



Universidade do Minho
Escola de Engenharia

Lisete Inês Queiroz Ribeiro

Categorização e Determinação dos Fatores de Risco de Doenças Coronárias

Dissertação de Mestrado

Mestrado integrado em Engenharia e Gestão de Sistemas de
Informação

Trabalho efetuado sob a orientação do

Professor Doutor Manuel Filipe Santos e

Professor Doutor Carlos Filipe Portela

Outubro 2018

Agradecimentos

O desenvolvimento desta dissertação foram momentos definitivos no meu percurso académico, tendo sido constituída por diversos altos e baixos que foram ultrapassados, não só por mim, mas principalmente pelas pessoas que me rodeavam.

Em primeiro lugar um agradecimento especial aos meus orientadores por todo suporte, paciência e principalmente a disponibilidade ao longo deste longo processo.

Aos meus pais que foram os principais formadores do meu ensino, que trabalharam muito e se esforçaram ainda mais para eu conseguir a formação num curso extraordinário.

Aos meus amigos, os verdadeiros, que sabem perfeitamente quem são, que me acompanharam desde do primeiro dia até agora, nesta jornada e que nunca me deixaram desistir. Um muito obrigada por todos os momentos memoráveis que me proporcionaram.

Ao meu namorado e melhor amigo que nunca me deixou desistir em momento algum, tendo sido o meu porto de abrigo para a realização desta dissertação, assim como todo o meu percurso académico.

Por fim, mas não menos importante, quero deixar um último agradecimento à Universidade do Minho e ao Departamento de Sistemas de Informação pela instituição fantástica que me permitiu crescer a nível pessoal e académico.

Resumo

As doenças coronárias, em particular os enfartes do miocárdio, são uma das principais causas de morte nos países desenvolvidos. Estas doenças contribuem para a perda de qualidade de vida em todo o mundo, sendo que a sua deteção precoce é muito importante para que se possa fazer um tratamento adequado. Isto, com o objetivo de reduzir a taxa de mortalidade em todo o mundo.

Este projeto está dividido em duas fases. Na primeira fase foram analisadas as doenças coronárias, bem como os fatores de risco associados. Para além disso foram revistos alguns conceitos importantes, como *Data Mining* e descoberta de conhecimento em base de dados. Nesta primeira fase foi desenvolvido o planeamento, assim como definidos os objetivos principais e estruturantes.

A segunda fase concentrou-se em dados fornecidos pelo projeto Deus ex Machine. Inicialmente estava previsto que os dados seriam provenientes do Centro Hospitalar do Porto e do Hospital da Senhora da Oliveira de Guimarães, no entanto, e como não foi possível a obtenção dos dados, estes foram provenientes do projeto DEM. Este projeto desenvolveu um conjunto de folhas de cálculo com toda a informação registada relacionados com a atividade física. Depois de analisados e tratados estes dados através de técnicas de *Data Mining*, é possível perceber como é que o exercício físico ajuda na prevenção de doenças cardiovasculares, tendo como base as alterações dos batimentos cardíacos.

Este projeto baseou-se na metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), sendo que está dividido em seis fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e implementação.

Foi desenvolvido um modelo de *Data Mining* capaz de perceber se o exercício físico poderá ajudar na prevenção de doenças cardiovasculares, tendo como base as alterações dos batimentos cardíacos. Este modelo obteve resultados superiores ao *threshold* definido de 85%.

Palavras chave: Doenças Coronárias, *Data Mining*, CRISP-DM.

Abstract

Heart diseases, in particular myocardial infarction, are one of the main causes of death in developed countries. These diseases contribute for a lower quality of life universally. Early detection is very important so an appropriate treatment can be made, aiming to reduce the mortality rate worldwide.

This project is divided in two phases. On the first phase, a research was conducted on heart diseases, as well as the associated risk factors. Furthermore, a few concepts were reviewed, such as Data Mining and knowledge discovery in databases. In this phase, the project plan was developed and the main goals and structuring objectives were defined.

The second phase concentrated on data, provided by the provided by the Deus ex Machine project. Initially it was foreseen that the data would come from the Hospital Center of Porto and the Hospital of Senhora da Oliveira of Guimarães, however, and since it was not possible to obtain the data, these came from the DEM project. This project developed a set of spreadsheets with all the information recorded related to physical activity. After analyzing and treating these data through Data Mining techniques, it is possible to see how physical exercise helps in the prevention of cardiovascular diseases, based on changes in heart rate.

This project was based on the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM is divided in six phases: business understanding, data understanding, data preparation, modeling, evaluation and implementation.

A data mining model was developed to understand whether physical exercise can help prevent cardiovascular diseases, based on changes in heart rate. This model obtained results superior to threshold defined of 85%, in most of the defined scenarios

Keywords: *heart diseases, Data Mining, CRISP-DM.*

1.	Enquadramento do trabalho proposto	1
1.1.	Introdução e Motivação	1
1.2.	Enquadramento	1
1.3.	Objetivos.....	3
1.4.	Estrutura do Documento	4
2.	Revisão de Literatura	5
2.1.	Doenças Coronárias.....	2
2.1.1.	Exemplos de doenças coronárias	2
2.1.2.	Fatores de risco	8
2.1.3.	Alguns dados estatísticos, em Portugal.....	13
2.2.	Descoberta de Conhecimentos em Base de Dados (DCBD).....	17
2.2.1.	Definição	17
2.2.2.	Fases do Processo DCBD	19
2.3.	<i>Data Mining</i>	22
2.3.1.	Definição	22
2.3.2.	Vantagens	24
2.3.3.	Desvantagens	25
2.3.4.	Ciclo de Vida	25
2.3.5.	Abordagens <i>Data Mining</i>	25
2.4.	Trabalhos Relacionados	36
3.	Abordagem Metodológica e Ferramentas Utilizadas.....	41
3.1.	Abordagens Metodológicas.....	41
3.1.1.	Design Science Research (DSR)	41

3.1.2.	Cross Industry Standard Process for Data Mining (CRISP-DM)	44
3.2.	Ferramentas Utilizadas	51
3.2.1.	Ferramentas de Tratamentos dos Dados.....	51
3.2.2.	Ferramentas de Data Mining	52
4.	Desenvolvimento	54
4.1.	Compreensão do problema.....	54
4.2.	Compreensão dos Dados.....	54
4.2.1.	Atributos de cada um dos ficheiros	57
4.2.2.	Análise Exploratória	63
4.3.	Preparação dos Dados.....	75
4.4.	Modelação	79
5.	Discussão dos Resultados.....	86
6.	Conclusão	88
6.1.	Síntese do Trabalho Realizado	88
6.2.	Riscos Verificados.....	89
6.3.	Trabalho futuro.....	90
	Referências Bibliográficas.....	91
	Anexo I – Trabalhos Relacionados	98
	Anexo II – Matrizes de Confusão da Abordagem 1	108
	Decision Tree, DT	108
	<i>Support Machine Vector, SVM</i>	109
	Naive Bayes, NB	110
	Anexo III - Matrizes de Confusão da Abordagem 1, Método Sampling	111
	<i>Decision Tree</i> , DT	111
	<i>Support Decison Vector</i> , SVM	112
	<i>Naive Bayes</i> , NB	113

Anexo IV - Matrizes de Confusão da Abordagem 2	114
<i>Decision Tree</i> , DT	114
<i>Support Decison Vector</i> , SVM	115
<i>Naive Bayes</i> , NB.....	116

Lista de Figuras

Figura 1 - Tipos de Doenças Coronárias	3
Figura 2 - Número de Óbitos por Doenças Isquémicas Cardíaca, em Portugal (2009-2013).	15
Figura 3 - Número de Óbitos por Enfarte Agudo do Miocárdio, em Portugal (2009-2013).	16
Figura 4 - Processo DCBD (Oded Maimon, 2005, p. 3).....	19
Figura 5 - Árvore de decisão: jogar badminton	27
Figura 6 - Exemplo de uma Rede Neuronal Artificial.....	27
Figura 7 - Separação Linear de duas Classes	29
Figura 8 - Fases da Metodologia Design Science Research	41
Figura 9 - Etapas da metodologia CRISP-DM	45
Figura 10 - Subfases da Fase de Compreensão do Negócio do CRISP-DM.....	46
Figura 11 - Subfase da Fase de Compreensão dos Dados do CRISP-DM	47
Figura 12 - Subfase da Fase de Preparação dos Dados do CRISP-DM	48
Figura 13 - Subfase da Fase de Modelação do CRISP-DM	48
Figura 14 - Subfase da Fase de Avaliação do CRISP-DM	49
Figura 15 - Subfases da Fase de Implementação do CRISP-DM	50
Figura 16 - Logotipo Talend	51
Figura 17 - Logotipo Weka	52
Figura 18 - Esquema de Ficheiros.....	55
Figura 19 - ETL de todas as Uniões.....	56
Figura 20 - ETL BP_UNION.....	57
Figura 21 - Estrutura do ficheiro Blood Pressure inicial e final	57
Figura 22 - Gráfico Global Pressão Arterial	64
Figura 23 - Gráfico Evolução Temporal da Pressão Arterial	65
Figura 24 - Gráfico Global Pedómetro.....	66
Figura 25 - Gráfico Evolução Temporal Pedómetro	67
Figura 26 - Gráfico Global Batimentos Cardíacos	68
Figura 27 - Gráfico Evolução Temporal Batimentos Cardíacos Diário.....	69
Figura 28 - Gráfico Global Batimentos Cardíacos Horários	70
Figura 29 - Gráfico Evolução Temporal Batimentos Cardíacos Horários.....	71
Figura 30 - Gráfico Global Oxímetro.....	71
Figura 31 - Gráfico Evolução Temporal Oxímetro	72

Figura 32 - Gráfico Global de Peso.....	73
Figura 33 - Gráfico Evolução Temporal de Peso	74
Figura 34 - ETL Média Batimentos Cardíacos.....	75
Figura 35 - Tarefa (job) de todas as integrações.....	76
Figura 36 - Tarefa Eliminar Linhas Repetidas	76
Figura 37 - Ficheiro Inicial e Final após transformações	77
Figura 38 - Matriz de Confusão C1 Abordagem 1	81
Figura 39 - Matriz de Confusão C1, Abordagem 1, Método <i>Sampling</i>	81
Figura 40 - Matriz de Confusão C1, Abordagem 2.....	82
Figura 41 - Matriz de confusão C1, DT, Abordagem 1	108
Figura 42 - Matriz de confusão C2, DT, Abordagem 1	108
Figura 43 - Matriz de confusão C3, DT, Abordagem 1	108
Figura 44 - Matriz de confusão C4, DT, Abordagem 1	108
Figura 45 - Matriz de confusão C5, DT, Abordagem 1	108
Figura 46 - Matriz de confusão C1, SVM, Abordagem 1.....	109
Figura 47 - Matriz de confusão C2, SVM, Abordagem 1.....	109
Figura 48 - Matriz de confusão C3, SVM, Abordagem 1.....	109
Figura 49 - Matriz de confusão C4, SVM, Abordagem 1.....	109
Figura 50 - Matriz de confusão C5, SVM, Abordagem 1.....	109
Figura 51 - Matriz de confusão C1, NB, Abordagem 1.....	110
Figura 52 - Matriz de confusão C2, NB, Abordagem 1.....	110
Figura 53 - Matriz de confusão C3, NB, Abordagem 1.....	110
Figura 54 - Matriz de confusão C4, NB, Abordagem 1.....	110
Figura 55 - - Matriz de confusão C5, NB, Abordagem 1	110
Figura 56 - Matriz de confusão C1, DT, Abordagem 1, Método Sampling.....	111
Figura 57 - Matriz de confusão C2, DT, Abordagem 1, Método Sampling.....	111
Figura 58 - Matriz de confusão C3, DT, Abordagem 1, Método Sampling.....	111
Figura 59 - Matriz de confusão C4, DT, Abordagem 1, Método Sampling.....	111
Figura 60 - Matriz de confusão C5, DT, Abordagem 1, Método Sampling.....	111
Figura 61 - Matriz de confusão C1, SVM, Abordagem 1, Método Sampling	112
Figura 62 - Matriz de confusão C2, SVM, Abordagem 1, Método Sampling	112
Figura 63 - Matriz de confusão C3, SVM, Abordagem 1, Método Sampling	112

Figura 64 - Matriz de confusão C4, SVM, Abordagem 1, Método Sampling.....	112
Figura 65 - - Matriz de confusão C5, SVM, Abordagem 1, Método Sampling.....	112
Figura 66 - Matriz de confusão C1, NB, Abordagem 1, Método Sampling.....	113
Figura 67 - Matriz de confusão C2, NB, Abordagem 1, Método Sampling.....	113
Figura 68 - Matriz de confusão C3, NB, Abordagem 1, Método Sampling.....	113
Figura 69 - Matriz de confusão C4, NB, Abordagem 1, Método Sampling.....	113
Figura 70 - Matriz de confusão C5, NB, Abordagem 1, Método Sampling.....	113
Figura 71 - Matriz de confusão C1, DT, Abordagem 2.....	114
Figura 72 - Matriz de confusão C2, DT, Abordagem 2.....	114
Figura 73 - Matriz de confusão C3, DT, Abordagem 2.....	114
Figura 74 - Matriz de confusão C4, DT, Abordagem 2.....	114
Figura 75 - Matriz de confusão C5, DT, Abordagem 2.....	115
Figura 76 - Matriz de confusão C1, SVM, Abordagem 2.....	115
Figura 77 - Matriz de confusão C2, SVM, Abordagem 2.....	115
Figura 78 - Matriz de confusão C3, SVM, Abordagem 2.....	115
Figura 79 - Matriz de confusão C4, SVM, Abordagem 2.....	116
Figura 80 - Matriz de confusão C5, SVM, Abordagem 2.....	116
Figura 81 - Matriz de confusão C1, NB, Abordagem 2.....	116
Figura 82 - Matriz de confusão C2, NB, Abordagem 2.....	116
Figura 83 - Matriz de confusão C3, NB, Abordagem 2.....	117
Figura 84 - Matriz de confusão C4, NB, Abordagem 2.....	117
Figura 85 - Matriz de confusão C5, NB, Abordagem 2.....	117

Lista de Tabelas

Tabela 1 - Dados sobre as mortes por Doenças do Aparelho Circulatório, em 2014	14
Tabela 2 - Dados sobre as mortes por Doenças Isquémicas do Coração, em 2014.....	14
Tabela 3 - Dados sobre as mortes por Enfarte Agudo do Miocárdio, em 2014.....	16
Tabela 4 - Componentes de DCBD.....	18
Tabela 5 - Matriz de Confusão.....	29
Tabela 6 - Estudos Considerados. Adaptado: (Kadi et al., 2017)	37
Tabela 7 - Requisitos Weka	52
Tabela 8 - Atributos do ficheiro da Pressão Arterial	59
Tabela 9 - Atributos do ficheiro do Pedómetro	59
Tabela 10 - Atributos do ficheiro dos Batimentos Cardíacos Diários	60
Tabela 11 - Atributos do ficheiro dos Batimentos Cardíacos Horários	61
Tabela 12 - Atributos do ficheiro do Oxímetro	61
Tabela 13 - Atributos do ficheiro do Questionário.....	62
Tabela 14 - Atributos do ficheiro do Peso	63
Tabela 15 - Registo de Informação, Abordagem 1.....	83
Tabela 16 - Registo de Informação, Abordagem 1, Método Sampling.....	84
Tabela 17 - Registo de Informação, Abordagem 2.....	85
Tabela 18 - Riscos Cometidos.....	89
Tabela 19 - Trabalhos Relacionados. Adaptado:(Kadi et al., 2017)	98

Listas de Abreviaturas, Siglas e Acrónimos

AD – Árvores de decisão	MSE – <i>Mean Square Error</i>
AUC – <i>Area Under Curve</i>	OMS – Organização Mundial de Saúde
CHP – Centro Hospitalar do Porto	RAE – <i>Relative Absolute Error</i>
CRISP-DM – <i>Cross Industry Standard Process for Data Mining</i>	RNA – Redes Neurais Artificiais
DCBD – Descoberta de Conhecimentos em Base de Dados	ROC – <i>Receiver Operating Characteristic</i>
DM – <i>Data Mining</i>	ROCCH – <i>ROC Convex Hull</i>
DSR – <i>Design Science Research</i>	RSE – <i>Relative Square Error</i>
FP – Falsos Positivos	SI – <i>Sistemas de Informação</i>
FN – Falsos Negativos	SVM – <i>Support Vector Machine</i>
HSOG – Hospital da Senhora da Oliveira de Guimarães	TI – Tecnologias de Informação
KD – <i>Knowledge Discovery</i>	WHF – <i>World Heart Federation</i>
KDD – <i>Knowledge Discovery from Database</i>	WHO – <i>World Health Organization</i>
MAE – <i>Mean Absolute Error</i>	VP – Verdadeiros Positivos
	VN – Verdadeiros Negativos

1. Enquadramento do trabalho proposto

1.1. Introdução e Motivação

Diariamente, em cada hospital, são recolhidas grandes quantidades de dados que, normalmente, não são tratados. A partir desses dados, poderiam ser descobertas informações ocultas, que ajudariam numa tomada de decisão eficaz. A descoberta de padrões e relacionamentos, muitas vezes ocultos, não são explorados atempadamente, pelo o profissional de saúde (Palaniappan & Awang, 2008).

Um importante desafio para as organizações de saúde é a prestação de serviços de qualidade a custos acessíveis. Para uma prestação de serviços adequada, é necessário diagnosticar os doentes de forma correta e administrar os tratamentos corretos. Um mau diagnóstico, pode levar a consequências muito graves para o doente. As organizações de saúde podem também reduzir os custos, minimizando os testes realizados, até perceber o que o doente tem. Isto pode acontecer através do uso adequado dos sistemas de informação, nos computadores dos profissionais de saúde (Shouman, Turner, & Stocker, 2012).

A maioria dos hospitais, atualmente, tem um sistema de informação para gerir os dados dos doentes. Estes sistemas são projetados para suportar a faturação, a gestão de *stock* e a gestão de estatísticas simples. As decisões clínicas são, muitas vezes, feitas pela intuição do médico ou pela sua experiência, em vez da utilização do conhecimento, existentes nos dados (Oyyathevan & Askarunisa, 2014). Esta prática leva a preconceitos indesejados, erros e custos médicos excessivos que afetam a qualidade do serviço prestado aos doentes. Assim, com técnicas avançadas de *Data Mining*, esta situação pode deixar de acontecer. O *Data Mining* oferece um conjunto de estratégias que podem ser ligadas aos dados, de forma a ajudar os especialistas na tomada de decisão.

1.2. Enquadramento

O coração é uma parte importante do corpo humano, que ajuda a bombear e a purificar o sangue. A deficiência deste músculo pode ser motivo do desequilíbrio do cérebro, insuficiência renal e até mesmo a morte instantânea (Rajalakshmi & Nirmala, 2016, p. 1).

Hoje em dia, as doenças cardiovasculares, consequentemente as doenças coronárias, são a maior causa de morte, em todo o mundo (WHO, 2014). Esta doença ataca tão rapidamente que um doente não consegue ser tratado atempadamente. Este diagnóstico é algo muito difícil por parte dos médicos, pois um diagnóstico errado faz com que o hospital e, principalmente, o médico percam reputação, pondo em risco a vida de um doente.

O número de mortes causadas por doenças coronárias está previsto chegar a 23,3 milhões em 2030 (Miranda, Irwansyah, Amelga, Maribondang, & Salim, 2016, p. 1). Segundo a World Health Organization (2014), 17,5 milhões de pessoas morreram devido a doenças cardiovasculares no ano de 2012. Nos Estados Unidos, as doenças cardíacas matam uma pessoa em cada 34 segundos (Soni, Ansari, Sharma, & Soni, 2011).

O termo doença cardíaca ou doença coronária, inclui uma vasta gama de condições que afetam o coração, resultando em várias doenças, incapacidades e até mesmo em morte. Existem quatro grandes grupos de doenças cardíacas: as doenças reumáticas cardíacas (consequência a longo prazo da febre reumática aguda), as doenças isquémicas cardíacas (quando uma determinada substância se acumula nos vasos sanguíneos, impedindo o fluxo normal de sangue), as doenças inflamatórias cardíacas (quando existe uma inflamação no coração, causada por agentes infecciosos, como vírus, bactérias, fungos e parasitas) e outras (incluindo os defeitos cardíacos congénitos e a insuficiência cardíaca).

O número de mortes devido a doenças cardíacas tem tendência a aumentar e com uma previsão atempada estes números poderão baixar pois, os doentes são tratados de forma correta e rápida.

O exercício físico é uma das formas de combater os problemas cardíacos. Com a prática de exercício o coração habitua-se às atividades, fazendo com que os batimentos cardíacos sejam mais constantes ao longo do tempo.

As pessoas que fazem atividade física e não acompanham os valores da frequência cardíaca podem estar a realizar os exercícios de forma errada, o que pode prejudicar o rendimento e o objetivo do treino. Além de não alcançarem os objetivos previstos, estes podem ser acometidos por alterações cardíacas como braquicardias (frequências cardíacas muito baixas) e taquicardias (frequências cardíacas muito elevadas), ou até situações mais graves como o enfarte agudo do miocárdio.

1.3. Objetivos

Este projeto é a primeira parte de um estudo, sendo o principal objetivo determinar de que modo as alterações dos batimentos cardíacos provocadas pelo exercício físico, são essenciais na prevenção das doenças cardíacas.

O projeto estava previsto basear-se em dados reais fornecidos pelo Hospital de Santo António do Porto e pelo Hospital da Senhora da Oliveira de Guimarães (HSOG). No entanto, e como não foi possível a obtenção desses dados, este projeto baseou-se em dados fornecidos pelo projeto Deus Ex Machine (DEM).

Inicialmente os objetivos deste projeto eram a elaboração de um modelo de previsão de doenças coronárias, através da análise e tratamento de um conjunto de dados registados pelo Centro Hospitalar do Porto e pelo Hospital da Senhora da Oliveira de Guimarães e prever os fatores de riscos das doenças coronárias, através de técnicas de *Data Mining*. Como os dados tiveram que ser alterados, estes objetivos foram refinados, sendo que o principal objetivo desta dissertação é perceber de que forma o exercício físico ajuda na prevenção das doenças cardíacas tendo como base as alterações dos batimentos cardíacos. Outro objetivo desta dissertação é obter modelos de *Data Mining* capazes de suportar o processo de tomada de decisão.

A questão de investigação foi desenvolvida, de modo a enquadrar os objetivos deste projeto de investigação. Assim, a questão de investigação é a seguinte:

“De que modo o Data Mining pode contribuir para perceber de que modo o exercício físico ajuda na prevenção de doenças coronárias?”

De modo a satisfazer estes objetivos e a responder à questão de investigação foi desenvolvido um modelo de data mining capaz de perceber se o exercício físico poderá ajudar na prevenção de doenças cardiovasculares, tendo como base as alterações dos batimentos cardíacos.

Por fim, este projeto contribuirá para um novo conhecimento útil na área dos Sistemas de Informação aplicado às doenças coronárias.

1.4. Estrutura do Documento

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

- **Enquadramento do trabalho proposto:** neste capítulo será apresentada uma breve introdução e motivação sobre o projeto; será realizado um breve enquadramento e apresentada de forma sintetizada a estrutura do documento;
- **Revisão de Literatura:** neste capítulo será apresentada a revisão de literatura realizada neste projeto, onde serão abordadas as definições de doenças coronárias, de descoberta de conhecimento de base de dados e o conceito de *Data Mining*. Inicialmente, foi realizado um pequeno enquadramento sobre as doenças coronárias, foram dados alguns exemplos de fatores de risco e alguns dados estatísticos em Portugal. Posteriormente, foi dada a definição de descoberta de conhecimento em base de dados, assim como as fases do processo. De seguida, foi dada uma definição de *Data Mining*, foram dadas algumas vantagens e desvantagens, assim como o seu ciclo de vida e as suas abordagens;
- **Abordagem Metodológicas e Ferramentas Utilizadas:** serão apresentadas as duas metodologias que irão servir de suporte a esta dissertação, o CRISP-DM e a DSR (*Design Science Research*), explicando o que irá ser feito em cada uma das destas. Para além disso serão apresentadas as ferramentas que foram utilizadas no desenvolvimento desta dissertação
- **Desenvolvimento:** neste capítulo são apresentadas todas as fases de desenvolvimento desta dissertação, tendo em conta o CRISP-DM.
- **Conclusão:** neste capítulo será realizada uma pequena conclusão sobre esta dissertação, assim como o que será expectável fazer nas próximas fases de desenvolvimento;
- **Referências Bibliográficas:** neste capítulo estão apresentadas todas as referências utilizadas para a elaboração desta pré-dissertação.

2. Revisão de Literatura

Neste capítulo é elaborado um levantamento e uma análise dos trabalhos que já foram desenvolvidos, em termos práticos. Estes trabalhos são semelhantes ao que se pretende que seja desenvolvido, neste projeto de dissertação. Para além disso, foram revistos os conceitos de *Data Mining* e de Descoberta de Conhecimento em Base de Dados. Foram dados alguns exemplos de doenças coronárias, como as doenças reumáticas cardíacas, as doenças isquémicas cardíacas, as doenças inflamatórias cardíacas e entre outras.

Para o desenvolvimento deste capítulo, foram realizadas várias pesquisas em várias bases de dados e respetivos motores de busca, destacando-se os seguintes:

- ScienceDirect;
- Springer;
- IEEE Xplore;
- Scopus;
- Google Scholar;
- B-on;
- ResearchGate.

Para além destas bases de dados, foram utilizados outros web sites, destacando os seguintes:

- Fundação Portuguesa de Cardiologia (www.fpcardiologia.pt);
- World Heart Federation (<http://www.world-heart-federation.org>);
- World Health Organization (www.who.int).

Para efetuar a pesquisa, nas bases de dados referidas, foram utilizados vários termos, tais como:

- “Doenças coronárias” e “*Coronary Diseases*”;
- “Doenças cardíacas” e “*Cardiac Diseases*”;
- “Doenças do coração” e “*Heart Diseases*”;
- “Fatores de Risco de Doenças do Coração” e “*Risk Factors Heart Disease*”;
- “Descoberta de Conhecimento em Bases de Dados” e “*Knowledge Discovery in Databases*”;
- “*Data Mining*”;
- “Metodologias *Data Mining*” e “*Data Mining Methodology*”.

Para a escolha dos inúmeros artigos encontrados foram considerados fatores como a relevância do autor na área, o ano em que foram publicados, assim como a reputação dos artigos publicados.

2.1. Doenças Coronárias

O coração humano é do tamanho de um punho, sendo o músculo mais forte do corpo humano. O coração de um ser humano começa a bater ainda no útero, muito antes do nascimento. A média de batimentos cardíacos é de 100 000 vezes por dia, ou seja, 2,5 bilhões de vezes em 70 anos. Em cada batimento é bombeado sangue para todo o corpo, sendo transportado pelas artérias, chegando aos capilares, em todos os tecidos celulares. Depois de entregar o oxigênio e os nutrientes a todas as células, são recolhidos os resíduos, voltando novamente ao coração através das veias. No fígado, estes resíduos são removidos. Este sistema é muito vulnerável a falhas, relacionadas com vários fatores, muitos dos quais podem ser tratados e evitados (Mackay, Judith, Mensah, 2004, p. 18).

As doenças coronárias acontecem quando as funções do coração apresentam falhas, causadas pelo fumo, a dieta pouco saudável ou o stress diário¹. De seguida, são apresentadas algumas doenças que afetam o bom funcionamento do coração.

2.1.1. Exemplos de doenças coronárias

A World Heart Federation (WHF) é uma colaboração entre várias fundações do coração e sociedades médicas espalhadas por todo o mundo. A WHF é reconhecida pela Organização Mundial de Saúde (OMS) ² como uma organização não-governamental, parceira na prevenção das doenças cardiovasculares (Bo Remenyi, Carapetis, Wyber, Taubert, & Mayosi, 2013, p. 285).

Assim, de acordo com a WHF³, as doenças cardiovasculares são divididas em seis grandes grupos: doenças reumáticas, hipertensivas, isquémicas, cerebrovasculares, inflamatórias e outras. Contudo, estes seis grupos estão relacionados com as doenças do coração, dos vasos sanguíneos e do cérebro. Este projeto está relacionado com as doenças do coração. Assim, foi realizada uma seleção das doenças que afetam diretamente o coração. Dessa seleção, surgiu o esquema apresentado na Figura 1. Este esquema, apresenta quatro grupos das doenças cardiovasculares: as doenças reumáticas cardíacas, as doenças isquémicas cardíacas, as doenças inflamatórias cardíacas e outros tipos de doenças. Estas doenças afetam diretamente o coração.

¹ <http://www.world-heart-federation.org/cardiovascular-health/heart-disease/>

² A organização mundial de saúde foi fundada em 7 de abril de 1948. Em 1996, um acordo entre a UNESCO, a OMS e a WHF, promoveu a cooperação em diferentes projetos, relacionados com as doenças cardíacas. <http://www.world-heart-federation.org/about-us/history/>

³ <http://www.world-heart-federation.org/cardiovascular-health/heart-disease/different-heart-diseases/>

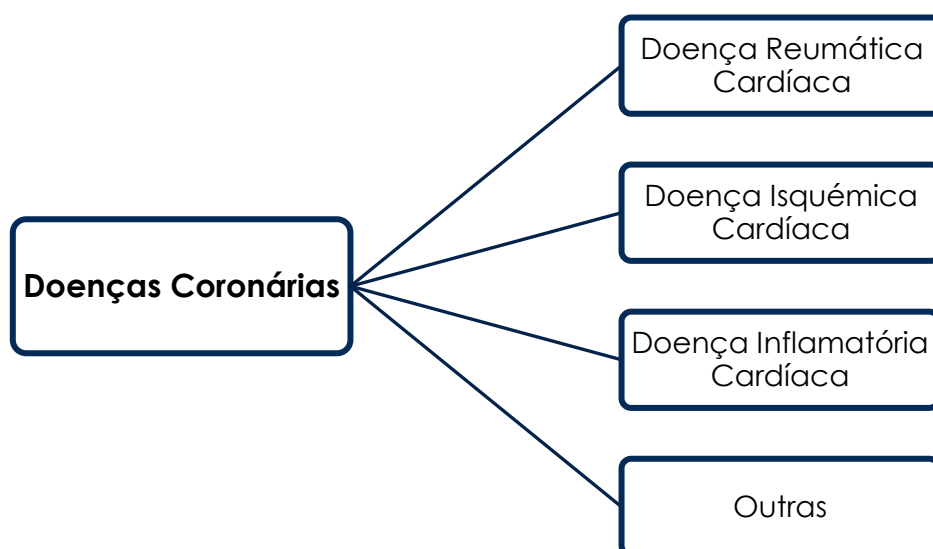


Figura 1 - Tipos de Doenças Coronárias

Para cada um destes conjuntos, existem outras doenças associadas que estão explicadas de seguida.

I. Doenças Reumáticas Cardíacas

As doenças reumáticas cardíacas são a única consequência, a longo prazo, da febre reumática aguda. Esta doença continua sem reduzir entre os países pouco desenvolvidos. Estima-se que, pelo menos, 15 milhões de pessoas estejam afetadas por esta doença em todo o mundo (B Remenyi et al., 2012, p. 297).

a. Doença Reumática Cardíaca

A doença reumática cardíaca é o resultado de um dano valvular, causado por uma resposta autoimune anormal a uma infeção estreptocócica do grupo A, num hospedeiro geneticamente suscetível (Marijon, Mirabel, Celermajer, & Jouven, 2012a, p. 953). As faringites, amigdalites, infeções no sangue, febres reumáticas, pneumonias e inflamações renais, são consideradas infeções estreptocócicas do grupo A⁴. Quando um hospedeiro é geneticamente suscetível significa que o hospedeiro não apresenta resistência a um determinado agente patogénico.

⁴ <http://www.manuaismsd.pt/?id=204&cn=1671>

Segundo Marijon *et al.* (2012b) a febre reumática aguda é um dos principais antecipadores desta doença. A doença reumática cardíaca é uma sequela da infecção causada pelos vírus estreptocócicos do grupo A, que não foi devidamente tratada, podendo afetar o coração.

