Universidade do Minho
Escola de Engenharia

Pamela de Figueiredo Coelho

Modelos de Data Mining como Serviço - Interpretação de Imagens

outubro de 2018

Modelos de Data Mining como Serviço - Interpretação de Imagens

Pamela Coelho

UMinho | 2018

Universidade do Minho
Escola de Engenharia

Pamela de Figueiredo Coelho

Modelos de *Data Mining* como Serviço – Interpretação de Imagens

Dissertação de Mestrado

Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação de:

Professor Doutor Carlos Filipe da Silva Portela

Professor Doutor Manuel Filipe Vieira Torres Santos

Outubro de 2018

DECLARAÇÃO

Nome: Pamela de Figueiredo Coelho

Endereço eletrónico: a71179@alunos.uminho.pt **Telefone:** 915342510

Cartão de Cidadão: 14630427

Título da dissertação: Modelos de Data Mining como Serviço – Interpretação de Imagens

Orientador:

Carlos Filipe da Silva Portela

Ano de conclusão: 2018

Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Declaro que concedo à Universidade do Minho e aos seus agentes uma licença não-exclusiva para arquivar e tornar acessível, nomeadamente através do seu repositório institucional, nas condições abaixo indicadas, a minha tese ou dissertação, no todo ou em parte, em suporte digital.

Declaro que autorizo a Universidade do Minho a arquivar mais de uma cópia da tese ou dissertação e a, sem alterar o seu conteúdo, converter a tese ou dissertação entregue, para qualquer formato de ficheiro, meio ou suporte, para efeitos de preservação e acesso.

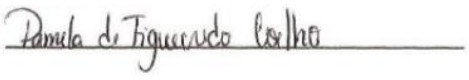
Retenho todos os direitos de autor relativos à tese ou dissertação, e o direito de a usar em trabalhos futuros (como artigos ou livros).

Concordo que a minha tese ou dissertação seja colocada no repositório da Universidade do Minho com o seguinte estatuto (assinale um):

1. Disponibilização imediata do conjunto do trabalho para acesso mundial;
2. Disponibilização do conjunto do trabalho para acesso exclusivo na Universidade do Minho durante o período de 1 ano, 2 anos ou 3 anos, sendo que após o tempo assinalado autorizo o acesso mundial.
3. Disponibilização do conjunto do trabalho para acesso exclusivo na Universidade do Minho.

Universidade do Minho, 22 de outubro de 2018

Assinatura:



AGRADECIMENTOS

Finalizado mais um capítulo da vida, nada mais importante que agradecer a quem contribuiu para que chegasse até aqui e a quem, de certa forma, quis fazer parte desta jornada.

Em primeiro lugar agradeço ao Professor Carlos Filipe da Silva Portela por todo o apoio e tempo disponibilizado a suportar qualquer dúvida, a qualquer momento. Apesar de não ter sido um caminho com muita interação, sabia que poderia contar com todo o seu auxílio. Deixo-lhe, desde já, um muito obrigada.

Em segundo lugar quero agradecer aos meus pais e aos meus irmãos que nunca desistiram e sempre estiveram lá para me incentivar a nunca desistir. Obrigada por tudo isso e muito mais. Quero agradecer também à minha tia Lucília com quem muito aprendi e a quem muito devo, ao longo deste percurso.

Por último, mas não menos importante, quero agradecer às pessoas que me acompanharam nesta vida académica e relembrar que ainda há muito para viver. Quero agradecer a quem partilha os dias comigo e sabe que nem sempre é fácil, mas é sempre a aprender. Quero agradecer a quem já esteve e continua a estar, a quem já esteve e foi embora e a quem ainda está para chegar. Como alguém um dia disse: Nada se cria, nada se perde, tudo se transforma.

A vocês, um gigantesco obrigada!

