técnica de CSC. Sendo que o objetivo é maximizar a métrica de sensibilidade, e revendo a fórmula apresentada na Secção 2.4.6, para que esta fique mais próxima do valor ideal, ou seja 1, as instâncias classificadas como Falsas Negativas têm de diminuir, idealmente ser próximas do valor 0.

Assim, definiu-se uma matriz de custo definida de 2 por 2, representada na Tabela 14, em que o peso relativo a FN foi alterado para o dobro.

Tabela 14 – Matriz de custo para a aplicação da técnica *Cost-sensitive Classification*

		Classe a prever	
		Positiva	Negativa
Classe real	Positiva	VP = 0	FN = 2
	Negativa	FP = 1	VN = 0

4.4.6 Reavaliação

Os resultados da aplicação da técnica de *Cost-sensitive Classification* para os diferentes modelos construídos estão apresentados na Tabela 15.

Com a alteração da matriz de custo como representada na Tabela 14 da secção anterior, as instâncias classificadas como Falsas Negativas foram prejudicadas na aplicação do algoritmo ao conjunto dos dados, ou seja, menos pacientes foram classificados incorretamente como sobreviventes.

Tabela 15 – Resultados da aplicação da técnica *Cost-sensitive Classification* ao Modelo 1 com os algoritmos *RBFClassifier* e *NaiveBayes*, ao Modelo 2 com os algoritmos *RandomForest* e *J48* e ao Modelo 3 com os algoritmos *NaiveBayes* e *RBFClassifier*, através de *Holdout sampling*

		Modelo 1		Modelo 2		Modelo 3	
Algoritmo / Métrica Avaliação		**RBF Classifier**	**NaiveBayes**	**Random Forest**	**J48**	**NaiveBayes**	**RBF Classifier**
Matriz de Confusão	VP	578	563	579	521	551	571
	VN	2942	3275	3923	4292	3278	3015
	FP	4586	4253	3605	3236	4250	4513
	FN	54	69	53	111	81	61
Exatidão		0,431	0,470	0,552	0,590	0,469	0,439
Sensibilidade		0,915	0,891	0,916	0,824	0,872	0,903
Especificidade		0,391	0,435	0,521	0,570	0,435	0,401
AUC		0,653	0,663	0,719	0,697	0,654	0,652

Comparando Tabelas 13 e 15 e relativamente ao Modelo 1 com o algoritmo *RBFClassifier*, o número de VP subiu de 469 para 578, o que significa que o modelo acertou cerca de 91% dos casos de pacientes falecidos. Por outro lado, o número de FN desceu de 163 para 54. Consequentemente, a métrica de sensibilidade subiu de 0,742 para 0,915, o que correspondeu a um aumento de 17,3% da aprendizagem sem a técnica CSC para a aprendizagem que usou a técnica. Pelo contrário, e como já era esperado já que sensibilidade e especificidade são métricas inversamente relacionáveis, a métrica especificidade desceu de 0,673 para 0,391, correspondendo a um decréscimo de 28,2%.

Dos resultados apresentados depois da aplicação da técnica de CSC, o Modelo 2 com a aplicação do algoritmo *RandomForest* apresentou o melhor comportamento no conjunto de dados desconhecido. Apresentou o valor mais elevado no que diz respeito à métrica de sensibilidade, cujo objetivo era maximizar, e é foi o modelo com menores valores de prejuízo nas restantes métricas, quando comparado com a avaliação sem a técnica de CSC.

De um modo geral, e comparativamente com os resultados da Secção de avaliação dos modelos (Secção 4.4.4), ou seja, modelos construídos sem a técnica de CSC, a métrica de sensibilidade subiu consideravelmente para todos os modelos testados, cumprindo-se o objetivo principal de identificação de um maior número de potenciais dadores. Por outro lado, o número de Falsos Positivos aumentou, o que significa que o número de pacientes falecidos incorretamente classificados aumentou. Esta situação

poderá levar a algum esforço da equipa do GCCT a desconsiderar pacientes que sobreviveram, e não faleceram como o modelo indicou. No entanto, esta situação é preferível à não identificação de pacientes potenciais dadores.

4.5 Conclusão e Trabalho Futuro

Este capítulo incidiu essencialmente na análise dos passos necessários para a construção de um modelo de DM cuja finalidade é a deteção de potenciais dadores de órgãos. A tolerância ao erro na saúde é reduzida, dado que envolve vidas humanas. Assim, e com base neste critério, foram definidos dois objetivos principais: maximizar a sensibilidade de modo a identificar o maior número de dadores possível, encontrando uma relação de compromisso entre sensibilidade e especificidade. Isto significa que se pretendeu aproximar a métrica de sensibilidade do valor ideal, diminuindo o número de casos FN (pacientes incorretamente classificados como sobreviventes), e reduzir o número de casos FP (pacientes incorretamente classificados como falecidos), de forma a atingir maior especificidade.

Partindo de um conjunto de dados do repositório de potenciais dadores, e utilizando a metodologia CRISP-DM, que demonstrou ser completa e adequada ao problema em questão, foram dados passos sucessivos na construção de um modelo de DM que respondesse às necessidades exigidas e cumprisse os objetivos propostos.

Durante as primeiras duas etapas do CRISP-DM, compreensão do negócio e compreensão dos dados, foi relevante o estudo e a assimilação do domínio do problema. Foi estudada a divisão do conjunto em 2 blocos, um de treino e um de teste, dada importância à representatividade dos mesmos para a correta modelação e previsão de dados novos e, ainda, levantada a necessidade de balanceamento do conjunto de dados.

Durante a terceira fase da metodologia adotada, a preparação dos dados, utilizaram-se três técnicas de pré-processamento: limpeza de informação inconsistente, criação de novos atributos com base em atributos existentes de modo a gerar informação útil e, ainda, a transformação de dados. O objetivo foi garantir a criação de modelos de DM precisos sobre dados consistentes, fiáveis e de qualidade. Ao mesmo tempo garantir

também a extração de conhecimento útil, proveniente da tabela fonte, através da manipulação de informação e consequente construção de novos atributos. Com a transformação, obteve-se melhor precisão e eficiência nos resultados da aplicação dos algoritmos de DM ao conjunto de dados.

Para ser possível uma correta validação dos resultados, fizeram-se dois blocos de testes. Primeiramente, e numa fase mais experimental, usando a técnica *k-Fold Cross Validation* e numa segunda fase usando a técnica *Holdout Sampling*. Para a primeira, utilizou-se o valor de k=10, de modo a dividir o conjunto de dados em dez partes iguais. O objetivo foi poder fazer dez iterações e o resultado final ser uma média das mesmas. Assim, garantiu-se que os resultados da aplicação dos algoritmos foram representativos do conjunto de dados. Por outro lado, e como a avaliação de um modelo implica a validação deste através de um conjunto aleatório de novas amostras, utilizou-se o *Holdout Sampling* para dividir $^2/_3$ dos dados para o conjunto de treino e $^1/_3$ para o conjunto de teste. Desta forma, obteve-se uma visão de como o modelo se irá comportar num conjunto de pacientes novos potenciais dadores.

Relativamente ao balanceamento dos dados, as técnicas abordadas incluíram *undersampling* e *oversampling*, sendo que para a primeira utilizaram-se as técnicas de *cluster undersampling* e RUS, e para a segunda utilizou-se a técnica SMOTE.

Como primeira conclusão, e com base nos resultados observados, é possível afirmar que a utilização de diferentes técnicas de balanceamento tem impacto nos resultados da aplicação dos algoritmos ao conjunto de dados. Entre o RUS e o *cluster undersampling,* ambas técnicas onde são removidas amostras do conjunto maioritário, o RUS obteve melhores resultados. Em teoria, a técnica que envolve a criação de *clusters* e a seleção das amostras da classe maioritária relevantes para a criação do conjunto de treino final, deveria obter melhor resultados do que a seleção destas amostras de forma aleatória. No entanto, e tal como é sugerido na literatura [107], a abordagem de *cluster undersampling* é apropriada para conjuntos de dados onde há confiança que as amostras das classes positiva e negativa definem corretamente as propriedades de uma classe positiva e negativa, respetivamente. No entanto, e em particular em conjuntos de dados clínicos, não há garantia que classe positiva e negativa reflitam as características reais de determinado registo. No que diz respeito ao SMOTE, técnica de *oversampling*, esta superou as duas técnicas de *undersampling* utilizadas.

As etapas seguintes, modelação, avaliação e implementação foram representativas nesta dissertação.

Como não é um tema novo e existem diversos trabalhos representativos da área, optou-se por fazer uma análise da literatura de forma a perceber quais os algoritmos de DM que obteriam melhores resultados no conjunto de dados disponível. Dos algoritmos selecionados, e tendo em conta que a métrica AUC, obtida a partir da análise da curva ROC, é usada como medida de qualidade do desempenho de um modelo, foi possível inferir os algoritmos que geraram os melhores modelos para o caso de estudo. Estes variaram conforme a técnica de balanceamento utilizada. As árvores de decisão (*J48* e *RandomForest*) obtiveram os valores de AUC mais elevados e, portanto, um melhor comportamento quando foi aplicado o SMOTE ao conjunto de treino. As redes neuronais artificiais (*RBFClassifier*) e a aprendizagem probabilística (*NaiveBayes*) obtiveram melhores resultados quando foram aplicadas técnicas de *undersampling*.

Na fase de otimização dos modelos, deu-se ênfase à classificação correta dos dados relativos a pacientes falecidos, sendo a métrica a maximizar a de sensibilidade. Utilizou-se a técnica de *Cost-sensitive Classification* de forma a diferenciar os custos associados a cada classe a prever num modelo de DM. Em problemas como o desta dissertação, em que o custo de perder um potencial dador de órgãos é superior a não identificar um paciente que sobreviveu, justificou-se a criação de uma matriz de custo, de 2 por 2, em que o peso relativo a FN foi alterado para o dobro, para que os resultados finais fossem influenciados e se atingisse o objetivo proposto. O modelo com melhores resultados na fase de otimização resultou da aplicação do algoritmo *RandomForest* ao conjunto de dados.

Por fim, pode dizer-se que foram encontradas soluções para o problema identificado, através da utilização de técnicas de DM, que permitirão a identificação mais eficiente de potencias dadores de órgãos pelo GCCT no CHP, e a consequente diminuição de doentes em fila de espera.

Como trabalho futuro, seria interessante alargar o número de atributos a considerar para a modelação, incluindo mais características relativas ao paciente, como a escala de coma a que foi associado, reatividade das pupilas, pressão arterial, análises de componentes do sangue, doenças identificadas, como diabetes, hipertensão, se é ou não fumador, entre outras. Depois desta nova seleção de atributos, seria importante a

aplicação de técnicas que escolham os atributos que mais impactam o desempenho dos algoritmos no conjunto de dados. Outra sugestão prende-se com a avaliação do impacto na aprendizagem da divisão do conjunto de dados, em treino e teste, usando diferentes percentagens, como por exemplo 50/50 ou 60/40. Sugere-se ainda a realização de um maior número de testes relacionado com o balanceamento dos dados, com a aplicação de técnicas híbridas e inovadoras da literatura, como por exemplo o *SmoteBoost* e o *RUSBoost* [108, 109]. Estas são técnicas que combinam SMOTE e RUS com o meta-algoritmo *boosting*, tendo este como objetivo a combinação de vários modelos com fraco desempenho para produzir um modelo com bom desempenho. Outra sugestão de trabalho futuro prende-se com a utilização dos algoritmos de aprendizagem com implementações diferentes, modificando os parâmetros de cada algoritmo e registando o seu impacto na tarefa de classificação. Seria também relevante manter um histórico dos pacientes que se tornaram efetivamente dadores de órgãos, para que seja possível o modelo de DM especializar-se em casos reais de sucesso e prever casos novos de forma mais eficaz. Por fim, abranger bases de dados de várias instituições de saúde seria outro ponto de melhoria relevante.