A febre reumática tem a capacidade de causar fibroses nas válvulas do coração, levando à incapacidade de bombear o sangue, seguidamente da insuficiência cardíaca e, conseqüentemente, à morte (Marijon et al., 2012b).

b. Doença Valvular Cardíaca

No coração existem quatro válvulas que protegem a saída de sangue, das quatro cavidades que o coração possui. As válvulas fazem com que o sangue flua na direção correta e não corra para trás.

Uma válvula deteriorada danifica o fluxo de sangue de duas maneiras: se a válvula não abrir completamente ou ficar rígida pode obstruir a passagem do fluxo de sangue (dá-se o nome de válvula rígida); ou se a válvula não fechar adequadamente, fazendo com que o sangue venha em sentido oposto (dá-se o nome de regurgitação).

Tanto a válvula rígida como a regurgitação podem exercer uma pressão extra no coração. No primeiro caso, o coração é obrigado a bater mais fortemente, para forçar o sangue a entrar nas cavidades. No segundo caso, o coração tem que bater mais fortemente, para conseguir bombear sangue suficiente no sentido suposto contra o sangue que flui no sentido oposto, através da válvula danificada.

Dependendo da válvula afetada, existirá uma acumulação de fluidos de sangue nos pulmões, tornozelos ou pernas.

Os problemas das válvulas cardíacas podem ser congénitos. Isto acontece quando algum fator durante a evolução fetal leva a válvula a apresentar uma irregularidade. (Foundation, 2014)

As válvulas podem ser danificadas pela febre reumática, por certos medicamentos ou por tratamentos de radiação contra o cancro ⁵.

⁵ <http://www.world-heart-federation.org/cardiovascular-health/heart-disease/different-heart-diseases/>

II. Doenças Isquêmicas Cardíacas

Isquemia significa um reduzido fornecimento de sangue. Assim, a doença isquêmica cardíaca é quando uma determinada substância se acumula nos vasos sanguíneos, impedindo o fluxo normal de sangue. A doença isquêmica cardíaca é uma das principais causas de mortalidade, em todo o mundo. Um importante fator de risco para estas doenças é os níveis elevados de colesterol de baixa densidade (considerado o colesterol “mau”) (Varbo, Benn, & Nordestgaard, 2014, p. 359).

a. Angina

A angina é uma dor ou desconforto transitório no tórax. Esta dor ou desconforto ocorre quando o coração não recebe sangue e oxigênio suficiente. A angina, geralmente ocorre durante alguns minutos. A principal causa desta doença é a aterosclerose (explicado de seguida), que se define como o acumulo de substâncias gordas nas artérias coronárias. Se as artérias estão obstruídas, o sangue não consegue fluir normalmente, causando a dor ou desconforto no peito. Esta doença pode ser reduzida com medicamentos, mudança do estilo de vida ou procedimentos médicos (M.Altinbas, B.Kaplan, K.Ucar, M.Ozkan, E.Karahacioglu, O.Er, 2004, p. 1).

b. Aterosclerose cardíaca

A aterosclerose pode afetar qualquer artéria do corpo, incluindo as artérias do coração, cérebro, braços e pernas, podendo causar diferentes doenças, tendo como bases as artérias afetadas.

A aterosclerose cardíaca é o acumular de matérias gordas na parede das artérias do coração, como o colesterol. Ao longo do tempo, estas matérias gordas endurecem e estreitam as artérias, reduzindo o fluxo sanguíneo. Isto pode levar a ataques cardíacos ou mesmo a morte (National Heart Lung and Blood Institute, 2009).

c. Doença da artéria coronária

A doença da artéria coronária é causada pela aterosclerose, ou seja, é a acumulação de substâncias gordas nas paredes das artérias coronárias. As artérias coronárias são responsáveis por transportar o oxigênio e outros nutrientes ao miocárdio. Caso as artérias coronárias estejam obstruídas, a passagem de sangue para o miocárdio será menor, causando ataques cardíacos, por exemplo. Esta

doença é resultante do ambiente, da dieta, do estilo de vida ou da história familiar (Bhatia, 2010, p. 24).

Ao longo do tempo, esta doença pode enfraquecer o músculo do miocárdio e levar a insuficiências cardíacas e arritmias. A insuficiência cardíaca ocorre quando o coração não consegue bombear o sangue para todas as partes do corpo, e as arritmias ocorrem quando o ritmo de batimentos cardíacos tem valores diferentes dos normais (National Heart, Lung, 2009).

Uma arritmia é definida como um ritmo cardíaco anormal, ou seja, alterações importantes na atividade cardíaca normal. Quando o batimento cardíaco é muito rápido é chamado de taquicardia. Por outro lado, quando o batimento cardíaco é muito lento dá-se o nome de bradicardia (Portela et al., 2014, p. 1).

d. Ataque cardíaco

Como já referido anteriormente, o coração recebe oxigénio e nutrientes através das artérias coronárias. Quando este fluxo de sangue é cortado, a diminuição de oxigénio e nutrientes pode causar danos duradouros ao coração. Quando este bloqueio é súbito, damos o nome de ataque cardíaco. A principal causa para o aparecimento de um ataque cardíaco é a aterosclerose. (WHO, 2005, p. 8,9)

e. Morte Súbita Cardíaca

A morte súbita cardíaca é definida como sendo a morte inesperada dentro de uma hora, após os inícios dos sintomas, ou dentro de 24 horas, a partir do momento em que o doente foi observado pela última vez e não apresentava quaisquer sintomas. A morte súbita está entre as principais causas de morte, nos Estados Unidos, estando responsável por mais 50 % de todas as doenças cardiovasculares fatais. As principais causas da morte súbita cardíaca é a cardiomiopatia hipertrófica e anomalias congénitas da artéria coronária (Steinberg, Laksman, & Krahn, 2016, p. 710).

III. Doenças inflamatórias cardíacas

Uma inflamação no coração, geralmente, é causada por agentes infecciosos, como vírus, bactérias, fungos e parasitas, ou por materiais tóxicos do ambiente, como água, alimentos, ar, gases tóxicos

e poluição. Os sintomas desta doença incluem dor torácica, angina (explicada anteriormente), falta de ar, inchaço nos pés e tornozelos ou fadiga (Lu, Sun, Liu, Zheng, & Zhang, 2015, p. 851).

a. Miocardite

A miocardite é uma infecção que ocorre no músculo cardíaco, causado por um vírus ou por outra doença imune. O vírus ataca o músculo do coração e resulta na sua inflamação (Lu et al., 2015, p. 851).

b. Pericardite

A pericardite é a infecção que ocorre no pericárdio. O pericárdio é o saco que envolve o coração e fornece lubrificação para o mesmo. Este saco protege o coração dentro da cavidade torácica. Esta infecção pode ser causada por tumores, insuficiência renal, ataque cardíaco, e por vírus ou bactérias (Lu et al., 2015, p. 852).

c. Endocardite

A Endocardite é uma infecção no endocárdio. O endocárdio é o forro interno do coração. Geralmente, esta infecção ocorre a pessoas que sofrem degeneração das válvulas ou que possuam uma válvula artificial. Os sintomas desta infecção podem se desenvolver durante um longo período de tempo ou se manifestar de repente, dependendo do “defeito” que o coração possui (Lu et al., 2015, p. 852).

d. Cardiomiopatia

A cardiomiopatia é uma doença progressiva do músculo cardíaco, que resulta de anomalias estruturais e funcionais do miocárdio, na ausência de doenças cardíacas congénitas, doenças valvulares cardíacas, doenças nas artérias coronárias e hipertensão. As cardiomiopatias podem ser (Sahan, Sahan, Karamanliouglu, Gul, & Tufekciouglu, 2016, p. 504):

- Dilatadas: Aumento do ventrículo esquerdo e redução do desempenho do miocárdio;
- Restritivas: Paredes ventriculares rígidas, dificultando o enchimento das cavidades do coração;
- Hipertróficas: Paredes cardíacas espessas e aumento do desempenho cardíaco.

IV. Outras

a. Defeitos Cardíacos Congénitos

Os defeitos cardíacos congénitos são as malformações das estruturas cardíacas presentes nos nascimentos. Estas malformações podem ser causadas por (Mendis, Puska, & Norrving, 2011, p. 3):

- Uma relação estreita de sangue entre os pais (consanguinidade);
- Infecções maternas (como a rubéola);
- Uso materno de drogas e álcool;
- Nutrição materna deficiente.

Em muitos casos, a causa é desconhecida. Os defeitos cardíacos congénitos incluem orifícios no septo do coração, válvulas anormais e anormalidades nas câmaras do coração (Mendis et al., 2011, p. 3).

b. Insuficiência Cardíaca

A insuficiência cardíaca acontece quando o coração não consegue bombear, de forma eficaz, o sangue por todo o corpo. A insuficiência cardíaca leva a uma baixa irrigação sanguínea no cérebro o que leva a um baixo teor de oxigénio e danos nas células nervosas. Além disto, a insuficiência cardíaca aumenta o risco de trombose, devido à obstrução de um ou mais vasos sanguíneos (de Bruijn & Ikram, 2014, p. 2).

Mais de 2 % da população é afetada pela insuficiência cardíaca, sendo que a sua prevalência aumenta com a idade. Esta doença está a diminuir nos EUA e na maioria dos países Europeus. Isto deve-se à melhor prevenção e ao grande avanço no tratamento da insuficiência cardíaca, nos últimos anos (Metra, Carubelli, Ravera, & Stewart Coats, 2016, p. 766).

2.1.2. Fatores de risco

De acordo com a Fundação Portuguesa de Cardiologia⁶, existe um conjunto de fatores que se designam de fatores de risco. Alguns destes não podem ser modificados, como a hereditariedade,

⁶ <http://www.fpcardiologia.pt/>

o sexo e a idade. Outros, podem ser modificados com medidas de estilo de vida e medicamentos. Os fatores de risco das doenças cardiovasculares, consequentemente das doenças coronárias, são o sedentarismo, hipertensão, tabagismo, stress, obesidade, diabetes e dislipidemia, sendo explicados de seguida.

I. Sedentarismo

A inatividade física é um grande problema na Europa. As crianças tornam-se menos ativas fisicamente e só alguns países é que têm a dose diária recomendada de exercício físico. Quando deixam a escola, metade dos adolescentes tornam-se inativos. Com o avançar da idade, devido ao trabalho, as pessoas ficam ainda mais inativas. Em 2008, 31.3% dos adultos de 15 anos ou mais eram insuficientemente ativos (World Heart Federation, 2012, p. 2). Posto isto, um estilo de vida sedentário está associado com a duplicação de risco de morte prematura e um risco ainda maior de doenças cardíacas.

A atividade física tem vários benefícios para os doentes cardíacos, resultando numa diminuição da mortalidade em 20-30 %. Os países desenvolvidos têm o dobro de inatividade física (devido à automação do trabalho e utilização dos veículos para o transporte) do que os países não desenvolvidos, tanto para os homens como para as mulheres (World Heart Federation, 2012, p. 2).

Na Europa, uma pequena parte das pessoas com doenças do coração, participa em atividades físicas. Portugal, faz parte dos países com menores índices de atividade física da europa, sendo um país muito exposto a doenças cardíacas⁷.

Um quarto da população idosa sofre de doenças cardíacas. As alterações a nível mental e fisiológicas, que vêm com o aumento da idade, podem contribuir para a inatividade física. Se a atividade física for, desde sempre, regular, pode retardar as mudanças relacionadas á idade, melhorando o funcionamento físico e atrasando os sintomas da doença cardíaca (I. Graham et al., 2007, p. 2393).

II. Hipertensão

Quando o coração bombeia sangue para os vasos sanguíneos, esse sangue é empurrado contra as paredes dos vasos, criando pressão arterial. O corpo precisa de pressão arterial para mover o

⁷ <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/sedentarismo/>

sangue para todas as partes do corpo, através das artérias. As artérias saudáveis são elásticas, aumentando de tamanho, tendo em conta a dureza que o sangue é empurrado contra as paredes. Para que as artérias continuem saudáveis, é necessário que a pressão arterial esteja dentro de um intervalo saudável (Arbor pharmaceuticals, 2014, p. 2).

Hipertensão significa que a pressão nas artérias é maior do que aquilo que deveria ser. Não existe causas para a maioria dos casos de pressão arterial. Esta doença é chamada de “assassino silencioso” (World Heart Federation, 2012, p. 1) pois, geralmente, não tem sinais nem sintomas, sendo que pode causar ataques cardíacos (National Heart Lung and Blood, 2016).

A pressão arterial é medida como dois números, escritos um sobre o outro, registados em milímetros por mercúrio (exemplo: 120/68 mm Hg). O número superior é a pressão sistólica (a pressão nas artérias quando o sangue sai) e o número inferior é a pressão diastólica (pressão nas artérias quando o coração está relaxado entre batimentos). A pressão arterial é elevada quando a pressão sistólica é superior a 140 e a pressão diastólica é superior a 90 (World Heart Federation, 2012, p. 1).

De acordo com a World Heart Federation (2012, p. 1), cerca de 1 bilhão de pessoas tem hipertensão e estima-se que, em 2025, 1,56 bilhões de adultos tenham hipertensão. Afirma ainda que, a hipertensão é a principal causa das doenças cardiovasculares e que é a causa mais importante de morte prematura em todo o mundo, sendo que este problema está a aumentar significativamente.

Em Portugal, existem cerca de dois milhões de hipertensos, sendo que 50% destes sofrem desta patologia, 25% estão medicados e 11% têm a tensão controlada⁸.

III. Tabagismo

Fumar entre 2 a 3 cigarros por dia aumenta o risco de doenças cardiovasculares, e consequentemente as doenças cardíacas, diminuindo a esperança de vida em 3 anos (Mehta, Wei, & Wenger, 2015, p. 3). O risco adicional de contrair uma doença cardíaca é medido pelo número de cigarros fumados diariamente. As pessoas que fumam mais de 40 cigarros por dia, têm dez vezes mais probabilidade de ter um enfarte do miocárdio do que as que não fumam (Scottish Intercollegiate Guidelines Network, 2007, p. 16).

⁸ <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/hipertensao/>

Os efeitos do tabagismo, nas doenças cardíacas, interagem diretamente com outros fatores de risco, como a idade, sexo, hipertensão arterial e diabetes (de Bruijn & Ikram, 2014, p. 4). O tabagismo contribui para o agravamento da aterosclerose.

Os homens fumadores têm três vezes mais de probabilidade de morrer entre os 45-65 anos e duas vezes mais probabilidade de morrer entre os 65-84, em comparação com os não-fumadores (Scottish Intercollegiate Guidelines Network, 2007, p. 16).

Na Europa, o consumo de tabaco é responsável por mais de um milhão de mortes anuais, sendo que em Portugal, o consumo de tabaco atinge os 20% a 26% da população, com predomínio de três homens e meio para cada mulher⁹.

Estima-se que o tabagismo cause cerca de 10% de todas as doenças cardiovasculares, sendo que o risco de desenvolvimento destas doenças é maior em mulheres fumadoras, homens jovens e pessoas obesas. Atualmente, existem cerca de 1 bilhão de pessoas fumadoras, em todo o mundo (World Heart Federation, 2012, p. 1).

IV. Stress

Muitos dos doentes cardíacos associam o stress do seu quotidiano à causa da sua doença. Embora o stress seja algo comumente usado, não tem uma definição precisa e não consegue ser prontamente medido. O stress está muito associado a pessoas que sofrem de depressão, ansiedade, ataques de pânico, isolamento social, falta de apoio social e acontecimentos crónicos na vida (Scottish Intercollegiate Guidelines Network, 2007, p. 43).

O stress não é a principal causa das doenças cardíacas. No entanto, pode contribuir para as agravar, uma vez que, faz subir a pressão arterial, dificulta a cicatrização, tornando as pessoas mais vulneráveis a ataques patogénicos exteriores¹⁰.

V. Obesidade

À medida que alguns dos fatores de risco cardíacos podem ter tendência a diminuir, a obesidade tende a aumentar, sendo que se está a tornar uma epidemia mundial, tanto nas crianças como nos adultos.

⁹ <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/tabagismo/>

¹⁰ <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/stress/>

Atualmente, estima-se que em todo o mundo mais de 1 bilhão de pessoas estão acima do peso e mais de 300 milhões estão obesas. Mais de um terço das crianças são obesas ou têm peso acima da média. (I. Graham et al., 2007, p. 2391)

A obesidade é um fator de risco cardíaco estreitamente ligado à dieta e à falta de atividade física. Existe obesidade quando há um desequilíbrio entre a ingestão de energia na dieta e o gasto da mesma. Cerca de 1,7 milhões de mortes (2,8%), em todo o mundo, são atribuídas ao baixo consumo de frutas e hortaliças. Em Portugal, quase metade da população apresenta excesso de peso e perto de um milhão de adultos sofre de obesidade¹¹. O consumo adequado destes alimentos, reduz o risco de doenças cardíacas (I. M. Graham, 2008, p. 90).

VI. Diabetes

Os diabetes são uma doença em que existe excesso de açúcar no sangue (glucose). Existem dois tipos de diabetes: O diabetes tipo 1 e o diabetes tipo 2. O diabetes tipo 1 é menos vulgar, as células do pâncreas deixam de produzir insulina, sendo que as pessoas necessitam de injetar insulina, para o resto da vida. O diabetes de tipo 2 é a forma de diabetes mais comum e o pâncreas é capaz de produzir insulina. Em Portugal, existe um aumento anual de 3 % no número de casos do diabetes tipo 2.¹²

Os homens com diabetes do tipo 2 têm um risco anual de doenças cardíacas de 2 a 4 vezes, enquanto que as mulheres têm um risco maior (3 a 5 vezes). Existe um estudo que estima que 15% dos ataques cardíacos, na Europa Ocidental, e 9% na Europa Central e Oriental são devidos a diabetes do tipo 2. Outro estudo, afirma que as mulheres com diabetes tinham 8.7 vezes mais de probabilidade de morrer devido a uma doença cardíaca, em relação àquelas que não tinham diabetes (I. M. Graham, 2008, p. 123).

VII. Dislipidemia

Dislipidemia é um termo usado para designar todas as anomalias quantitativas ou qualitativas dos lípidos (gordura) do sangue. Este é um dos importantes fatores de risco da aterosclerose, a principal causa de morte nos países desenvolvidos, incluindo Portugal.¹³

¹¹ <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/obesidade/>

¹² <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/diabetes/>

¹³ <http://www.fpcardiologia.pt/saude-do-coracao/factores-de-risco/dislipidemia/>

A Dislipidemia inclui:

- Colesterol de baixa densidade, considerado colesterol “mau”;
- Colesterol de alta densidade, considerado colesterol “bom”;
- Triglicéridos.

O excesso de calorias no corpo é convertido em triglicéridos e armazenados em células de gordura, em todo o corpo. O colesterol de baixa densidade é depositado nas paredes das artérias do coração, provocando aterosclerose cardíaca. Os triglicéridos e o colesterol elevado aumentam o risco de doenças cardíacas. Globalmente, o colesterol elevado é estimado em causar 2,6 milhões de mortes em todo o mundo (Mendis et al., 2011, p. 42).

2.1.3. Alguns dados estatísticos, em Portugal

De acordo com o Instituto Nacional de Estatística (INE) (2016, p. 92), em 2014, existiu um total de 33288 número de mortes devido a doenças do aparelho circulatório, em Portugal. Observaram-se mais números mortes em mulheres do que em homens, 17713 e 14575, respetivamente.

O número de óbitos por esta causa corresponde a 30.7% da mortalidade total, em Portugal. Representando 27.3% do total de óbitos de homens e 34.2% no caso das mulheres.

Em 2014, a idade média de morte, por doenças do aparelho circulatório, foi de 80.8 anos. Para os homens, a idade média de morte foi 77.7 anos e 83.3 anos no caso das mulheres. Estas informações estão apresentadas na Tabela 1. Para além disso, ainda estão apresentadas informações relativas a algumas faixas etárias.

O INE afirma ainda que o Alentejo Central, o Alentejo Oeste e Viseu apresentam os valores mais elevados verificados no país, 33,7%, 33,6% e 33,6%, respetivamente. Estes valores têm em conta as mortes por doenças cardíacas em relação à mortalidade total do país. A região do Algarve registou o valor mais baixo (26,8%). A idade média de morte mais elevada foi observada na região da Beira Baixa (82,9 anos), sendo que a mais reduzida registou-se na Região Autónoma da Madeira (78,8 anos) (Estatística, 2016, p. 92).

Tabela 1 - Dados sobre as mortes por Doenças do Aparelho Circulatório, em 2014. Adaptado:(Estatística, 2016, p. 94)

			
Total de óbitos (nº)	33288	17713	14575
Idade média de morte (nº anos)	80,8	83,3	77,7
Proporção de óbitos* (%)	30,7	34,2	27,3
Óbitos com menos de 65 anos (nº)	2992	791	2201
Óbitos com 65 e mais anos (nº)	29295	16922	12373
Óbitos com menos de 70 anos (nº)	4482	1290	3192
Óbitos com 75 e mais anos	25452	15492	9960

*Percentagem em relação ao total de óbitos

Em 2014, verificou-se um total de 7456 mortes devido a doença isquémica do coração. Registou-se 4178 óbitos em homens e 3278 em mulheres. Não se registaram óbitos com idades inferiores a 5 anos, nesse ano. O número de óbitos por esta causa corresponde a 7,1% da mortalidade total, em Portugal. Representando 7,8% do total de óbitos de homens e 6,3% no caso das mulheres.

A idade média de morte por doenças isquémicas cardíacas é de 77.3 anos. Para as mulheres e para os homens, a idade média de morte é de 82.0 e 73.6, respetivamente. A faixa etária onde ocorrem mais mortes, por esta causa, é nos idosos com mais de 65 anos.

Estas informações estão presentes na Tabela 2. Para além disto, estão presentes informações sobre os óbitos em algumas faixas etárias (Estatística, 2016, p. 95).

Tabela 2 - Dados sobre as mortes por Doenças Isquémicas do Coração, em 2014. Adaptado:(Estatística, 2016, p. 97)

			
Total de óbitos (nº)	7456	3278	4178
Idade média de morte (nº anos)	77,3	82,0	73,6
Proporção de óbitos* (%)	7,1	6,3	7,8
Óbitos com menos de 65 anos (nº)	1274	204	1070
Óbitos com 65 e mais anos (nº)	6182	3074	3108
Óbitos com menos de 70 anos (nº)	1818	357	1461
Óbitos com 75 e mais anos	4954	2695	2259

*Percentagem em relação ao total de óbitos

De acordo com a Direção Geral de Saúde (2015, p. 15), entre 2009 e 2013 houve uma variação negativa de 8,28%, tal como apresentado na Figura 2. Isto significa que entre 2009 e 2013 houve uma diminuição dos números de óbitos por doenças isquémicas cardíacas de 8%. Contudo, tal como apresentado na Tabela 2, em 2014, houve um aumento de quase 1000 óbitos, comparando com o ano anterior. Sendo que, 2014 foi o ano (desde 2009) em que houve o maior número de mortes por doença isquémica cardíaca.

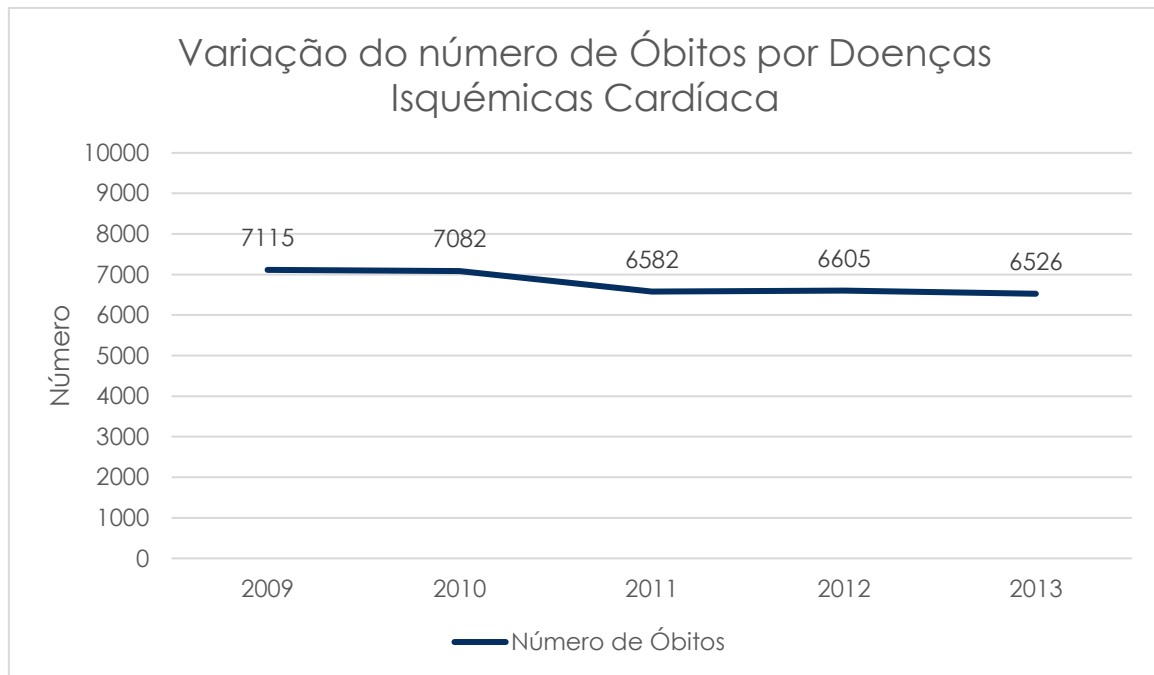


Figura 2 - Número de Óbitos por Doenças Isquémicas Cardíaca, em Portugal (2009-2013). Adaptado:(Direção-Geral da Saúde, 2015, p. 15)

Em 2014, apurou-se um total 4619 mortes devido ao enfarte agudo do miocárdio, vulgarmente chamado de ataque cardíaco. Registou-se 2617 óbitos em homens e 2002 em mulheres. Não se registaram óbitos com idades inferiores a 5 anos, sendo o número de óbitos mais elevados a partir dos 75 anos. O número de óbitos por ataques cardíacos corresponde a 4,4% da mortalidade total, em Portugal, representando 4,9% do total de óbitos de homens e 3,9% nas mulheres.

Estas informações estão apresentadas na Tabela 3. Para além disso, estão apresentadas outras informações, como a idade média de morte e o número de mortes por algumas faixas etárias (Estatística, 2016, p. 100).

Tabela 3 - Dados sobre as mortes por Enfarte Agudo do Miocárdio, em 2014. Adaptado:(Estatística, 2016, p. 100)

			
Total de óbitos (nº)	4619	2002	2617
Idade média de morte (nº anos)	76,4	81,2	72,8
Proporção de óbitos* (%)	4,4	3,9	4,9
Óbitos com menos de 65 anos (nº)	861	146	715
Óbitos com 65 e mais anos (nº)	3758	1856	1902
Óbitos com menos de 70 anos (nº)	1225	251	974
Óbitos com 75 e mais anos	2944	1605	1339

*Percentagem em relação ao total de óbitos

Segundo a Direção Geral de Saúde (2015, p. 16), entre 2009 e 2013 existiu uma variação negativa 7,98%, tal como apresentado na Figura 3. Isto significa que entre 2009 e 2013 houve uma diminuição dos números de óbitos por doenças isquémicas cardíacas de 8%. Contudo, tal como apresentado na Tabela 2, em 2014, houve um aumento de quase 400 óbitos, em comparação a 2013.

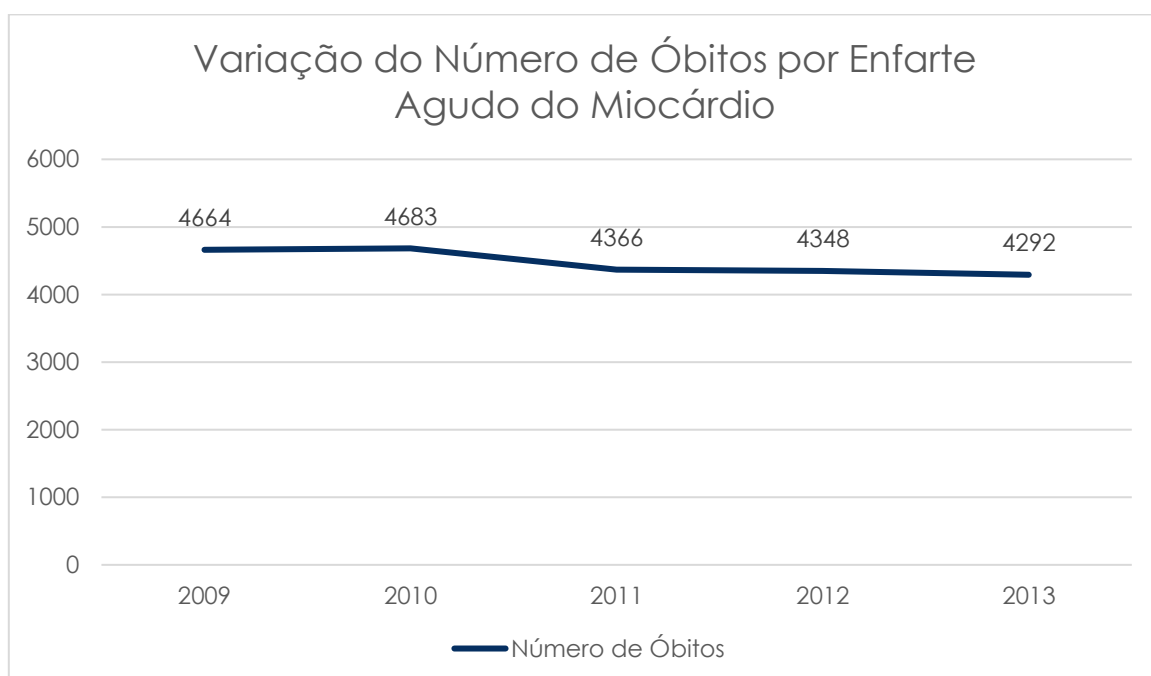


Figura 3 - Número de Óbitos por Enfarte Agudo do Miocárdio, em Portugal (2009-2013). Adaptado: (Direção-Geral da Saúde, 2015, p. 16)

2.2. Descoberta de Conhecimentos em Base de Dados (DCBD)

2.2.1. Definição

A área de conhecimento *Knowledge Discovery in Databases* (KDD), que em português significa “Descoberta de Conhecimentos em Base de Dados (DCBD)”, foi apresentada formalmente em 1989 como um conceito alargado para a procura de conhecimento nos dados. Foi definida como sendo o processo de identificação e reconhecimento de padrões, de forma automática, em base de dados (M. F. Santos & Azevedo, 2005, p. 18).

Existem várias definições na literatura para esta descoberta de conhecimento em base de dados.

- DCBD é uma análise automática, exploratória e capaz de modelar grandes repositórios de dados. DCBD é o processo organizado de identificar padrões válidos, novos, úteis e compreensíveis a partir de um grande conjunto de dados (Maimon & Rokach, 2005, p. 1);
- Segundo Maletic e Marcus (2005, p. 1), DCBD é um processo organizado de identificação de padrões válidos, úteis, novos e compreensíveis de um grande e complexo conjunto de dados;
- Para Marcin Mironczuk (2015, p. 1), DCBD é uma especialização do processo de *Knowledge Discovery* (KD), sendo este definido como o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e interpretáveis. Segundo este autor, quando se adiciona a frase “*in databases*” a KD, enfatiza-se o facto de os dados estarem armazenados em base de dados, assim como as informações úteis para o utilizador e os algoritmos de processamento dos dados;
- DCBD é o processo de obtenção de informações, através do *Data Mining*, e a interpretação dessas informações de forma a criar conhecimento. Estas informações têm que ser válidas, novas, úteis e compreensíveis pelos seres humanos (Miller & Han, 2009, p. 3).

Com base nestas definições pode-se concluir que a descoberta de conhecimento de base de dados é um processo não trivial de identificar padrões novos, úteis, interpretáveis e válidos de um grande conjunto de dados.

Existe um conjunto de componentes que estão presentes em todas as definições de DCBD, apresentados na Tabela 4. Os componentes apresentados nessa tabela são: dados, padrão, válidos, novos, úteis e interpretáveis. A Tabela 4 apresenta, também, uma explicação sobre cada um dos componentes.

Tabela 4 - Componentes de DCBD

Componentes	Explicação
Dados	Os dados são um conjunto de casos de uma base de dados.
Padrão	Um padrão é uma expressão que descreve um subconjunto de casos, numa determinada linguagem.
Validos	Os padrões descobertos têm que ser válidos, ou seja, devem possuir algum grau de confiança.
Novos	Um padrão deverá fornecer informações novas sobre os dados.
Úteis	Os padrões descobertos devem ser utilizados.
Interpretáveis	Os padrões têm que ser interpretáveis com o objetivo de ser compreendidos pelos humanos.