RESUMO

O volume de dados produzidos pelas diversas organizações, de inúmeras áreas, tem vindo a aumentar de forma acentuada ao longo dos anos. A capacidade humana não permite a análise eficiente desta quantidade de elementos, pelo que é necessário desenvolver sistemas que possibilitem a descoberta de conhecimento. As imagens são componentes que contêm informação relevante para áreas como a medicina. As imagens raio X ou imagens de lesões na pele suportam o diagnóstico de um profissional de saúde. Contudo, um profissional de saúde nem sempre consegue ser preciso neste processo. Deste modo, os sistemas de apoio à decisão vieram suportar o conhecimento necessário para a adoção da decisão acertada.

No âmbito do projeto *Deux ex Machina* foi desenvolvido um protótipo designado *Pervasive Data Mining Engine*. Esta plataforma permite efetuar, em tempo-real, processos de classificação e regressão. Neste momento, este protótipo não permite a realização de processos de análise, classificação e indexação de imagens, de modo percetível. Assim sendo, um artefacto de análise de imagens provenientes do projeto foi elaborado, com o intuito da sua posterior integração na plataforma.

A metodologia *Cross Industry Standard Process for Data Mining*, enquadrada com a metodologia *Design Science Research*, permitiu a exploração de modelos preditivos de classificação, como as *Convolutional Neural Networks*. Já na exploração de modelos descritivos de *clustering*, o método *k-means* foi explorado. No total foram desenvolvidos três modelos. O primeiro com acuidade de 91%, o segundo com acuidade de 44% e o terceiro com a demonstração da aplicação do método de *clustering*.

O presente documento corresponde ao relatório de Dissertação, onde, para além da exploração e apresentação de conceitos relevantes ao tema do projeto, também apresenta uma componente prática de exploração de modelos de *Data Mining* aplicados a imagens de lesões na pele.

PALAVRAS - CHAVE

Data Mining, *Image Mining*, *Image Mining* na Saúde, Classificação, *Clustering*, Sistemas de Apoio à Decisão, *Pervasive Data Mining Engine, Convolutional Neural Networks,* Melanomas

Abstract

The volume of data produced by several organizations in different areas has been increasing sharply over the years. The human capacity does not allow an efficient analysis of this quantity of records, that is why it is necessary to develop systems that allow the automatic discovery of knowledge. Images are components that contain relevant information to areas such as medicine. An X-ray image, as well as a skin lesion image, supports the diagnosis of a health professional. However, a health professional cannot be always precise in this process. Therefore, the decision support systems have come to bear the knowledge necessary for the right decision's adoption.

Under the project Deux ex Machina, a prototype called Pervasive Data Mining Engine was developed. This platform allows real-time processing of classification and regression tasks. At this moment, this prototype does not allow the implementation of processes of analysis, classification and indexing of images, in a perceptible way. Therefore, an image analysis artifact was developed for later integration into the platform.

The methodology Cross Industry Standard Process for Data Mining, linked with the Design Science Research, allowed the implementation of predictive classification models, such as Convolutional Neural Networks. In the exploration of descriptive models of clustering, the k-means method was explored. In total, three models were performed. The first with accuracy of 91%, the second with accuracy of 44% and the third with a demonstration of the clustering application method.

This document corresponds to the Dissertation report, where in addition to the exploration and presentation of concepts relevant to the project theme, it also presents a practical component of exploration of Data Mining models applied to images of skin lesions.

Keywords

Data Mining, Image Mining, Medical Image Mining, Classification, Clustering, Decision Support Systems, Pervasive Data Mining Engine, Convolutional Neural Networks, Melanoms