5. Plataforma de Apoio à Decisão Clínica no Transplante de Órgãos

5.1 Introdução

A transplantação de órgãos é o melhor tratamento para salvar vidas na fase final da falência de órgãos. Qualquer pessoa, ao falecer, é um potencial dador de órgãos ou tecidos para transplante, desde que, em vida, não se tenha manifestado contra esta possibilidade, nomeadamente através de inscrição no Registo Nacional de Não Dadores (RENNDA).

Segundo a Direção Europeia para a Qualidade dos Medicamentos e Cuidados de Saúde (EQDM) do Conselho da Europa, e como já anteriormente apresentado na Secção 1.2, em 2017 mais de 144 000 pacientes faziam parte das listas de espera dos Estados-Membros da União Europeia. Ainda relativamente a 2017, 6518 pacientes morreram enquanto aguardavam por um transplante. Morrem, por dia, 18 doentes em lista de espera por não haver órgãos disponíveis. Por hora, são acrescentados às listas de espera europeias cerca de 6 novos pacientes, como relatado no Boletim de Transplantes [9]. Estes números representam a verdadeira dimensão das necessidades dos doentes. O fator mais crítico continua a ser a oferta de órgãos para transplantação.

A deteção precoce de potenciais dadores é o ponto de partida para a transplantação. A única maneira de garantir que não se perdem potenciais dadores é poder identificar e monitorizar possíveis dadores individualmente em hospitais ou áreas geográficas relevantes. A definição de um sistema de avaliação que identifique todas as mortes em hospitais que tenham o potencial de contribuir para a doação de órgãos é crucial.

Neste sentido, e no âmbito deste projeto de dissertação, surgiu este caso de estudo que consistiu na readaptação da plataforma Web de apoio à decisão clínica, o Organite, atualmente implementada no CHP. Nas próximas secções estão descritos todos os passos envolvidos no projeto.

5.2 Definição do Problema e Objetivos da Solução

A implementação de aplicações Web para integrar informação e outros recursos de várias fontes de dados, tornando-os disponíveis através de uma interface simples para o utilizador é, hoje em dia, uma necessidade obrigatória nas organizações de saúde. Estas aplicações conseguem dar suporte às equipas durante o dia-a-dia de trabalho, ajudando no processo de tomada de decisões.

Assim, e aliando os benefícios das TI ao interesse de melhorar a qualidade dos serviços no CHP, surge o Organite, uma plataforma Web de apoio à decisão clínica. Esta plataforma já se encontra implementada atualmente no CHP, tendo surgido no âmbito de projetos de dissertação de Engenheiros Biomédicos da Universidade do Minho [110, 111]. O objetivo desta plataforma é representar a informação de forma a que a equipa do GCCT consiga identificar facilmente fatores chave e desempenhar ações orientadas a um objetivo. O resultado deverá apresentar uma solução de alto nível, pronta para responder a objetivos previamente delineados e providenciar um sistema de apoio à decisão operacional na área do transplante de órgãos.

Apesar das vantagens e dos benefícios indubitáveis que a utilização desta ferramenta de BI pelos profissionais de saúde traz na prestação de cuidados de saúde, a plataforma não tem sido utilizada no seu pleno potencial. Assim sendo, possíveis atualizações na sua arquitetura e nas suas funcionalidades podem ser necessárias de modo a incentivar de novo o seu uso e, assim, redivulgar a solução de TI junto aos profissionais de saúde, bem como impulsionar a sua contínua manutenção e expansão no futuro.

A finalidade principal deste caso de estudo residiu na adaptação da plataforma de BI Organite para ir de encontro às necessidades identificadas pela equipa do GCCT no CHP. Deste modo, e após a definição do problema, foram delineados os seguintes objetivos a cumprir no processo de concretização da solução proposta:

- Estudo da arquitetura e funcionalidades da plataforma atual de BI, de modo a identificar fraquezas a transformar em pontos de melhoria que tragam valor aos prestadores de cuidados de saúde;
- Escolha das metodologias e tecnologias envolvidas para o desenho e desenvolvimento das novas funcionalidades na plataforma de BI;

- Criação e integração de um sistema de notificações de eventos neurológicos adversos na plataforma de BI;
- Criação e implementação de um sistema de dados clínicos persistente na plataforma de BI;

O propósito principal deste caso de estudo residiu, assim, na implementação de funcionalidades que trouxessem melhoria à plataforma Organite e consequentemente contribuíssem para a prestação de cuidados de saúde de qualidade.

5.3 Desenho e Desenvolvimento

5.3.1 Arquitetura BI da Plataforma

Nesta Secção estão descritas todas as considerações envolvidas no desenho da arquitetura da plataforma Organite, atualmente em utilização pela equipa do GCCT no CHP.

No que diz respeito às fontes de dados, e, portanto, às origens dos dados que vão suportar o sistema, estas são sistemas de bases de dados operacionais (sistemas OLTP) do CHP. O DW projetado e desenhado para a plataforma Organite foi construído recorrendo à execução de tarefas de *Extract, Transform and Load* (ETL) sobre os sistemas OLTP, seguindo a metodologia de Kimball. Assim, as fontes de dados do Organite incluem dados de diferentes sistemas operacionais, responsáveis por alimentar o DW.

A criação e gestão do repositório de dados em memória está fora do âmbito desta dissertação, apenas são descritos alguns conceitos de arquitetura para maior clareza e facilidade de entendimento da plataforma por parte do leitor.

Das bases de dados disponíveis no CHP, descritas na Secção 2.1, o PCE revelou ser a fonte de dados mais relevante, sendo o SClinico e o SI utilizado no Serviço de Urgência (SU) fontes de dados auxiliares. A inclusão da fonte de dados relacionada com o sistema de emergência é justificada devido ao facto de um elevado número de pacientes com lesões neurológicas graves darem entrada no hospital através do SU. Estes sistemas são

interoperáveis já que interagem com um *web service*, disponibilizado pela plataforma AIDA, que armazena os registos resultantes dos mesmos (Secção 2.1.1).

O DW representa o repositório analítico onde a informação, previamente preparada, é armazenada na base de dados de potenciais dadores de órgãos e disponibilizada para os utilizadores finais. Este preparamento prévio, ou seja, a pesquisa e os filtros da informação clínica, a consolidação e o processo ETL, é conduzido por um conjunto de procedimentos, em linguagem *Java*. É importante referir que estes procedimentos estão otimizados para poderem operar em tempo real e garantir a atualização diária da informação.

Os procedimentos *Java* responsáveis por inserir e atualizar dados no repositório assim como detetar a ocorrência de eventos críticos para notificação à equipa do GCCT, estão integrados na plataforma AIDA, mais especificamente na máquina *hsa-siima*.

O conjunto de dados pré-processado e agregado que representa os potenciais dadores identificados na plataforma, provém da informação da tabela ORGANITE_REPOSITORY, constituída pelos seguintes campos:

- *ID_SEQ*: identificador único e incremental;
- *ID_EPIS_EXT*: identificador para cada episódio clínico. Um episódio clínico é definido como um conjunto de eventos registados desde que o paciente dá entrada no hospital. Os episódios podem ter o mesmo ID se pertencerem a módulos distintos;
- *COD_DIAG*: código ICD-9-CM do diagnóstico registado. Para um episódio podem ser associados diferentes códigos de diagnóstico, cada um deles representando uma nova linha;
- *NOTES*: descrição do diagnóstico pela equipa médica e/ou da equipa de enfermagem.
- *DT_DIAG*: data do diagnóstico;
- *DT_FOUND*: data de entrada do diagnóstico no sistema;
- *NUM_ORD_PROFESSIONAL*: número sequencial único que identifica o médico que regista o diagnóstico;
- *MODULO:* modulo clínico do episódio;
- *BD:* base de dados de onde provém a informação;

- *NUM_PROCESSO*: número de processo do cliente; é zero quando o paciente dá entrada no hospital pela primeira vez;
- *NUM_SEQUENCIAL*: número sequencial do paciente; nunca é nulo;
- *COD_ESPECIALIDADE*: código da especialidade onde se deu o episódio;
- *DES_ESPECIALIDADE*: descrição da especialidade onde se deu o episódio;
- *DTA_NASCIMENTO*: data de nascimento do paciente;
- *SEXO*: género do paciente;
- *ESTADO*: condição clínica do paciente: entrada, urgências, internado, alta, falecido e desconhecido;
- *DIAG_RECENTE*: identifica o episódio mais recente e os episódios antigos dos pacientes.

Sobre este conjunto de dados, utilizam-se técnicas de análise computacionais automáticas para extração de conhecimento: múltiplas instruções SQL para manipular os dados de acordo com as necessidades de análise do utilizador final e modelos otimizados de *Data Mining* para contribuir no processo de tomada de decisão.

5.3.2 Arquitetura Web da Plataforma

O Organite segue uma arquitetura REST, descrita previamente na Secção 3.2.1, sendo esta uma arquitetura Web cliente-servidor que particiona as tarefas dos servidores e dos clientes.

O servidor atua principalmente como um repositório central para fornecer o código inicial do lado do cliente que é executado no *browser*, armazenar dados gerados pelo utilizador numa base de dados e ainda gerir a autenticação dos utilizadores inscritos na plataforma. A aplicação do lado do cliente inclui a interface do utilizador, gerida por controladores da aplicação no *browser*.

O sucesso e a ampla adoção deste tipo de aplicações é atribuível, em parte, ao *design* de interfaces de utilizador intuitivas, bem como ao aumento da utilização de abordagens de desenvolvimento de aplicações da Web onde mais código é transferido do servidor para o cliente [112]. Ao mover a aplicação para o *browser*, o Organite fornece uma experiência mais interativa e responsiva ao utilizador, atualizando diretamente o

conteúdo da página da Web a partir do *browser* e, portanto, exigindo que menos conteúdo seja gerado pelo servidor.

Esta funcionalidade é adquirida pela utilização de comunicações *Asynchronous JavaScript* e *XML* (*AJAX*), que permite que os dados originais sejam solicitados diretamente do servidor, formatados usando HTML e, de seguida, atualizados na página Web, sendo que todas as ações decorrem no *browser*. Usando esta abordagem, o tempo de resposta da interface é menor, solicitando, formatando e injetando conteúdo diretamente, sem exigir uma atualização da página da Web atual ou a navegação para uma nova página.