As melhorias contínuas em Tecnologias de Informação (TI) e a sua adoção no controlo de processos faz com que exista uma grande quantidade de dados para serem recolhidos. O DCBD é uma resposta a esta elevada quantidade de dados, pois este termo engloba princípios e técnicas de estatística, aprendizagem de máquinas, reconhecimento de padrões, pesquisa numérica e visualização científica, para acomodar estas grandes quantidades de dados (Miller & Han, 2009, p. 3,4).

Muitos autores confundem o termo *Data Mining* com o termo DCBD (M. F. Santos & Azevedo, 2005, p. 19). No entanto, neste projeto, os termos DCBD e *Data Mining* são usados em sentidos diferentes; DCBD refere-se ao processo como um todo, enquanto que *Data Mining* é o componente central deste processo (explicado nos capítulos seguintes).

DCBD é um processo iterativo, visto que pode existir recuos a etapas anteriores, e iterativo, uma vez que necessita da participação do utilizador sempre que é necessário a tomada de decisão (M. Y. Santos & Ramos, 2006, p. 104). As diversas fases do processo de DCBD (descritas na secção 2.2.2) estão apresentadas na Figura 4.

2.2.2. Fases do Processo DCBD

O processo de DCBD (descoberta de conhecimento de base de dados) (Figura 4) é composto por várias fases, sendo um processo iterativo e interativo. Cada fase é iterativa pois, se necessário, é possível retroceder para fases anteriores, para se ajustar determinados pormenores. Este processo é iterativo uma vez que necessita do utilizador sempre que é necessária a tomada de decisão. Não existe uma fórmula ou uma maneira de saber qual o caminho mais correto em cada uma das etapas. Assim, é necessária uma compreensão profunda do processo, bem como as diferentes possibilidades que existem em cada uma das etapas (Maimon & Rokach, 2005, p. 2).

Este processo começa com a determinação dos objetivos finais a serem atingidos. Depois de agrupar os dados de forma organizada, passa-se à limpeza dos mesmos, através de um pré-processamento dos dados. Posteriormente, estes dados são guardados num formato correto, que facilita as técnicas de *Data Mining*. Tendo em conta os objetivos a serem atingidos, é escolhida uma técnica de *Data Mining*. Por fim, é gerado um relatório de descobertas, que são validadas e interpretadas, permitindo obter conhecimento (M. F. Santos & Azevedo, 2005).

A Figura 4 representa graficamente o processo de descoberta de conhecimento em base de dados. Na imagem estão representadas as várias fases deste processo: Compreensão do domínio e objetivos da DCBD (Domain Understanding and KDD Goals), Seleção e Adição (Selection and Addition), Pré-processamento (Preprocessing), Transformação (Transformation), Data Mining, Avaliação, Interpretação e Descoberta de Conhecimento (Evaluation, Interpretation and Discovered Knowledge).

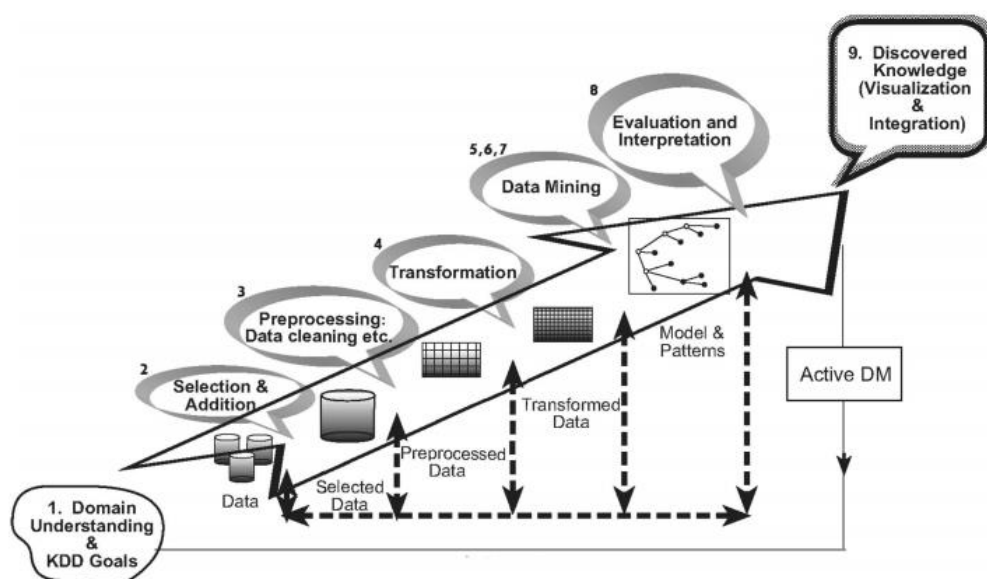


Figura 4 - Processo DCBD (Oded Maimon, 2005, p. 3)

I. Compreensão do domínio e objetivos da DCBD (*Domain Understanding and KDD Goals*)

Esta é a etapa preparatória de todo o processo. Nesta etapa é preparado todo o processo, de modo a entender o que deve ser feito em cada uma das fases, bem como as decisões que se poderão tomar (sobre os algoritmos e as representações, por exemplo). As pessoas que trabalham com este processo, estão encarregadas de definir e entender os objetivos do utilizador final e o ambiente em que este ocorrerá. Sendo o processo de descoberta de conhecimento um processo iterativo, poderá existir uma revisão ou ajuste desta etapa em fases posteriores (Oded Maimon, 2005, p. 2). Entendidos os objetivos de DCBD, inicia-se a seleção e recolha de um conjunto de dados ou variáveis necessárias, conforme definido na próxima etapa.

II. Seleção e Adição (*Selection and Addition*)

Depois de definidos os objetivos do processo de KDD, os dados utilizados neste processo deverão ser determinados. Os dados são selecionados de acordo com alguns critérios, sendo definidos subconjuntos de dados (M. F. Santos & Azevedo, 2005, p. 20). Estes dados serão armazenados em repositórios. O principal objetivo da seleção de dados é limitar o espaço de pesquisa, eliminando atributos que não têm qualquer interesse no processo (atributos com carácter informativo, como nomes e números de contribuinte) (M. Y. Santos & Ramos, 2006, p. 105).

Esta etapa é muito importante porque o *Data Mining* aprende e descobre novos padrões a partir dos dados disponíveis. Se alguns dos atributos importantes estão a faltar, então, é provável, que todo o estudo possa falhar (Oded Maimon, 2005, p. 3).

III. Pré-processamento (*Preprocessing*)

Nesta etapa, a confiabilidade dos dados é melhorada. Existe uma limpeza dos dados, onde os valores inconsistentes são tratados e onde se pensa em soluções para a omissão de dados (M. F. Santos & Azevedo, 2005, p. 20). Estes problemas nos dados ocorrem, porque existem deficiências na introdução dos mesmos (M. Y. Santos & Ramos, 2006, p. 105). Consiste numa fase onde existe a remoção de ruídos ou de erros (Oded Maimon, 2005, p. 3).

A verificação destas inconsistências permite que se identifique os valores que não fazem sentido, tendo em conta o problema onde estão inseridos.

IV. Transformação (*Transformation*)

Nesta etapa, a pesquisa torna-se mais eficiente. Existe uma diminuição do número de linhas ou colunas a analisar, que, conseqüentemente, diminui o espaço de pesquisa (M. Y. Santos & Ramos, 2006, p. 106).

Esta etapa é crucial para o sucesso de todo o processo de DCBD, mas é uma etapa que depende do projeto em que está inserido (Oded Maimon, 2005, p. 3).

A forma como os dados são tratados e agregados influencia diretamente os resultados obtidos na fase de *Data Mining*. Esta fase pode ser repetida várias vezes ao longo de todo o processo, de forma a melhorar os resultados obtidos (M. Y. Santos & Ramos, 2006, p. 106).

Depois de completadas as etapas anteriores, as etapas posteriores estão relacionadas com o *Data Mining*, onde o principal foco está nos algoritmos selecionados para cada projeto (Oded Maimon, 2005, p. 3).

V. *Data Mining*

Segundo Oded Mainon (2005, p. 3) nesta etapa, inicialmente é necessário escolher o tipo de abordagem a usar: Classificação, Regressão ou Clustering. Esta escolha depende dos objetivos inicialmente definidos, bem como das fases anteriores.

Nesta fase, possivelmente, será necessário empregar o algoritmo várias vezes para que se obtenha um resultado considerado satisfatório.

VI. Avaliação, Interpretação e Descoberta de Conhecimento (*Evaluation, Interpretation and Discovered Knowledge*)

Os resultados deste processo de descoberta de conhecimento podem ser visualizados de diferentes maneiras. Contudo, estas visualizações devem ser de fácil interpretação e criteriosas, de modo a identificar a necessidade de retornar a qualquer uma das fases anteriores (M. F. Santos & Azevedo, 2005, p. 21). Esta fase centra-se principalmente na compreensibilidade e utilidade do modelo criado.

Nesta fase, são avaliados e interpretados os padrões obtidos, tendo em conta os objetivos definidos inicialmente (Oded Maimon, 2005, p. 3). Existem falhas no processo quando os modelos obtidos

não satisfazem as necessidades do utilizador (M. Y. Santos & Ramos, 2006, p. 107). Neste caso, é aconselhado o retrocesso a fases anteriores para alterar as decisões tomadas.

Os padrões identificados são interpretados como conhecimento, sendo este conhecimento documentado para uso posterior no suporte à tomada de decisão.

2.3. *Data Mining*

Data Mining é uma das fases do processo de descoberta de conhecimento (DCBD). Tem como objetivo a procura de relacionamentos, padrões e modelos que estão presentes ou implícitos nos dados.

Esta área está em grande evolução e expansão, sendo que o grande volume de dados disponíveis podem-se tornar fonte de conhecimento útil e com aplicabilidade em diferentes domínios (M. F. Santos & Azevedo, 2005).

A acessibilidade e abundância de informações torna o *Data Mining* uma questão de considerável importância e necessidade (Azevedo & Santos, 2011, p. 68).

2.3.1. Definição

A área de *Data Mining* pode ser usada sem o indivíduo se aperceber. O simples facto de se inscrever para obter um cartão de um determinado supermercado, o indivíduo está a criar dados. Estes dados são armazenados nos computadores das organizações. Posteriormente, estes são analisados e são descobertos padrões, ou seja, indicadores de interesses, hábitos e comportamentos dos indivíduos que se inscreveram para obter o cartão do supermercado (North, 2012, p. 3).

Todos os sistemas automatizados geram, de alguma forma, dados para fins de diagnóstico ou análise. Isto resulta em grandes quantidades de dados, que tem atingido a ordem dos 1.0×10^{15} bytes ou 1.0×10^{18} bytes. Esta quantidade de dados é um resultado direto dos avanços da tecnologia e da informatização de todos os aspetos da vida moderna (Aggarwal, 2015, p. 2). Para além disso, este grande aumento de dados, deve-se à diminuição dos custos do armazenamento destes dados.

Inicialmente, os dados eram analisados, tendo em conta folhas de cálculo, e eram processados através de métodos tradicionais, como os histogramas e a estatística. Mas, com a evolução dos sistemas de informação e o aumento do volume de dados, estes métodos tornaram-se inexecutáveis.

Estes métodos são incapazes de encontrar padrões, pois só utilizam métodos estatísticos. No sentido de fazer face a estas dificuldades encontradas pelos métodos tradicionais, surgiu o *Data Mining*. O *Data Mining* tem como objetivo extrair conhecimento, padrões e tendências de uma forma autónoma ou semiautónoma (M. F. Santos & Azevedo, 2005, p. 1).

A área de *Data Mining* fornece técnicas e ferramentas, permitindo que grandes quantidades de dados sejam analisadas automaticamente.

Existem várias definições de *Data Mining*, sendo que nenhuma está mais correta que outra, pois todas convergem para o mesmo sentido. De seguida, serão apresentadas algumas dessas definições.

- *Data Mining* é o estudo da recolha, limpeza, processamento, análise e aquisição de informações úteis a partir dos dados. O termo *Data Mining* é usado, de forma genérica, para descrever os diferentes aspetos de processamento de dados, pois existem várias formulações e representações dos dados (Aggarwal, 2015, p. 1);
- *Data Mining*, permite que as pessoas encontrem e interpretem determinados padrões nos dados, ajudando-os a tomar decisões melhores, de forma a beneficiar melhor os seus clientes (North, 2012, p. 3);
- *Data Mining* consiste na procura de relacionamentos, padrões ou modelos que estão implícitos nas grandes quantidades de dados armazenados (M. Y. Santos & Ramos, 2006, p. 127);
- *Data Mining* consiste na aplicação de métodos e técnicas a grandes quantidades de dados, de forma a encontrar padrões, com o intuito de gerar e descobrir conhecimento (M. F. Santos & Azevedo, 2005, p. 10).

Depois de agregar todas as opiniões dos autores referidos, considera-se que *Data Mining* é um processo de extração de informação, desconhecida, a partir de grandes quantidades de dados. Esta informação é processada através de técnicas e métodos de análise de dados.

Data Mining aplica métodos estatísticos e lógicos a um grande conjunto de dados. Esses métodos podem ser usados para categorizar os dados, ou ser usados para criar modelos preditivos. A categorização dos dados inclui a identificação de características semelhantes num grande conjunto de dados analisados. Os modelos preditivos transformam essas identificações em expectativas sobre as quais podemos tomar decisões. Por exemplo, o proprietário de uma estância de ski pode prever a data de abertura da instância com base nas previsões de chegada de neve (North, 2012, p. 15).

O *Data Mining* ajuda a entender como os elementos do negócio se relacionam uns com os outros, fornecendo pistas sobre decisões que se podem tomar para que o negócio seja executado de forma mais clara e de forma a gerar mais receita. Por exemplo, o *Data Mining* pode ajudar um gestor a identificar onde pode cortar nos custos, sem danificar a estrutura da organização e onde os gastos trazem melhor retorno para a mesma (Brown, 2014, p. 9).

As técnicas de *Data Mining* tornam-se indispensáveis a várias áreas, devido à grande necessidade de analisar o conhecimento que os dados podem trazer a uma organização (M. F. Santos & Azevedo, 2005, p. 12).

É importante reconhecer que, o *Data Mining* não fornece respostas a todas as perguntas, e nem se pode esperar que os modelos de previsão produzam resultados que se tornarão realidade (North, 2012, p. 15).

2.3.2. Vantagens

São inúmeras as vantagens que existem na implementação de ferramentas de *Data Mining* numa organização, contudo, destacamos as seguintes (M. F. Santos & Azevedo, 2005, p. 8):

- Sem ferramentas de *Data Mining*, as organizações apenas podiam fazer consultas tradicionais em base de dados, utilizando *queries*, que apenas lidavam com questões do género “qual foi o meu rendimento total nos últimos cinco anos?”. Com o *Data Mining* as consultas passam a ser muito mais alargadas, mais rápidas e rentáveis, em comparação com as consultas tradicionais, que são mais lentas e dispendiosas. Assim, com o *Data Mining*, já se pode responder a questões do género “O que é provável acontecer ao número de vendas no próximo mês? Porquê?”;
- Com o *Data Mining* é possível a deteção de padrões e tendências nos dados, através do uso de algoritmos ou mecanismos de pesquisa (existe uma grande variedade de algoritmos e técnicas de análise), isto permite um conhecimento adicional acerca do negócio;
- Existem várias abordagens para a análise dos dados: quando já se sabe o que se vai pesquisar e apenas se confirmam hipóteses (abordagem top-down) ou quando não se sabe o que se vai pesquisar e os sistemas geram hipóteses de forma autónoma (abordagem bottom-up);

- O *Data Mining* ajuda na tomada de decisão, onde se evitam contextos que não são desejados para a organização. A organização consegue prever tendências futuras, tendo em conta os padrões encontrados nos dados.

2.3.3. Desvantagens

Para além das inúmeras vantagens do uso das ferramentas de *Data Mining*, existem também algumas desvantagens. O facto de existir uma grande variedade de abordagens poderá se tornar uma desvantagem, pois inicialmente é necessário perceber quais as vantagens e desvantagens de utilizar determinada abordagem, tendo em conta os objetivos do projeto. É necessário aplicar a melhor abordagem e técnica para que não hajam resultados falsos, de indução em erro.

As ferramentas a utilizar também poderão se tornar uma desvantagem. Para a escolha da ferramenta é necessária uma análise das existentes no mercado, para perceber qual a ferramenta que melhor se encaixa nos objetivos do projeto. Existem ferramentas de *Data Mining* muito complexas, onde é necessário que o utilizador tenha um conhecimento em áreas como programação, *Data Mining* e estatística (M. F. Santos & Azevedo, 2005).

2.3.4. Ciclo de Vida

Para uma melhor compreensão de todos os processos incluídos no *Data Mining* é necessário analisar o seu ciclo de vida. A Figura 9 representa o ciclo de vida de um projeto de *Data Mining*, conhecido como CRISP-DM (*Cross Industry Standard Process for Data Mining*), sendo constituído por seis fases. Este ciclo de vida será explicado de seguida, ao pormenor, 3.1.2. Cross Industry Standard Process for Data Mining (CRISP-DM)

2.3.5. Abordagens *Data Mining*

I. Classificação

A classificação supõe que existe um conjunto de objetos, caracterizados por um atributo, que pertencem a classes diferentes. O objetivo é construir modelos que atribuem a classificação correta a objetos que anteriormente não foram classificados (Shearer et al., 2000, p. 20).

Segundo Manuel Filipe Santos e Carla Azevedo (2005, p. 14), “a classificação corresponde a encontrar uma função que associa um caso a uma classe dentro de diversas classes discretas (quando os elementos são números inteiros) de classificação, de forma a classificar um novo objeto de acordo com um modelo de classificação”.

Por exemplo, pode-se estabelecer que um mamífero é definido por estas características: beber leite e ser vertebrado. Então, todos os animais que se regem por estas características são classificados como mamíferos.

Entre as técnicas mais comuns na abordagem de classificação encontram-se (Ramageri, 2010, p. 302):

- Árvores de decisão;
- Redes Neurais artificiais;
- Classificação Bayesiana;
- Support Vector Machines (SVM).

As **árvores de decisão** (AD) são uma das técnicas mais intuitivas e frequentemente usadas, em *Data Mining*, sendo fáceis de configurar, do ponto de vista de um analista, e fáceis de interpretar, do ponto de vista do utilizador. As AD são utilizadas para separar um conjunto de dados em classes (normalmente, sim ou não (1 ou 0)) (Elkan, 2010, p. 64).

A AD é uma representação de um conjunto de regras, que classificam a instância, desde do nó de raiz até às folhas (nós terminais). Cada nó da árvore é um teste para as variáveis (atributos da instância), sendo que cada ramo desse nó corresponde aos valores possíveis dessa variável (M. F. Santos & Azevedo, 2005, p. 45).

A Figura 5 mostra o funcionamento de uma AD. Por exemplo, se estiver a chover não se joga badminton. Por outro lado, se o céu está nublado e não existe vento, então pode-se jogar badminton.

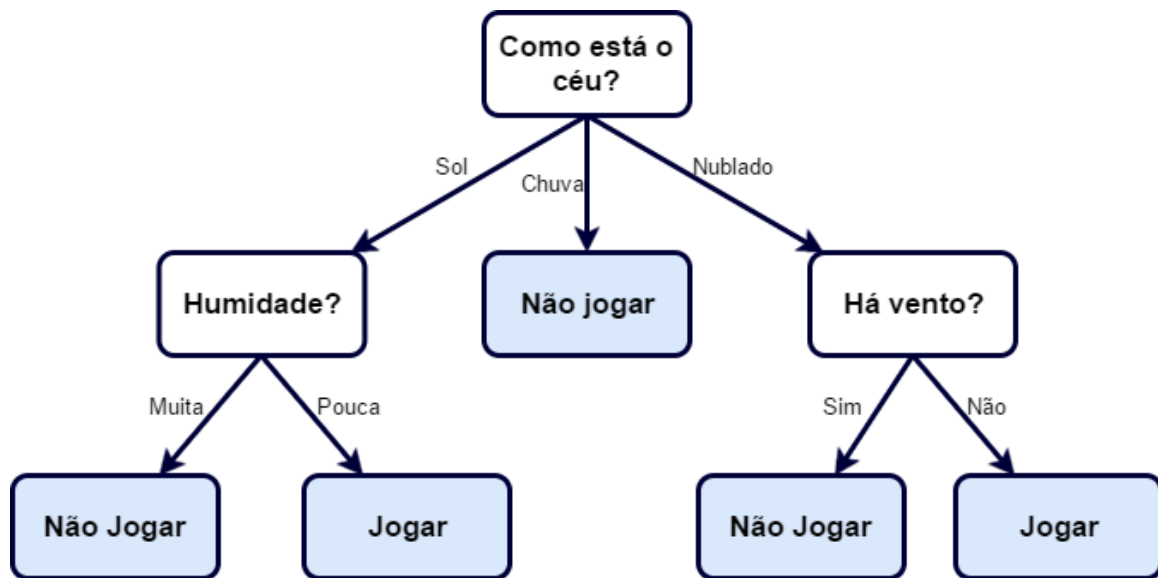


Figura 5 - Árvore de decisão: jogar badminton. Adaptado: (M. F. Santos & Azevedo, 2005, p. 46)

As **Redes Neurais Artificiais (RNA)** são um conjunto de elementos com processamento simples, como nós, unidades ou neurónios artificiais. As RNA possuem um grande número de interligações, sendo a estrutura muito semelhante ao cérebro humano. Uma RNA é uma estrutura computacional baseada em neurónios (unidades de processamento) ligados entre si e organizados em grupos (camadas) (M. F. Santos & Azevedo, 2005, p. 53).

A Figura 6 é um exemplo de uma RNA, onde as diferentes camadas são dispostas paralelamente. A primeira camada é a camada de entrada e a última é a camada de saída. As camadas intermédias são designadas camadas escondidas.

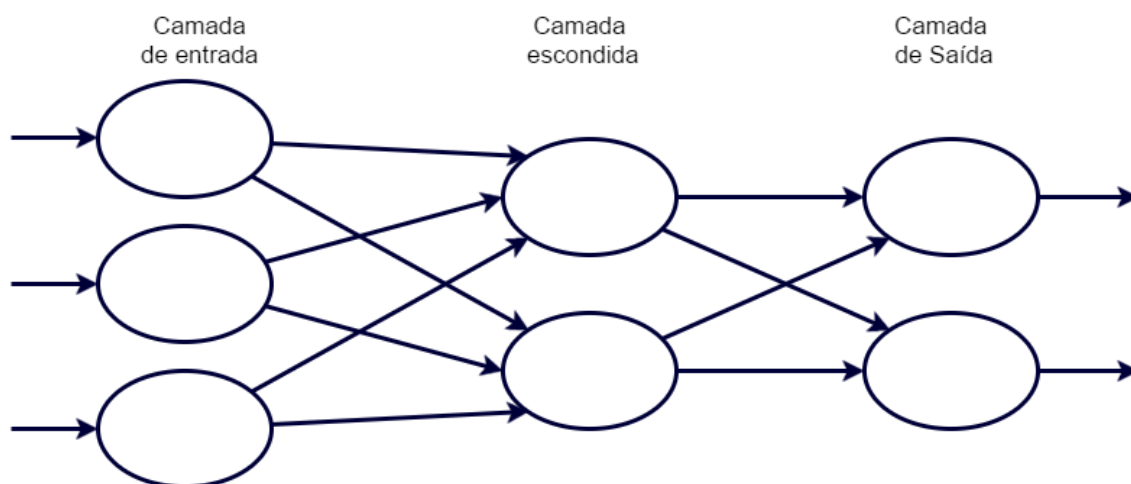


Figura 6 - Exemplo de uma Rede Neuronal Artificial. Adaptado: (M. F. Santos & Azevedo, 2005, p. 57)

A **classificação Bayesiana** baseia-se no pressuposto que a classificação de padrões é expressa em termos probabilísticos. A teoria de Bayes fornece uma estrutura para classificar padrões em classes. Estas classes são baseadas nas probabilidades dos padrões e as nas suas características (Cios, Swiniarski, Pedrycz, & Kurgan, 2007, p. 307). Segundo Manuel Filipe Santos e Carla Azevedo (2005, p. 83), esta teoria é descrita como “uma forma de organizar as regras de probabilidade para atualizar a probabilidade a priori, ou confiança, dada a nova informação, resultando na confiança a posteriori”. O teorema de Bayes pode ser expressado pela seguinte expressão:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Onde:

- $P(H|X)$ representa a probabilidade a posteriori;
- $P(H)$ representa a probabilidade a priori;
- $P(X|H)$ representa a função densidade de probabilidade (a probabilidade da classe H);
- $P(X)$ representa a função densidade da probabilidade incondicional.

As **Support Vector Machines** (SVM) são um método de classificação de dados lineares e não lineares. Esta técnica utiliza um mapeamento não linear para transformar os dados de treino numa dimensão superior. Dentro desta dimensão, é procurado um hiperplano ótimo de separação linear, surgindo uma fronteira de decisão que separa os tuplos das diferentes classes (Jiawei, Kamber, Han, Kamber, & Pei, 2012a, p. 408). Na Figura 7 está representada a separação linear de dois conjuntos de dados, sendo que existe um número infinito de linhas de separação (fronteiras de decisão) e hiperplanos.

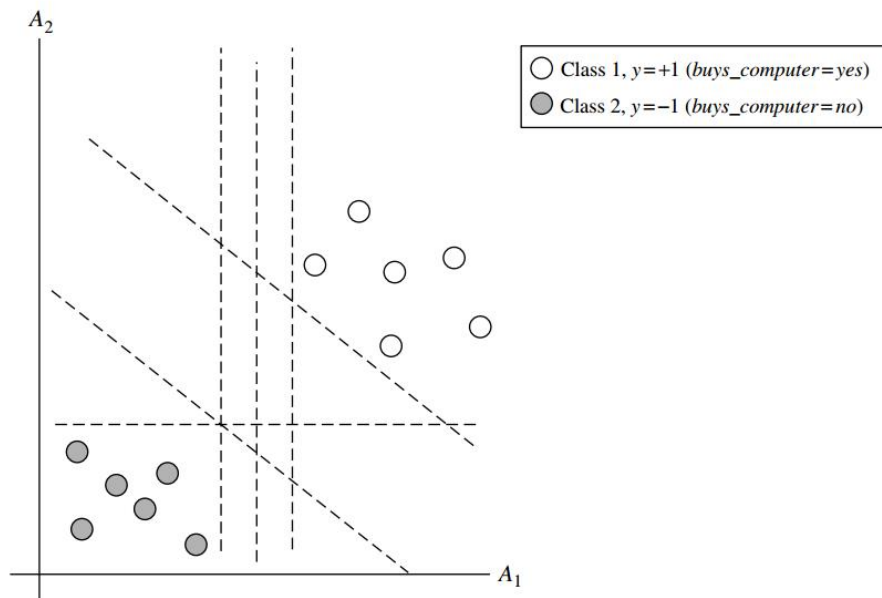


Figura 7 - Separação Linear de duas Classes. Fonte:(Jiawei et al., 2012a, p. 409)

Após gerados os modelos de *Data Mining*, torna-se necessário avaliar o seu desempenho. Segundo Manuel Filipe Santos e Carla Azevedo (M. F. Santos & Azevedo, 2005, p. 93), a técnica mais utilizada é a **matriz de confusão**, para problemas de classificação. A matriz de confusão permite calcular taxas de erro e definir a curva de *Receiver Operating Characteristic* (ROC).

Na Tabela 5 está apresentada uma matriz de confusão. Nas linhas são colocados os valores reais (Real A e Real B), nas colunas estão colocadas as previsões (Previsão A e Previsão B) efetuadas pelo modelo. Uma matriz de confusão indica o número de classificações corretas versus as previsões realizadas para cada classe.

Tabela 5 - Matriz de Confusão

Classe	Previsão A	Previsão B
Real A	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
Real B	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

Existem, essencialmente, quatro indicadores que resumem as possibilidades de erros e acertos (M. F. Santos & Azevedo, 2005, p. 94):

- **Verdadeiros Positivos (VP)**: correspondem ao número de exemplos positivos classificados corretamente;

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

Da matriz de confusão podem ser retiradas muitas métricas, das quais destacam-se as seguintes (M. F. Santos & Azevedo, 2005, p. 95):

- **Acuidade**: corresponde à percentagem da soma dos números de VP com os FP sobre o número total de casos analisados, sendo representada pela fórmula seguinte.

$$acc = \frac{VP + VN}{n} * 100(\%)$$

- **Sensibilidade**: consiste na percentagem de amostras negativas que são classificadas corretamente sobre a soma dos verdadeiros positivos com os falsos negativos, sendo representado pela fórmula seguinte.

$$sens = \frac{VP}{VP + FN} * 100(\%)$$

- **Especificidade**: consiste na percentagem de amostras negativas que são classificadas corretamente sobre a soma dos verdadeiros negativos com os verdadeiros positivos, sendo representado pela fórmula seguinte.

$$spec = \frac{VN}{VN + FP} * 100(\%)$$

- **Cobertura**: consiste na percentagem da soma das amostras positivas com o número de amostras positivas mal classificadas sobre o número total de casos analisados, sendo representado pela fórmula seguinte.

$$cov = \frac{VP + VN}{n} * 100(\%)$$

- **Confiança positiva**: consiste na percentagem de amostras positivas que são classificadas corretamente sobre o total da soma das amostras positivas com o número de amostras positivas mal classificadas, sendo representado pela fórmula seguinte.

$$prel = \frac{VP}{VP + FP} * 100(\%)$$

- **Confiança negativa**: consiste na percentagem de amostras negativas que são classificadas corretamente sobre o total da soma das amostras negativas com o número de amostras negativas mal classificadas, sendo representado pela fórmula seguinte.

$$nrel = \frac{VN}{VN + FN} * 100(\%)$$

- **Taxa de erro**: corresponde à percentagem da soma de falsos positivos com falsos negativos sobre o número total de casos analisados, sendo representado pela fórmula seguinte.

$$erro = \frac{FP + FN}{n} * 100(\%)$$

A curva de ROC permite estabelecer a relação entre a sensibilidade e a especificidade, sendo apropriado quando existem duas classes. Esta curva permite a avaliação do desempenho de um classificador. Existem duas técnicas frequentemente utilizadas, a fim de fornecer uma comparação do desempenho dos diferentes classificadores: AUC (Area Under Curve) e ROCCH (ROC Convex Hull). A AUC é uma métrica de desempenho obtida através do cálculo da área por baixo da curva de ROC. A ROCCH permite declarar um subconjunto de classificadores como potencialmente ótimos (M. F. Santos & Azevedo, 2005, p. 97).

II. **Regressão**

Consiste em encontrar uma previsão de um valor de uma variável contínua, tendo em conta os valores de outras variáveis, ou seja, trata-se de encontrar uma função para a previsão de uma variável (M. F. Santos & Azevedo, 2005, p. 15).

Por exemplo, prever a velocidade do vento como uma função de temperatura, humidade e pressão atmosférica.

Entre as técnicas mais comuns na abordagem de regressão encontram-se (Ramageri, 2010):

- Regressão linear;
- Regressão não linear.

Na regressão linear, os dados são modelados usando uma linha reta. A regressão linear é a forma mais simples de regressão. A regressão linear modela uma variável Y (variável de resposta), como uma função linear de outra variável aleatória X (variável de previsão), sendo que pode ser representada por esta função (Jiawei, Kamber, Han, Kamber, & Pei, 2012b, p. 30).

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

Onde:

- α representa a interseção com Y;
- β representa o declive da reta.

Na regressão não linear, a relação entre as variáveis de previsão e as variáveis de resposta não apresentam um comportamento linear. Um exemplo de regressão não linear é quando estas variáveis são modeladas como uma função polinomial. Assim, são adicionados termos polinomiais ao modelo linear básico, podendo, através da transformação das variáveis, converter o modelo não linear em modelo linear (Jiawei et al., 2012b, p. 31).

Depois de gerados os modelos de *Data Mining*, tendo em conta uma abordagem regressiva, é necessário avaliar o seu desempenho. O principal objetivo de um método de regressão é conceber o melhor modelo segundo uma medida da estimativa do erro (M. F. Santos & Azevedo, 2005, p. 98). Esse erro pode ser calculado tendo em conta métricas como: *Mean Absolute Error* (MAE), *Relative Absolute Error* (RAE), *Mean Square Error* (MSE) e *Relative Square Error* (RSE).