Lista de Abreviaturas, Siglas e Acrónimos

ACC - *Accuracy*

BI – *Business Intelligence*

BMP – *Windows Bitmap*

CMYK – *Cyan, Magenta, Yellow and Black color model*

CNN – *Convolutional Neural Networks*

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

DEM – *Deux ex Machina*

DL – *Deep Learning*

DM – *Data Mining*

DSRM – *Design Science Research Methodology*

DWT – *Discrete Wavelet Transform*

EPS – *Encapsulated PostScript*

ETL – *Extraction, Transformation and Loading*

GIF – *Graphics Interchange Format*

HSV – *Hue, Saturation and Value color model*

IDS *Group – Intelligent Data Systems Group*

IM – *Image Mining*

IoT – *Internet of Things*

JPEG – *Joint Photographic Experts Group*

KDD – *Knowledge Discovery from Data*

OLAP – *Online Analytical Processing*

PACS – *Picture Archiving and Communication Systems*

PBM – *Portable Bitmap Format*

PDF – *Portable Document Format*

PDME – *Pervasive Data Mining Engine*

PNG – *Portable Network Graphics*

RGB – *Red, Green and Blue color model*

SVG – *Scalable Vectorial Graphics*

TIFF – *Tagged Image File Format*

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O presente projeto encontra-se dividido em duas fases principais de elaboração. A primeira fase consiste na compreensão dos objetivos do projeto e na recolha de informação sobre o tema a explorar - interpretação de imagens com recurso a modelos de *Data Mining*. Já a segunda fase consiste na aplicação do conhecimento adquirido para criação de modelos classificativos e descritivos, que, tal como o nome indica, permitam classificar e agrupar imagens de melanomas.

Assim sendo, neste capítulo será apresentada o enquadramento e motivação para a execução do projeto; os objetivos a alcançar, com o desenvolvimento do mesmo; a metodologia escolhida para procura de informação sobre o tema; e, por último, a estrutura do relatório.

1.1. Enquadramento e Motivação

A quantidade de dados, provenientes de diversas áreas e/ou atividades, tem vindo a aumentar ao longo dos anos. A necessidade de trabalhar os dados para perceção de padrões que proporcionem informação útil tem sido uma constante.

Uma vertente dos dados obtidos são as imagens. Ao contrário dos dados popularmente explorados, as imagens diferem no tipo e na forma como são adquiridas e armazenadas. Deste modo, o tratamento de imagens é um domínio de importância elevada para o reconhecimento de padrões significantes para aquisição de conhecimento. Na área da medicina, a análise de um raio X ainda é efetuada pelo profissional de saúde responsável. Segundo um estudo relativo à deteção de pneumonia, através de uma imagem raio X do peito (Rajpurkar et al., 2017), foi desenvolvido um algoritmo que permite diagnosticar pneumonia com maior precisão do que radiologistas experientes. A utilização de bases de dados que contenham uma enorme quantidade de imagens relacionadas com a existência de anomalias, como um osso partido, permitirão o diagnóstico com maior certeza. Isto porque, a comparação de uma imagem *input*, com os exemplos existentes, permitirá um *output* – o diagnóstico. Esta comparação será possível, pois as imagens são constituídas por pixéis que de acordo com as suas características, nomeadamente cor, textura, localização, entre outras; possibilitarão a deteção de possíveis anomalias.

A organização responsável pelo projeto é a *Intelligent Data Systems Group* (IDS), do *Algoritmi Research Centre*, sediada no campus de Azúrem da Universidade do Minho. O principal objetivo da organização é proporcionar a investigação em diversas áreas como *Adaptative Business Intelligence*, *Intelligent Decision Support Systems*, *Data Mining*, entre outros domínios. As

investigações pretendem abordar problemas complexos, em tempo-real, distribuídos e online, para sustentação de problemas emergentes. Por conseguinte, no âmbito do projeto de investigação *Deux ex Machina* (DEM) foi elaborado um protótipo de uma plataforma que permite, em tempo real, facilitar o modo como são desenvolvidos os modelos de DM. Esta plataforma, denominada *Pervasive Data Mining Engine* (PDME), está preparada para executar processos integrais de classificação e regressão, mas não está desenvolvida ao ponto de efetuar processos de análise, classificação ou indexação de imagens, de forma compreensível. Assim sendo, foi elaborada uma componente de análise de um conjunto de imagens de lesões da pele. A PDME está concebida para proporcionar um acesso remoto aos utilizadores, ou seja, esta poderá ser acedida em qualquer momento e em qualquer lugar. O artefacto desenvolvido poderá, possivelmente num trabalho futuro, ser trabalhado de forma à posterior integração na plataforma.