A aplicação Web encontra-se implementada no CHP num servidor alojado na máquina *hsa-aida18.* O seu acesso é possível através de qualquer dispositivo na intranet através do endereço *http://172.21.201.151/organite/*.

Os componentes da aplicação Web podem ser divididos em duas partes: *back-end* e *front-end*. Estes dois distinguem a separação entre a camada de acesso a dados - *back-end* - e a camada de apresentação – *front-end*.

Back-end

A arquitetura do lado do servidor para o Organite foi criada usando a micro *framework* minimalista *Flask*, descrita na Secção 3.2.1.

No caso concreto desta dissertação, faz sentido continuar a usar *Flask* como *framework* de *back-end*, já que é flexível e adaptável às necessidades. É uma boa estratégia para projetos desta dimensão que têm oportunidade de melhoria contínua nos anos subsequentes.

Numa arquitetura MVC, o Flask cobre V e C, ou seja, Vista e Controlador. O Flask não fornece uma camada de Modelo integrada pronta para utilização. Esta camada é adaptada às necessidades do projeto e do utilizador [92]. Na plataforma Organite, o Modelo é constituído pelo *Object Relation Mapper SQLALchemy*.

Object Relation Mapping (ORM) é uma técnica de mapeamento de parâmetros de objetos para a estrutura das tabelas do servidor. Os ORM permitem que as aplicações façam a gestão de uma base de dados usando entidades de alto nível, como classes,

objetos e métodos, convertendo-as em comandos de bases de dados [92]. Em aplicações *Flask*, o *SQLAlchemy* é utilizado para executar operações CRUD numa base de dados, fornecendo a eficácia e a flexibilidade do SQL. *Flask-SQLAlchemy* é a extensão do Flask que suporta o *SQLAlchemy* [93].

Acedendo à plataforma Organite, o fluxo da informação inicia com um pedido do utilizador por determinado conteúdo, inserindo um URL. O Controlador recebe esta solicitação e o Modelo retorna os dados necessários, através da execução de *queries* SQL que enviam pedidos de consulta às bases de dados. O Controlador organiza os resultados e envia-os para a Vista, que disponibiliza estes dados num formato específico (JSON) para que estejam acessíveis na camada *front-end* para exploração.

Front-end

O *front-end* é a parte do código de uma aplicação que é geralmente visível para os utilizadores finais, sendo assim uma interface que os convida a interagir com a aplicação. É acedido diretamente pelo utilizador para receber ou utilizar recursos de *back-end*, permitindo fazer pedidos ao sistema de informação subjacente. No desenvolvimento de *front-end*, existem muitas bibliotecas e estruturas disponíveis para tornar a codificação mais fácil, como o *AngularJS* (uma estrutura *JavaScript*), D3.js (uma biblioteca *JavaScript*) e *Bootstrap* (uma estrutura front-end), descritas na Secção 3.2.1.

Nesta tese, a framework usada especificamente para descrever o comportamento *front-end* é o AngularJS. Tudo no *front-end* é escrito principalmente em *HyperText Markup Language* (HTML), CSS (*Cascading Style Sheets*) e *JavaScript*.

Da mesma forma que o *back-end*, o *front-end* também assume um padrão de design MVC. Como o lado do cliente é abstraído do lado do servidor, o *AngularJS* assume a responsabilidade de recuperar o conteúdo das APIs REST que entregam o conteúdo. De modo a sincronizar dados entre o lado do servidor e o lado do cliente, o *AngularJS* possui uma interface simples para aceder à REST API usando comunicação *Asynchronous JavaScript and XML* (AJAX). Os recursos providenciados são armazenados localmente no *browser* como modelos. Quando o utilizador altera um valor de um parâmetro, apenas a versão local é modificada até que o utilizador solicite que a informação seja guardada no servidor. Para guardar a informação no servidor, os dados armazenados localmente

são enviados para a REST API em formato JSON, usando o método HTTP apropriado (*POST* para criar, *PUT* para atualizar). Os dados no formato JSON são, seguidamente, manipulados em *JavaScript* de forma a serem moldados e, posteriormente, representados e visualizados.

A interface do utilizador é construída numa série de vistas em *AngularJS*. Cada vista tem como base um *template* HTML que define como determinado objeto é apresentado na página Web. Definindo uma vista para componente da interface, as vistas podem ser reutilizadas para múltiplos objetos do mesmo tipo.

O *layout* geral do Organite foi desenvolvido usando uma estrutura de *front-end* do *Bootstrap*, elegante e intuitiva. Inclui código *CSS* e *JavaScript* que fornece uma linha de *layout* base para elementos de página Web comuns, como tabelas, barras de navegação, botões e caixas de texto.

Os gráficos gerados no *browser* usam a biblioteca JavaScript D3 (*Data-Driven Documents*) [113]. O D3 fornece uma abordagem funcional para conectar dados diretamente a elementos da página. Por meio de *data binding*, o D3 pode atualizar automaticamente os elementos do gráfico em resposta a alterações nos objetos de dados subjacentes, o que facilita visualizações dinâmicas e interativas.

5.3.3 Implementação de um Sistema de Dados Persistente

O profissional de saúde, no dia-a-dia, acede à plataforma Organite para monitorizar os pacientes identificados como potenciais dadores. Dado que o volume de dados apresentado é significativo, a análise diária é exaustiva e pode tornar-se inconclusiva.

De forma a facilitar esta análise, adicionou-se a funcionalidade de edição à lista de diagnósticos atuais, sendo para isso necessário implementar um sistema de dados persistente. Assim, e após uma análise, o profissional de saúde pode modificar o estado da análise e, consequentemente a cor de uma linha, referente a um determinado paciente para um dos cinco estados:

- Sem critérios: quando identificado que já não é um potencial dador, a linha correspondente ao doente fica vermelha;

- Analisado: quando o paciente está a ser acompanhado porque pode efetivamente ser um dador, cor amarela;
- Dador: quando se trata de um paciente que foi dador, cor verde;
- Não referenciado: quando o paciente não foi referenciado, mas pode eventualmente ser um potencial dador, cor azul;
- Não analisado: quando o paciente ainda não foi analisado, cor branca.

A coluna de anotações permite ainda ao utilizador adicionar ou editar notas acerca da análise do episódio em questão.

A tabela disponibiliza um modo de edição que ativa um *dropdown* na coluna estado da análise e disponibiliza uma caixa de texto na coluna anotações. O utilizador tem a possibilidade de modificar estes campos e gravar ou descartar as alterações.

A Figura 15 ilustra um esquema que representa a comunicação entre o utilizador, o cliente (AngularJS), o servidor (Flask) e a base de dados numa tarefa de edição pelo utilizador. Para cada pedido do cliente, está representado o método HTTP associado e o URL de forma a demonstrar a utilização da REST API.

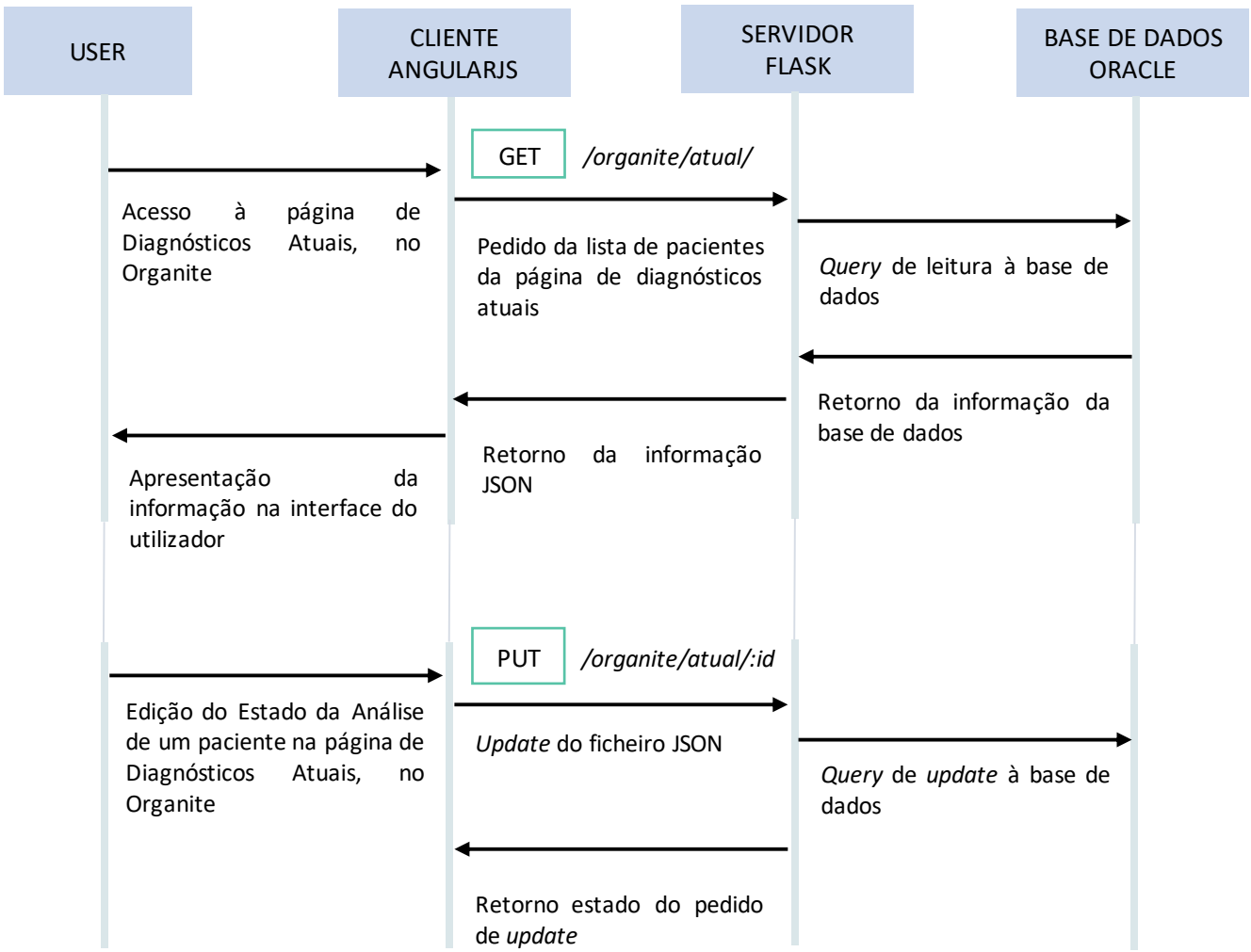


Figura 15 – Comunicação entre utilizador, cliente, servidor e base de dados numa tarefa de edição da tabela disponibilizada na página de Diagnósticos Atuais na plataforma Organite.

O utilizador depois de fazer *login* na plataforma *Organite* acede à página de diagnósticos atuais. O cliente, AngularJS, faz um pedido ao servidor, através da REST API, com o método GET, à informação que contém a lista de pacientes que o cliente requisitou (*/organite/atual/*). Através de uma *query* de leitura à base de dados, a informação requerida é retornada, em formato JSON. Os dados no formato JSON são, seguidamente, manipulados em *JavaScript* de forma a serem moldados e, posteriormente, representados e visualizados.