Em regressão, a MSE é a medida principal mais utilizada, pois tende a ser a medida mais fácil de manipular matematicamente. Esta medida de desempenho é calculada tendo em conta a seguinte expressão, onde p representa os valores previstos das instâncias de teste e a os valores atuais (Witten, Frank, & Hall, 2011, p. 177):

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

A MAE é uma medida de desempenho alternativa, onde são medidos os erros individuais. Esta métrica tende a exagerar nos *outliers*, ou seja, tende a exagerar nas instâncias cujo o erro de previsão é maior do que os outros. Esta medida é calculada tendo em conta a seguinte expressão,

onde p representa os valores previstos das instâncias de teste e a os valores atuais (Witten et al., 2011, p. 177):

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

O **RSE** é uma métrica um pouco diferente, uma vez que, neste caso, o erro é formado em relação ao que deveria ter sido caso se utilizasse um preditor simples. O preditor simples é a média dos valores reais dos dados de treino. O RSE pega no erro quadrático total e normaliza-o, dividindo pelo erro quadrático total do preditor padrão. É definido pela seguinte expressão, onde p representa os valores previstos das instâncias de teste e a os valores atuais (Witten et al., 2011, p. 177):

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

$$\bar{a} = \frac{1}{n} \sum_i a_i$$

O **RAE** é uma métrica normalizada da mesma maneira que o RSE, sendo definido como o erro total absoluto. Esta medida é calculada tendo em conta a seguinte expressão, onde p representa os valores previstos das instâncias de teste e a os valores atuais (Witten et al., 2011, p. 177):

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

$$\bar{a} = \frac{1}{n} \sum_i a_i$$

III. Segmentação (Clustering)

A segmentação permite identificar grupos homogêneos de objetos em que cada grupo é uma classe, assim, os dados que os grupos contêm devem apresentar algum tipo de semelhança, relativamente a outros que pertencem à mesma classe.

Permite a identificação de um conjunto de categorias ou segmentos para descrever os dados, sendo que a segmentação assegura a similaridade entre os grupos da mesma classe é alta e a similaridade entre grupos de diferentes classes é baixa (M. F. Santos & Azevedo, 2005, p. 16).

Existem muitos métodos de segmentação, sendo difícil fornecer uma categorização nítida dos mesmos. Isto acontece porque as diferentes categorias podem-se sobrepor, sendo que um método pode ter recursos de várias categorias. No entanto, em 2012, o autor Jiawei *et al.* (2012b, p. 450) conseguiu agrupar os métodos de segmentação em categorias. Assim, os principais métodos de classificação podem ser agrupados nas seguintes categorias (Jiawei et al., 2012a, p. 450):

- **Métodos de partição**: um conjunto de elementos, um método de partição constrói partições dos dados, onde cada partição representa um cluster e tem que possuir um elemento ou mais. Cada elemento apenas pode pertencer a uma partição;
- **Métodos hierárquicos**: um método hierárquico cria uma decomposição hierárquica de um conjunto de dados, sendo classificado de duas maneiras: bottom-up e top-down. A abordagem bottom-up começa com cada objeto a formar um grupo isolado, onde, sucessivamente, os objetos próximos uns dos outros fundem-se e formam um só (o nível mais alto da hierarquia). A abordagem top-down começa com todos os elementos de um cluster, sendo que, sucessivamente, esse cluster é dividido em clusters menores. Finaliza quando um cluster possuir apenas um elemento;
- **Métodos baseados em densidade**: estes métodos podem ser usados para filtrar o ruído ou *outliers* e descobrir clusters de forma arbitrária. Estes métodos dividem um conjunto de objetos em múltiplos clusters exclusivos, ou numa hierarquia de clusters. Tipicamente, os métodos baseados em densidade, apenas consideram clusters exclusivos;
- **Métodos baseados em Grid**: estes métodos quantificam o espaço do objeto num número finito de células, que formam uma estrutura Grid, sendo que todas as operações de segmentação são realizadas nesta estrutura. A principal vantagem deste método é a rapidez de processamento, que é independente do número de objetos e dependente apenas do número de células da estrutura Grid.

Para avaliar os resultados dos modelos de *Data Mining* baseados na segmentação, utilizam-se dois grandes fatores: a separabilidade e a *compactness*. A separabilidade mede a diferença que existe entre os clusters, pois estes devem estar amplamente separados. Existem três abordagens comuns que medem a distância entre dois grupos diferentes: distância entre o membro mais próximo do aglomerado, distância entre os membros mais distantes e a distância entre os centros dos

aglomerados. A *compactness* avalia o quão próximo o membro de um determinado cluster está em relação a outro. Uma medida comum de *compactness* é a variância. Para avaliar estes dois fatores utilizam-se duas métricas: *Davies-Bouldin Index* e o *Dunn Index*. Estas métricas fazem uma avaliação interna dos clusters (Kovács, Legány, & Babos, 2005).

IV. Associação ou Dependência

A associação ou dependência pretende encontrar um modelo que descreva dependências entre variáveis e cria um modelo de associações. As associações surgem quando várias ocorrências estão ligadas a um único evento, ou seja, podem ser identificados através de padrões e tendências.

Geralmente, a associação é usada em aplicações onde se procura identificar dados que possam ser colocados juntos do mesmo pacote de negociações (M. F. Santos & Azevedo, 2005, p. 17).

V. Sumarização

A sumarização utiliza métodos para encontrar uma descrição compacta para um subconjunto de dados. As técnicas de sumarização são aplicadas à análise exploratória de dados e à geração automática de relatórios (M. F. Santos & Azevedo, 2005, p. 17).

VI. Visualização

O método de visualização é responsável pela visualização gráfica dos resultados de uma análise, sejam estes os resultados intermédios ou finais. Os dados são apresentados sob a forma de diagramas e gráficos, de forma explícita, que entre si têm complexidade. Quanto melhor for a descrição de um conjunto de informações, melhor será a compreensão do domínio em que está inserido (M. F. Santos & Azevedo, 2005, p. 17).

Por exemplo, histogramas, gráficos de dispersão, gráficos de curvas, gráficos 3D, mapas sensoriais, gráficos de superfície, entre muitas outras possibilidades.

VII. Deteção de desvios (*outliers*)

A deteção de desvios identifica alterações significativas nos dados a partir de previsões anteriores ou valores normativos. Dentro de um conjunto de dados, a deteção de desvios localiza possíveis anormalidades dos dados.

Por exemplo, determinado grupo de pessoas costuma ir às compras depois de passar no banco, mas na semana passada foi ao cinema (M. F. Santos & Azevedo, 2005, p. 17).

2.4. Trabalhos Relacionados

Em 2017, foi realizada uma revisão de artigos que aplicam técnicas de *Data Mining* em Cardiologia. Esta revisão foi publicada no *International Journal of Medical Informatics*. Os artigos revistos são compreendidos entre 1 de janeiro de 2000 a 31 de dezembro de 2015. Foram identificados 2392 artigos, sendo que após alguns critérios de exclusão foram analisados 149 artigos. Neste estudo, existe uma tabela que resume o score dado a cada um desses artigos (Kadi, Idri, & Fernandez-aleman, 2017, p. 25). Essa tabela está apresentada em anexo (Tabela 19), sendo que apresenta uma estrutura com dez colunas. Na primeira coluna estão apresentados os títulos dos artigos selecionados. A segunda coluna, indica a técnica de *Data Mining* utilizada (RQ1). A terceira coluna, indica o desempenho geral das técnicas selecionadas (RQ3). A avaliação dos resultados é composta por 7 colunas, respondendo às seguintes questões:

- Primeira coluna (A): o estudo apresenta evidências empíricas que são analisadas quantitativamente ou qualitativamente? As opções de resposta são: quantitativamente (1); qualitativamente (0).
- Segunda coluna (B): o estudo apresenta um desenho experimental bem detalhado e justificado? As opções de resposta são: sim (1); parcialmente (0.5); não (0).
- Terceira coluna (C): o estudo mede e relata a taxa de acuidade do método experimental? As opções de resposta são: sim (1); não (0).
- Quarta coluna (D): o estudo apresenta uma comparação entre os métodos empíricos e outros método? As opções de resposta são: sim (1); não (0).
- Quinta coluna (E): O estudo analisa de forma bem explícita os benefícios e limitações do estudo? As opções de resposta são: sim (1); parcialmente (0.5); não (0).

- Sexta coluna (F): O estudo foi publicado de forma estável e reconhecida? As respostas variam entre 0 e 2.
- Sétima coluna (Score): soma das colunas de A à F.

Depois de analisados todos os artigos presentes na Tabela 19 apenas foram considerados os artigos parecidos com este projeto recentes. Uma vez que ainda existia uma grande quantidade de artigos para analisar, foram definidos alguns critérios de seleção:

- Estudo quantitativo;
- Possuir um score maior ou igual a 6.5;
- Artigos datados desde 2009.

Após aplicados os critérios de seleção surge, então a Tabela 6. Nesta tabela estão apresentados os 6 artigos que obtiveram melhor score. Tanto a Tabela 6 como a tabela em anexo (Tabela 19), apresentam a mesma estrutura.

Tabela 6 - Estudos Considerados. Adaptado: (Kadi et al., 2017)

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal (Mohebbi & Ghassemian, 2012)	SVM	93.1	1	1	1	1	0.5	2	6.5
Discrimination Power of Short-Term Heart Rate Variability Measures for CHF Assessment (Pecchia, Melillo, Sansone, & Bracale, 2011)	CART	79.3	1	1	1	1	0.5	2	6.5
Support Vectors Machine-based identification of heart valve diseases using heart sounds (Maglogiannis, Loukis, Zafiropoulos, & Stasis, 2009)	SVM	93.4	1	1	1	1	1	2	7
Ventricular fibrillation and tachycardia classification using a machine learning approach (Li, Rajagopalan, & Clifford, 2014)	SVM	98.1	1	1	1	1	1	1.5	6.5
Multiclass maximum margin clustering via immune evolutionary algorithm for automatic diagnosis of electrocardiogram arrhythmias (Zhu, Ding, & Hao, 2014)	Clustering	97.4	1	1	1	1	0.5	2	6.5
Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images (Acharya et al., 2013)	K-NN	86	1	1	1	1	0.5	2	6.5
A heart disease recognition embedded system with fuzzy cluster algorithm (de Carvalho, Moreno, Pimenta, Crepaldi, & Cintra, 2013)	F-Clust	92	1	1	1	1	0.5	2	6.5

Em 2014, existe um artigo, que propõe um algoritmo de classificação utilizando um método de aprendizagem mecânica, SVM. Foram utilizadas três base de dados: a base de dados da associação americana do coração, a base de dados da taquicardia ventricular da universidade de Creighton e a base de dados da arritmia ventricular maligna. Foram testadas diferentes métricas e diferentes tamanhos de janela, variando de 1 a 10. Obteve-se uma acuidade de 98.1%, uma sensibilidade de 98.4% e uma especificidade de 98%, nos dados de treino, com uma janela de tamanho 5 (Li et al., 2014).

Um artigo, também em 2014, propõe um método de segmentação baseado no algoritmo de clustering de margem máxima e no algoritmo evolutivo imunológico para diagnóstico de arritmias. Foram analisados cinco tipos de arritmias e foram utilizados três tipos de indicadores de avaliação de desempenho, para avaliar o efeito do método de segmentação. Os três indicadores eram a sensibilidade, especificidade e acuidade. Este método refletiu melhor desempenho em termos de capacidade de procura e de capacidade de convergência, comprovando a sua eficácia na deteção de arritmias (Zhu et al., 2014).

Para detetar eficientemente as doenças da artéria coronária, a partir de eletrocardiogramas, foram exploradas algumas técnicas de *Data Mining* baseadas nos computadores. Assim, neste estudo, foi desenvolvida e apresentada uma dessas técnicas, para a classificação de casos normais e casos afetados com a doença. Foram extraídas muitas características dos eletrocardiogramas. Estes eletrocardiogramas pertenciam a uma base de dados com 400 casos normais e 400 casos com a doença da artéria coronária. Com o classificador Gaussian Mixture Model, os resultados obtidos foram os que tiveram maior acuidade, sensibilidade e especificidade (Acharya et al., 2013).

Outro artigo, apresenta um sistema de identificação de doenças cardíacas. Este sistema oferece uma redução do processamento de um eletrocardiograma, sendo que o seu objetivo é análise dos sinais cardíacos. Os eletrocardiogramas são aplicados no sistema que executa uma filtragem e, em seguida, utiliza o algoritmo Gufstafson-Kessel fussy para a classificação do sinal cardíaco. A classificação indica doenças cardíacas comuns, utilizando a base de dados europeia de eletrocardiogramas. Os resultados provam que o sistema pode realizar deteções de doenças cardíacas, num conjunto pequeno de dados (de Carvalho et al., 2013).

Para além deste estudo, existe outro que se considera importante referir, que tem como objetivo a prevenção dos doentes com arritmias cardíacas através de técnicas de *Data Mining*. Este estudo,

realizado em 2014, permite ajudar os clínicos a agir de forma proativa, ajudando na tomada de decisão. Como variáveis de entrada foram consideradas as adquiridas na admissão do doente e as variáveis complementares, como os sinais vitais, resultados laboratoriais e as terapêuticas, coletadas por hora. Foram considerados oito cenários diferentes pela combinação de diferentes variáveis dos dados do doente. O modelo mais sensível obteve um resultado de 95.58% e o mais exato um resultado de 84.89% (Portela et al., 2014).

3. Abordagem Metodológica e Ferramentas Utilizadas

Neste capítulo são apresentadas todas as abordagens utilizadas para o desenvolvimento desta dissertação, assim como, as ferramentas principais de modelação e transformação dos dados.

3.1. Abordagens Metodológicas

A elaboração deste projeto de investigação está assente em duas metodologias: *Cross industry Standard Process for Data Mining* (CRISP-DM) e *Design Science Research* (DSR) como metodologia de investigação. O CRISP-DM é a metodologia usada na a base de descrição deste trabalho, sendo apresentado em detalhe na secção 3.1.2..

3.1.1. Design Science Research (DSR)

Esta metodologia é utilizada para a realização de pesquisas no âmbito dos Sistemas de Informação (SI). Tem como objetivo o desenvolvimento de artefactos que podem ser aplicados por profissionais de SI em projetos semelhantes (Peffer, Tuunanen, Rothenberger, & Chatterjee, 2007, p. 11). Na Figura 8 é apresentado o processo de metodologia de DSR. Inicialmente, o utilizador tem que definir o problema e mostrar a sua importância. De seguida, tem que definir quais os melhores artefactos a utilizar e escolher o que faz mais sentido, tendo em conta o problema a resolver. Por fim, é necessário avaliar se a solução é eficiente e eficaz e comunicar os problemas através de publicações.

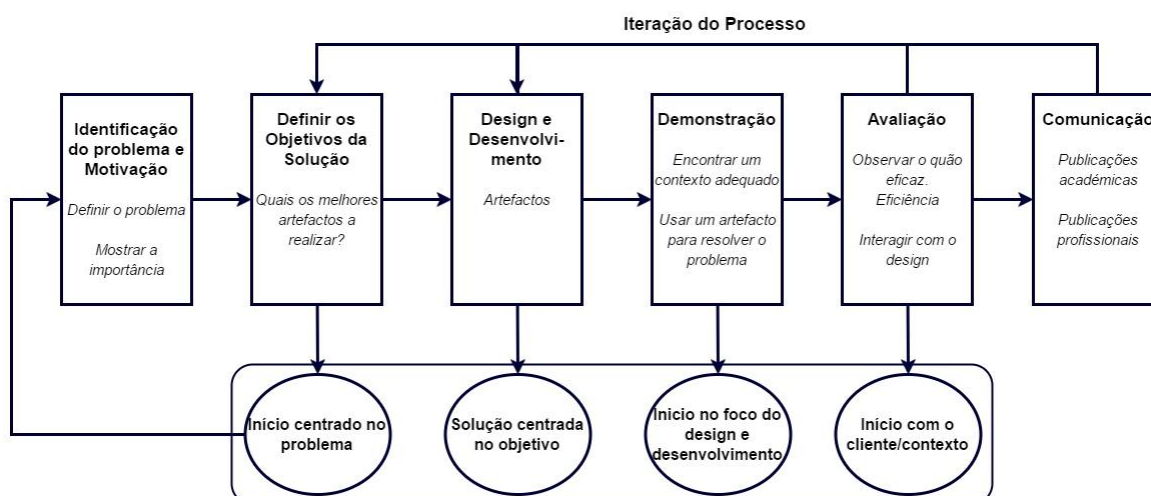


Figura 8 - Fases da Metodologia Design Science Research. Adaptado: (Peffer et al., 2007)

I. Identificação do problema e motivação

Nesta fase, é necessário definir o problema de pesquisa e uma solução, devidamente explicada. Uma vez que a solução do problema será utilizada para desenvolver um artefacto, pode ser útil atomizar o problema conceitualmente. Assim, a solução consegue capturar a complexidade do problema. Ao justificar o valor da solução, motiva o pesquisador e o público da pesquisa a procurar a solução e aceitar os resultados. Da mesma forma, ajuda o público da pesquisa a perceber o raciocínio do pesquisador sobre o problema. Os recursos necessários para esta atividade são o conhecimento do estado do problema e a importância da sua solução (Peppers et al., 2007, p. 12).

II. Definir os objetivos da solução

Nesta fase, é necessária uma definição dos objetivos da solução, tendo em conta a definição do problema e conhecimento daquilo que é possível e viável. Os objetivos podem ser quantitativos, por exemplo, condições em que uma solução desejável seria melhor que as outras atuais, ou qualitativos, por exemplo, uma descrição de como um novo artefacto deverá suportar soluções para problemas não abordados até agora. Os objetivos devem ser inferidos racionalmente a partir da especificação do problema. O recurso necessário para esta atividade é o conhecimento do estado do problema e das soluções atuais existentes, assim como, a sua eficácia (Peppers et al., 2007, p. 12).

III. Design e desenvolvimento

Esta fase, consiste na criação dos artefactos, sendo construções, métodos, modelos, instâncias ou novas propriedades de recursos técnicos, sociais e informacionais. Um artefacto pode ser qualquer objeto projetado onde uma contribuição de pesquisa está incorporada no design. Nesta fase, é necessário determinar a funcionalidade desejada do artefacto e a sua arquitetura e, posteriormente, a criação real do artefacto. O recurso necessário para esta atividade é o conhecimento da teoria que pode ser levado a uma solução (Peppers et al., 2007, p. 13).

IV. Demonstração

Nesta fase é demonstrado o uso do artefacto para resolver uma ou mais instâncias do problema, através do seu uso na experiência, simulações, caso de estudos, provas ou outras atividades apropriadas. O conhecimento de como utilizar o artefacto para resolver o problema, é o recurso necessário para esta atividade (Peppers et al., 2007, p. 13).

V. Avaliação

Nesta fase, é observado e medido o quão bem o artefacto suporta a solução do problema, sendo que são comparados os objetivos da solução com os resultados, efetivamente, observados na fase anterior. Dependendo da natureza do artefacto ou do problema, a avaliação pode assumir várias formas. Pode incluir: a comparação das funcionalidades do artefacto com os objetivos da solução, medidas objetivas de desempenho quantitativo, resultados de pesquisas de satisfação, feedback dos clientes ou simulações. Conceitualmente, a avaliação pode incluir qualquer evidência empírica apropriada ou provas lógicas. No final desta etapa os pesquisadores decidem se querem voltar à fase do design e desenvolvimento. Isto pode acontecer, com vista a melhorar a eficácia do artefacto ou para deixar mais melhorias para projetos subsequentes (Peppers et al., 2007, p. 13).

VI. Comunicação

Nesta fase, é comunicado a importância do problema, a utilidade do artefacto, bem como, o rigor do design e a sua efetividade para os pesquisadores e outros profissionais. Em pesquisas académicas, os pesquisadores podem usar a estrutura deste processo para estruturar o artigo, assim como a estrutura nominal de um processo de pesquisa experimental (definição de problema, revisão de literatura, desenvolvimento de hipóteses, recolha dos dados, análise, resultados, discussão e conclusão). A comunicação requer conhecimento da cultura disciplinar (Peppers et al., 2007, p. 14).

No caso deste projeto de dissertação, inicialmente o problema consistia na dificuldade em prever atempadamente um doente com doenças coronárias, tendo em conta os fatores de risco das doenças cardiovasculares. Devido a uns problemas de atraso na entrega dos dados por parte do HSO (Hospital Santo António) e HSOG (Hospital Senhora da Oliveira, Guimarães) e para combater a ausência de dados, foram fornecidos outros conjuntos de dados. Assim sendo, o problema agora

consiste na dificuldade em prever as variações nos batimentos cardíacos de um determinado utilizador. A solução passa por desenvolver um modelo de previsão que apresente de forma atempada e eficiente se determinado doente poderá ter uma variação dos batimentos cardíacos. Isto com o objetivo de perceber se as alterações dos batimentos cardíacos, causados pelo exercício físico, podem ajudar na prevenção de doenças cardíacas.

Depois de desenvolvido este modelo, serão realizados testes para perceber se os resultados vão de encontro com os objetivos definidos e interpretados os resultados, representando o fim deste projeto de dissertação.

3.1.2. Cross Industry Standard Process for Data Mining (CRISP-DM)

Esta é uma metodologia de *Data Mining* que descreve abordagens usadas pelos especialistas para resolver diversos problemas.

O CRISP-DM foi concebido nos finais de 1996, pelo consórcio formado pelas empresas NCR (Estados Unidos da América e Dinamarca), DaimlerChrysler AG (Alemanha), SPSS Inc. (Estados Unidos da América) e OHRA (Grupo Bancário Holandês) (Chapman et al., 2000, p. 1).

Este modelo consiste em seis fases, com setas indicando as dependências mais importantes e que ocorrem com mais frequência, entre as fases. A sequência das fases não se segue rigorosamente, uma vez que é possível retroceder a uma fase anterior sempre que se desejar. O resultado de cada fase, indica qual a próxima fase a ser executada (IBM, 2011, p. 1).

O círculo externo, representado na Figura 9, simboliza a natureza cíclica de um projeto de *Data Mining*. Isto significa que, um projeto de *Data Mining* não termina quando a solução é implementada, uma vez que o projeto pode desencadear novas questões de negócio, iniciando um novo ciclo (Chapman et al., 2000, p. 10).

O CRISP-DM é o mais usual em problemas que envolvam *Data Mining*. Segundo o KDnuggets (2014), o CRISP-DM tem uma taxa de preferência de 43%, entre os profissionais da área.

Manuel Filipe Santos e Carla Azevedo (2005, p. 34) dizem que “a metodologia CRISP-DM é extremamente completa e documentada, uma vez que as suas fases estão devidamente organizadas, estruturadas e definidas, permitindo que o projeto possa ser facilmente compreendido ou revisto”.

O CRISP-DM divide o ciclo de vida de um projeto de *Data Mining* em seis fases: Compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e implementação. Sendo estas fases explicadas de seguida.

A Figura 9 representa, de forma esquematizada, as fases de um projeto de *Data Mining* tendo em conta o CRISP-DM.

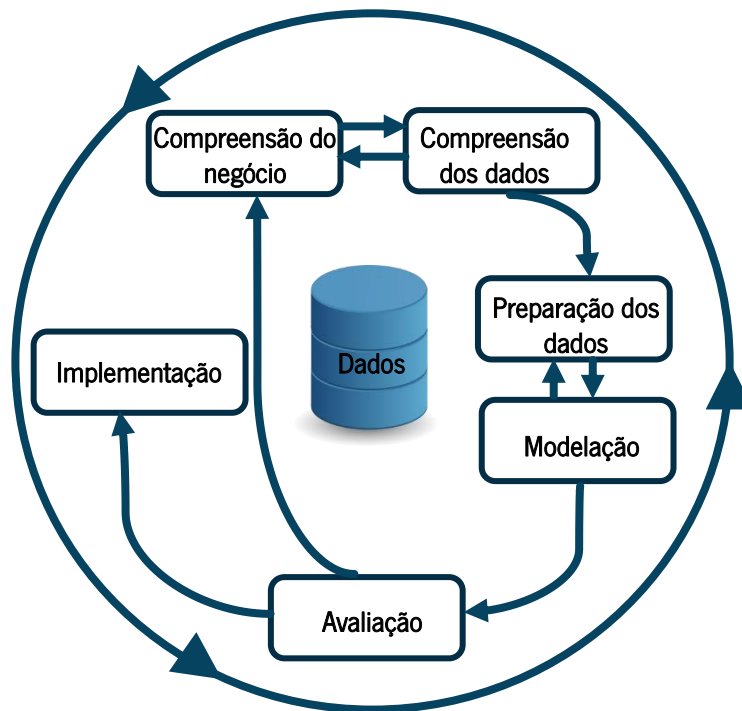


Figura 9 - Etapas da metodologia CRISP-DM. Adaptado (IBM, 2011, p. 9)

I. Compreensão do negócio

A primeira fase de um projeto de *Data Mining* centra-se na compreensão dos objetivos e requisitos do projeto, tendo o negócio como perspetiva. Este conhecimento é convertido numa definição de um problema de *Data Mining* e num plano inicial. Este plano inicial é criado com o intuito de atingir os objetivos de negócio (Chapman et al., 2000, p. 10).

Inicialmente, num projeto de *Data Mining*, é necessário perceber a necessidade de realizar um projeto deste tipo, compreender a perspetiva do problema e descobrir fatores que podem influenciar os resultados (M. F. Santos & Azevedo, 2005, p. 27).

O principal elemento de um estudo de *Data Mining* é saber para que serve esse estudo. Para isso, é necessário perceber quais são os objetivos de negócio tendo em conta o estudo a ser realizado.

Nesta fase, deve ser estabelecido um orçamento para servir de apoio ao estudo (Olson & Delen, 2008, p. 11).

Na Figura 10 estão apresentadas as subfases da compreensão do negócio: inicialmente é necessário determinar os objetivos de negócio e avaliar a situação atual, sendo que, posteriormente, é necessário definir os objetivos de negócio e produzir o plano de projeto.

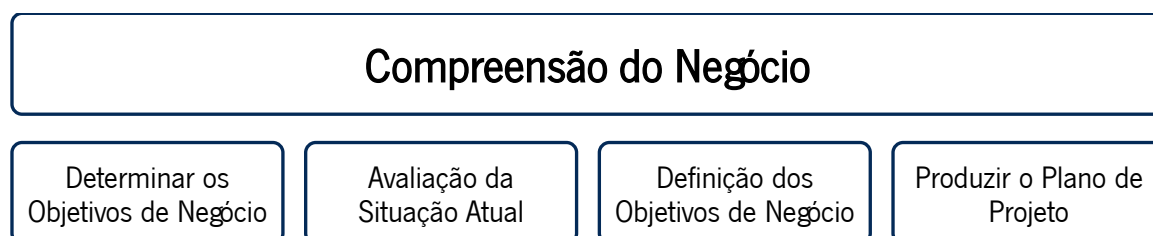


Figura 10 - Subfases da Fase de Compreensão do Negócio do CRISP-DM. Adaptado: (M. F. Santos & Azevedo, 2005, p. 28)

Tendo em conta este projeto, nesta fase foi explorado o projeto DEM (*Deus ex Machine*), de forma a definir os objetivos e os resultados atingidos.

II. Compreensão dos Dados

Esta fase inicia-se com a recolha inicial dos dados e prossegue com a sua análise, de forma a identificar problemas de qualidade dos dados e descoberta de subconjuntos interessantes, para formar hipótese sobre informações escondidas (Chapman et al., 2000, p. 10).

Existem três questões a serem consideradas na seleção de dados. A primeira questão consiste em criar uma descrição concisa e clara do problema. A segunda questão passa pela identificação dos dados relevantes para a descrição do problema. Por fim, mas não menos importante, as variáveis selecionadas para os dados relevantes devem ser independentes. Isto quer dizer que as variáveis não contêm informações sobrepostas. Isto torna mais fácil a descoberta de padrões pelos algoritmos de *Data Mining* (Olson & Delen, 2008, p. 12).

Na Figura 11 estão apresentadas as subfases da compreensão dos dados: recolha inicial dos dados, descrição dos dados, exploração dos dados e verificação da qualidade dos dados. Estas subfases estão explicadas ao longo desta subsecção.

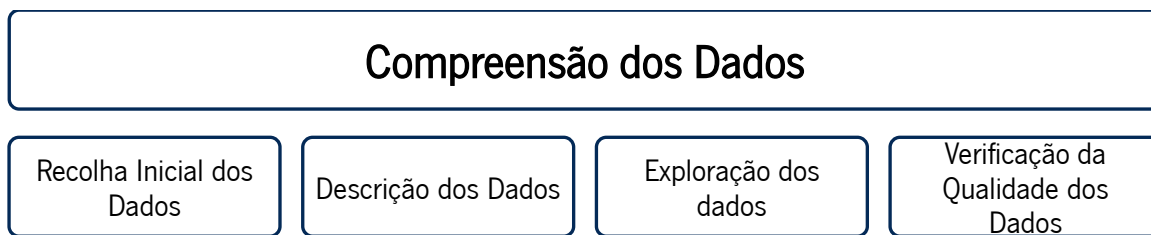


Figura 11 - Subfase da Fase de Compreensão dos Dados do CRISP-DM. Adaptado: (M. F. Santos & Azevedo, 2005, p. 29)

De forma resumida e tendo em conta este projeto de dissertação, nesta fase, foram analisados os dados fornecidos pelo projeto DEM, com o objetivo da descrição de cada um dos atributos, assim como a descoberta de problemas nos dados;

III. Preparação dos dados

Esta fase, envolve todas as atividades necessárias para a construção do conjunto de dados final, aquele que será usado na ferramenta de modelação, a partir do conjunto inicial dos dados. Esta fase é realizada várias vezes, até que se obtenham resultados considerados bons. A preparação dos dados inclui a seleção de atributos, bem com a transformação e limpeza dos dados (Chapman et al., 2000, p. 11).

O objetivo desta fase é limpar os dados selecionados para uma melhor qualidade dos mesmos. A limpeza dos dados consiste em filtrar, agregar e preencher valores omissos. Através da filtragem dos dados, os dados são examinados, de modo a descobrir *outliers* e redundância. Os *outliers* são causados por muitas razões, tais como erros humanos ou técnicos. Os *outliers* são valores que diferem muito da maioria dos dados. A redundância consiste em informações registadas várias vezes de maneira diferente. Através da agregação de dados, as dimensões dos dados serão reduzidas, obtendo-se informações agregadas. Embora o conjunto de dados tenha um volume pequeno, as informações permanecerão (Olson & Delen, 2008, p. 12).

Na Figura 12 estão apresentadas as subfases da preparação dos dados: seleção dos dados, limpeza dos dados, construção dos dados, integração dos dados e formatação dos dados. Estas subfases estão descritas ao longo desta subsecção.

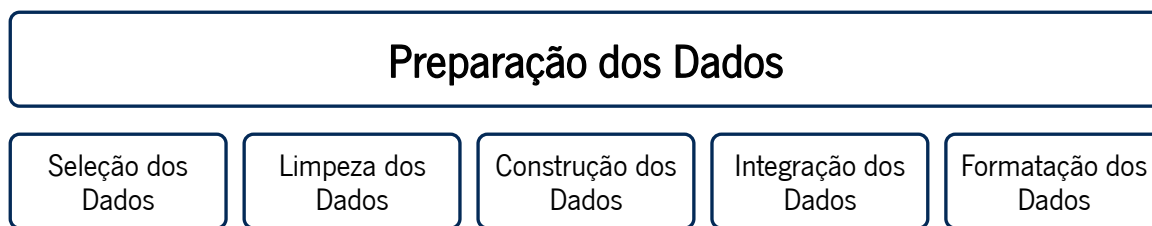


Figura 12 - Subfase da Fase de Preparação dos Dados do CRISP-DM. Adaptado: (M. F. Santos & Azevedo, 2005, p. 30)

Neste projeto de dissertação e para esta fase, foram realizadas um conjunto de tarefas para o tratamento dos dados, como a seleção dos atributos a utilizar, assim como a transformação e limpeza dos dados.