Os modelos de *Data Mining* aplicados, por si só, a este domínio, não conseguem alcançar os resultados satisfatórios esperados para o processamento da imagem (Zahradnikova, Duchovicova, & Schreiber, 2015). Assim sendo, com recurso a *packages* de processamento de imagem *EBImage*, a *packages* para modelação e avaliação, como o *TensorFlow* e o *Keras*, foi possível explorar técnicas de DM aplicadas a imagens de lesões na pele. No desenvolvimento da componente e exploração de todo o processo, a ferramenta R (*open-source tool*) foi manipulada com o auxílio das funcionalidades do *Anaconda Navigator*.

1.2. Objetivos

Ao longo desta dissertação conceitos como Inteligência Artificial, Estatísticas, Reconhecimento, *Data Mining*, *Machine Learning*, *Image Mining*, Processamento de Imagem, Recuperação de Imagem, entre outros, foram abordados. Com isto pretendeu-se que o devido conhecimento fosse adquirido, para que o projeto decorresse da melhor e mais coerente forma possível.

Com base nos objetivos e resultados alcançados pretendeu-se, para o presente projeto de dissertação, expor um contributo científico de resposta à seguinte questão de investigação:

- De que forma os modelos de *Data Mining* suportam a interpretação de imagens?

De acordo com o ponto 1.1 Enquadramento e Motivação, os objetivos principais deste projeto de dissertação são os seguintes:

- Explorar e analisar *datasets* de imagens provenientes do projeto DEM;
- Aplicar métodos de classificação e *clustering*.

Relativamente aos objetivos secundários do projeto de dissertação, podem ser enumerados os seguintes:

- Explorar a ferramenta R e a plataforma PDME;
- Adquirir conhecimento de novos algoritmos inteligentes para a identificação de padrões.

No capítulo 6 Conclusões, os objetivos foram novamente referidos de forma a perceber se foram ou não alcançados no término do projeto.

1.3. Metodologia de Investigação

No âmbito da pesquisa científica, a metodologia de investigação utilizada foi a *Design Science Research Methodology* (DSRM). A missão desta metodologia, segundo Van Aken, é desenvolver conhecimento que permita a construção e desenho de artefactos – ou seja, solucionar problemas de construção -, ou melhorar a performance de entidades existentes – isto é, solucionar problemas de melhoria (citado de Van Aken, 2004, p.224).

De acordo com a aplicação da metodologia DSRM, pretendeu obter-se resposta às seguintes questões (Dresch, Lacerda, & Antunes, 2015):

- Qual o problema que se pretende resolver com a invenção, ou desenvolvimento, de um novo artefacto premeditado?
- Qual o benefício da sua execução?
- Para quem?
- Quão significante seria solucionar o problema?

Assim sendo, e de acordo com o trabalho realizado, a resposta à primeira questão corresponde ao problema da interpretação de imagens recorrendo a modelos de *Data Mining*. Ou seja, pretendeu-se explorar e adquirir conhecimento sobre o processo de análise, classificação e *clustering* de imagens clínicas; com base na aplicação de modelos inteligentes. A elaboração do presente projeto de dissertação permitiu a criação de uma componente a ser integrada, numa fase seguinte a este trabalho, na plataforma PDME. Com isto, a plataforma evolui na capacidade de processar modelos de classificação e *clustering*, de imagens, em tempo-real. O tema da dissertação surgiu no âmbito do projeto de investigação *Deux ex Machina* (DEM) - da organização

Intelligent Data Systems Group (IDS) - e, como tal, o relatório foi elaborado com o intuito de satisfazer a necessidade do mesmo. A componente permite servir de base à integração de novas funcionalidades de análise à ferramenta PDME. Desta forma, a solução é muito relevante para o projeto DEM, pois permite a obtenção de uma ferramenta mais íntegra de análise de dados em tempo-real .