De seguida, e ainda no mesmo caso de uso, o utilizador ativa o modo de edição e altera o valor de um parâmetro. Apenas a versão local é modificada até que o utilizador dê a instrução de guardar para solicitar que a informação seja guardada no servidor.

Para guardar a informação no servidor, os dados armazenados localmente são enviados para a REST API, em formato JSON, usando o método HTTP PUT para atualizar o ID único associado à edição do utilizador da informação que contém a lista de pacientes (*/organite/atual/:id*). É feita uma *query* de *update* à base de dados para refletir a modificação efetuada, enviando-se posteriormente ao cliente um *status* do pedido de *update*.

Assim, se porventura um novo utilizador entrar na plataforma ou fizer *refresh* na sua página de diagnósticos atuais, já irá ver a última edição e gravação submetida.

5.3.4 Módulo de Notificações

A equipa do GCCT além das funções diárias no hospital relacionadas com o cuidado dos pacientes, tem uma preocupação adicional de acompanhar e estudar a evolução de um paciente que sofreu eventos neurológicos devastadores. O objetivo é perceber se poderá evoluir para morte cerebral, causa mais frequente para os dadores de órgãos. Se assim acontecer, é consultado o RENNDA e existe uma conversa com a família para se estudar a viabilidade do paciente se tornar dador de órgãos. Toda esta dinâmica, numa situação ideal, começa a ser posta em prática com alguma antecedência, de forma a facilitar e a tornar mais eficaz o posterior processo de tratamento, transporte e transplantação de órgãos num dador compatível, identificado pelas Unidades de Colheita e de Transplantação.

Se a equipa do GCCT for alertada cedo para determinado diagnóstico de um paciente, a monitorização é feita de forma eficaz e o processo é agilizado. Já existe atualmente implementado um sistema que envia uma mensagem para o número de telefone e email associados à inscrição dos utilizadores na plataforma Organite. O aviso sinaliza situações que apresentam eventos adversos em relatórios recentes de tomografia computorizada crânio-encefálica (TAC), identificando o número sequencial do paciente, o número do exame, o episódio, o módulo e a data de deteção.

Para terem acesso a dados mais detalhados sobre os pacientes identificados, a equipa do GCCT acede à plataforma Organite e procura manualmente pelo paciente em questão no separador dos diagnósticos atuais. Sendo que tanto o número sequencial

como o do episódio não são números únicos, o utilizador poderá ver mais do que uma linha associada à sua pesquisa.

O conjunto de passos para conseguir monitorizar um paciente em específico não é simples nem intuitivo para o utilizador, tornando a sua experiência de interação com a plataforma negativa.

Assim, e perante a necessidade identificada, foi projetada e desenvolvida uma solução de melhoria do sistema de notificações, que inclui o alerta de notificações na plataforma Web, permitindo a interação dos utilizadores com as mesmas. Na Figura 16 encontra-se representada a arquitetura do sistema de notificações.

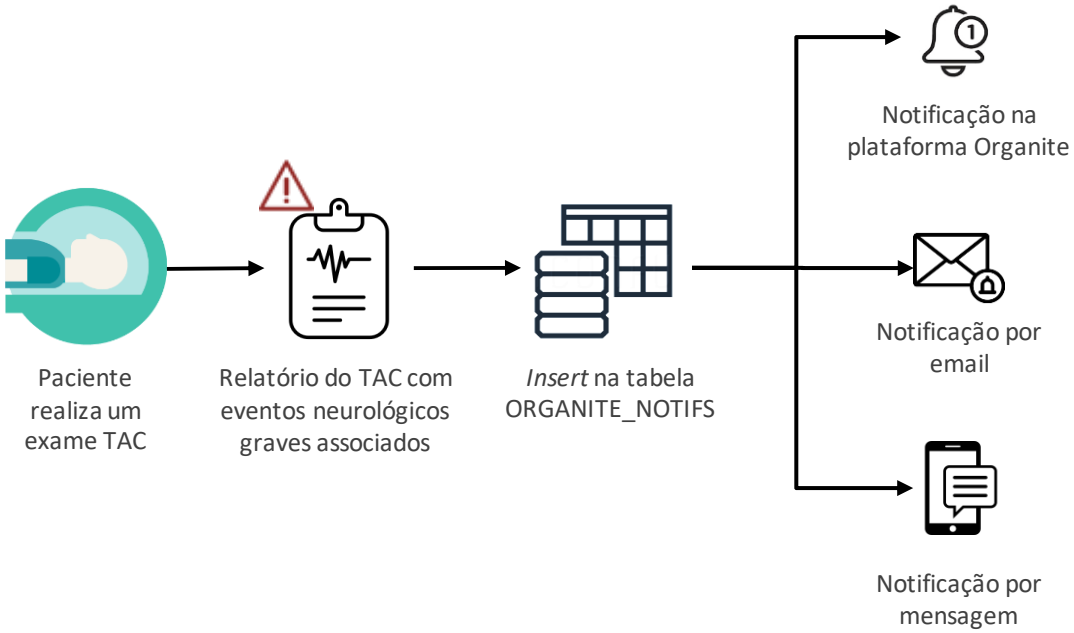


Figura 16 – Arquitetura do sistema de notificações.

As novas funcionalidades permitem que, depois de um utilizador iniciar sessão na plataforma, receba de imediato uma notificação *pop-up*, ou seja, uma janela que surge automaticamente, com o número de eventos identificados e aos quais deve ser prestada atenção. Através de um clique na notificação, o utilizador é encaminhado para uma página onde é apresentada uma lista de notificações.

Os utilizadores podem aceder à página da lista total de notificações de várias formas. Clicando na notificação *pop-up* que surge no ecrã, clicando na opção "Notificações" do menu de navegação que surge do lado esquerdo ou ainda aquando do clique no ícone do sino, no canto superior direito da página. Relativamente a esta última opção, um

dropdown surge aquando do clique do utilizador no ícone e é exibida uma lista com as novas notificações com detalhe ao número sequencial dos pacientes. O número sequencial encontra-se clicável para ser possível entrar na página de detalhe do paciente selecionado.

A página da lista total de notificações contém uma tabela com a mesma informação que é enviada na mensagem e no email. No entanto, esta informação encontra-se clicável para ser possível aceder mais rapidamente ao exame crânio-encefálico ou ao detalhe do paciente.

As primeiras linhas da tabela corresponderão sempre às mensagens não lidas e, portanto, com maior prioridade. A tabela encontra-se também ordenada por data, das notificações mais antigas às mais recentes. Permite ao utilizador marcar uma notificação como lida e também não lida, ou ainda eliminar caso determinado evento identificado não seja relevante. Para ser visualmente percetível, as notificações não lidas estão salientadas a negrito.

Nesta página é ainda possível fazer uma pesquisa, através de *search boxes*, do exame, número sequencial do paciente e da data de deteção do evento neurológico.

É relevante explicar a parte técnica do sistema de notificações implementado.

Quando um exame é sinalizado porque contém uma palavra ou expressão da lista previamente definida, é adicionada uma linha a uma tabela de notificações (ORGANITE_NOTIFS) com as colunas *Estado_Notif* e *Historico*, do tipo *boolean*, ambas preenchidas com o valor "0". A primeira coluna foi adicionada à tabela original com o intuito de distinguir as notificações lidas das notificações não lidas pelos utilizadores. Assim, o valor "0" corresponde a notificações novas, ou seja, a notificações não lidas, e o valor "1" corresponde a notificações já lidas. Relativamente à segunda coluna, o objetivo é identificar as notificações que os utilizadores descartam na plataforma, fazendo estas parte de um histórico de notificações para uma possível análise futura.

Assim, quando o sistema é iniciado, a aplicação do lado do servidor é responsável por extrair dados da tabela de notificações e armazenar a informação num ficheiro JSON. A *query* que o sistema faz à base de dados exclui o histórico de notificações, já que este não traz informação útil, pelo menos imediata, para o utilizador.

Do lado do cliente, e já com base no ficheiro JSON, é feita uma contagem das novas notificações, através de um filtro ao conjunto nome/valor '*Estado_Notif:0*'. O resultado

é armazenado numa variável e apresentado na notificação *pop-up* quando o utilizador inicia sessão.

A notificação *pop-up* é gerada através da utilização do *Bootstrap Notify*, versão 3.0.0, uma ferramenta que permite criar mensagens a serem exibidas depois da execução de ações do utilizador. As mensagens são facilmente personalizáveis através de diferentes cores, por exemplo, que indicam aviso, sucesso ou erro, tornando melhor a interação entre utilizador e plataforma [114].

Já na página da lista total de notificações, quando o utilizador marca uma notificação como lida, o ficheiro JSON é atualizado e é feito um *update* da linha correspondente na tabela ORGANITE_NOTIFS da base de dados, passando o valor da coluna *Estado_Notif* para "1". Se o utilizador pretender marcar uma notificação já lida como não lida, o procedimento é o mesmo, sendo que a coluna *Estado_Notif* é atualizada para o valor "0".

É ainda possível descartar uma notificação não relevante para que não surja na página das notificações. Quando esta ação é executada, a coluna *Historico* é atualizada para o valor "1". Assim a notificação só é eliminada visualmente na plataforma, já que o ficheiro JSON deixa de a englobar, mas não é eliminada permanentemente na tabela da base de dados.

5.3.5 Interface do Utilizador

O Organite tem várias secções que permitem ao utilizador fazer diferentes tipos de análise conforme a sua necessidade. A *homepage*, representada na Figura 17, permite ao utilizador iniciar sessão se já tiver acesso garantido à plataforma, registar-se se estiver a aceder à plataforma pela primeira vez ou ainda aceder ao painel das estatísticas, sendo este um painel sem restrição em termos de privacidade, podendo ser acedido por várias equipas do CHP.

Figura 17 – Homepage da plataforma Organite.

O *dropdown* no canto superior da página permite ao utilizador anónimo escolher a opção de iniciar sessão ou registar-se. A página de início de sessão (Figura 18) é idêntica à *homepage*, diferindo apenas na *form* que surge para preenchimento do número mecanográfico do profissional de saúde e da *password* escolhida no momento de registo. A página de registo apresenta igualmente uma *form* para preenchimento com campos adicionais para permitir o registo: número mecanográfico, nome, email, telemóvel, *password*. Todos os campos são de preenchimento obrigatório, não sendo permitido ao utilizador submeter o pedido sem os preencher de forma correta.