IV. Modelação

Nesta fase, várias técnicas de modelação são selecionadas e aplicadas, e os seus parâmetros são ajustados de forma a otimizar os resultados, tendo em conta o estudo. Diferentes modelos são criados para o mesmo tipo de problema de *Data Mining* (Chapman et al., 2000, p. 11).

A modelação de dados é onde a aplicação de *Data Mining* é usada para conceber resultados para várias situações. Dependendo do tipo de dados, existem várias abordagens que podem ser aplicadas. (Olson & Delen, 2008, p. 15). Na secção 2.3.5 são abordados, ao pormenor, as várias abordagens e técnicas de *Data Mining*.

Na Figura 13 estão apresentadas as subfases da Modelação: Seleção das Técnicas de Modelação, Criação de uma Conceção de Teste, Construção dos Modelos e Revisão dos Modelos.

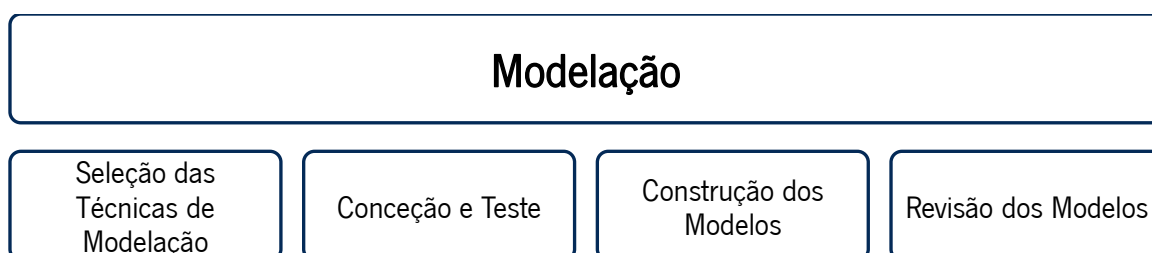


Figura 13 - Subfase da Fase de Modelação do CRISP-DM. Adaptado: (M. F. Santos & Azevedo, 2005, p. 31)

De forma resumida e tendo em conta este projeto de dissertação, nesta fase foram selecionadas as técnicas de modelação utilizadas assim como foram aplicadas as técnicas aos modelos de dados.

V. Avaliação

Nesta fase do projeto, existe uma construção de um modelo que tem bons resultados, tendo em conta a análise dos dados. Assim, é necessário rever todas as fases anteriores, para ter a certeza de que o modelo atinge adequadamente os objetivos de negócio (Chapman et al., 2000, p. 11).

Para interpretar corretamente os padrões é importante escolher uma ferramenta de visualização adequada. Muitas ferramentas de visualização estão disponíveis, incluindo gráficos de pizza, gráficos de dispersão e histogramas. Quanto mais fácil for a interpretação gráfica, mais fácil será para os utilizadores entenderem. Uma boa interpretação destes padrões, leva a decisões de negócio produtivas (Olson & Delen, 2008, p. 18).

Na Figura 14 estão apresentadas as subfases da Avaliação: Avaliação dos resultados, revisão do processo e determinação dos próximos passos. Estas subfases estão explicadas ao longo desta subsecção.

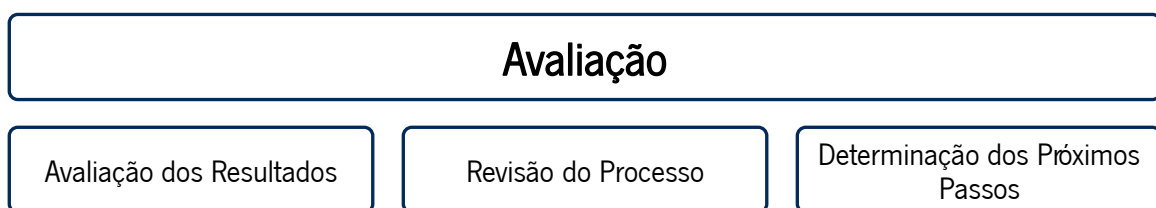


Figura 14 - Subfase da Fase de Avaliação do CRISP-DM. Adaptado: (M. F. Santos & Azevedo, 2005, p. 32)

Nesta fase foram analisados todos os resultados obtidos nas fases anteriores, assim como as técnicas e métricas de avaliação aplicadas aos modelos de *Data Mining*.

Depois de executados cada um dos modelos foi feita uma revisão de modo a tentar melhorar os resultados obtidos.

VI. Implementação

A criação do modelo não significa o fim do projeto. O conhecimento adquirido terá de ser organizado e apresentado de alguma forma, para que o utilizador final o possa utilizar. Dependendo dos requisitos, esta fase pode ser simples, como a criação de um relatório, ou bastante complexa, como implementar todo o processo de *Data Mining* (Chapman et al., 2000, p. 11).

O comportamento do utilizador final muda ao longo do projeto e o que era correto na seleção dos dados pode já ter mudado. Se ocorrerem mudanças significativas, o conhecimento criado já não é verdadeiro. Assim, é importante que o comportamento do utilizador final seja monitorizado ao longo da fase de implementação (Olson & Delen, 2008, p. 19).

Na Figura 15 estão apresentadas as subfases da Implementação. Estas subfases são o Planeamento da Avaliação dos Resultados, o Planeamento da Monitorização e Manutenção, a Produção do Relatório Final e a Revisão do Projeto, sendo estas subfases explicadas ao longo desta subsecção.

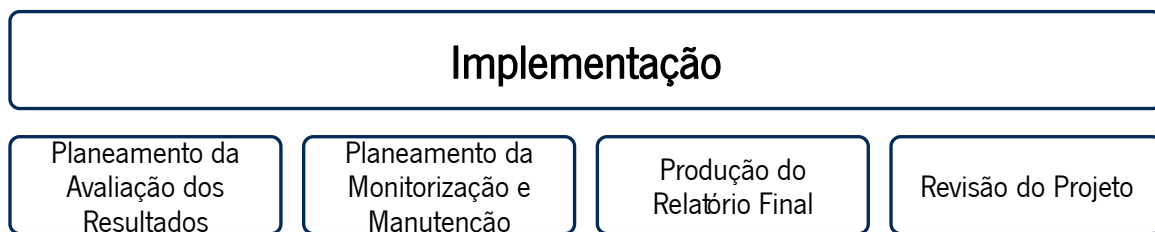


Figura 15 - Subfases da Fase de Implementação do CRISP-DM. Adaptado: (M. F. Santos & Azevedo, 2005, p. 33)

Esta última fase não será executada neste projeto de investigação.

3.2. Ferramentas Utilizadas

Neste capítulo é apresentado um pequeno resumo de todas as ferramentas utilizadas ao longo deste projeto de dissertação. Este capítulo é dividido em duas secções:

- Ferramentas de Tratamento de Dados – Ferramentas utilizadas no tratamento de dados e para a realização do ETL (*Extract, Transforming and Loading*);
- Ferramentas de *Data Mining* (DM) – Ferramentas utilizadas para a modelação dos vários cenários propostos;

Estas duas secções serão explicadas de seguida.

3.2.1. Ferramentas de Tratamentos dos Dados

Nesta secção serão apresentadas todas as ferramentas utilizadas no tratamento de dados, assim como na elaboração do ETL. Para a elaboração do tratamento dos dados foi utilizado o Talend.



Figura 16 - Logotipo Talend

O Talend é um software open source que permite às empresas avaliar a qualidade dos dados contidos nas bases de dados e decidir quais ações devem ser tomadas para corrigir dados errados ou incompletos.

O Talend atende às seguintes necessidades do utilizador:

- Descoberta de metadados - que identificam a estrutura das bases de dados que precisam ser analisados.
- Definição de estatísticas - que definem as estatísticas e métricas que precisam ser medidas em cada item de dados.
- Resultados e gráficos, que facilitam a visualização dos resultados e avaliam o nível de qualidade dos dados¹⁴.

¹⁴ <https://www.talend.com/>

3.2.2. Ferramentas de Data Mining

Nesta secção serão apresentadas as ferramentas de *Data Mining* (DM) utilizadas neste projeto de dissertação. Estas ferramentas foram responsáveis por executar os vários modelos propostos, com o objetivo de saber qual o modelo que apresenta melhores resultados.



Figura 17 - Logotipo Weka

Para a modelação das técnicas de DM foi utilizado o software Weka.

O WEKA é um conjunto de algoritmos de machine learning para tarefas de data mining. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou podem ser chamados através de código Java. Neste projeto, os algoritmos são aplicados diretamente a um conjunto de dados. O WEKA contém ferramentas de pré-processamento de dados, classificação, regressão, agrupamento e regras de associação e visualização¹⁵.

Weka é um software open source emitido sob a *GNU General Public License*.

Os únicos requisitos para este software é a versão do Java. A Matriz XXX mostra a versão mínima necessária para executar uma versão weka específica, bem como, a versão WEKA e Java utilizados neste projeto.

Tabela 7 - Requisitos Weka¹⁶

		JAVA				
		1.4	1.5	1.6	1.7	1.8
WEKA	<3.4.0	X	X	X	X	X
	3.4.x	X	X	X	X	X
	3.5.x	3.5.0 à 3.5.2	>3.5.2	X	X	X
	3.6.x		X	X	X	X
	3.7.x		3.7.0	>3.7.0	>3.7.13	X

¹⁵ <https://www.cs.waikato.ac.nz/ml/weka/>

¹⁶ <https://www.cs.waikato.ac.nz/ml/weka/requirements.html>

		JAVA				
		1.4	1.5	1.6	1.7	1.8
	3.8.x				3.8.0 à 3.8.1	>3.8.1
	3.9.x				3.9.0 à 3.9.1	>3.9.1

4. Desenvolvimento

4.1. Compreensão do problema

As fontes de estudo como livros e bibliografias diversas foram bastante essenciais para perceber o mundo das doenças cardíacas, assim como quais as várias doenças associadas ao coração. Apesar deste projeto de dissertação ter mudado o seu contexto inicial, o estado de arte revelou-se importante, uma vez que vários conceitos estudados foram aplicados a este novo contexto.

Inicialmente o problema consistia na dificuldade em prever atempadamente um doente com doenças coronárias, tendo em conta os fatores de risco das doenças cardiovasculares. Devido a uns problemas de atraso na entrega dos dados por parte do HSO (Hospital Santo António) e HSOG (Hospital Senhora da Oliveira, Guimarães) e para combater a ausência de dados, foram fornecidos outros conjuntos de dados. Assim sendo, e devido à mudança de contexto, o objetivo deste estudo passa por perceber como o exercício físico ajuda no combate das doenças cardiovasculares, tendo como base as alterações dos batimentos cardíacos.

Hoje em dia, as pessoas são sedentárias não usufruindo do exercício físico da melhor maneira. Com este estudo pretende-se visualizar se, através do exercício, os utilizadores têm ou não variações dos batimentos cardíacos, podendo ajudar na prevenção de várias doenças associadas à não prática de exercício físico.

Assim, como este modelo, os profissionais de saúde, assim como os utilizadores podem ficar conscientes no quão importante é o exercício físico na prevenção das doenças cardiovasculares.

Para que esta dissertação fosse de fácil leitura, foi construído um glossário de termos apresentado no início deste documento, com algumas siglas técnicas.

4.2. Compreensão dos Dados

Inicialmente o conjunto de dados para o desenvolvimento deste projeto, correspondia a dados fornecidos pelo Hospital de Santo António e pelo Hospital Senhora da Oliveira em Guimarães. Devido a uns problemas, estes não foram entregues atempadamente. Para combater este problema, foram

fornecidos outros conjunto de dados, com o objetivo de perceber como o exercício físico ajuda no combate das doenças cardiovasculares.

A compreensão dos dados é uma tarefa importante, uma vez que na mesma se identifica o significado de cada um dos atributos disponíveis na base de dados.

Foram fornecidas 56 folhas de cálculo, sendo estes correspondentes a 7 utilizadores. Com o objetivo de uma análise mais fácil a estes conjuntos de dados, foi desenvolvido um processo ETL (*Extract, Transforming and Loading*) que identificasse o tipo de utilizador, com o objetivo de se poder identificar a quem pertence um determinado conjunto de dados. Este identificador, foi também uma maneira de conseguir fazer a ligação entre todos os ficheiros. Para além do identificador criado, a ligação entre todos os ficheiros deu-se tendo em conta o *timestamp*.

Parte dos ficheiros, eram dados retirados de um *smartwatch* utilizado neste teste (MIO).

Antes de integrar todos os ficheiros num único, foi necessário agrupar todos os ficheiros que continham a mesma informação num único ficheiro. Esta união deu-se tendo em conta a Figura 18, sendo que a azul mais forte estão apresentados os novos ficheiros criados e a azul mais clara estão apresentados os ficheiros originais deste projeto de dissertação. Por exemplo, o ficheiro *BLOOD PRESSURE*, contém informação dos seguintes ficheiros originais: *Blood Pressure Senior1*; *Blood Pressure Senior2*; *Blood Pressure Senior3*; *Blood Pressure Senior4*; *Blood Pressure Utilizador1*; *Blood Pressure Utilizador2*; *Blood Pressure Utilizador3*



Figura 18 - Esquema de Ficheiros

Para isto, foi necessário o desenvolvimento das seguintes tarefas (*jobs*):

- BP_UNION
- MIO_DA_UNION
- MIO_HR_DA_UNION
- MIO_HR_HO_UNION
- OXI_UNION
- QUEST_UNION
- WEI_UNION

A Figura 19 representa um ETL geral, criado com o objetivo de executar todos os jobs anteriormente descritos todos ao mesmo tempo.

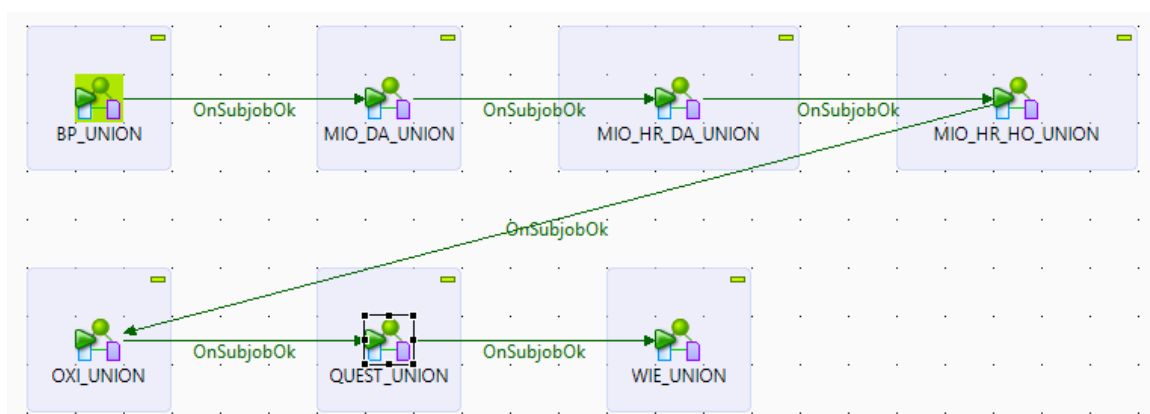


Figura 19 - ETL de todas as Uniãoes

Cada um destes jobs, representados na Figura 19, foi desenvolvido com o objetivo de juntar a informação comum dos diferentes utilizadores num único ficheiro, ficando com um total de 7 ficheiros diferentes. Tal como apresentado na Figura 20, para cada um dos ficheiros foi acrescentado um identificador, sendo representado este processo do tmap1 ao tmap7. Os identificadores para cada um dos ficheiros foram dados segundo as seguintes regras:

- Caso o ficheiro diga respeito ao sénior 1 é adicionado o identificador 'S1'
- Caso o ficheiro diga respeito ao sénior 2 é adicionado o identificador 'S2'
- Caso o ficheiro diga respeito ao sénior 3 é adicionado o identificador 'S3'
- Caso o ficheiro diga respeito ao sénior 4 é adicionado o identificador 'S4'
- Caso o ficheiro diga respeito ao utilizador 1 é adicionado o identificador 'U1'
- Caso o ficheiro diga respeito ao utilizador 2 é adicionado o identificador 'U2'
- Caso o ficheiro diga respeito ao utilizador 3 é adicionado o identificador 'U3'

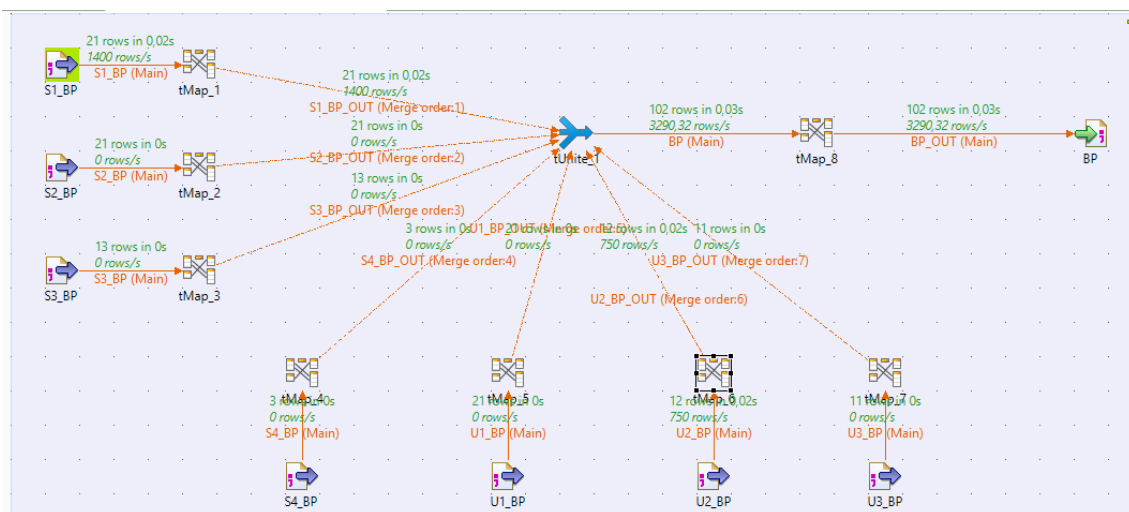


Figura 20 - ETL BP_UNION

O componente responsável pela junção dos ficheiros é o tunite, representado a azul na Figura 20. O output, representado com o nome BP na Figura 20, é um único ficheiro excel com os dados que, neste caso, dizem respeito à pressão arterial (*Blood Pressure*) dos vários utilizadores.

O ficheiro excel de output exemplificativo da Figura 20, está apresentado na Figura 21 b). A Figura 21 a) está apresentado a estrutura inicial do ficheiro.

	A	B	C	D		A	B	C	D	E	F
1	TIMESTAMP	SYSTOLE	DIASTOLE	HR		1	TIMESTAMP	SYSTOLE	DIASTOLE	HR	IDENTIFICADOR
2						2					
3	1,50E+12	163	89	76		3	1,50E+12	163	89	76	S1
4	1,50E+12	158	88	69		4	1,50E+12	158	88	69	S1
5	1,50E+12	142	88	81		5	1,50E+12	142	88	81	S1
6	1,50E+12	142	77	62		6	1,50E+12	142	77	62	S1
7	1,50E+12	146	82	64		7	1,50E+12	146	82	64	S1
8	1,50E+12	141	83	64		8	1,50E+12	141	83	64	S1
9	1,50E+12	129	81	63		9	1,50E+12	129	81	63	S1
10	1,50E+12	140	79	63		10	1,50E+12	140	79	63	S1

a)

b)

Figura 21 - Estrutura do ficheiro Blood Pressure inicial e final

4.2.1. Atributos de cada um dos ficheiros

Os atributos de cada um dos ficheiros diz respeito aos ficheiros resultantes da Figura 18. Assim, os ficheiros resultantes desta figura são os seguintes:

- Pressão Arterial (*Blood Pressure.csv*) – Neste ficheiro estão apresentados todos os atributos representativos da pressão arterial, como a diástole, sistole e HR (Batimentos Cardíacos) de cada um dos utilizadores.

- Pedómetro (*MIO Daily.csv*) – Neste ficheiro são apresentados a quantidade de passos de cada um dos utilizadores, tendo em conta os dias de medição.
- Batimentos Cardíacos Diários (*MIO HR Daily.csv*) – Neste ficheiro são apresentadas as estatísticas de batimentos cardíacos diários dos vários utilizadores do *smartwatch*.
- Batimentos Cardíacos Horários (*MIO HR Hourly.csv*) – Neste ficheiro são apresentadas as estatísticas de batimentos cardíacos horários dos vários utilizadores do *smartwatch*. Estas medições foram realizadas, no máximo, durante dois dias.
- Oxímetro (*Oximeter.csv*) – Neste ficheiro são apresentadas as medições da concentração de oxigénio dos vários utilizadores.
- Questionário (*Questionnaire.csv*) – Neste ficheiro são apresentadas as respostas dos vários utilizadores às questões realizadas durante os vários dias de teste.
- Peso (*Weight.csv*) – Neste ficheiro são apresentados os pesos medidos ao longo das várias medições dos vários participantes deste teste.

Para cada um dos ficheiros foi desenvolvida uma tabela que apresenta uma breve descrição sobre os atributos, assim como os intervalos de valores possíveis. Para além disso, é apresentado a percentagem de nulos, distintos e nº de registos.

Os nomes presentes nos atributos, são os nomes originais presentes nas folhas de cálculo.

O identificador do Doente está em texto, portanto não dá para definir o mínimo e o máximo. Assim, foi atribuído um N/A (Não Atribuído) a este atributo.

i. Pressão Arterial

O ficheiro *Blood Pressure.csv* contém um total de 95 registos, sendo que na Tabela 8 estão apresentados os atributos deste ficheiro, bem como uma breve descrição dos mesmos e intervalos de valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 8 - Atributos do ficheiro da Pressão Arterial

Atributo	Descrição	Mínimo	Máximo	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milesegundos.	1496830 000000	1503312 686889	0	88
Systole	Mede o período de contração muscular das camaras do coração.	82	185	0	58
Diastole	Mede o período de relaxamento muscular do coração.	41	157	0	51
HR	Mede os batimentos cardíacos	35	82	0	31
Identificador	Identificador do Doente	N/A	N/A	0	7

ii. Pedómetro

O ficheiro *MIO Daily.csv* contém um total de 125 registos, sendo que na Tabela 9 estão apresentados os atributos considerados para o ficheiro de passos medidos pelo *smartwatch*, bem como uma breve descrição dos mesmos e intervalos de valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 9 - Atributos do ficheiro do Pedómetro

Atributo	Descrição	Mínimo	Máximo	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milesegundos.	148046 0000000	150051 0000000	0	53
Steps	Passos efetuados nesse dia de medição	0	12406	0	67
Date_str	Data no formato DD/MM/YYYY.	29/11/2016	20/07/2017	0	60
Identificador	Identificador do Doente	N/A	N/A	0	7

iii. Batimentos Cardíacos Diários

O ficheiro *MIO HR Daily.csv* contém um total de 86 registos, sendo que na Tabela 10 estão apresentados os atributos considerados para o ficheiro dos batimentos cardíacos diários registados pelo *smartwatch*, bem como uma breve descrição dos mesmos e intervalos de valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 10 - Atributos do ficheiro dos Batimentos Cardíacos Diários

Atributo	Descrição	Mínimo	Máximo	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milesegundos.	149687 0000000	1503183 600000	0	51
MIN	Mínimo de batimentos cardíacos	0	69	0	9
MAX	Máximo de Batimentos Cardíacos	0	212	0	48
MEAN	Média de Batimentos Cardíacos	0	77	0	29
MEDIAN	Mediana de Batimentos Cardíacos	0	78	0	21
STD_DEV	Desvio padrão de Batimentos Cardíacos	0	46	0	25
IDENTIFICA DOR	Identificador do Doente	N/A	N/A	0	7

iv. Batimentos Cardíacos Horários

O ficheiro *MIO HR Hourly.csv* contém um total de 105 registos, sendo que na Tabela 11 estão apresentados os atributos considerados para o ficheiro dos batimentos cardíacos diários registados pelo *smartwatch*, bem como uma breve descrição dos mesmos e intervalos de valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 11 - Atributos do ficheiro dos Batimentos Cardíacos Horários

Atributo	Descrição	Mínimo	Máximo	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milissegundos.	149683 0000000	1503262 800000	0	105
MIN	Minimo de batimentos cardíacos	0	69	0	24
MAX	Maximo de Batimentos Cardíacos	0	212	0	43
MEAN	Média de Batimentos Cardíacos	0	77	0	25
MEDIAN	Mediana de Batimentos Cardíacos	0	78	0	22
STD_DEV	Desvio padrão de Batimentos Cardíacos	0	46	0	24
IDENTIFICADOR	Identificador do Doente	N/A	N/A	0	7

v. Oxímetro

O ficheiro *Oximeter.csv* contém um total de 91 registos, sendo que na Tabela 12 estão apresentados os atributos considerados para o ficheiro da concentração de oxigénio no sangue, bem como uma breve descrição dos mesmos e intervalos de valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 12 - Atributos do ficheiro do Oxímetro

Atributo	Descrição	Mínimo	Máximo	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milissegundos.	149683 0000000	150331 0000000	0	54
Spo2	Medição do nível de Oxigénio no Sangue	87	99	0	10
Identificador	Identificador do Doente	N/A	N/A	0	7

vi. Questionário

O ficheiro *Questionnaire.csv* contém um total de 94 registos, sendo que na Tabela 13 estão apresentados os atributos considerados para o ficheiro de questões efetuadas aos diferentes utilizadores, bem como uma breve descrição dos mesmos e os valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 13 - Atributos do ficheiro do Questionário

Atributo	Descrição	Valores Possíveis	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milisegundos.	1497430000000, 1497590000000 (...)	0	86
Q1	Desde de a ultima avaliação teve algum desmaio?	Sim; Não	0	2
Q2	Durante a noite, quando deitado, sentiu mais falta de ar?	Sim; Não	0	2
Q3	Durante a realização das suas atividades diárias, sentiu mais falta de ar?	Sim; Não	0	2
Q4	No final do dia de ontem, sentiu os pés mais inchados?	Sim; Não	0	2
Q5	Como classifica o seu estado de saúde hoje?	Bem; Muito bem	0	2
Identificador	Identificador do Doente	S1, S2, S3, S4, U1, U2, U3	0	7

vii. Peso

O ficheiro *Weight.csv* contém um total de 89 registos, sendo que na Tabela 14 estão apresentados os atributos considerados para o ficheiro do peso, bem como uma breve descrição dos mesmos e intervalos de valores possíveis. Para além disso é apresentado a percentagem de nulos e o nº de registos distintos de cada um dos registos.

Tabela 14 - Atributos do ficheiro do Peso

Atributo	Descrição	Mínimo	Máximo	% Nulos	Nº Distintos
Timestamp	O dia e hora em que o utilizador efetuou a medição. Tempo expresso em milissegundos.	1496826183157	1503311954227	0	87
WEIGHT_KG_FLOAT	Peso do utilizador no dia das medições, em quilograma	13.93	105.52	0	85
Identificador	Identificador do Doente	N/A	N/A	0	7

4.2.2. Análise Exploratória

De forma a facilitar o processo de análise dos dados, foram utilizados os gráficos dinâmicos presentes no Excel, para cada um das folhas de cálculo. Assim, as folhas de cálculo analisadas nesta secção foram as seguintes:

- Pressão Arterial (*Blood Pressure.csv*);
- Pedómetro (*MIO Daily.csv*);
- Batimentos Cardíacos Diários (*MIO HR Daily.csv*);
- Batimentos Cardíacos Horários (*MIO HR Hourly.csv*);
- Oxímetro (*Oximeter.csv*);
- Questionário (*Questionnaire.csv*);
- Peso (*Weight.csv*).

i. Pressão Arterial

O primeiro ficheiro a ser analisado é *Blood Pressure.csv*. Para este ficheiro foi realizada uma análise global, tendo em conta as médias de cada um dos atributos, assim como análise temporal.

Na Figura 22 está apresentada uma análise global das médias dos dados apresentados. Estão apresentadas as médias da Sístole, diástole e HR (Batimentos Cardíacos).

Tendo em conta a análise da Figura 22 pode-se concluir que, em média, o utilizador 3 (U3) possui os batimentos cardíacos mais elevados (média de 71.8). Por oposição o utilizador 1 (U1) possui os batimentos cardíacos mais baixos, em média. Em relação à sístole e à Diástole também é U3 que possui mais elevada e o U3 que a possui mais baixa. Tendo em conta a Sístole é o Sênior 3 (S3) que a possui mais elevada, em comparação com os restantes participantes.

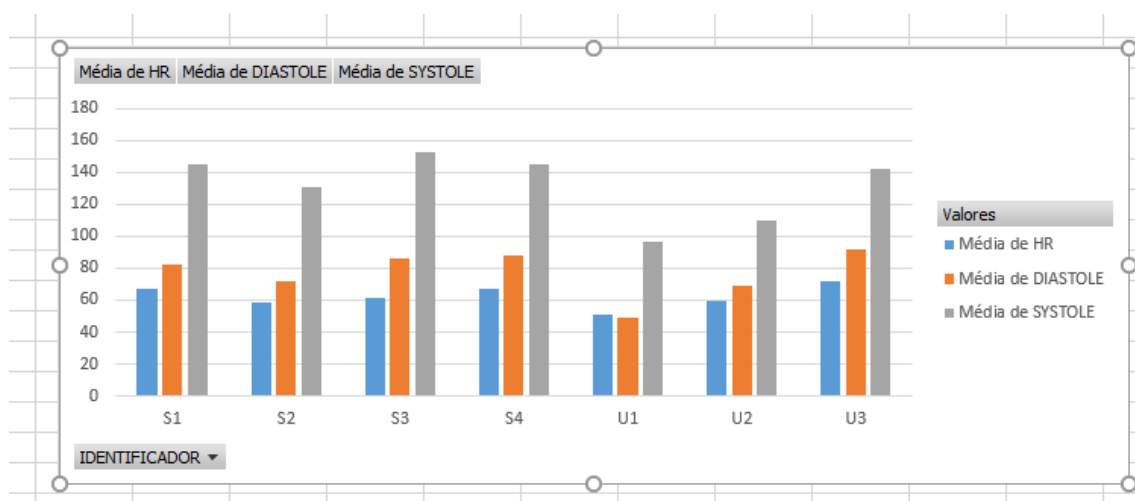


Figura 22 - Gráfico Global Pressão Arterial

Na Figura 23 está apresentada a evolução da sístole, diástole e batimentos cardíacos dos participantes do estudo ao longo do tempo de análise.



Figura 23 - Gráfico Evolução Temporal da Pressão Arterial

Depois de análise desta figura, verifica-se que este sénior 1 (S1) e o utilizador 3 possuem uma variação da sístole, diástole e batimentos cardíacos constantes ao longo do tempo de medição, não existindo uma grande variação dos mesmos.

Verifica-se também que o sénior 2 (S2) não possui uma sístole e diástole constante, uma vez que existem muitas variações ao longo do tempo. Tendo em conta os batimentos cardíacos (HR), representados a cinzento, verifica-se que o sénior 2 (S2) apresenta apenas uma pequena variação, sendo constantes nas restantes medições.

Tal como acontece no sénior 2, o sénior 3 (S3) e o utilizador 2 (U2) também não possuem a sístole e diástole constantes, existindo algumas variações ao longo do tempo de análise. Tendo em conta os batimentos cardíacos (HR), verifica-se que o sénior 3 apresenta duas variações significativas, mantendo-se constante ao longo das várias medições, enquanto que o utilizador 2 apresenta uma variação de batimentos cardíacos constante ao longo do tempo.

O utilizador 1 (U1) apresenta os três atributos com bastantes variações ao longo do tempo de análise.

O sénior 4 (S4) apenas possui duas medições, tornando-se esta uma análise inconclusiva.

ii. Pedómetro

O ficheiro seguinte a ser analisado é *MIO Daily.csv*. Para este ficheiro foi realizada uma análise global, tendo em conta as médias de passos dados de cada um dos utilizadores ao longo deste teste, assim como análise temporal dos mesmos.

Na Figura 24 está apresentada a média geral de passos de cada um dos utilizadores ao longo das várias medições. Tendo em conta a análise da Figura 24, verifica-se que o utilizador que apresenta uma maior quantidade de passos é o Sénior 1 (uma média de 5233,2 passos). Por oposição é o sénior 4 que possui uma média de passos dados mais baixa (71 passos).