No contexto desta dissertação, tal como referido anteriormente, um dos objetivos principais compreendeu o desenvolvimento de uma componente de análise, e exploração, de *datasets* de imagens relativas a diagnóstico e terapêutica. Esta componente corresponde ao artefacto produzido, em conformidade com a metodologia DSRM. Este artefacto permite obter resposta à questão de investigação mencionada no ponto 1.2 Objetivos e insere-se no projeto de investigação DEM. O produto final será, posteriormente, aproveitado para futuras investigações e integrado na plataforma PDME, que proporcionará o acesso, em tempo-real, a modelos de DM a aplicar a variados processos. Os artefactos premeditados, no caso dos sistemas de informação, incluem sistemas, métodos, metodologias, procedimentos, práticas, teorias e outras tecnologias para problemas específicos (Dresch et al., 2015). Existem três tipos de artefactos: artefactos do tipo produto ou aplicação; artefactos relativos a metodologias, processos ou intervenções; e, por último, artefactos que englobam os dois contextos (Gregor & Jones, 2007). Neste caso, o artefacto será uma componente de integração em *software*, ou seja, insere-se no primeiro tipo de artefactos.

Segundo Alan Hevner (2007), a *framework* de investigação nos sistemas de informação foca três ciclos de investigação: o ciclo de relevância, o ciclo do projeto e o ciclo do rigor.

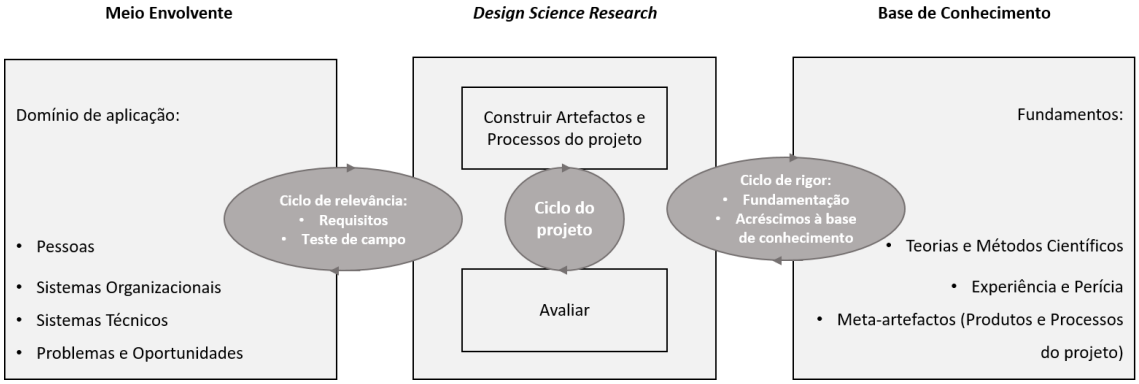


Figura 1 - Ciclos da metodologia DSRM (adaptado de Hevner, 2007*)*

Com base na Figura 1 é possível averiguar as ligações e posicionamento dos três ciclos da metodologia DSRM. O ciclo de relevância une o meio contextual, do projeto de investigação, com as atividades da metodologia. O ciclo de rigor conecta as atividades da metodologia com a base de conhecimento dos fundamentos científicos, da experiência e da perícia que orienta o projeto de

investigação. O ciclo do projeto permite a interação entre as atividades principais de construção e avaliação dos artefactos e processos do projeto de investigação.

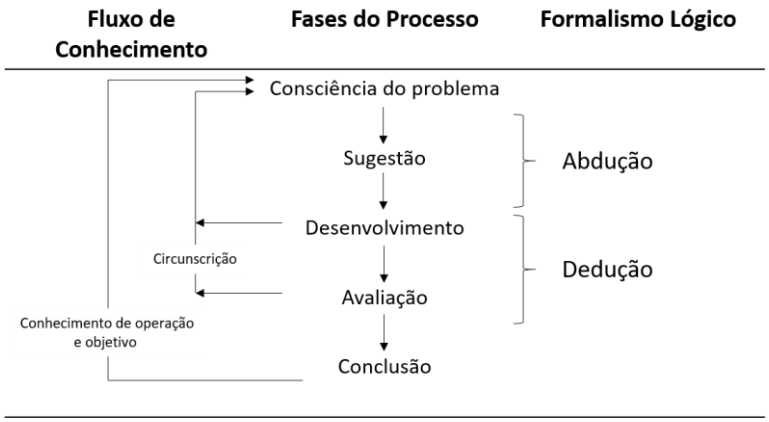


Figura 2 - Raciocínio presente no ciclo do projeto (adaptado de Vaishnavi & William Kuechler, 2007*)*