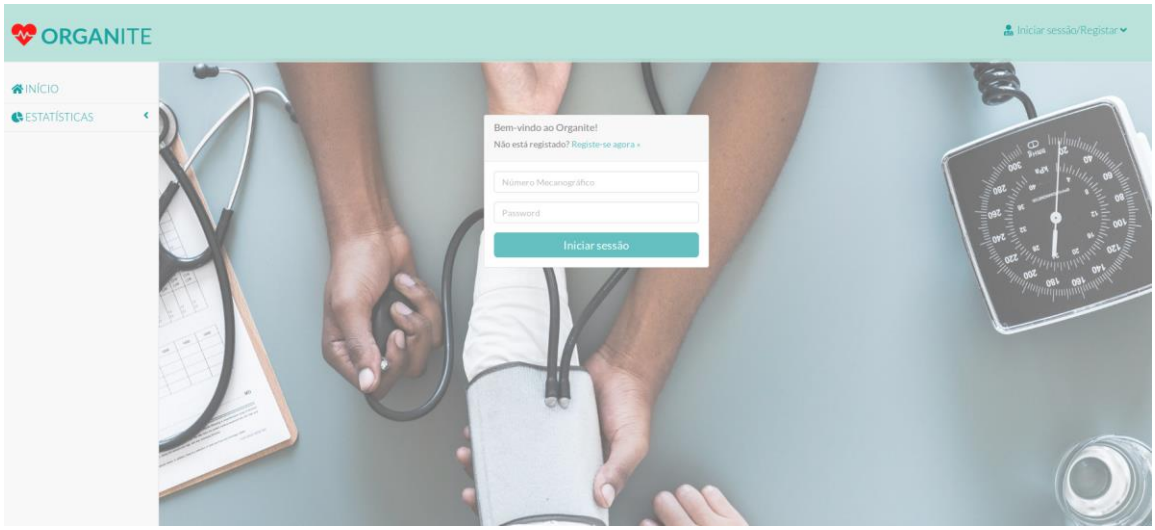


Figura 18 – Página de início de sessão do utilizador na plataforma Organite.

Se o campo relativo ao email não for preenchido corretamente surge uma mensagem *pop-up* que obriga à correção da sintaxe do campo preenchido. O mesmo acontece para o caso do número de telemóvel. Os caracteres inseridos têm obrigatoriamente de ser números. Estas verificações são relevantes para que as notificações da plataforma possam ser comunicadas por diferentes meios e alertem eventos adversos relevantes ao profissional de saúde.

Quando o utilizador clica no botão "Submeter", é enviado um email para um administrador da plataforma responsável por aprovar o acesso do utilizador. O utilizador vê uma mensagem num *pop-up* que o avisa que o pedido de registo foi submetido e que irá receber um email quando a conta for aprovada. O administrador entra na plataforma e acedendo à página "Gestão de Utilizadores", vê uma lista de utilizadores ordenados de forma descendente pela data de adesão. Os utilizadores novos, ou seja, que ainda não foram aprovados, distinguem-se pela cor da linha distinta das restantes. Além de permitir fazer a aprovação da nova conta, pode ser dado um de dois tipos de privilégios: utilizador ou administrador. Um utilizador pode navegar pela plataforma, mas não lhe é permitido fazer nenhuma gestão da mesma. O administrador tem todos os privilégios.

Após o utilizador fazer o *login* com as suas credenciais, é redirecionado para uma *homepage* com mais informação. Existem macro indicadores que dizem quantos pacientes novos existem com eventos neurológicos adversos e quantos pacientes existentes ainda não foram analisados. Está também apresentada informação estatística relativa ao total de pacientes identificados no último dia, na última semana e no último mês. Clicando nos macro indicadores, o utilizador é redirecionado para a página de diagnósticos atuais, e clicando nas estatísticas, o utilizador é redirecionado para a página de análise gráfica de dados, agregada por vários períodos temporais e domínios de informação. Aquando o *login* surge um *pop-up* no canto superior da página com informação sobre o número de novos eventos neurológicos adversos, como se encontra representado na Figura 19. Clicando na mensagem, o utilizador é redirecionado para a página de notificações.

Figura 19 – *Pop-up* de notificações relativas a eventos neurológicos adversos aquando do início da sessão de um utilizador.

À esquerda da *homepage* e visível em toda a plataforma quando o utilizador tem sessão iniciada, encontra-se a navegação dos menus principais que o Organite disponibiliza: início, diagnósticos, que inclui atual, histórico e previsão, estatísticas e notificações.

No canto superior direito existem dois *dropdowns*, também acessíveis através de qualquer página do Organite quando o utilizador tem sessão iniciada. O primeiro é referente ao módulo das notificações e o segundo ao perfil do utilizador. Clicando nas notificações, surge uma lista de números sequenciais relativos aos últimos 10 pacientes com eventos neurocríticos não lidos. Estes são clicáveis e redirecionam o utilizador para a página de detalhe do paciente em questão. Acedendo ao detalhe de um paciente, o utilizador tem acesso a dados relevantes do diagnóstico do paciente, nomeadamente a localização no hospital, exames realizados, características gerais e acesso direto ao PCE para análise mais detalhada do perfil clínico.

No final da lista, existe ainda o botão "ver mais" que abre a página das notificações onde é possível ver todas as notificações lidas e não lidas. O segundo *dropdown* dá acesso ao perfil do utilizador, à página de gestão de utilizadores (se o utilizador for um administrador), à lista de códigos ICD9 e permite que o utilizador faça *logout* da plataforma.

Relativamente ao menu lateral de navegação do Organite, este permite o acesso às páginas de principal interesse da plataforma: diagnósticos, estatísticas e notificações.

O módulo dos diagnósticos divide-se em três tipos de análise: atual, histórico e previsão.

A página dos diagnósticos atuais, representada na Figura 20, apresenta uma tabela com informação dos pacientes potenciais dadores dos últimos seis meses. As colunas da tabela incluem o número sequencial do paciente, o número de processo, o ID do episódio, anotações associadas ao episódio, a especialidade, a data do diagnóstico, o estado do paciente, o estado da análise e observações. As últimas duas colunas da tabela, estado da análise e observações, são colunas alteráveis, criadas com o intuito de ajudar os profissionais de saúde a organizar e agilizar a sua análise.

Todas as colunas da tabela permitem o *sort* ascendente ou descendente da informação. A predefinição quando o utilizador acede à página permite ver os pacientes com diagnósticos mais recentes no topo da tabela, ou seja, esta encontra-se ordenada por data de diagnóstico de forma descendente. A tabela tem também paginação, permitindo navegar entre os registos dos últimos seis meses.

A parte superior da tabela inclui um conjunto de filtros de pesquisa para exploração de dados na tabela. Assim, os utilizadores podem fazer uma pesquisa por número sequencial, número de processo, ID do episódio, data do diagnóstico e estado da análise (cor).

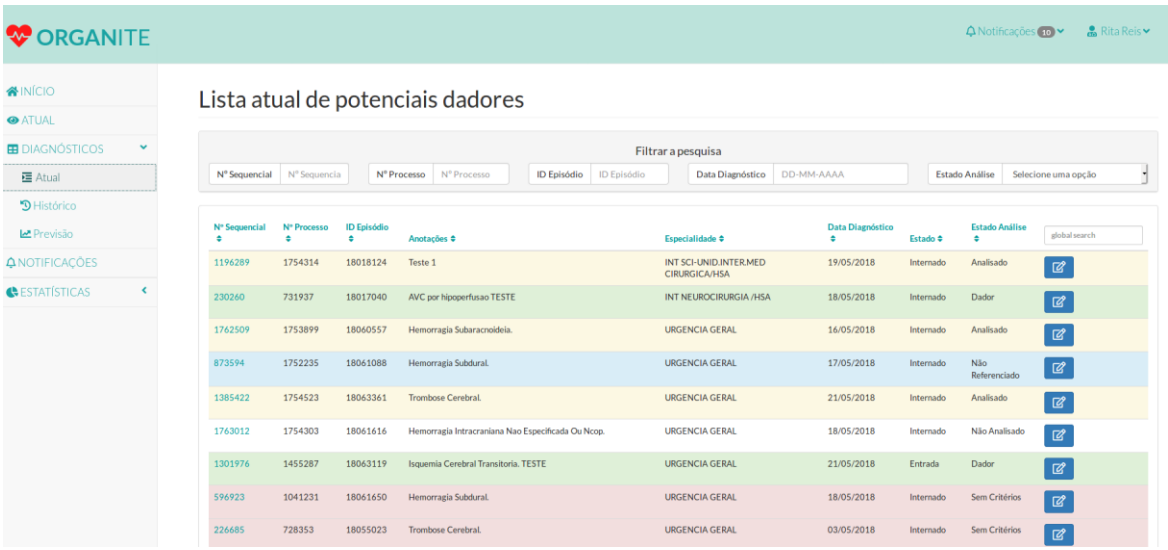


Figura 20 – Página de diagnósticos atuais da plataforma Organite.

Informação clínica mais antiga encontra-se na página do histórico dos diagnósticos, assim como informação relativa a pacientes cujo estado da análise é "Sem Critérios" ou "Dador", pois deixa de fazer sentido ser incluída na página dos diagnósticos atuais depois da análise estar concluída.

Na página de previsão dos diagnósticos, é apresentada uma lista de pacientes com um prognóstico associado assim como a probabilidade desse prognóstico estar correto. Os pacientes cujo prognóstico é 'Não Sobrevive' e com maior percentagem associada requerem monitorização mais próxima pela equipa do GCCT, já que têm maior probabilidade de se tornarem dadores de órgãos.

Esta previsão é feita através de um modelo de *Data Mining,* explorado e detalhado no caso de estudo da Secção 6. Na Figura 21, é apresentado um protótipo da implementação do modelo de DM na plataforma Organite.

Figura 21 – Protótipo da implementação do modelo de previsão de *Data Mining* na plataforma Organite.

5.4 Conclusão e Trabalho Futuro

O Organite demonstra a utilização de práticas web standard e modernas que criam uma interface de utilizador interativa. Em termos clínicos é uma ferramenta cujo propósito é de utilização diária por um membro da equipa do GCCT, de forma a

monitorizar pacientes sinalizados com eventos neurológicos graves sem ter de deslocar-se de uma forma constante à cama destes pacientes. Providencia um nível de interação com o utilizador simples e intuitivo, facilitando a procura de episódios e pacientes específicos, a análise da informação por período temporal, atual e histórica, a modificação do estado desta análise, a adição de notas pessoais por evento sinalizado. Através do sistema de notificações implementado, é mais rápida a identificação dos pacientes com eventos neurológicos adversos a serem analisados. Esta rápida identificação leva a que o processo de doação de órgãos inicie cedo e potencia o sucesso de uma possível transplantação.

No geral, a plataforma Organite disponibiliza as ferramentas necessárias para ser uma ferramenta de trabalho diária eficaz, que contribui positivamente para a deteção de um maior número de potenciais dadores de órgãos.

Em termos de pontos de melhoria, a implementação do protótipo de *Data Mining* na plataforma Organite é essencial para a melhor monitorização dos pacientes e consequente diminuição da não identificação de potenciais dadores.

Identifica-se também como sugestão a inclusão de um filtro temporal à semana e ao mês nas tabelas de diagnósticos, para facilitar a pesquisa de pacientes. Seria também interessante acrescentar uma página onde os utilizadores possam adicionar comentários, *feedback* de análises, avisos, visíveis para toda a comunidade de utilizadores da plataforma. O objetivo é criar dinamismo de análise, contribuir para trabalho conjunto e eficiente e não individual e exaustivo.

Relativamente ao sistema de notificações, seria relevante existir um *log* de modificações dos utilizadores, já que as notificações são disponibilizadas a nível geral na plataforma e não a nível individual por conta de utilizador. Assim, o primeiro utilizador a aceder à plataforma, poderá marcar a notificação como lida e fazer a sua análise. Os utilizadores seguintes já poderão ver a análise do primeiro utilizador, assim como já irão verificar que a notificação foi marcada como lida.