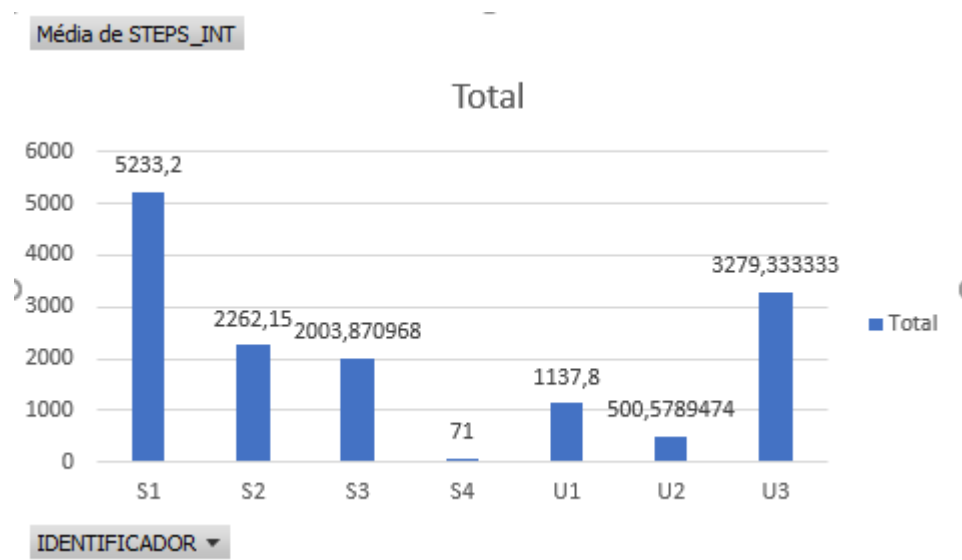


Figura 24 - Gráfico Global Pedómetro

Na Figura 25 é apresentado a evolução temporal de passos dados por cada um dos participantes. Todos os participantes não têm uma variação constante, sendo que o nº de passos dados é diferente em todas as medições.

O sénior 4 não está aqui apresentado, uma vez que este possui apenas uma medição.

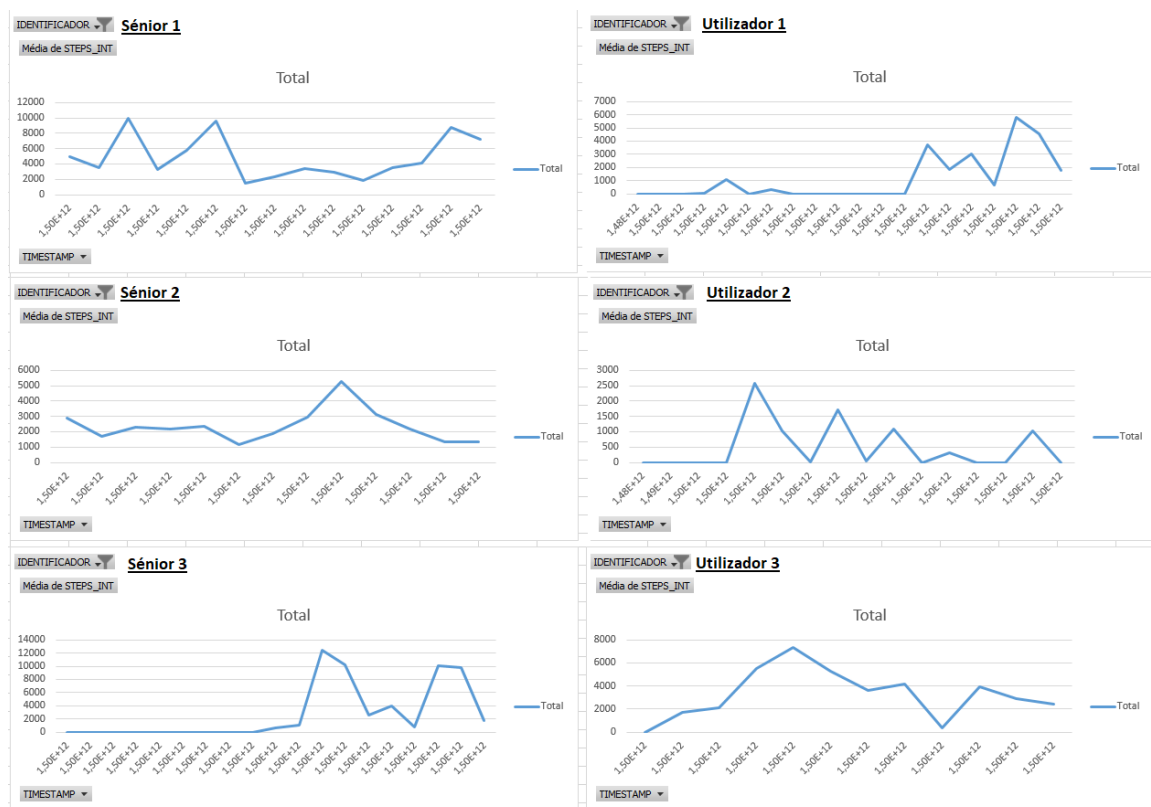


Figura 25 - Gráfico Evolução Temporal Pedómetro

iii. Batimentos Cardíacos Diários

O ficheiro seguinte a ser analisado é *MIO HR Daily.csv*. Para este ficheiro foi realizada uma análise global, tendo em conta as médias do Mínimo, Máximo e Média de Batimentos Cardíacos de cada um dos utilizadores ao longo deste teste, assim como análise temporal destes atributos.

Na Figura 26 estão apresentadas as médias do mínimo, máximo e média de batimentos cardíacos registadas pelo smartwach ao longo de um determinado dia. Assim, verifica-se que o participante que apresenta uma média máxima (MAX) de batimentos cardíacos é o utilizador 3 (U3) e que o participante sénio 1 (S1) apresenta uma média máxima de batimentos cardíacos mais baixos.

Verifica-se também que o Sénior 2 (S2) apresenta uma média mínima (MIN) de batimentos cardíacos mais baixa e o utilizador 3 apresenta uma média mínima de batimentos cardíacos mais elevada. O sénior 2, sénior 3, utilizador 1 e utilizador 2 não apresentam valores para a média mínima de batimentos cardíacos, uma vez que os resultados estão todos a zero, por má utilização do smartwatch.

Tendo em conta a média (MEAN), analisa-se que o utilizador 2 e utilizador 3 apresentam uma média mais elevada. Por oposição, o sénior 1 apresenta uma média mais baixa de batimentos cardíacos registados ao longo do dia.

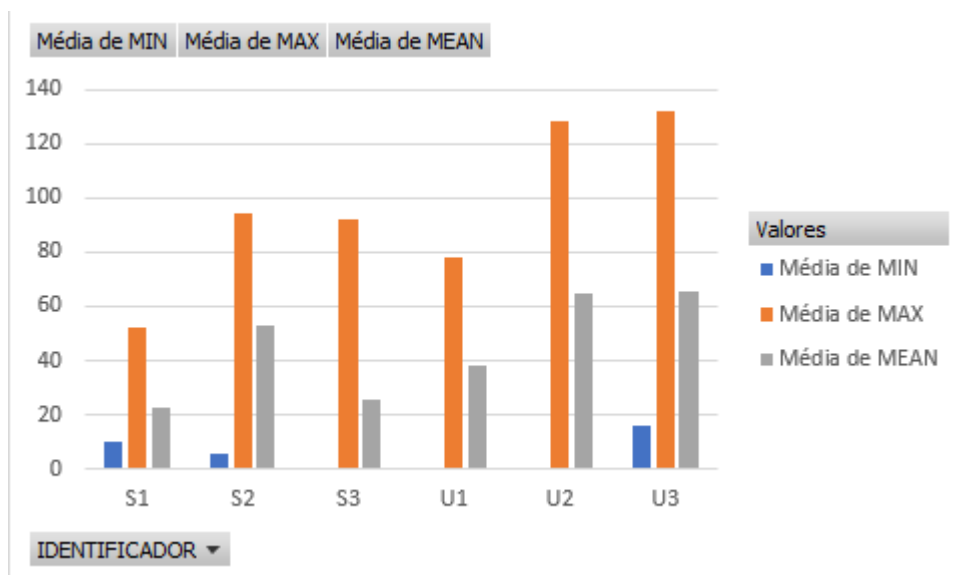


Figura 26 - Gráfico Global Batimentos Cardíacos

Depois da análise da Figura 27 pode-se analisar que à exceção do Utilizador 2, todos os participantes não possuem uma variação constante ao longo das várias medições de mínimo, Máximo e média de batimentos cardíacos.

Quando o máximo, mínimo e média é zero, significa que os utilizadores não utilizaram da melhor maneira o aparelho de medição.

O utilizador 2 possui uma variação constante ao longo das várias medições, sendo que apenas possui uma variação significativa da média máxima numa determinada medição.

O sénior 4 não possui medições, razão pela qual que não está apresentado na Figura 27.

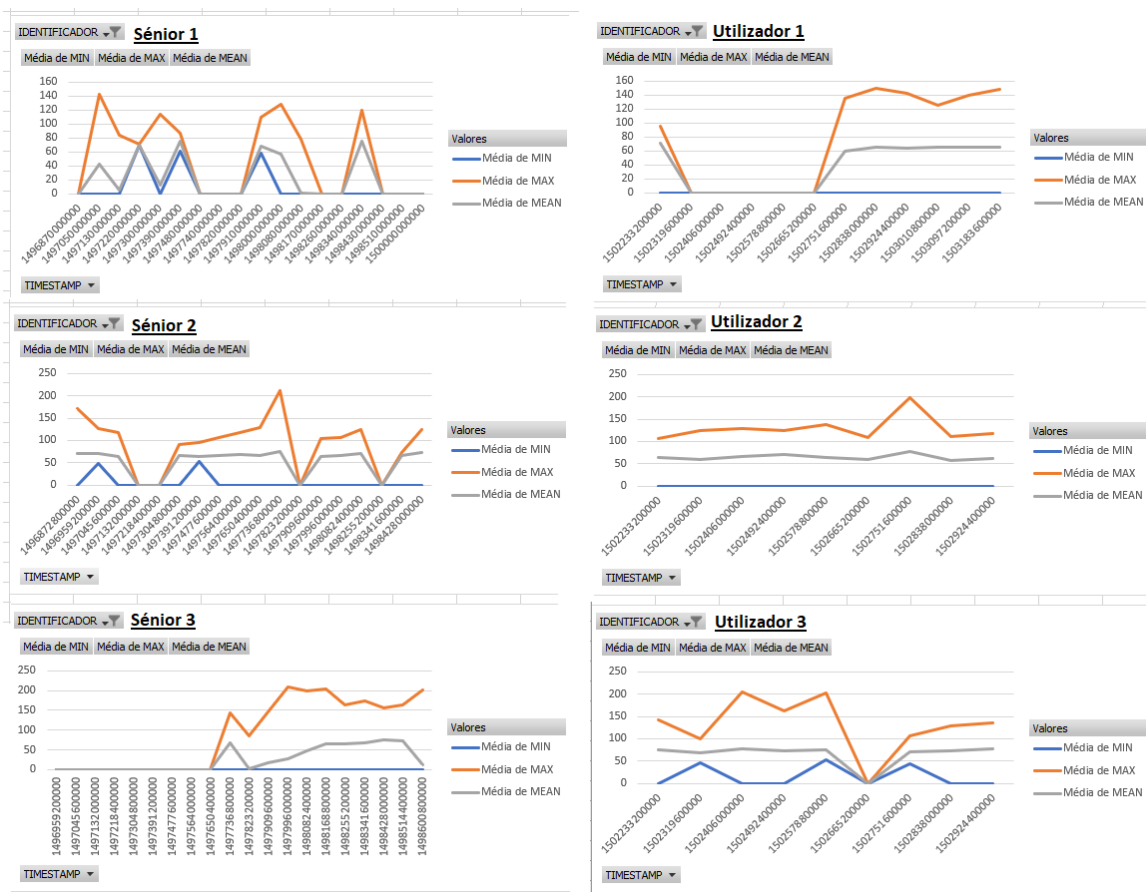


Figura 27 - Gráfico Evolução Temporal Batimentos Cardíacos Diário

iv. Batimentos Cardíacos Horários

O ficheiro seguinte a ser analisado é *MIO HR Hourly.csv*. Para este ficheiro foi realizada uma análise global, tendo em conta as médias do Mínimo, Máximo e Média de Batimentos Cardíacos de cada um dos utilizadores ao longo deste teste, assim como análise temporal destes atributos.

Na Figura 28 está representado o gráfico global da média do mínimo, máximo e média de batimentos cardíacos. Tendo em conta a análise da figura, verifica-se que o sénior 4 (S4) apresenta uma maior média máxima de batimentos cardíacos horários e o utilizador 1 (U1) apresenta uma media mais baixa em comparação com os restantes participantes.

Pode-se analisar também que o utilizador 1 (U1) apresenta uma média mais baixa de mínimos batimentos cardíacos, enquanto que o sénior 4 (S4) apresenta uma maior média mínima.

Tendo em conta a média de batimentos cardíacos (MEAN) é o sénior 4 e o utilizador 1 que apresenta uma média mais alta e mais baixa, respetivamente.

O sénior 1 apresenta as medições horárias todas a zero, razão pela qual não apresenta qualquer tipo de valor na Figura 28.

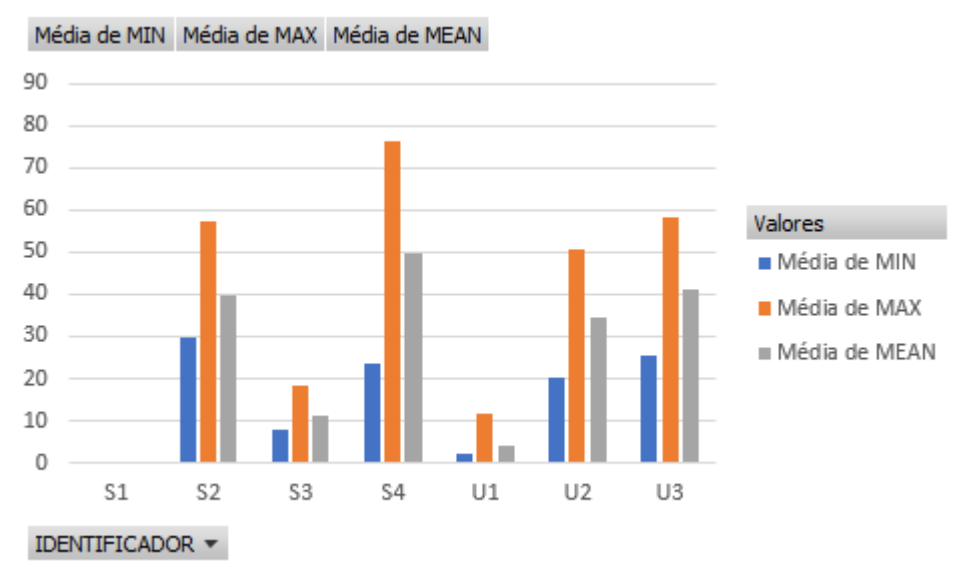


Figura 28 - Gráfico Global Batimentos Cardíacos Horários

Na Figura 29 está apresentada a evolução temporal dos batimentos cardíacos horária ao longo das várias medições.

Depois da análise da Figura 29 verifica-se que os participantes tiveram uma grande dificuldade na utilização do aparelho, uma vez que existem muitas medições a zero.

Analisa-se também que nas horas medidas, os utilizadores não têm uma média, mínimo e máximo de batimentos cardíacos constantes.

O Sénior 1 não está aqui apresentado uma vez que possui os valores todos a zero.

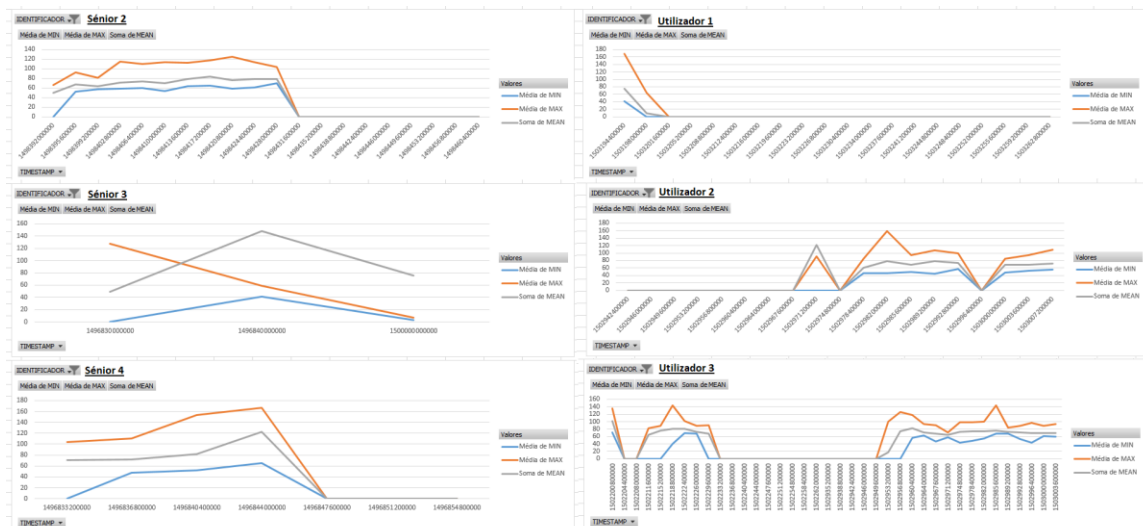


Figura 29 - Gráfico Evolução Temporal Batimentos Cardíacos Horários

v. Oxímetro

O ficheiro seguinte a ser analisado é *Oximeter.csv*. Para este ficheiro foi realizada uma análise global, tendo em conta a média de concentração de oxigénio no sangue de cada um dos utilizadores ao longo deste teste, assim como uma análise temporal.

Na Figura 30 está apresentado a média de concentração de oxigénio de sangue de cada um dos utilizadores. Depois da análise da Figura 30, verifica-se que o utilizador 3 (U3) possui uma maior concentração média de oxigénio no sangue. Opostamente, o sênior 2 (S2) possui uma concentração média de oxigénio no sangue muito mais baixa.

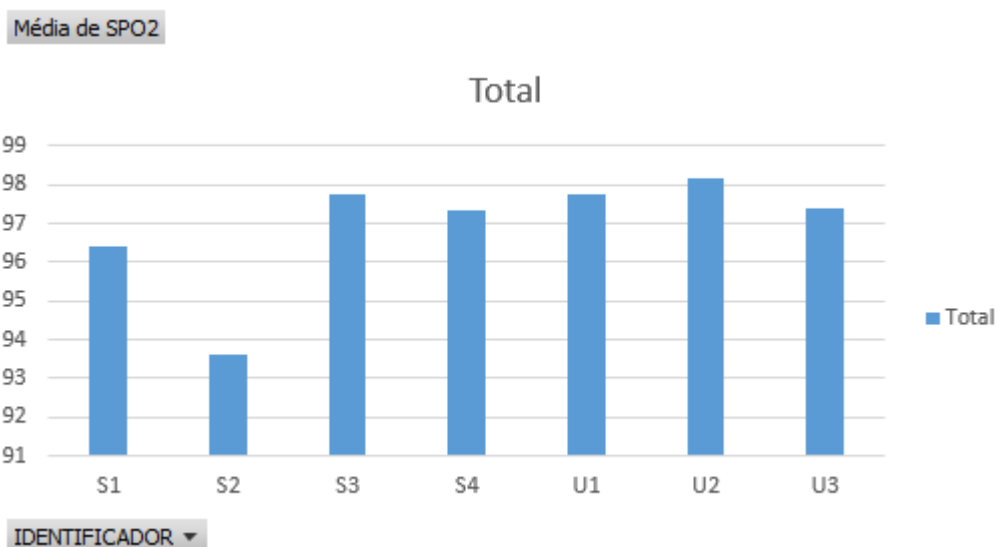


Figura 30 - Gráfico Global Oxímetro

Na Figura 31 está apresentado a evolução temporal da concentração de oxigénio no sangue dos vários participantes no estudo.

O sénior 1 não está presente neste estudo, uma vez que apenas apresenta uma medição. O sénior 4 apresenta duas medições, o que torna a análise inconclusiva.

Os restantes participantes não apresentam uma variação de oxigénio no sangue ao longo das várias medições.



Figura 31 - Gráfico Evolução Temporal Oxímetro

vi. Questionário

O ficheiro seguinte a ser analisado é *Questionnaire.csv*. Para este ficheiro não foi possível realizar uma análise gráfica dos dados, uma vez que os dados apresentados estão em texto.

vii. Peso

O ficheiro seguinte a ser analisado é *Weight.csv*. Para este ficheiro foi realizada uma análise global, tendo em conta a média de peso de cada um dos utilizadores ao longo deste teste, assim como a análise temporal destes atributos.

Na Figura 32 está apresentado a média de peso de cada um dos participantes deste estudo.

Depois da análise da figura, verifica-se que o utilizador 1 (U1) é o que possui uma média mais elevada de peso. Por oposição, o sénior 3 (S3) é o que possui menos peso. De realçar que todos os utilizadores estão na faixa dos 81 até aos 102 quilos.

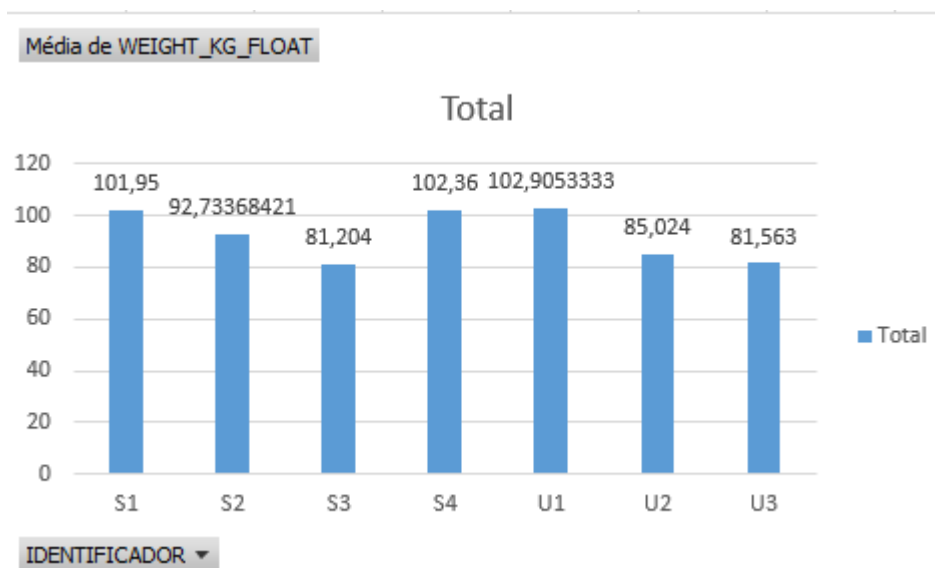


Figura 32 - Gráfico Global de Peso

Depois da análise da evolução do peso ao longo das várias medições, representado na Figura 33, verifica-se que o sénior 1 (S1) apresenta uma maior variação do peso ao longo das várias medições. Enquanto que os restantes participantes apresentam uma evolução constante.

O sénior 4 não está apresentado na Figura 33, uma vez que apenas apresenta uma única medição do peso.

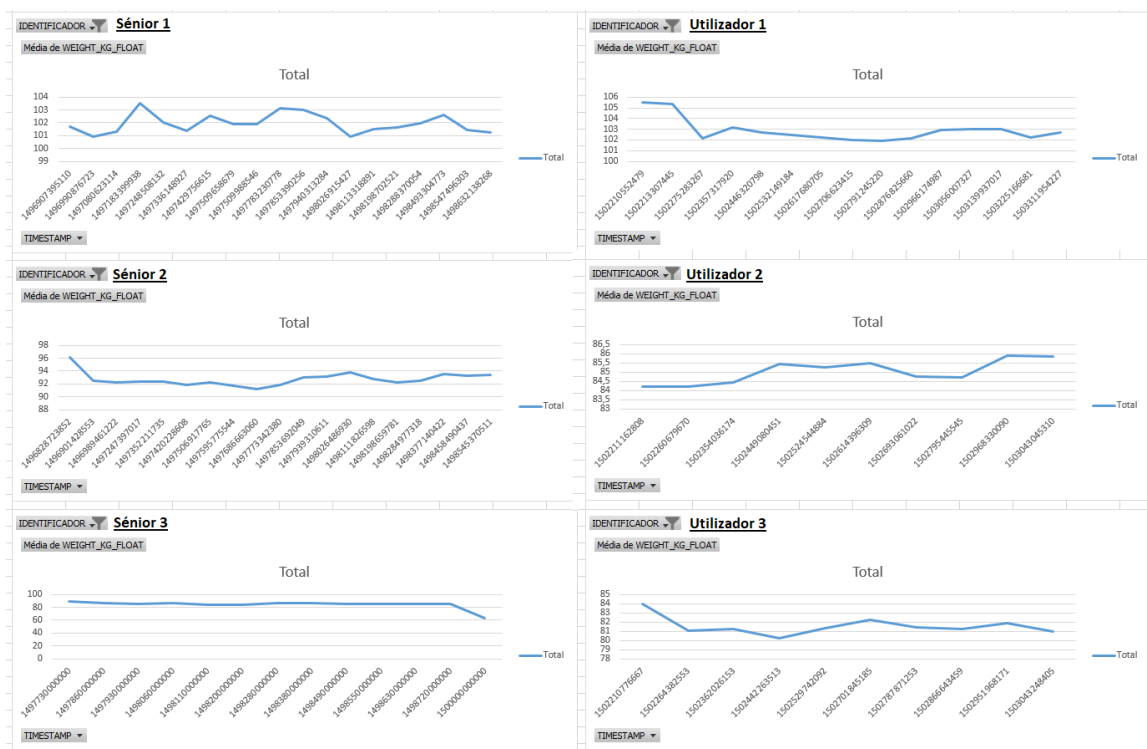


Figura 33 - Gráfico Evolução Temporal de Peso

4.3. Preparação dos Dados

Para que fosse possível calcular a variação dos batimentos cardíacos foi necessário calcular a média dos batimentos cardíacos de cada um dos utilizadores. Para tal, foi desenvolvido um pequeno ETL, representado na Figura 34, com o objetivo de calcular de forma automática a média de batimentos cardíacos de cada um dos utilizadores

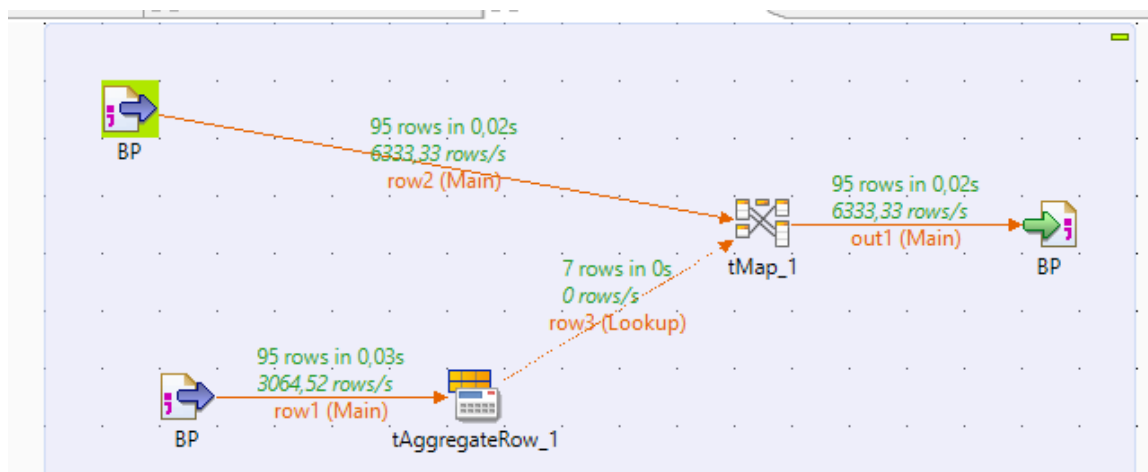


Figura 34 - ETL Média Batimentos Cardíacos

Posteriormente foi necessário unificar a base de dados num único ficheiro, sendo que, para isso, foi necessária a integração dos vários ficheiros. Para se fazer modelação em Data Mining, é necessário que os dados para execução dos modelos estejam numa única fonte de dados.

Assim, para se fazer uma correta integração e não se perder qualquer tipo de informação relevante foram necessários a criação das seguintes tarefas (*jobs*):

- INTEGRATION_BP
- INTEGRATION_WEI
- INTEGRATION_MIO_DA
- INTEGRATION_MIO_HR_DA
- INTEGRATION_REQUEST
- INTEGRATION_OXI
- INTEGRATION_ALL

O ficheiro correspondente ao MIO HR horário, não foi integrado, uma vez que apenas continha informação de dois dias no máximo, o que fazia com que a informação ficasse ainda com mais dados repetidos.

Para que as tarefas criadas anteriormente fossem executadas todas umas a seguir às outras, foi necessário a criação de uma tarefa que executasse todas as tarefas mencionadas anteriormente. Esta tarefa está apresentada na Figura 35.

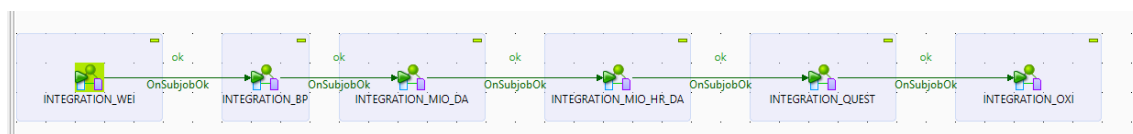


Figura 35 - Tarefa (job) de todas as integrações

Depois de integrados todos os ficheiros, existia um total de 72545 registos. Como, certamente, existiam linhas repetidas, foi desenvolvido outra tarefa que elimina as linhas repetidas. Esta tarefa está representada na Figura 36.

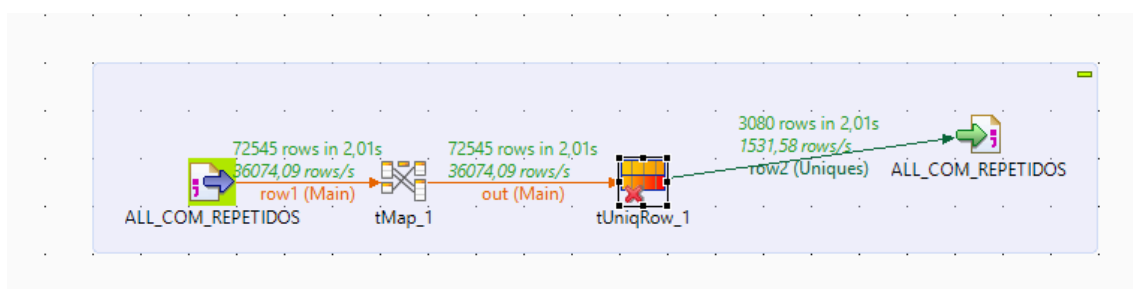


Figura 36 - Tarefa Eliminar Linhas Repetidas

Depois deste job ser executado, apenas ficaram no ficheiro cerca de 3080 registos, ou seja cerca de 4,3% do ficheiro inicial de Integrações.

Posteriormente, foram eliminadas as linhas que continham bastante informação nula, assim como as colunas. Todas as colunas e linhas que continham mais de 50% de informação nula foram eliminadas. Também foram eliminadas as colunas do TIMESTAMP e do IDENTIFICADOR, uma vez que estas não são uma boa informação para se analisar num processo de modelação de Data Mining.

Sendo assim, as colunas eliminadas foram:

- MIN
- MAX
- MEAN
- MEDIAN
- STD_DEV
- ST_DEC

- ST_QUART
- LAST_QUARTSAMPLE_SIZ
- TIMESTAMP
- IDENTIFICADOR

Depois de eliminadas as linhas com uma grande percentagem de informação nula, ficaram no ficheiro cerca de 2927 registos, ou seja cerca de 95% do ficheiro produzido na Figura 36.