6. PROVA DE CONCEITO

Uma prova de conceito permite demonstrar, na prática, os conceitos, as metodologias e as tecnologias envolvidas na elaboração de determinado projeto, de forma a validar a solução proposta através da prova da sua viabilidade e utilidade.

A análise SWOT (*Strengths, Weaknesses, Opportunities and Threats*), previamente descrita na Secção 3.3, é uma ferramenta utilizada para estruturar um planeamento estratégico, promovendo uma análise dos pontos fortes e fracos (fatores internos), assim como das oportunidades e ameaças (fatores externos).

No âmbito desta dissertação, a análise SWOT consiste em avaliar a plataforma Organite de forma a perceber quais os seus pontos fortes e pontos fracos, assim como as oportunidades e ameaças que poderá enfrentar, sendo assim possível traçar diretrizes para estratégias futuras.

Os principais pontos fortes identificados são:

- Elevada interoperabilidade ao permitir a recolha de informação de vários SIH;
- Elevada escalabilidade e adaptabilidade pelo facto de se tratar de um sistema modular que possibilita implementar melhorias sem ser necessário alterar todo o sistema;
- Centralização de informação relacionada com a doação de órgãos;
- Disponibilização near-real time de informação de potenciais dadores;
- Possibilidade de alteração de registos pelo utilizador, permitindo uma análise personalizada da informação clínica;
- Capacidade de previsão do diagnóstico dos pacientes e consequente potencialidade na doação de órgãos;
- Capacidade de notificar (por email, mensagem e na plataforma Web) o profissional de saúde aquando da identificação de novos potenciais dadores;
- Localização em tempo real dos pacientes no CHP;
- Elevada usabilidade pela interface intuitiva e de fácil utilização (user-friendly);
- Segurança e confidencialidade dos dados assegurada por mecanismos de autenticação;

- Não requer a utilização de ferramentas externas de BI para a visualização dos dados em tabelas, diagramas e gráficos.

No que diz respeito às suas limitações, ou seja, pontos fracos, podem ser apontados:

- Necessária ligação à intranet do CHP, não existindo de momento possibilidade de aceder à informação do exterior;
- Não existência de redundância;
- Indisponibilidade de dados relativos a pacientes que foram efetivamente dadores de órgãos.

Por outro lado, e tendo em conta fatores externos à plataforma, é possível identificar as seguintes oportunidades:

- Disponibilização de uma ferramenta de apoio à decisão clínica na identificação de possíveis dadores de órgãos;
- Aumento do número total de doações;
- Redução do erro médico;
- Extração de conhecimento do conjunto de dados clínicos armazenado nos SIH;
- Recurso a novas tecnologias de forma a melhorar a qualidade da informação nos serviços de saúde;
- Progressivo aumento da interoperabilidade em meio hospitalar.

Em último lugar, para além de oportunidades, os fatores externos também representam por vezes ameaças, tais como:

- Alterações no esquema organizacional dos SIH podem resultar num funcionamento imprevisível da plataforma;
- Dependência da informação disponibilizada pelos SIH, podendo originar problemas de disponibilização e atualização dos dados na plataforma;
- Difícil aceitação e adaptação por parte dos profissionais de saúde a novos tecnologias;

Posto isto, está provada a viabilidade, utilidade e usabilidade da plataforma Organite e é possível encarar o futuro com perspetivas otimistas, uma vez que os pontos fortes associados às oportunidades apresentadas, suplantam as possíveis ameaças e pontos fracos identificados.

7. CONCLUSÃO E TRABALHO FUTURO

Esta dissertação encerra-se com uma breve conclusão, onde se destacam as principais contribuições alcançadas no âmbito do desenvolvimento e da exploração de ferramentas informática para o apoio à decisão e à prática clínica em unidades hospitalares.

Sendo que a deteção precoce de potenciais dadores é o ponto de partida para a transplantação, a definição de um sistema de avaliação que identifique as mortes em hospitais que tenham o potencial de contribuir para a doação de órgãos é crucial.

A equipa do Gabinete de Coordenação de Colheita e Transplantação (GCCT) do Centro Hospital do Porto (CHP) beneficia do Organite, plataforma inteligente do âmbito deste projeto de dissertação, à qual pode aceder via intranet, e explorar dados *near-real time* dos pacientes que dão entrada no hospital e são sinalizados como potenciais dadores de órgãos.

O profissional de saúde é alertado via mensagem no telemóvel e também via email de que existem novos eventos neurológicos críticos a ser analisados. Acedendo à plataforma Organite, recebe prontamente uma notificação do número de pacientes urgentes a monitorizar. Acedendo ao detalhe de um paciente, tem acesso a dados relevantes do diagnóstico do paciente, nomeadamente a localização no hospital, exames realizados, características gerais e acesso direto ao Processo Clínico Eletrónico (PCE) para análise mais detalhada do perfil clínico. Da lista urgente de pacientes notificados, e de forma ao profissional de saúde perceber qual a ação necessária no imediato, pode aceder à Secção de previsão, onde é apresentada uma lista ordenada de forma descendente pela maior probabilidade dos pacientes se tornarem potenciais dadores. Os pacientes com maior percentagem associada vão requerer uma monitorização mais próxima pela equipa do GCCT.

Assim, a situação do potencial dador é controlada e o tempo despendido entre a colheita, a recuperação e a transplantação, parâmetros cruciais para uma transplantação bem-sucedida, diminui.

7.1 Principais Contribuições

Com a realização deste projeto de dissertação foram criados vários artefactos na área das Tecnologias de Informação, isto é, foram desenhadas e desenvolvidas soluções que respondem a necessidades identificadas pelos profissionais de saúde do Centro Hospitalar do Porto (CHP). Assim, a concretização deste projeto incluiu a projeção, desenho, desenvolvimento e implementação de artefactos bem-sucedidos de TI, com base num conjunto de metodologias, *Design Science Research* como metodologia de investigação, CRISP-DM como metodologia de modelação em *Data Mining*, e ainda em *frameworks* tecnológicas, como *Flask e AngularJS.*

Relativamente à *Questão de Investigação nº1*, e de forma a melhorar o sistema de apoio à decisão clínica implementado, o Organite, fez-se um levantamento das necessidades de melhoria junto da equipa do Gabinete de Coordenação de Colheita e Transplantação (GCCT), no Centro Hospitalar do Porto (CHP). Os resultados do levantamento incluíram pontos de melhoria relacionados com a experiência do utilizador, nomeadamente na permissão da inserção e modificação de valores associados a um paciente, à complexidade da informação apresentada sem detalhe explicativo, à dificuldade na procura de um paciente, à falibilidade do sistema de notificações. Optou-se pela manutenção das tecnologias de base envolvidas na construção da plataforma, no *back-end Flask,* no *front-end AngularJS,* havendo atualização das funcionalidades para as versões recentes e complemento com novas funcionalidades, como o *Bootstrap Notify,* para responder à melhoria de requisitos funcionais. Também contribuíram para a melhoria da prestação de cuidados de saúde, a criação e implementação de um sistema de notificações na plataforma Web e de um sistema de dados clínicos persistente. A primeira funcionalidade permitiu diminuir o tempo de procura de um paciente pelo profissional de saúde, acelerando o processo de análise, e a segunda permitiu a interação do utilizador com a informação clínica, permitindo edição, inserção e remoção de dados. O novo desenho da plataforma Organite tornou-a *user-friendly,* com características inovadoras, cuja arquitetura pode ser facilmente adaptada a novas bases de dados e, portanto, nova informação útil e explorável pelos profissionais de saúde.

Um dos principais objetivos desta dissertação, levantados pela *Questão de Investigação nº 2,* é a identificação do maior número de dadores possível, no menor período de tempo. Para isso, recorreu-se ao estudo de técnicas de *Data Mining* (DM) e análise dos passos necessários para a construção de um modelo cuja finalidade é a determinação do prognóstico de um paciente de um repositório de potenciais dadores.

Definiram-se dois objetivos principais na construção do modelo de DM: maximizar a métrica estatística de sensibilidade de modo a identificar o maior número de dadores possível, encontrando uma relação de compromisso entre esta e a métrica estatística especificidade. Isto significa que se pretendeu aproximar a métrica de sensibilidade do valor ideal, diminuindo o número de casos Falsos Negativos (pacientes incorretamente classificados como sobreviventes), e reduzir o número de casos Falsos Positivos (pacientes incorretamente classificados como falecidos), de forma a atingir maior especificidade.

A comparação de vários algoritmos de DM de diferentes famílias, sob as mesmas condições, é uma das contribuições relevantes deste trabalho, já que permitiu aumentar o conhecimento sobre o potencial destes algoritmos e a sua utilização em problemas semelhantes. O meta-algoritmo *RandomForest* revelou ser o algoritmo mais adequado para a resolução deste problema, quando combinado com técnicas de *Cost-sensitive Classification*.

A construção de modelos de DM com diferentes técnicas de balanceamento e algoritmos de DM associados levou a um conjunto alargado de resultados para análise. Note-se a importância do conhecimento e consolidação do mesmo no impacto que as técnicas de balanceamento podem surtir na construção de um modelo sobre um conjunto de dados desequilibrado, cujo objetivo é a previsão de uma variável binária. A técnica SMOTE (*Synthetic Minority Over-sampling Technique*) foi a técnica de balanceamento com melhor desempenho nos testes.

Adicionalmente, a metodologia CRISP-DM revelou-se consistente e adequada, pois permitiu uma melhoria progressiva dos resultados ao longo do projeto.

Face aos objetivos propostos, conjunto de dados usado, conjunto de testes realizados e face ao exposto no capítulo 6, é possível dizer que foram encontradas soluções para o problema identificado, que permitirão a identificação mais eficiente de potencias

dadores de órgãos pelo GCCT no CHP, e a consequente diminuição de doentes em fila de espera.

De forma a provar a viabilidade, utilidade e usabilidade das soluções de TI definidas ao longo deste projeto de dissertação, e respondendo à *Questão de Investigação nº3,* aplicou-se a metodologia da Prova de Conceito através da utilização de uma análise SWOT (*Strengths, Weaknesses, Opportunities and Threats*). Garantiu-se, assim, que as ferramentas desenhadas e implementadas apresentam elevada escalabilidade, são úteis, intuitivas e incluem informação estruturada, traduzindo-se em conteúdo clínico de valor.

7.2 Trabalho Futuro

A plataforma desenvolvida poderá evoluir os seus componentes (*Data Warehouse* construído, *back-end* e *front-end* da aplicação Web, modelo de *Data Mining*), sendo capaz de integrar atualizações e adições de módulos sem afetar todo o sistema. Assim, e de forma a responder à *Questão de Investigação nº3*, analisaram-se propostas de alteração e de melhorias possíveis.

De forma a providenciar um sistema de apoio à decisão mais robusto, a plataforma deverá evoluir nos procedimentos de recolha de informação, incluindo mais características relativas ao paciente, como a escala de coma a que foi associado (Escala de Coma de Glasgow (GSC)), reatividade das pupilas, pressão arterial, análises de componentes do sangue, doenças identificadas, como diabetes, hipertensão, se é ou não fumador, entre outras.

A recolha de informação adicional beneficiará o modelo de *Data Mining* que prevê se um paciente se irá tornar ou não num potencial dador de órgãos, já que os resultados obtidos se vão aproximar mais da realidade clínica. Depois desta nova seleção de informação, seria importante a aplicação de técnicas que escolham os atributos que mais impactam o desempenho dos algoritmos no conjunto de dados. Exemplos destas técnicas poderiam incluir a remoção de atributos correlacionados entre si, a determinação do peso de cada um dos atributos na classificação de uma amostra e a remoção de redundâncias.

Outras sugestões para a melhoria do modelo de DM prendem-se com a avaliação do impacto na aprendizagem da divisão do conjunto de dados, a aplicação de técnicas híbridas no balanceamento dos dados e a utilização dos algoritmos de aprendizagem com implementações diferentes.

A manutenção de um histórico dos pacientes que se tornaram efetivamente dadores de órgãos é um ponto crucial de melhoria, para que seja possível o modelo de DM especializar-se em casos reais de sucesso e prever casos novos de forma mais eficaz.

Abranger bases de dados de várias instituições de saúde seria outro ponto de melhoria relevante, apesar de se entender a possível dificuldade desta melhoria dados os aspetos burocráticos e de sensibilidade na manipulação deste tipo de informação.

Em termos de pontos de melhoria no que diz respeito ao *front-end* do Organite, a implementação do protótipo de *Data Mining* é essencial para a melhor monitorização dos pacientes e consequente diminuição da não identificação de potenciais dadores. Identifica-se também como sugestão a inclusão de um filtro temporal à semana e ao mês nas tabelas de diagnósticos, para facilitar a pesquisa de pacientes. Seria também interessante acrescentar uma página onde os utilizadores possam adicionar comentários, *feedback* de análises, avisos, visíveis para toda a comunidade de utilizadores da plataforma. O objetivo é criar dinamismo de análise, contribuir para trabalho conjunto e eficiente e não individual e exaustivo.

8. BIBLIOGRAFIA

[1] Oztekin, A. (2010). Data mining-based survival analysis and simulation modeling for lung transplant (Doctoral dissertation, Oklahoma State University).

[2] Associação Portuguesa de Insuficiências Renais (2017). Doação e Transplantação de Órgãos e Tecidos (Documento original: Kidney Health Australia).

[3] Robalo, R. (2016). Cuidados de enfermagem à pessoa em situação neurocrítica, potencial dadora de órgãos/tecidos (Doctoral dissertation, Instituto Politécnico de Setúbal. Escola Superior de Ciências Empresariais).

[4] Comité Europeu para a Transplantação de Órgãos. Direção Europeia da Qualidade dos Medicamentos e Cuidados de Saúde (EDQM) (2013). Guia para a qualidade e segurança dos órgãos para transplantação. França.

[5] Coordenação Nacional de Transplantação (2018). Doação e Transplantação de Órgãos Atividade Nacional 2012-2017.

[6] Sociedade Portuguesa de Transplantação. (2014). Retrieved March 22, 2018, from SPT website: *http://www.spt.pt/site/desktop/indice-5.php*

[7] Global Observatory on Donation and Transplantation - GODT. (2016). Retrieved March 25, 2018, from GODT website: *http://www.transplant-observatory.org/summary/*

[8] Doação e Transplantação de Órgãos. (2017). Retrieved April 5, 2018, from SNS website: *https://www.sns.gov.pt/monitorizacao-do-sns/doacao-e-transplantacao-de-orgaos-2/*

[9] Council of Europe (2018). The European Day for Organ Donation and Transplantation. Retrieved April 5, 2018, from: *https://www.edqm.eu/sites/default/files/factsheet_organ_tissue_cell_donation_eodd_2018.pdf*

[10] El-Sappagh, S. H., & El-Masri, S. (2014). A distributed clinical decision support system architecture. Journal of King Saud University-Computer and Information Sciences, 26(1), 69-78.

[11] Buntin, M. B., Burke, M. F., Hoaglin, M. C., & Blumenthal, D. (2011). The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health affairs, 30(3), 464-471.

[12] Bardhan, I. R., & Thouin, M. F. (2013). Health information technology and its impact on the quality and cost of healthcare delivery. Decision Support Systems, 55(2), 438-449.

[13] Lee, J., McCullough, J. S., & Town, R. J. (2013). The impact of health information technology on hospital productivity. The RAND Journal of Economics, 44(3), 545-568.

[14] Kellermann, A. L., & Jones, S. S. (2013). What it will take to achieve the as-yet-unfulfilled promises of health information technology. Health affairs, 32(1), 63-68.

[15] Cresswell, K. M., & Sheikh, A. (2015). Health information technology in hospitals: current issues and future trends. Future Hospital Journal, 2(1), 50-56.

[16] Ketikidis, P., Dimitrovski, T., Lazuras, L., & Bath, P. A. (2012). Acceptance of health information technology in health professionals: An application of the revised technology acceptance model. Health informatics journal, 18(2), 124-134.

[17] Payne, T. H. (2000). Computer decision support systems. Chest, 118(2), 47S-52S.

[18] Lee, H. W., Ramayah, T., & Zakaria, N. (2012). External factors in hospital information system (HIS) adoption model: a case on malaysia. Journal of medical systems, 36(4), 2129-2140.

[19] Chen, R. F., & Hsiao, J. L. (2012). An investigation on physicians' acceptance of hospital information systems: a case study. International journal of medical informatics, 81(12), 810-820.

[20] Peixoto, H., Santos, M., Abelha, A., & Machado, J. (2012). Intelligence in Interoperability with AIDA. In International Symposium on Methodologies for Intelligent Systems (pp. 264-273). Springer, Berlin, Heidelberg.

[21] J. Machado, A. Abelha, J. Neves, and M. Santos, "Ambient Intelligence in Medicine" in IEEE 2006 Biomedical Circuits and Systems Conference Healthcare Technology, BioCAS 2006, pp. 94–97, 2006.

[22] Gomes, P., Paiva, N., & Simões, B. (2009). Análise da viabilidade económica das aplicações SAM e SAPE. Lisbon, Portugal: Universidade Nova.

[23] Aggelidis, V. P., & Chatzoglou, P. D. (2012). Hospital information systems: Measuring end user computing satisfaction (EUCS). Journal of biomedical informatics, 45(3), 566-579.

[24] Lu, C. H., Hsiao, J. L., & Chen, R. F. (2012). Factors determining nurse acceptance of hospital information systems. CIN: Computers, Informatics, Nursing, 30(5), 257-264.

[25] Kushniruk, A. W., Bates, D. W., Bainbridge, M., Househ, M. S., & Borycki, E. M. (2013). National efforts to improve health information system safety in Canada, the United States of America and England. International journal of medical informatics, 82(5), e149-e160.

[26] Cardoso, L., Marins, F., Portela, F., Abelha, A., & Machado, J. (2014). Healthcare interoperability through intelligent agent technology. Procedia Technology, 16, 1334-1341.

[27] Cardoso, L., Marins, F., Portela, F., Santos, M., Abelha, A., & Machado, J. (2014). The next generation of interoperability agents in healthcare. International journal of environmental research and public health, 11(5), 5349-5371.

[28] Abelha, A., Analide, C., Machado, J., Neves, J., Santos, M., & Novais, P. (2007, September). Ambient intelligence and simulation in health care virtual scenarios. In Working Conference on Virtual Enterprises (pp. 461-468). Springer, Boston, MA.

[29] Arnott, D., & Pervan, G. (2016). A critical analysis of decision support systems research revisited: the rise of design science. In Enacting Research Methods in Information Systems (pp. 43-103). Palgrave Macmillan, Cham.

[30] Pestotnik, S. L., Classen, D. C., Evans, R. S., & Burke, J. P. (1996). Implementing antibiotic practice guidelines through computer-assisted decision support: clinical and financial outcomes. Annals of internal medicine, 124(10), 884-890.

[31] Musen, M. A., Middleton, B., & Greenes, R. A. (2014). Clinical decision-support systems. In Biomedical informatics (pp. 643-674). Springer, London.

[32] Butler, C. E., Noel, S., Hibbs, S. P., Miles, D., Staves, J., Mohaghegh, P., ... & Murphy, M. F. (2015). Implementation of a clinical decision support system improves compliance with restrictive transfusion policies in hematology patients. Transfusion, 55(8), 1964-1971.

[33] Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., ... & Suh, K. S. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. Journal of clinical bioinformatics, 5(1), 4.

[34] Turban, E., Sharda, R., & Delen, D. (2010). Decision Support and Business Intelligence Systems (required). Google Scholar.

[35] Bonney, W. (2013). Applicability of business intelligence in electronic health record. Procedia-Social and Behavioral Sciences, 73, 257-262.

[36] Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. Communications of the ACM, 54(8), 88-98.

[37] Santos, M. Y., & Ramos, I. (2006). Business Intelligence: Tecnologias da informação na gestão de conhecimento. FCA-Editora de Informática, Lda.

[38] Mettler, T., & Vimarlund, V. (2009). Understanding business intelligence in the context of healthcare. Health informatics journal, 15(3), 254-264.

[39] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

[40] Hočevar, B., & Jaklič, J. (2010). Assessing benefits of business intelligence systems–a case study. Management: journal of contemporary management issues, 15(1), 87-119.

[41] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[42] Al-Debei, M. M., & Avison, D. (2010). Developing a unified framework of the business model concept. European Journal of Information Systems, 19(3), 359-376.

[43] Santos, V., & Belo, O. (2013). Modeling ETL data quality enforcement tasks using relational algebra operators. Procedia Technology, 9, 442-450.

[44] Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). The data warehouse lifecycle toolkit. John Wiley & Sons.

[45] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

[46] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[47] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[48] Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

[49] Maimon, O. Z., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook (2 ed.). New York, USA: Springer Science Business Media, Inc.

[50] Colak, I., Sagiroglu, S., & Yesilbudak, M. (2012). Data mining and wind power prediction: A literature review. Renewable Energy, 46, 241-247.

[51] Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance.

[52] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. USA: SPSS Inc.

[53] Vercellis, C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. Politecnico di Milano, Italy: A John Wiley and Sons, Ltd., Publication.

[54] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121–167, 1998.

[55] Wei-Yin Loh and Yu-Shan Shih. Split selection methods for classification trees. Statistica sinica, 7(4):815–840, 1997.

[56] Oztekin, A., Delen, D., & Kong, Z. J. (2009). Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. International journal of medical informatics, 78(12), e84-e96.

[57] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[58] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees (Wadsworth, Belmont, CA). ISBN-13, 978-0412048418.

[59] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large databases, VLDB (Vol. 1215, pp. 487-499).

[60] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[61] Osuna, E., Freund, R., & Girosi, F. (1997, June). Training support vector machines: an application to face detection. In cvpr (Vol. 97, No. 130-136, p. 99).

[62] Bramer, M. (2007). Principles of data mining (Vol. 180). London: Springer.

[63] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

[64] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

[65] Ling, C. X., & Sheng, V. S. Cost-Sensitive Learning and the Class Imbalance Problem. 2011. Encyclopedia of Machine Learning: Springer.

[66] Elkan, C. (2001, August). The foundations of cost-sensitive learning. In International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.