Após eliminar as colunas o ficheiro ficou com apenas com as seguintes colunas:

- SYSTOLE
- DIASTOLE
- HR_BP
- HR_AVG
- STEPS_INT
- SPO2
- WEIGHT_KG_FLOAT

A Figura 37 representa a estrutura inicial de dados (em cima), assim como a estrutura final do ficheiro após as alterações mencionadas ao longo desta secção.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	TIMESTAMP	SYSTOLE	DIASTOLE	HR_BP	HR_AVG	STEPS_IN	MIN	MAX	MEAN	MEDIAN	STD_DEV	_ST_DEC	LAST_DEC	_ST_QUA	LAST_QU	Column1	SPO2	Q1	Q2	Q3	Q4	Q5	WEIGHT_KG_FLOAT	
2	28/06/2017																						101.27	
3	27/06/2017																						101.46	
4	26/06/2017																						102.64	
5	24/06/2017																						101.97	
6	23/06/2017																						101.62	
7	22/06/2017																						101.5	
8	21/06/2017																						100.91	
9	20/06/2017																						102.38	
10	19/06/2017																						102.99	
11	18/06/2017																						103.13	
12	15/06/2017																						101.89	
13	15/06/2017																						101.89	
14	14/06/2017																						102.54	
15	13/06/2017																						101.38	

	A	B	C	D	E	F	G	H
1	SYSTOLE	DIASTOLE	HR_BP	HR_AVG	STEPS_INT	SPO2	WEIGHT_KG_FLOAT	
2	157	86	57	61		98	84.9	
3	163	89	76	67		98	85.59	
4	164	88	56	61		98	85.59	
5	158	88	69	67		94	85.09	
6	158	88	69	67		94	85.09	
7	158	88	69	67		98	85.09	
8	158	88	69	67		98	85.09	
9	151	78	56	61		94	85.09	
10	151	78	56	61		94	85.09	
11	151	78	56	61		98	85.09	
12	151	78	56	61		98	85.09	
13	142	88	81	67		98	85.77	
14	142	88	81	67		98	85.77	
15	152	87	75	61		98	85.77	

Figura 37 - Ficheiro Inicial e Final após transformações

Para que fosse possível prever a variação dos batimentos cardíacos foi necessário a criação de uma nova coluna que apresenta a variação tendo em conta duas abordagens:

- Abordagem 1
 - Esta abordagem apenas apresenta dois valores (1,2). É apresentado o valor 1 quando o valor médio dos batimentos cardíacos é menor do que os batimentos cardíacos; é apresentado o valor 2 quando o valor médio dos batimentos cardíacos é maior ou igual do que os batimentos cardíacos.
- Abordagem 2
 - Esta abordagem apresenta quatro valores (1,2,3,4)
 - O valor 1 é apresentado quando a média de batimentos (HR_AVG) cardíacos é igual aos batimentos cardíacos(HR_BP) + 1 e é igual aos batimentos cardíacos + 2

$$(HR_AVG == (HR_BP + 1) \ || \ HR_AVG == (HR_BP + 2))$$
 - O valor 2 é apresentado quando a média de batimentos (HR_AVG) cardíacos é igual aos batimentos cardíacos(HR_BP) - 1 e é igual aos batimentos cardíacos - 2

$$(HR_AVG == (HR_BP - 2) \ || \ (HR_AVG) == (HR_BP) - 1)$$
 - O valor 3 é apresentado quando a média de batimentos (HR_AVG) cardíacos é menor ou igual aos batimentos cardíacos(HR_BP) + 2

$$(row2.HR_AVG <= (row2.HR_BP + 2))$$
 - O valor 4 é apresentado quando a média de batimentos (HR_AVG) cardíacos é maior ou igual aos batimentos cardíacos(HR_BP) - 2

$$(row2.HR_AVG >= (row2.HR_BP - 2))$$

Assim, foram elaborados dois ficheiros distintos, com o objetivo de executar os modelos e verificar se existem alterações nos resultados.

Depois disso, a estrutura final se encontra otimizada para correr os modelos de data mining no software escolhido para tal.

4.4. Modelação

Após aplicadas todas as alterações na fase de preparação dos dados, foram então, utilizadas as ferramentas seleccionadas para aplicação de técnicas de Data Mining.

Para a realização da modelação, foi necessária a criação de diferentes cenários através dos atributos dos ficheiros gerados na preparação dos dados. O primeiro cenário criado abarca todas as variáveis do ficheiro. Os restantes cenários foram construídos através de testes. Primeiro tirou-se a sístole e a diástole e ficaram os restantes atributos, posteriormente colocou-se novamente a sístole e a diástole, mas desta vez apenas possuía um dos outros atributos, para os diferentes cenários. No final obtiveram-se 5 cenários com os diferentes grupos de variáveis que posteriormente foram submetidos aos algoritmos de classificação.

As variáveis utilizadas na modelação de dados foram as seguintes:

- Sístole (SYS)
- Diástole (DIAS)
- Passos (STEPS)
- Oxigénio no sangue (SPO2)
- Peso (WEI)
- Variação de Batimentos cardíacos (CLASS)

Estes atributos foram organizados nos seguintes Cenários:

- C1 = {SYS, DIAS, STEPS, SPO2, WEI, CLASS}
- C2 = {STEPS, SPO2, WEI, CLASS}
- C3 = {SYS, DIAS, WEI, CLASS}
- C4 = {SYS, DIAS, STEPS, CLASS}
- C5 = {SYS, DIAS, SPO2, CLASS}

Como explicado no final do capítulo anterior foi necessária a criação de duas abordagens diferentes:

- **Abordagem 1** onde existem duas classes distintas.
 - Classe 1: onde a média é menor do que o valor dos batimentos cardíacos
 - Classe 2: onde a média é maior ou igual do que o valor dos batimentos
- **Abordagem 2** onde existem quatro classes distintas
 - Classe 1: onde a média de batimentos (HR_AVG) cardíacos é igual aos batimentos cardíacos(HR_BP) + 1 e é igual aos batimentos cardíacos + 2

- Classe 2: onde a média de batimentos (HR_AVG) cardíacos é igual aos batimentos cardíacos(HR_BP) - 1 e é igual aos batimentos cardíacos - 2
- Classe 3: onde a média de batimentos (HR_AVG) cardíacos é menor ou igual aos batimentos cardíacos(HR_BP) + 2
- Classe 4: onde quando a média de batimentos (HR_AVG) cardíacos é maior ou igual aos batimentos cardíacos(HR_BP) - 2

Os algoritmos selecionados para a modelação foram SVM (Support Vector Machine), DT (Decision Trees), NB (Naive Bayes).

Assim, foram, inicialmente, executados um total de 30 Modelos: 15 modelos por cada uma das abordagens, sendo apresentados na seguinte tupla:

- $DMM = \{5 \text{ cenários}, 3 \text{ técnicas de classificação}, 2 \text{ abordagens}\} = 5 \cdot 3 \cdot 2 = 30 \text{ modelos}$

De seguida, é apresentado um exemplo que mostra o modelo para o cenário 1, usando DT para a abordagem 1.

- $DMM = \{C5, DT, \text{Abordagem 1}\}$

O método de classificação utilizado foi o *10 Fold CV* que consiste em dividir os o conjunto total de dados em 10 subconjuntos do mesmo tamanho e, a partir disso, um subconjunto é utilizado para testes e os restantes para treino.

Para a avaliação destes modelos é definido um *threshold* de 85 %, sendo que os resultados são considerados bons se o valor for superior ao *threshold* definido.

• Aplicação de Técnicas de Classificação

Os cenários criados anteriormente, foram introduzidos no WEKA. Inicialmente, é escolhida a fonte de dados, num formato .ARFF, onde são selecionados os atributos pretendidos, previamente tratados e preparados para este processo no weka. Posteriormente é escolhido o modelo de classificação que se deseja aplicar ao conjunto de dados.

Após a construção de cada cenário e aplicação dos algoritmos de Classificação, foram obtidos diversos resultados, como a matriz de confusão e curva de ROC.

A Figura 38 apresenta a matriz de confusão para o cenário 1 (C1), da **abordagem 1** e para a classificação Decision Tree (DT).

O C1 é constituído por todas as variáveis do ficheiro. A abordagem 1 diz respeito à abordagem que possui duas classes.

De realçar, que as restantes matrizes de confusão dos cenários propostos correspondentes a esta abordagem, estão apresentadas no Anexo II – Matrizes de Confusão da Abordagem 1.

```

a      b      <-- classified as
21     20 |      a = 1
12 2874 |      b = 2

```

Figura 38 - Matriz de Confusão C1 Abordagem 1

Para esta abordagem foi realizado o cálculo de quatro Métricas: Acuidade, Sensibilidade, Especificidade e Taxa de Erro. Assim, tendo em conta o cenário referido anteriormente, foram obtidos os seguintes valores:

- Acuidade: $\left(\frac{21+2874}{2874+12+21+20}\right) * 100 = 98,91\%$
- Sensibilidade: $\left(\frac{2874}{2874+12}\right) * 100 = 99,58\%$
- Especificidade: $\left(\frac{21}{21+20}\right) * 100 = 51,22\%$
- Taxa de erro total = $100 - 98,91 = 1,09\%$

De modo a equilibrar os valores do target, foram replicados os dados que apresentavam a porção mais pequena, através de um método de *sampling*

Na abordagem 1 foram replicados os dados correspondentes à classe 2. Depois dessa replicação é apresentada a matriz de confusão na Figura 39. Os dados foram replicados 135 vezes, passando de 21 dados para 2870.

As restantes matrizes de confusão desta abordagem estão apresentadas no Anexo III - Matrizes de Confusão da Abordagem 1, Método Sampling

```

a      b      <-- classified as
2870     0 |      a = 1
46 2840 |      b = 2

```

Figura 39 - Matriz de Confusão C1, Abordagem 1, Método *Sampling*

Assim, tendo em conta o cenário referido anteriormente e a replicação dos dados, foram obtidos os seguintes valores:

- Acuidade: $\left(\frac{2870+2840}{2870+2840+46}\right) * 100 = 99,20\%$
- Sensibilidade: $\left(\frac{2840}{2840+46}\right) * 100 = 98,41\%$
- Especificidade: $\left(\frac{2870}{2870}\right) * 100 = 100\%$
- Taxa de erro total = $100 - 99,20 = 0,8\%$

A Figura 40 apresenta a matriz de confusão para o cenário 1 da abordagem 2 da DT. Para a **abordagem 2** apenas foi possível o cálculo da acuidade e taxa de erro, uma vez que esta abordagem possui quatro classes.

As restantes matrizes de confusão dos outros cenários estão apresentadas no Anexo IV - Matrizes de Confusão da Abordagem 2

a	b	c	d	<-- classified as
1418	2	4	5	a = 1
2	2	5	3	b = 2
3	3	1423	9	c = 3
15	2	6	25	d = 4

Figura 40 - Matriz de Confusão C1, Abordagem 2

Assim sendo, a acuidade e taxa de erro apresentam os seguintes cálculos:

- Acuidade: $\left(\frac{1418+2+1423+25}{1418+2+4+5+2+2+5+3+3+3+1423+9+15+2+6+25}\right) * 100 = 97,98\%$
- Taxa de erro: $100 - 97.98 = 2.02\%$

Como os valores dos target apresentam valores muito dispares, para esta abordagem não foram replicados os valores das diferentes targets.

Assim sendo, iniciou-se a modelação da **primeira abordagem**, através da qual foram criados 30 modelos:

- 5 cenários, 3 técnicas de classificação, 2 métodos de sampling = $5 \times 3 \times 2 = 30$ modelos

Inicialmente, foram executados os modelos e técnicas correspondentes ao método sem *sampling*. Depois de executados todos os modelos no WEKA, os dados das métricas foram registados na Tabela 15.

Tabela 15 - Registo de Informação, Abordagem 1

	Técnica	Acuidade	Sensibilidade	Especificidade	Taxa de Erro
C1	DT	98,91	99,58	51,22	1,09
	SVM	98,60	100	0	1.4
	NB	95,38	96,02	51,22	4,62
C2	DT	98,73	99,79	24,39	1,27
	SVM	98,60	100	0	1.4
	NB	94,5	95,84	0	5,5
C3	DT	98,91	99,58	51,22	1,09
	SVM	98,60	100	0	1.4
	NB	98,08	98,75	51,22	1,92
C4	DT	98,43	99,34	34,15	1,57
	SVM	98,60	100	0	1.4
	NB	98,02	96,02	43,90	1,98
C5	DT	98,77	99,55	43,90	1,23
	SVM	98,60	100	0	1.4
	NB	98,02	96,02	43,90	1,98

Depois da análise desta tabela, pode-se aferir que todos os modelos possuem bons resultados. Pode-se concluir também que a métrica SVM possui sempre o mesmo resultado, independentemente do cenário. Conclui-se também que o cenário 1 e o cenário 3 têm exatamente

os mesmos resultados. O que significa que o modelo apenas considera para a análise os atributos definidos no cenário 3.

Depois de analisados os resultados, passou-se à execução do método com a repetição dos dados – Método de *Sampling*. A Tabela 16 apresenta os resultados obtidos, depois da execução do modelo no weka.

Tabela 16 - Registo de Informação, Abordagem 1, Método Sampling

	Técnica	Acuidade	Sensibilidade	Especificidade	Taxa de Erro
C1	DT	99,20	98,41	100	0,80
	SVM	79,33	97,27	59,9	20,67
	NB	93,38	94,07	92,68	6,62
C2	DT	67,06	99,79	34,15	32,94
	SVM	70,92	100	42	29,08
	NB	65,95	95,15	36,59	34,05
C3	DT	98,18	98,37	100	0,82
	SVM	79,69	98,65	60,63	20,30
	NB	62,49	33,06	92,09	37,51
C4	DT	99,92	98,65	100	0,68
	SVM	79,83	98,58	60,98	20,17
	NB	84,38	91,27	77,46	15,62
C5	DT	99,29	98,58	100	0,71
	SVM	79,69	98,65	60,62	20,31
	NB	88,76	96,02	43,90	11,24

Depois da análise desta tabela, pode-se concluir que o cenário 2 é o que possui piores resultados em todas as técnicas de DM. Tendo em conta os que possuem melhores resultados, o cenário 4 é o que possui os resultados melhores, em comparação com os restantes.

Posteriormente, procedeu-se à modelação **da segunda abordagem**, através do qual foram criados 15 modelos:

- 5 cenários, 3 técnicas de classificação = $5 \times 3 = 15$

Como nesta abordagem não existiu a replicação dos dados, apenas existiu um método, sendo que os resultados foram registados na Tabela 17.

Esta tabela não apresenta informação para a sensibilidade e especificidade, uma vez que esta abordagem possui mais do que duas classes.

Tabela 17 - Registo de Informação, Abordagem 2

	Técnica	Acuidade	Taxa de Erro
C1	DT	97,98	2,02
	SVM	97,23	2,76
	NB	84,42	15,58
C2	DT	46,98	53,02
	SVM	48,51	51,49
	NB	43,29	56,71
C3	DT	98,02	1,98
	SVM	97,23	2,77
	NB	87,5	12,5
C4	DT	97,64	2,36
	SVM	97,23	2,77
	NB	87,43	12,57
C5	DT	97,64	2,36
	SVM	97,23	2,77
	NB	87,43	12,57

Depois da análise desta tabela conclui-se que estes cenários possuem resultados bons, uma vez que são resultados que ultrapassam o *threshold* definido de 85 %, em exceção o cenário 2 que apresenta resultados inferiores a 50%.

O melhor cenário será o cenário um, uma vez que possui resultados superiores, em comparação com os restantes cenários.

5. Discussão dos Resultados

Depois de realizadas todas as fases da recolha de dados, análise e tratamento dos dados, bem como a modelação e obtenção de resultados, segue-se a discussão dos resultados, com o objetivo de confronta-los com os objetivos inicialmente propostos para esta dissertação.

Os dados para esta dissertação foram explorados de modo a compreender a sua necessidade no processo de Data Mining (DM), bem como para aferir erros ou omissões que precisassem de correção. Esta foi uma das fases mais longas pois requereu que fossem verificados todos os tipos de anomalias possíveis.

Para ajudar na execução deste projeto foi utilizado o WEKA e recorreu-se a técnicas de classificação. Relativamente aos métodos utilizados, foi utilizado *Cross Industrial Standard Process for Data Mining* (CRISP-DM) para a aplicação de técnicas de DM.

A técnica de Classificação foi aplicada em duas abordagens: na primeira foi utilizada a previsão para duas classes (classe 1 onde os batimentos cardíacos são menores do que a média de batimentos cardíacos e a classe 2 onde os batimentos cardíacos são maiores ou iguais do que a média de batimentos cardíacos) e na segunda foi utilizada uma previsão para quatro classes (classe 1 onde a média de batimentos cardíacos é igual aos batimentos cardíacos + 1 e é igual aos batimentos cardíacos + 2; classe 2 onde a média de batimentos cardíacos é igual aos batimentos cardíacos - 1 e é igual aos batimentos cardíacos - 2; classe 3 onde a média de batimentos cardíacos é menor ou igual aos batimentos cardíacos + 2; classe 4: onde quando a média de batimentos cardíacos é maior ou igual aos batimentos cardíacos - 2)

Para cada abordagem foram utilizadas as seguintes técnicas de classificação: Decision Tree (DT), Naive Bayes (NB) e Support Vector Machines (SVM).

Para cada modelo foram realizados testes de métricas: Sensibilidade, Acuidade, Especificidade e Taxa de erro. Para a segunda abordagem só foi possível realizado o teste de acuidade devido ao facto de ser constituída por 4 classes.

Foram executados os modelos e cenários correspondentes da primeira abordagem e conclui-se que todos os modelos apresentam bons resultados, uma vez que os valores de acuidade apresentam um valor superior a 90%, sendo este valor superior ao *threshold* definido de 85%.

Contudo, a especificidade de todos os modelos da abordagem 1 apresentam uma especificidade inferior a 55%.

De maneira de melhorar estes resultados, principalmente de especificidade, foram realizadas replicas da *target* que apresentava uma quantidade menor de valores, sendo que os dados foram replicados 132 vezes.

Depois da análise destes dados verificou-se que acuidade já não apresenta, em todos os cenários e modelos, um valor superior a 90%. Contudo, em comparação com a abordagem 1 sem *sampling* os valores da técnica de DT é superior em todos os casos. A especificidade aumentou, apresentando uma especificidade superior a 55% em maior parte dos casos analisados.

Depois disso, com o objetivo de verificar se existe a possibilidade dos resultados apresentarem melhores resultados foi executada a abordagem 2 que possui quatro classes.

Depois da análise conclui-se que esta abordagem obteve resultados mais baixos do que a abordagem 1. Contudo para todos os cenários, com exceção ao cenário 2, foram obtidos resultados superiores a 85%.

Os resultados obtidos permitiram oferecer uma nova visão para a comunidade científica, uma vez que são superiores ao *threshold* definido de 85%, sendo relevante no domínio dos Sistemas de Apoio à Decisão (SAD) e Data Mining (DM), pois poderá ser uma maneira de perceber como exercício físico ajuda na prevenção das doenças cardíacas. Estes resultados, permitem, também, uma ação preventiva por parte dos médicos, para que os seus doentes sejam mais ativos.

6. Conclusão

6.1. Síntese do Trabalho Realizado

Esta dissertação tinha como objetivo principal a elaboração de um modelo de previsão de doenças coronárias, tendo em conta os dados registados pelo Centro Hospitalar do Porto e pelo Hospital da Senhora da Oliveira de Guimarães. Devido a um problema nos atrasos dos dados, e de forma a combater a falta de dados, o contexto deste projeto teve que ser alterado. Assim, os objetivos tiveram que ser refinados de acordo com os dados recebidos.

Numa primeira fase foram analisados dados recolhidos de exercício físico, com o objetivo de perceber se exercício é um fator na prevenção de doenças cardíacas.

Todos os modelos desenvolvidos trazem resultados que permitem identificar que a pratica de exercício físico é um fator importante na vida das pessoas, podendo ajudar na prevenção de doenças de cardíacas. Quando as pessoas praticam exercício a sua pressão arterial fica mais estabilizada, fazendo com que não tenham uma tensão arterial tão elevada, contribuindo assim para a prevenção de doenças relacionadas com o coração.

Através deste modelo os profissionais de saúde podem ajudar os seus doentes a terem uma vida mais ativa, mostrando que se fizerem exercício físico, as alterações dos batimentos cardíacos ficam mais estabilizados, assim como a pressão arterial, podendo ajudar na prevenção das doenças cardíacas.

Assim, conclui-se que os objetivos propostos foram concluídos com sucesso. O facto de terem sido cumpridos os objetivos e os resultados esperados, permitiu a criação de conhecimento útil para a comunidade científica.

A questão de investigação foi desenvolvida, de modo a enquadrar os objetivos principais deste projeto de investigação. Assim, a questão de investigação era a seguinte:

“De que modo o Data Mining pode contribuir para perceber de que modo o exercício físico ajuda na prevenção de doenças coronárias?”

Depois da análise do modelo desenvolvido em Data Mining, pode-se concluir que esta questão de investigação foi respondida, uma vez que este modelo produziu resultados capazes de perceber que o exercício físico é um dos fatores mais importantes para a prevenção de Doenças Cardíacas.

6.2. Riscos Verificados

Como qualquer projeto, existem riscos associados, sendo que para este projeto, os principais riscos cometidos foram:

Tabela 18 - Riscos Cometidos

Risco	Prob.	Imp.	Seri.	Consequências	Ação Atenuante
Má gestão do projeto	3	4	12	Atraso no desenvolvimento do projeto	Fazer um estudo de como fazer uma gestão de projeto adequada; Utilização de ferramentas adequadas
Perda de informação do Projeto	2	5	10	Refazer algumas tarefas; Atraso no desenvolvimento do projeto	Realização de backups frequentes do trabalho realizado.
Atraso no desenvolvimento do projeto	2	4	8	Atraso no desenvolvimento do projeto	Controlo mais apertado do desenvolvimento do projeto; Cumprimento das datas estipuladas no plano de projeto.
Problemas com as ferramentas	1	3	3	Atraso na realização das tarefas	Pesquisa de informação e consulta de tutoriais sobre as ferramentas.
Legenda: Prob. - Probabilidade (1-5) Imp. – Impacto (1-5) Seri. – Seriedade (Prob. X Imp.)					

6.3. Trabalho futuro

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

1. Determinar novos atributos e novos dados que possam ser inseridos na modelação para obter novos modelos e melhores resultados;
2. Aplicar diferentes técnicas de Data Mining e novos cenários.

Referências Bibliográficas

- Acharya, U. R., Sree, S. V., Muthu Rama Krishnan, M., Krishnananda, N., Ranjan, S., Umesh, P., & Suri, J. S. (2013). Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. *Computer Methods and Programs in Biomedicine*, 112(3), 624–632. <https://doi.org/10.1016/j.cmpb.2013.07.012>
- Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-14142-8>
- Arbor pharmaceuticals. (2014). Understanding and Managing High Blood Pressure. *Journal of American Heart Association*, 1–16.
- Azevedo, A., & Santos, M. F. (2011). *A Perspective on Data Mining Integration with Business Intelligence. Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains*. Retrieved from <http://www.igi-global.com>
- Bhatia, S. K. (2010). Biomaterials for clinical applications. *Biomaterials for Clinical Applications*, 1–283. <https://doi.org/10.1007/978-1-4419-6920-0>
- Brown, M. S. (2014). (For Dummies) Meta S. Brown-Data Mining For Dummies-Wiley Publishing Inc. (2014).pdf. Retrieved from www.wiley.com
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76. <https://doi.org/10.1109/ICETET.2008.239>
- Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). Supervised Learning: Statistical Methods. In *Data Mining: A Knowledge Discovery Approach* (pp. 307–379). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-36795-8_11
- de Bruijn, R. F. A. G., & Ikram, M. A. (2014). Cardiovascular risk factors and future risk of Alzheimer's disease. *BMC Medicine*, 12:1 Artic, 1–9. <https://doi.org/Artn 130\rDoi 10.1186/S12916-014-0130-5>
- de Carvalho, H. H., Moreno, R. L., Pimenta, T. C., Crepaldi, P. C., & Cintra, E. (2013). A heart disease recognition embedded system with fuzzy cluster algorithm. *Computer Methods and Programs in Biomedicine*, 110(3), 447–454. <https://doi.org/10.1016/j.cmpb.2013.01.005>
- Direção-Geral da Saúde. (2015). Portugal: Doenças Cérebro-Cardiovasculares em Números - 2015. *Saúde Em Números*, 92. <https://doi.org/ISSN: 2183-0681>

- Elkan, C. (2010). *Predictive analytics and data mining*. Retrieved from <http://cseweb.ucsd.edu/users/elkan/255/dm.pdf>
- Estatística, I. N. de. (2016). *Causas de Morte - 2014*. (I. P. Instituto Nacional de Estatística, Ed.). Lisboa. Retrieved from www.ine.pt
- Foundation, B. H. (2014). Heart Valve Disease. Retrieved from http://www.medicinenet.com/heart_valve_disease/page4.htm
- Graham, I., Atar, D., Borch-Johnsen, K., Boysen, G., Burell, G., Cifkova, R., ... Zampelas, A. (2007). †European guidelines on cardiovascular disease prevention in clinical practice: executive summaryFourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by re. *European Heart Journal*, 28(19), 2375. <https://doi.org/10.1093/eurheartj/ehm316>
- Graham, I. M. (2008). Risk Factors & Cardiovascular Disease. *Dialogues in Cardiovascular Medicine*, 13(2), 77–152.
- IBM. (2011). IBM SPSS Modeler CRISP-DM Guide, 53.
- Jiawei, H., Kamber, M., Han, J., Kamber, M., & Pei, J. (2012a). *Data Mining: Concepts and Techniques*. San Francisco, CA, itd: Morgan Kaufmann. <https://doi.org/10.1016/B978-0-12-381479-1.00001-0>
- Jiawei, H., Kamber, M., Han, J., Kamber, M., & Pei, J. (2012b). *Data Mining: Concepts and Techniques*. San Francisco, CA, itd: Morgan Kaufmann. <https://doi.org/10.1016/B978-0-12-381479-1.00001-0>
- Kadi, I., Idri, A., & Fernandez-aleman, J. L. (2017). International Journal of Medical Informatics Knowledge discovery in cardiology: A systematic literature review. *International Journal of Medical Informatics*, 97, 12–32. <https://doi.org/10.1016/j.ijmedinf.2016.09.005>
- KDnuggets. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved January 5, 2017, from <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>
- Kovács, F., Legány, C., & Babos, A. (2005). Cluster Validity Measurement Techniques. *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence, 2006*, 1–11. <https://doi.org/10.7547/87507315-91-9-465>

- Li, Q., Rajagopalan, C., & Clifford, G. D. (2014). Ventricular fibrillation and tachycardia classification using a machine learning approach. *IEEE Transactions on Biomedical Engineering*, 61(6), 1607–1613. <https://doi.org/10.1109/TBME.2013.2275000>
- Lu, L., Sun, R., Liu, M., Zheng, Y., & Zhang, P. (2015). The Inflammatory Heart Diseases: Causes, Symptoms, and Treatments. *Cell Biochemistry and Biophysics*, 72(3), 851–855. <https://doi.org/10.1007/s12013-015-0550-7>
- M. Altınbaş, B. Kaplan, K. Ucar, M. Ozkan, E. Karahacıoğlu, O. Er, C. K. and I. (2004). 5-FU plus leucovorin (Mayo's Regimen) for concurrent chemoradiotherapy after surgery in patients with resectable locally advanced gastric carcinoma: Preliminary results of a phase II study. *Journal of Clinical Oncology*, 22(4196), 1–8.
- Mackay, Judith, Mensah, G. A. (2004). Types of Cardiovascular Diseases. *The Atlas of Heart Disease and Stroke*, 18–19. Retrieved from http://www.who.int/topics/cerebrovascular_accident/en/
- Maglogiannis, I., Loukis, E., Zafiropoulos, E., & Stasis, A. (2009). Support Vectors Machine-based identification of heart valve diseases using heart sounds. *Computer Methods and Programs in Biomedicine*, 95(1), 47–61. <https://doi.org/10.1016/j.cmpb.2009.01.003>
- Maimon, O., & Rokach, L. (2005). Introduction to knowledge discovery in database. *Data Mining and Knowledge Discovery Handbook*, 1–17. Retrieved from <http://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf>
- Marijon, E., Mirabel, M., Celermajer, D. S., & Jouven, X. (2012a). Rheumatic heart disease. *The Lancet*, 379(9819), 953–964. [https://doi.org/10.1016/S0140-6736\(11\)61171-9](https://doi.org/10.1016/S0140-6736(11)61171-9)
- Marijon, E., Mirabel, M., Celermajer, P. D. S., & Jouven, P. X. (2012b). Rheumatic heart disease, 1–12.
- Mehta, P. K., Wei, J., & Wenger, N. K. (2015). Ischemic heart disease in women: A focus on risk factors. *Trends in Cardiovascular Medicine*, 25(2), 140–151. <https://doi.org/10.1016/j.tcm.2014.10.005>
- Mendis, S., Puska, P., & Norrving, B. (2011). Global Atlas on cardiovascular disease prevention and control. *World Health Organization, Geneva*, 164. https://doi.org/NLM_classification:WG_120
- Metra, M., Carubelli, V., Ravera, A., & Stewart Coats, A. J. (2016). Heart failure 2016: still more questions than answers. *International Journal of Cardiology*, 227, 766–777.

<https://doi.org/10.1016/j.ijcard.2016.10.060>

Miller, H. J., & Han, J. (2009). Geographic data mining and knowledge discovery An Overview. *Geographic Data Mining and Knowledge Discovery*, 1–26.
<https://doi.org/doi:10.1201/9781420073980.ch1>

Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of Cardiovascular Disease Risk ' s Level for Adults Using Naive Bayes Classifier, *22*(3), 196–205.

Mironczuk, M. (2015). The cascading knowledge discovery in databases process in the information system development. *2015 2nd International Conference on Computer Science, Computer Engineering, and Social Media, CSCESM 2015*, (November).
<https://doi.org/10.1109/CSCESM.2015.7331873>

Mohebbi, M., & Ghassemian, H. (2012). Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal. *Computer Methods and Programs in Biomedicine*, *105*(1), 40–49.
<https://doi.org/10.1016/j.cmpb.2010.07.011>

National Heart, Lung, and B. I. (2009). Coronary heart disease. *BMJ (Clinical Research Ed.)*, *303*(6804), 701–704. <https://doi.org/10.1161/CIRCULATIONAHA.114.011675>

National Heart Lung and Blood. (2016). What Is High Blood Pressure? *American Heart Association*, 1–13.

National Heart Lung and Blood Institute. (2009). At a Glance: Atherosclerosis.

North, M. (2012). *Data Mining for the Masses. Computer*. Retrieved from <http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/DataMiningForTheMasses.pdf%5Cnhttps://sites.google.com/site/dataminingforthemasses/>

Oded Maimon, L. R. (2005). *Data Mining and Knowledge Discovery Handbook. Data Mining and Knowledge Discovery Handbook*. https://doi.org/10.1007/0-387-25465-x_2

Olson, D. L., & Delen, D. (2008). 2 Data Mining Process. *Springer*.

Oyyathevan, S., & Askarunisa, D. A. (2014). AN EXPERT SYSTEM FOR HEART DISEASE PREDICTION USING DATA MINING TECHNIQUE :, *1*(4), 1–6.