[67] Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review.

[68] He, H., & Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, (9), 1263-1284.

[69] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

[70] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[71] Jan Larsen and Cyril Goutte. On optimal data split for generalization estimation and model selection. In Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop., pages 225–234. IEEE, 1999.

[72] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

[73] Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical chemistry, 39(4):561–577, 1993.

[74] Tom Fawcett. An introduction to ROC analysis. Pattern recognition letters, 27(8):861–874, 2006.

[75] James A Hanley. Characteristic (ROC) curvel. Radiology, 743:29–36, 1982.

[76] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

[77] Von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 28(1), 75-105.

[78] March, S. T., & Storey, V. C. (2008). Design science in the information systems discipline: an introduction to the special issue on design science research. MIS quarterly, 32(4), 725-730.

[79] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

[80] March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision support systems, 15(4), 251-266.

[81] Sheng, Q. Z., Qiao, X., Vasilakos, A. V., Szabo, C., Bourne, S., & Xu, X. (2014). Web services composition: A decade's overview. Information Sciences, 280, 218-238.

[82] Fielding, R. T., & Taylor, R. N. (2000). Architectural styles and the design of network-based software architectures (Vol. 7). Doctoral dissertation: University of California, Irvine.

[83] Rodriguez, A. (2008). Restful web services: The basics. IBM developerWorks, 33, 18.

[84] Belqasmi, F., Glitho, R., & Chunyan, F. (2011). RESTful web services for service provisioning in next generation networks: a survey. IEEE Communications Magazine.

[85] Jadhav, M. A., Sawant, B. R., & Deshmukh, A. (2015). Single page application using angularjs. International Journal of Computer Science and Information Technologies, 6(3), 2876-2879.

[86] Seshadri, S., & Green, B. (2014). AngularJS: Up and Running: Enhanced Productivity with Structured Web Apps. " O'Reilly Media, Inc.".

[87] Williamson, K. (2015). Learning AngularJS: A Guide to AngularJS Development. " O'Reilly Media, Inc.".

[88] Seshadri, S., & Green, B. (2014). AngularJS: Up and Running: Enhanced Productivity with Structured Web Apps. " O'Reilly Media, Inc.".

[89] Pereira, S., Portela, F., Santos, M. F., Machado, J., & Abelha, A. (2015, September). Predicting Preterm Birth in Maternity Care by Means of Data Mining. In Portuguese Conference on Artificial Intelligence (pp. 116-121). Springer, Cham.

[90] Data-driven documents. Retrieved September 2, 2018 from *https://d3js.org/*.

[91] Nvd3. Retrieved September 3, 2018 from: *http://nvd3.org/*.

[92] Maia, I. (2015). Building Web Applications with Flask. Packt Publishing Ltd.

[93] Grinberg, M. (2018). Flask web development: developing web applications with python. " O'Reilly Media, Inc.".

[94] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). Citeseer.

[95] Gonçalves, J. M., Portela, F., Santos, M. F., Silva, Á., Machado, J., Abelha, A., & Rua, F. (2014). Real-time Predictive Analytics for Sepsis Level and Therapeutic Plans in Intensive Care Medicine. International Journal of Healthcare Information Systems and Informatics (IJHISI), 9(3), 36-54.

[96] Schmidt, B. (2006). Proof of principle studies. Epilepsy research, 68(1), 48-52.

[97] Hay, G. J., & Castilla, G. (2006, July). Object-based image analysis: strengths, weaknesses, opportunities and threats (SWOT). In Proc. 1st Int. Conf. OBIA (pp. 4-5).

[98] Ghazinoory, S., Abdi, M., & Azadegan-Mehr, M. (2011). SWOT methodology: a state-of-the-art review for the past, a framework for the future. Journal of business economics and management, 12(1), 24-48.

[99] Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.

[100] Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation, 115(5), 654-657.

[101] Dag, A., Oztekin, A., Yucel, A., Bulur, S., & Megahed, F. M. (2017). Predicting heart transplantation outcomes through data analytics. Decision Support Systems, 94, 42-52.

[102] Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. Decision Support Systems, 74, 150-161.

[103] Lu, H. Y., Li, T. C., Tu, Y. K., Tsai, J. C., Lai, H. S., & Kuo, L. T. (2015). Predicting long-term outcome after traumatic brain injury using repeated measurements of Glasgow Coma Scale and data mining methods. Journal of medical systems, 39(2), 14.

[104] Yoo, K. D., Noh, J., Lee, H., Kim, D. K., Lim, C. S., Kim, Y. H., ... & Kim, Y. S. (2017). A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study. Scientific reports, 7(1), 8904.

[105] Gandhi, M., & Singh, S. N. (2015, February). Predictions in heart disease using techniques of data mining. In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE) (pp. 520-525). IEEE.

[106] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[107] Rahman, M. M., & Davis, D. (2013, July). Cluster based under-sampling for unbalanced cardiovascular data. In Proceedings of the World Congress on Engineering (Vol. 3, pp. 3-5).

[108] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1), 185-197.

[109] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003, September). SMOTEBoost: Improving prediction of the minority class in boosting. In European conference on principles of data mining and knowledge discovery (pp. 107-119). Springer, Berlin, Heidelberg.

[110] Fernandes, B. D. P. (2016). Real-time healthcare intelligence in organ transplantation (Master dissertation, Universidade do Minho).

[111] Torres, L. P. P. (2015). Plataforma de suporte à decisão médica para transplante de órgãos baseada em Business Intelligence (Master dissertation, Universidade do Minho).

[112] Walker, J. D., & Chapra, S. C. (2014). A client-side web application for interactive environmental simulation modeling. Environmental Modelling & Software, 55, 49-60.

[113] Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. IEEE transactions on visualization and computer graphics, 17(12), 2301-2309.

[114] Bootstrap Notify. Retrieved November 27, 2018 from: *http://bootstrap-notify.remabledesigns.com/*

ANEXO A – GLOSSÁRIO

Arquitetura cliente-servidor é uma arquitetura Web para aplicações de rede, que particiona tarefas pelos servidores e pelos que solicitam determinado conteúdo ou serviço, os clientes.

Business Intelligence é um processo tecnológico que tem como objetivo analisar grandes quantidades de dados não estruturados, proporcionando uma forma de obter informação útil e apresentável em relatórios e *dashboards*.

Dashboard consiste na apresentação visual dos dados, de forma a que os utilizadores tenham um conhecimento abrangente e rápido sobre informação relevante.

Data Mart é um subconjunto de dados de um *Data Warehouse*, atendendo necessidades específicas do negócio.

Data Mining é um processo computacional que permite identificar padrões em grandes conjuntos de dados, utilizando variados métodos de sistemas de aprendizagem e extração de conhecimento.

Escala de coma de Glasgow consiste numa escala neurológica que permite medir/avaliar o nível de consciência de uma pessoa que tenha sofrido um traumatismo crânio-encefálico.

Framework é uma estrutura de suporte definida para auxiliar a organização e o desenvolvimento de um projeto de software, podendo incluir programas de suporte, bibliotecas de código, *scripts* ou outros *softwares* para auxiliar o desenvolvimento de um projeto.

Indicadores correspondem a uma forma de simplificar e sintetizar fenómenos complexos através da sua quantificação (parâmetro, medida ou valor).

Matriz de confusão de uma determinada hipótese oferece uma medida efetiva de um modelo de *Data Mining*, ao mostrar o número de classificações corretas *versus* as classificações previstas para cada classe, sobre um conjunto de exemplos.

Morte cerebral é definida como a cessação irreversível das funções de todas as estruturas neurológicas intracranianas de um paciente.

Processo Clínico Eletrónico consiste na recolha e no registo de um conjunto de dados clínicos de um paciente, nomeadamente informações sobre o seu estado de saúde, histórico clínico, exames realizados, cirurgias, entre outras, em formato digital.

Query é um pedido de informação a uma base de dados ou conjunto de tabelas.

User-friendly define aplicações, ferramentas, sistemas ou processos que sejam intuitivos e de fácil utilização e interação.

ANEXO B – PUBLICAÇÕES

B.1 Business Intelligence for Nutrition Therapy

Autores: Rita Reis, Ana Mendonça, Diana Lisandra Azevedo, Hugo Peixoto e José Machado

Livro/Editora: Next-Generation Mobile and Pervasive Healthcare Solutions, pp. 203-2018, IGI Global

Ano: 2018

***Abstract**:* The assessment of health status in communities throughout the world is a massive information technology challenge. Data warehousing provides a flexible environment to support the business management and serve as an integrated repository for data. With the addition of models and analytic tools that have the potential to provide actionable information resources and support effective problem identification, critical decision-making, and strategy formulation, implementation, and evaluation. Of particular interest are the factors of influence like the patient's height or weight and its impact on processes and results. A multidimensional process is a way to discover health care processes according to certain factors of influence. This study aims to implement a data warehousing environment for decision support, in the context of nutrition evaluation, to integrate data obtained from a health care facility. This paper highlights the implementation of Business Intelligence in health care settings allows searching and interpreting stored information to support decisions concerning people's life.

B.2 Machine Learning in Nutritional Follow-up Research

Autores: Rita Reis, Hugo Peixoto, José Machado e António Abelha

Livro/Editora: Open Computer Science, Volume 7 (1), pp. 41-45, De Gruyter Academic Publishing

Ano: 2017

***Abstract**:* Healthcare is one of the world's fastest growing industries, having large volumes of data collected on a daily basis. It is generally perceived as being 'information

rich' yet 'knowledge poor'. Hidden relationships and valuable knowledge can be discovered in the collected data from the application of data mining techniques. These techniques are being increasingly implemented in healthcare organizations in order to respond to the needs of doctors in their daily decision-making activities. To help the decision-makers to take the best decision it is fundamental to develop a solution able to predict events before their occurrence. The aim of this project was to predict if a patient would need to be followed by a nutrition specialist, by combining a nutritional dataset with data mining classification techniques, using WEKA machine learning tools. The achieved results showed to be very promising, presenting accuracy around 91%, specificity around 97% and precision about 95%.

B.3 A Case-Based Reasoning Approach to GBM Evolution

Autores: Ana Mendonça, Joana Pereira, Rita Reis, Victor Alves, António Abelha, Filipa Ferraz, João Neves, Jorge Ribeiro, Henrique Vicente, José Neves

Conferência: International Conference on Computational Collective Intelligence (ICCCI 2018)

Livro/Editora: Computational Collective Intelligence, pp 489-498, Springer International Publishing

Ano: 2018

***Abstract**: GlioBastoma Multiforme* (GBM) is an aggressive primary brain tumor characterized by a heterogeneous cell population that is genetically unstable and resistant to chemotherapy. Indeed, despite advances in medicine, patients diagnosed with GBM have a median survival of just one year. *Magnetic Resonance Imaging* (*MRI*) is the most widely used imaging technique for determining the location and size of brain tumors. Indisputably, this technique plays a major role in the diagnosis, treatment planning, and prognosis of GBM. Therefore, this study proposes a new *Case Based Reasoning* approach to problem solving that attempts to predict a patient's GBM volume after five months of treatment based on features extracted from MR images and patient attributes such as age, gender, and type of treatment.