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining

- techniques. *2008 IEEE/ACS International Conference on Computer Systems and Applications*.
<https://doi.org/10.1109/AICCSA.2008.4493524>
- Pecchia, L., Melillo, P., Sansone, M., & Bracale, M. (2011). Discrimination power of short-term heart rate variability measures for CHF assessment. *IEEE Transactions on Information Technology in Biomedicine*, 15(1), 40–46. <https://doi.org/10.1109/TITB.2010.2091647>
- Peppers, 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), 45–78. <https://doi.org/10.2753/MIS0742-1222240302>
- Portela, F., Filipe, M., Silva, Á., Hospitalar, C., Rua, F., Hospitalar, C., ... Arrhythmia, A. C. (2014). Preventing Patient Cardiac Arrhythmias by Using Data Mining Techniques, (December), 8–10.
- Rajalakshmi, K., & Nirmala, K. (2016). Heart disease prediction with mapreduce by using weighted association classifier and K-Means. *Indian Journal of Science and Technology*, 9(19). <https://doi.org/10.17485/ijst/2016/v9i19/93827>
- Ramageri, M. (2010). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering*, 1(4), 301–305.
- Remenyi, B., Carapetis, J., Wyber, R., Taubert, K., & Mayosi, B. M. (2013). Position statement of the World Heart Federation on the prevention and control of rheumatic heart disease. *Nature Reviews. Cardiology*, 10(5), 284–92. <https://doi.org/10.1038/nrcardio.2013.34>
- Remenyi, B., Wilson, N., Steer, A., Ferreira, B., Kado, J., Kumar, K., ... Carapetis, J. (2012). World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease—an evidence-based guideline. *Nat Rev Cardiol*, 9(5), 297–309. <https://doi.org/10.1038/nrcardio.2012.7>
- Sahan, E., Sahan, S., Karamanliouglu, M., Gul, M., & Tufekciouglu, O. (2016). The MOGE(S) classification. *Herz*, 41(6), 503–506. <https://doi.org/10.1007/s00059-015-4394-0>
- Santos, M. F., & Azevedo, C. (2005). *Data Mining: Descoberta de conhecimento em base de dados* (09–2005th ed.). FCA.
- Santos, M. Y., & Ramos, I. (2006). *Business Intelligence: Tecnologias da Informação na Gestão de conhecimento* (2ª Edição). FCA.
- Scottish Intercollegiate Guidelines Network. (2007). Risk estimation and the prevention of

- cardiovascular disease. (SIGN Guideline No 97), (February), 1–76. Retrieved from [papers2://publication/uuid/3C0E50C8-460C-497B-9876-E01D1296291E](https://pubs2://publication/uuid/3C0E50C8-460C-497B-9876-E01D1296291E)
- Shearer, C., Watson, H. J., Grecich, D. G., Moss, L., Adelman, S., Hammer, K., & Herdlein, S. a. (2000). The CRIS-DM model: The New Blueprint for Data Mining. *Journal of Data Warehousing* 14, 5(4), 13–22. Retrieved from www.spss.com%5Cnwww.dw-institute.com
- Shouman, M., Turner, T., & Stocker, R. (2012). Using data mining techniques in heart disease diagnosis and treatment. *2012 Japan-Egypt Conference on Electronics, Communications and Computers*. <https://doi.org/10.1109/JEC-ECC.2012.6186978>
- Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, 17(8), 43–48. <https://doi.org/10.5120/2237-2860>
- Steinberg, C., Laksman, Z. W. M., & Krahn, A. D. (2016). Sudden cardiac death: A reappraisal. *Trends in Cardiovascular Medicine*, 26(8), 709–719. <https://doi.org/10.1016/j.tcm.2016.05.006>
- Varbo, A., Benn, M., & Nordestgaard, B. G. (2014). Remnant cholesterol as a cause of ischemic heart disease: Evidence, definition, measurement, atherogenicity, high risk patients, and present and future treatment. *Pharmacology and Therapeutics*, 141(3), 358–367. <https://doi.org/10.1016/j.pharmthera.2013.11.008>
- WHO. (2005). Heart Attacks and Strokes Don ' t be a victim. *WHO Library Cataloguing in Publication Data*, 8–10. <https://doi.org/ISBN 92 4 154672 7>
- WHO. (2014). Global status report on noncommunicable diseases 2014. *World Health*, 176. <https://doi.org/ISBN 9789241564854>
- Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. Complementary literature None. <https://doi.org/0120884070, 9780120884070>
- World Heart Federation. (2012). Cardiovascular Disease Risk Factors. *Cardiology Journal*, (April 2012), 6–9. <https://doi.org/10.1016/j.amjcard.2005.09.041>
- Zhu, B., Ding, Y., & Hao, K. (2014). Multiclass maximum margin clustering via immune evolutionary algorithm for automatic diagnosis of electrocardiogram arrhythmias. *Applied Mathematics and*

Computation, 227, 428–436. <https://doi.org/10.1016/j.amc.2013.11.028>

Anexo I – Trabalhos Relacionados

Tabela 19 - Trabalhos Relacionados. Adaptado:(Kadi et al., 2017)

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Review of <i>Data Mining</i> classification models in cardiovascular disease diagnosis (2011)	-	-	-	-	-	-	-	-	-
Review of heart disease prediction system using <i>Data Mining</i> and hybrid intelligent techniques (2013)	-	-	-	-	-	-	-	-	-
A <i>Data Mining</i> approach for prediction of heart disease using neural networks (2012)	NN	99.2	1	1	1	0.5	0	0	3.5
A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease (2012)	DT	93.2	1	1	1	0	0	1.5	4.5
An association rule mining-based methodology for automated detection of ischemic ECG beats (2006)	AssRule	87	1	1	1	1	0.5	1.5	6
Artificial neural network models for prediction of cardiovascular autonomic dysfunction in general Chinese population (2013)	ANN	-	1	1	0	0	1	1	4
Assessment of the risk factors of coronary heart events based on <i>Data Mining</i> with decision trees (2010)	C4.5	75	1	1	1	0	0.5	2	5.5
Association rule analysis for the assessment of the risk of coronary heart events (2009)	Apriori	-	1	1	1	0	0	0.5	3.5
Cardiac arrhythmia classification using neural networks with selected feature (2013)	NN	87.7	7	7	7	0.5	0	0	3.5
Classification of cardiac abnormalities using heart rate signals (2004)	ANN	85	1	1	1	0	0.5	0	3.5
Classification of heart disease using K-nearest neighbor and genetic algorithm (2013)	K-NN+GA	95	1	1	1	1	0	0	4
Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability (2013)	CART	85.4	1	0.5	1	1	1	0	4.5
Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules (2011)	FM	-	1	1	1	1	0	0	4
Biomedical (cardiac) <i>Data Mining</i> : extraction of significant patterns for predicting heart condition (2014)	K-means	-	1	1	1	1	0.5	0	4.5
Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches (2012)	C5	97.5	1	1	1	0	1	2	6
Heart disease prediction system using associative classification and genetic algorithm (2012)	G4	98	1	1	1	1	0	0	4

Titulo	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Intelligent heart disease prediction system using probabilistic neural network (2013)	PNN	94.6	1	1	1	0	0.5	0	3.5
Entropy-based <i>Data Mining</i> on the example of cardiac arrhythmia suppression (2014)	FM	-	1	1	1	0	0.5	1.5	5
Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal (2008)	SVM	98.5	1	1	1	1	0.5	1.5	6
Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal (Mohebbi & Ghassemian, 2012)	SVM	93.1	1	1	1	1	0.5	2	6.5
A new method for classification of ECG arrhythmias using neural network with adaptive activation function (2010)	NN	98.1	1	1	1	0	0	1.5	4.5
Combination <i>Data Mining</i> methods with new medical data to predicting outcome of coronary heart disease (2007)	SVM	92.1	1	1	1	0.5	0	0	3.5
Diagnosis of cardiovascular abnormalities from compressed ECG: a <i>Data Mining</i> Based approach (2011)	FM	97	1	1	1	0	0	2	5
Sparse representation-based heartbeat classification using independent component analysis (2012)	BN	98.3	1	1	1	0	0	1.5	4.5
A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in Type 2 Diabetes patients (2013)	K-NN+GA	96	1	1	1	0	0	0.5	3.5
Survival prediction in patients undergoing ischemic cardiopathy (2009)	ANN	56.2	1	1	1	1	0	1.5	5.5
Cardiovascular abnormality detection method using cardiac sound characteristic waveform with data clustering technique, International Conference on Control (2007)	FC-means	-	0	1	1	0.5	1	0	3.5
Cardiovascular abnormality detection method using cardiac sound characteristic waveform with data clustering technique, International Conference on Control (2007)	MLP-NN	95	1	1	1	1	0	0	4
Development of evolutionary <i>Data Mining</i> algorithms and their applications to cardiac disease diagnosis (2012)	K-means	90	1	1	1	1	0	1.5	5.5
Genetic neural network based <i>Data Mining</i> in prediction of heart disease using risk factos (2013)	NN	96.2	1	1	1	0	0	0.5	3.5
Association rule discovery with fuzzy decreasing support on Syndrome	FM	-	1	1	1	0.5	0	0	3.2

Titulo	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Differentiation and medication in coronary heart disease (2010)									
Effective diagnosis of coronary artery disease using the rotation forest ensemble method (2012)	ANN	91.2	1	1	1	1	0	1.5	5.5
Design of a fuzzy-based decision support system for coronary heart disease diagnosis (2012)	GA	86	1	1	1	1	0	1.5	5.5
Predictive models of prolonged stay after Coronary Artery Bypass surgery (2012)	LR	76.6	1	1	1	1	1	0	5
<i>Data Mining</i> approach for coronary artery disease screening (2011)	ID3	72.6	1	1	1	0	1	0.5	4.5
Heart disease classification using neural network and feature selection (2011)	ANN	89.5	1	1	1	0	0.5	0	3.5
Effective diagnosis of heart disease through neural networks ensembles (2009)	NN	98	1	1	1	1	0	1.5	5.5
Automated diagnosis of Coronary artery disease based on <i>Data Mining</i> and fuzzy modeling (2008)	DT+FM	80	1	1	1	1	0.5	2	6.5
Discrimination Power of Short-Term Heart Rate Variability Measures for CHF Assessment (Pecchia et al., 2011)	CART	79.3	1	1	1	1	0.5	2	6.5
Early detection of decompensation conditions in heart failure patients by knowledge discovery: the HEARTFAID approach (2008)	DT	92	1	1	1	0	0	1.5	4.5
Decision tree for adjuvant right ventricular support in patients receiving a left ventricular assist device (2012)	DT	85	1	1	1	1	0	0	4
Classification of the electrocardiogram signals using supervised classifiers and efficient features (2010)	NN	97.1	1	1	1	1	0	2	6
Detection of cardiac arrhythmia in electrocardiograms using adaptive feature extraction and modified support vector machines (2012)	SVM	98.9	1	1	1	1	0	1.5	5.5
Classification of Ventricular arrhythmia using a support vector machine based on morphological features (2013)	SVM	93.1	1	1	1	0	0	0.5	3.5
Using <i>Data Mining</i> techniques in heart disease diagnosis and treatment (2012)	-	-	-	-	-	-	-	0	-
A comparison of four <i>Data Mining</i> models: bayes, neural network, SVM and decision trees in identifying syndromes in Coronary heart disease (2007)	MLP-NN	89.2	1	1	1	1	0	0.5	4.5
Using decision tree for diagnosing heart disease patients (2011)	DT	77.9	1	1	1	1	0	0	4

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Heart sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals (2009)	ANN	94.4	1	1	1	0	0	1	4
Coronary artery disease prediction method using linear and nonlinear feature of heart rate variability in three recumbent postures (2009)	BN	85.7	1	1	1	1	0	1	5
Association rule mining to detect factors which contribute to heart disease in males and females (2013)	Apriori	99.3	1	1	1	0	0	1.5	4.5
Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization (2009)	ANN	91.2	1	1	1	0	0	1.5	4.5
Comparing performances of logistic regression, classification and regression tree (2008)	CART	78.7	1	1	1	0	0	1.5	4.5
Application of principal component analysis to ECG signals for automated diagnosis of cardiac health (2012)	LS-SVM	98.1	1	1	1	1	0	1.5	5.5
Exploring time series retrieved from cardiac implantable devices for optimizing patient follow-Up (2008)	FM	-	1	1	1	0	0	1.5	4.5
A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine (2010)	GA	79.1	1	1	1	1	0	1.5	4.5
A <i>Data Mining</i> approach for Coronary heart disease prediction using HRV features and carotid arterial wall thickness BMEI 2008 (2008)	SVM	-	1	1	1	1	0	0	4
An empirical study on applying <i>Data Mining</i> techniques for the analysis and prediction of heart disease (2013)	-	-	-	-	-	-	-	0	-
Support Vectors Machine-based identification of heart valve diseases using heart sounds (Maglogiannis et al., 2009)	SVM	93.4	1	1	1	1	1	2	7
Automatic detection of heart disease using na artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing (2007)	K-NN	87	1	1	1	1	0	1.5	5.5
Intelligent and effective heart disease prediction system using weighted associative classifiers (2011)	Ass Rule	75.8	7	7	7	0	1	0	4

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Ventricular fibrillation and tachycardia classification using a machine learning approach (Li et al., 2014)	SVM	98.1	1	1	1	1	1	1.5	6.5
Predicting the analysis of heart disease symptoms using medicinal <i>Data Mining</i> methods (2013)	C4.5	92	1	1	1	1	0	0	4
A <i>Data Mining</i> approach for diagnosis of coronary artery disease (2013)	NN	88.1	1	1	1	1	0	2	6
Intelligent heart disease prediction system using <i>Data Mining</i> techniques (2008)	DT	94.9	1	1	1	0	1	0.5	4.5
Heartbeat time series classification with support vector machines (2009)	SVM	93	1	1	1	0	0	2	5
Prediction of persistence of combined evidence-based cardiovascular medications in patients with acute coronary syndrome after hospital discharge using neural networks (2011)	NN	86.7	1	1	1	1	1	0	5
Association rule discovery with the train and test approach for heart disease prediction (2006)	Ass Rule	90	1	1	1	0	0	2	5
Heart sound screening in real-time assistive environments through MCMC Bayesian <i>Data Mining</i> (2014)	BN	89.2	1	1	1	1	0.5	1	5.5
Diagnosis of valvular heart disease through neural networks ensembles (2009)	NN	97.4	1	1	1	1	0	2	6
Effective analysis and predictive model of stroke disease using classification methods (2012)	NN	97.8	1	1	1	1	0	0	4
Knowledge discovery with classification rules in a cardiovascular dataset (2005)	GA+DT	95.3	0	1	1	1	0	2	5
Remote health monitoring of heart failure with <i>Data Mining</i> via CART method on HRV features (2011)	CART	96.3	1	1	1	1	0	1.5	5.5
Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning (2010)	K-NN	76	1	1	1	1	0	1.5	5.5
Unsupervised classification of ventricular extrasystoles using bounded clustering algorithms and morphology matching (2007)	-	99.7	1	1	1	1	0	0	4
Missing attribute value prediction based on artificial neural network and rough set theory (2008)	MLP-NN	73.3	1	1	1	1	0	0	4
Hybrid intelligent modeling schemes for heart disease classification (2014)	ANN	83.9	0	1	1	1	0.5	2	5.5

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Classification of heart rate data using artificial neural network and fuzzy equivalence relation (2003)	ANN	85	1	1	1	0	0	2	5
Heart diseases prediction based on ECG signals classification using a genetic-fuzzy system and dynamical model of ECG signals (2014)	FM+GA	98.6	1	1	1	0	0	1	4
Heartbeat classification using disease-specific feature selection (2014)	SVM	88.3	1	1	1	1	0	1	5
PcHD: personalized classification of heartbeat types using a decision tree (2014)	DT	99	1	1	1	1	0	1	5
Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system (2014)	FM	73.8	0	1	1	1	0.5	2	5.5
Electrocardiogram beat classification using empirical mode decomposition and multiclass directed acyclic graph support vector machine (2014)	SVM	98.9	1	1	1	1	0	1	5
Ventricular fibrillation and tachycardia classification using a machine learning approach (2014)	SVM	98.1	1	1	1	1	1	1.5	6.5
Hybrid classification engine for cardiac arrhythmia cloud service in elderly healthcare managemen (2014)	SVM+DT	85.7	1	1	1	1	0	1	5
Association rules mining on heart failure differential treatment based on the improved firefly algorithm (2014)	ARM	-	0	1	1	1	0.5	0	3.5
Multiclass maximum margin clustering via immune evolutionary algorithm for automatic diagnosis of electrocardiogram arrhythmias (Zhu et al., 2014)	Clustering	97.4	1	1	1	1	0.5	2	6.5
Prediction and management of readmission risk for congestive heart failure (2014)	-	-	1	1	0.5	0	0.5	0.5	3.5
Adaptive mining prediction model for content recommendation to coronary heart disease patients (2014)	FM+DT	69.2	0	1	1	1	0.5	1	4.5
Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images (Acharya et al., 2013)	K-NN	86	1	1	1	1	0.5	2	6.5
Identifying functional modules for coronary artery disease by a prior knowledge-based approach (2014)	NN	-	1	1	1	0	1	1	5
Prediction and diagnosis of cardio vascular disease—a critical survey (2014)	-	-	-	-	-	-	-	0	-

Titulo	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Expert knowledge integration in the <i>Data Mining</i> process with application to cardiovascular risk assessment (2015)	DT	78.5	1	1	1	1	0	0.5	4.5
<i>Data Mining</i> for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data (2015)	BN	83.1	1	1	1	0	1	2	6
Cardiac magnetic resonance image-based classification of the risk of arrhythmias in post-myocardial infarction patients (2015)	K-NN	83.3	1	1	1	1	0	1.5	5.5
Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques (2015)	SVM	98.9	1	1	1	1	0	2	6
Effect of multiscale PCA de-noising in ECG beat classification for diagnosis of cardiovascular diseases (2015)	SVM	99.9	1	1	1	1	0	1	5
A restricted Boltzmann machine based two-lead electrocardiography classification (2015)	LR	98.8	1	1	1	1	0.5	0.5	5
A multi-wavelet optimization approach using similarity measures for electrocardiogram signal classification (2015)	ANN	98.7	1	11	1	1	0	1	5
Assessment of cardiovascular risk based on a data-driven knowledge discovery approach (2015)	DT	78.8	1	1	1	1	0	0.5	4.5
Clustering of physical activities for quantified self and mHealth applications (2015)	HCA	82.4	1	1	0	1	0.5	0.5	4
Computational intelligence technique for early diagnosis of heart disease (2015)	NB	86.3	1	1	1	1	0	0	4
Predicting the likelihood of heart failure with a multi level risk assessment using decision tree (2015)	DT	86.5	1	1	1	1	0	0	4
Ventricular fibrillation and tachycardia classification using a machine learning approach (2015)	SVM+GA	98.1	1	1	1	1	0	0.5	4.5
BagMOOV: a novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting (2015)	SVM	84.7	1	1	1	1	0.5	1	5.5
Evaluating a Decision Tree-Based Approach for Cardiovascular Dysautonomias Diagnosis (2015)	C4.5	98.5	1	1	1	1	1	0	5
Analytical study of heart disease prediction comparing with different algorithms (2015)	-	-	-	-	-	-	-	-	-
Automatic classifier based on heart rate variability to identify fallers among hypertensive subjects (2015)	RF	67.3	0	1	1	1	0.5	0	3.5

Titulo	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
A web based cardiovascular disease detection system (2015)	C4.5	98.2	1	1	0.5	1	0	1.5	5
A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system (2015)	FM	84	1	1	1	1	0	0	4
Applications of <i>Data Mining</i> methods in the integrative medical studies of coronary heart disease: progress and prospect (2014)	-	-	-	-	-	-	-	-	-
Use of <i>Data Mining</i> techniques to determine and predict length of stay of cardiac patients (2013)	SVM	96	0	1	1	1	0.5	0	3.5
Coronary artery disease detection using a fuzzy-boosting PSO approach (2014)	PSO	85.7	1	1	1	1	0	1	5
Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes (2013)	DT	-	0	1	1	1	0.5	2	5.5
Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods (2012)	RF	82.6	0	0.5	0.5	1	0.5	1.5	4
Bayesian learning for cardiac SPECT image interpretation (2002)	BN	92.4	0	1	1	1	0	1.5	4.5
Automated ischemic beat classification using genetic algorithms and multicriteria decision analysis (2004)	FM	91	0	1	1	1	0	1.5	4.5
Fast clustering algorithm for large ECG data sets based on CS theory in combination with PCA and K-NN methods (2014)	K-NN	99.9	1	1	1	1	0.5	0.5	5
Multistage approach for clustering and classification of ECG data (2013)	NB	82.6	1	1	1	1	0	2	6
A heart disease recognition embedded system with fuzzy cluster algorithm (de Carvalho et al., 2013)	F-Clust	92	1	1	1	1	0.5	2	6.5
Using n-gram analysis to cluster heartbeat signals, BMC Med. Inform (2012)	K-means	91	1	1	1	1	0.5	1	5.5
Comparison of seven approaches for holter ECG clustering and classification (2007)	PNN	93.1	1	1	1	1	0	0.5	4.5
Modified ant colony clustering method in long-term electrocardiogram processing (2007)	Ant colony	79	1	1	1	1	0	0.5	4.5
Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines (2006)	SVM	81.2	0	1	1	1	0	1.5	4.5

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Support vector machine-based expert system for reliable heartbeat recognition (2004)	SVM	98	1	1	1	1	0	1.5	5.5
ECG beat recognition using fuzzy hybrid neural network (2001)	FM	96	1	0.5	1	1	0	1.5	5
A machine learning system to improve heart failure patient assistance (2014)	NN	84.7	0	1	1	1	0	1.5	4.5
Prediction of hospitalization due to heart diseases by supervised learning methods (2015)	LR	-	0	1	1	1	0.5	2	5.5
Modelling of inquiry diagnosis for coronary heart disease in Traditional Chinese Medicine by using multi-label learning (2010)	K-NN	76	0	1	1	1	0	2	5
Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images (2012)	DT	88.9	0	1	1	1	0	2	5
Non-invasive risk stratification of coronary artery disease: an evaluation of some commonly used statistical classifiers in terms of predictive accuracy and clinical usefulness (2009)	LR	98	0	1	1	1	0	1	4
Knowledge discovery with classification rules in a cardiovascular dataset (2005)	GA	95.3	0	1	1	1	0	2	5
Predicting coronary artery disease with medical profile and gene polymorphisms data (2007)	MLP-NN	86.2	0	1	1	1	0	1	4
Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression (2000)	CART	69.1	1	1	1	1	0	1.5	5.5
Combining multiple ECG features does not improve prediction of defibrillation outcome compared to single features in a large population of out-of-hospital cardiac arrests (2015)	NN	80	0	1	1	1	0.5	1	4.5
Machine learning approach for sudden cardiac arrest prediction based on optimal heart rate variability features (2014)	SVM	96.3	1	1	1	1	0.5	1	5.5
ECG arrhythmia classification using a probabilistic neural network with a feature reduction method (2013)	PNN	99.7	1	1	1	1	0	1.5	5.5
Comparison of prediction model for cardiovascular autonomic dysfunction using artificial neural network and logistic regression analysis (2013)	LR	67	0	1	1	1	0.5	2	5.5
A neural network approach to predicting outcomes in heart failure using cardiopulmonary exercise testing (2013)	ANN	-	0	1	1	1	0.5	2	5.5

Título	Classificação		Avaliação da qualidade						
	RQ1	RQ3(%)	A	B	C	D	E	F	Score
Robust algorithm for arrhythmia classification in ECG using extreme learning machine (2009)	ELM	98.7	1	1	1	1	0.5	1	5.5
Classification of electrocardiogram signals with support vector machines and particle swarm optimization (2008)	PSO-SVM	89.7	1	1	1	1	0.5	2	6.5
Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods (2008)	SVM	98.6	1	1	1	1	0	1.5	5.5
A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery—Part II: an illustrative example (2007)	BN	81.5	0	1	1	1	0	1	4
Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks (2006)	ANN	-	1	1	1	1	0.5	2	6.5
New algorithm of mortality risk prediction for cardiovascular patients admitted in intensive care unit (2015)	MLP-NN	96.4	1	1	1	1	0.5	1	5.5
High efficient system for automatic classification of the electrocardiogram beats (2011)	SVM	95.8	1	1	1	1	0	2	6
Arrhythmia detection and classification using morphological and dynamic features of ECG signals (2010)	SVM	99.6	1	1	1	1	0	0.5	4.5
Accurate prediction of coronary artery disease using reliable diagnosis system (2012)	RF	96.4	0	1	1	1	0	1.5	4.5
Prediction models for early risk detection of cardiovascular event (2012)	LR	56	0	1	1	1	0	1.5	4.5
A multi-stage automatic arrhythmia recognition and classification system (2011)	K-NN	85.6	1	1	1	1	0	1	5

Anexo II – Matrizes de Confusão da Abordagem 1

Decision Tree, DT

```
a    b    <-- classified as
21   20 |    a = 1
12 2874 |    b = 2
```

Figura 41 - Matriz de confusão C1, DT, Abordagem 1

```
a    b    <-- classified as
10   31 |    a = 1
6 2880 |    b = 2
```

Figura 42 - Matriz de confusão C2, DT, Abordagem 1

```
a    b    <-- classified as
21   20 |    a = 1
12 2874 |    b = 2
```

Figura 43 - Matriz de confusão C3, DT, Abordagem 1

```
a    b    <-- classified as
14   27 |    a = 1
19 2867 |    b = 2
```

Figura 44 - Matriz de confusão C4, DT, Abordagem 1

```
a    b    <-- classified as
18   23 |    a = 1
13 2873 |    b = 2
```

Figura 45 - Matriz de confusão C5, DT, Abordagem 1

Support Machine Vector, SVM

```
a    b    <-- classified as
0    41 |    a = 1
0 2886 |    b = 2
```

Figura 46 - Matriz de confusão C1, SVM, Abordagem 1

```
a    b    <-- classified as
0    41 |    a = 1
0 2886 |    b = 2
```

Figura 47 - Matriz de confusão C2, SVM, Abordagem 1

```
a    b    <-- classified as
0    41 |    a = 1
0 2886 |    b = 2
```

Figura 48 - Matriz de confusão C3, SVM, Abordagem 1

```
a    b    <-- classified as
0    41 |    a = 1
0 2886 |    b = 2
```

Figura 49 - Matriz de confusão C4, SVM, Abordagem 1

```
a    b    <-- classified as
0    41 |    a = 1
0 2886 |    b = 2
```

Figura 50 - Matriz de confusão C5, SVM, Abordagem 1

Naive Bayes, NB

```
a    b    <-- classified as
21   20 |    a = 1
115 2771 |    b = 2
```

Figura 51 - Matriz de confusão C1, NB, Abordagem 1

```
a    b    <-- classified as
0    41 |    a = 1
120 2766 |    b = 2
```

Figura 52 - Matriz de confusão C2, NB, Abordagem 1

```
a    b    <-- classified as
21   20 |    a = 1
36 2850 |    b = 2
```

Figura 53 - Matriz de confusão C3, NB, Abordagem 1

```
a    b    <-- classified as
18   23 |    a = 1
35 2851 |    b = 2
```

Figura 54 - Matriz de confusão C4, NB, Abordagem 1

```
a    b    <-- classified as
18   23 |    a = 1
35 2851 |    b = 2
```

Figura 55 -- Matriz de confusão C5, NB, Abordagem 1

Anexo III - Matrizes de Confusão da Abordagem 1, Método Sampling

Decision Tree, DT

```
a    b    <-- classified as
2870  0 |    a = 1
  46 2840 |    b = 2
```

Figura 56 - Matriz de confusão C1, DT, Abordagem 1, Método Sampling

```
a    b    <-- classified as
980 1890 |    a = 1
  6 2880 |    b = 2
```

Figura 57 - Matriz de confusão C2, DT, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2870  0 |    a = 1
  47 2839 |    b = 2
```

Figura 58 - Matriz de confusão C3, DT, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2870  0 |    a = 1
  39 2847 |    b = 2
```

Figura 59 - Matriz de confusão C4, DT, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2870  0 |    a = 1
  41 2845 |    b = 2
```

Figura 60 - Matriz de confusão C5, DT, Abordagem 1, Método Sampling

Support Decison Vector, SVM

```
a    b    <-- classified as
1719 1151 |    a = 1
39 2847 |    b = 2
```

Figura 61 - Matriz de confusão C1, SVM, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2870    0 |    a = 1
1674 1212 |    b = 2
```

Figura 62 - Matriz de confusão C2, SVM, Abordagem 1, Método Sampling

```
a    b    <-- classified as
1740 1130 |    a = 1
39 2847 |    b = 2
```

Figura 63 - Matriz de confusão C3, SVM, Abordagem 1, Método Sampling

```
a    b    <-- classified as
1750 1120 |    a = 1
41 2845 |    b = 2
```

Figura 64 - Matriz de confusão C4, SVM, Abordagem 1, Método Sampling

```
a    b    <-- classified as
1740 1130 |    a = 1
39 2847 |    b = 2
```

Figura 65 -- Matriz de confusão C5, SVM, Abordagem 1, Método Sampling

Naive Bayes, NB

```
a    b    <-- classified as
2660 210 |    a = 1
171 2715 |    b = 2
```

Figura 66 - Matriz de confusão C1, NB, Abordagem 1, Método Sampling

```
a    b    <-- classified as
1050 1820 |    a = 1
140 2746 |    b = 2
```

Figura 67 - Matriz de confusão C2, NB, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2643 227 |    a = 1
1932 954 |    b = 2
```

Figura 68 - Matriz de confusão C3, NB, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2223 647 |    a = 1
252 2634 |    b = 2
```

Figura 69 - Matriz de confusão C4, NB, Abordagem 1, Método Sampling

```
a    b    <-- classified as
2264 606 |    a = 1
41 2845 |    b = 2
```

Figura 70 - Matriz de confusão C5, NB, Abordagem 1, Método Sampling

Anexo IV - Matrizes de Confusão da Abordagem 2

Decision Tree, DT

```
      a      b      c      d  <-- classified as
1418      2      4      5 |      a = 1
      2      2      5      3 |      b = 2
      3      3 1423      9 |      c = 3
      15      2      6     25 |      d = 4
```

Figura 71 - Matriz de confusão C1, DT, Abordagem 2

```
      a      b      c      d  <-- classified as
621      0 808      0 |      a = 1
      0      0      8      4 |      b = 2
696      1 741      0 |      c = 3
      0      4     31     13 |      d = 4
```

Figura 72 - Matriz de confusão C2, DT, Abordagem 2

```
      a      b      c      d  <-- classified as
1418      2      4      5 |      a = 1
      2      4      3      3 |      b = 2
      3      3 1422     10 |      c = 3
      15      2      6     25 |      d = 4
```

Figura 73 - Matriz de confusão C3, DT, Abordagem 2

```
      a      b      c      d  <-- classified as
1418      2      4      5 |      a = 1
      3      0      2      7 |      b = 2
      3      2 1418     15 |      c = 3
      16      2      8     22 |      d = 4
```

Figura 74 - Matriz de confusão C4, DT, Abordagem 2

	a	b	c	d	<-- classified as
1418	2	4	5		a = 1
3	0	4	5		b = 2
3	2	1422	11		c = 3
16	4	10	18		d = 4

Figura 75 - Matriz de confusão C5, DT, Abordagem 2

Support Decision Vector, SVM

	a	b	c	d	<-- classified as
1426	0	3	0		a = 1
3	0	9	0		b = 2
18	0	1420	0		c = 3
26	0	22	0		d = 4

Figura 76 - Matriz de confusão C1, SVM, Abordagem 2

	a	b	c	d	<-- classified as
615	0	814	0		a = 1
5	0	7	0		b = 2
633	0	805	0		c = 3
16	0	32	0		d = 4

Figura 77 - Matriz de confusão C2, SVM, Abordagem 2

	a	b	c	d	<-- classified as
1426	0	3	0		a = 1
3	0	9	0		b = 2
18	0	1420	0		c = 3
26	0	22	0		d = 4

Figura 78 - Matriz de confusão C3, SVM, Abordagem 2

	a	b	c	d	<-- classified as
1426	0	3	0		a = 1
3	0	9	0		b = 2
18	0	1420	0		c = 3
25	0	23	0		d = 4

Figura 79 - Matriz de confusão C4, SVM, Abordagem 2

	a	b	c	d	<-- classified as
1426	0	3	0		a = 1
3	0	9	0		b = 2
18	0	1420	0		c = 3
25	0	23	0		d = 4

Figura 80 - Matriz de confusão C5, SVM, Abordagem 2

Naive Bayes, NB

	a	b	c	d	<-- classified as
1086	7	272	64		a = 1
1	0	4	7		b = 2
5	10	1368	55		c = 3
16	1	14	17		d = 4

Figura 81 - Matriz de confusão C1, NB, Abordagem 2

	a	b	c	d	<-- classified as
485	6	882	56		a = 1
0	0	12	0		b = 2
594	8	780	56		c = 3
2	0	44	2		d = 4

Figura 82 - Matriz de confusão C2, NB, Abordagem 2

	a	b	c	d	<-- classified as
1131	0	281	17		a = 1
1	0	4	7		b = 2
5	1	1413	19		c = 3
16	1	14	17		d = 4

Figura 83 - Matriz de confusão C3, NB, Abordagem 2

	a	b	c	d	<-- classified as
1131	0	281	17		a = 1
1	0	8	3		b = 2
7	0	1413	18		c = 3
17	1	15	15		d = 4

Figura 84 - Matriz de confusão C4, NB, Abordagem 2

	a	b	c	d	<-- classified as
1131	0	281	17		a = 1
1	0	8	3		b = 2
7	1	1413	17		c = 3
17	1	15	15		d = 4

Figura 85 - Matriz de confusão C5, NB, Abordagem 2