No ciclo do projeto, segundo Vaishnavi e Kuechler (2007), existe um raciocínio que se enquadra nesse processo. Com base na Figura 2 é possível compreender as fases desse raciocínio que demonstram a evolução do ciclo do projeto.

Desta forma o processo da metodologia DSRM, de acordo com Peffers, Tuunanen, Rothenberger e Chatterjee (2008), pode ser visualizado na Figura 3.

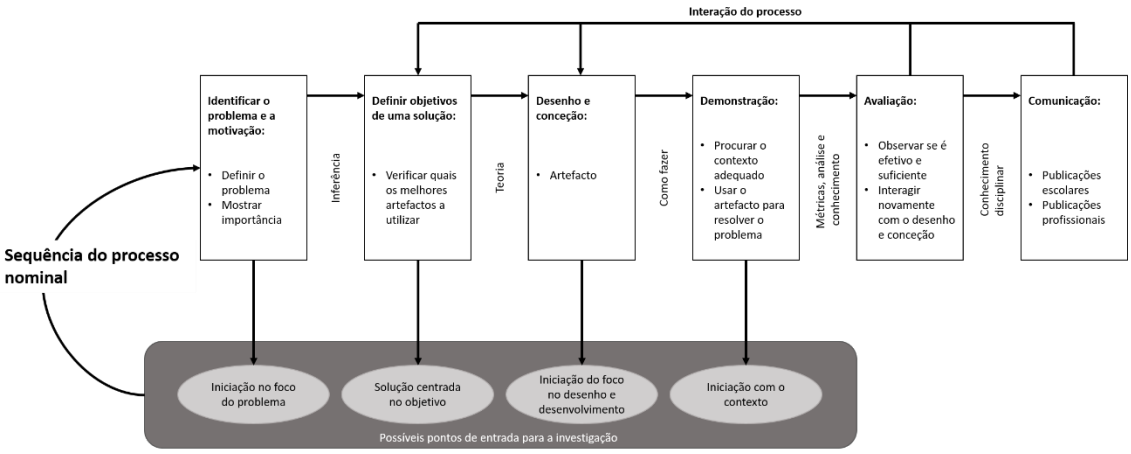


Figura 3 - Processo de DSRM (adaptado de Peffers et al., 2008*)*

O processo de interação da metodologia DSRM consiste na sequência do seguinte conjunto de passos:

1. Identificar o problema e motivação: o objetivo desta fase passa pela definição do problema inerente, de investigação, e pela justificação do valor de uma solução. Com o intuito da justificação do valor da solução, o investigador tem de estar motivado a averiguar a solução e a aceitar os seus resultados. Assim foi possível compreender o raciocínio integrante à

interpretação do problema. Os recursos necessários correspondem ao conhecimento do estado do problema e à importância da solução. No caso desta dissertação, esta etapa coincide com a revisão da literatura efetuada sobre os conceitos relevantes à execução da mesma.

2. Definir objetivos de uma solução: nesta fase os objetivos da solução são apresentados, tendo em conta a definição e consciencialização do que é possível e viável. Caso a solução expectável seja melhor do que as existentes, o objetivo é classificado como quantitativo. No entanto, se a descrição das espectativas do novo artefacto for apoiar soluções para problemas não abordados, o objetivo é classificado como qualitativo. Nesta etapa, recursos como o conhecimento do estado dos problemas e de atuais soluções existentes são preponderantes. Na presente dissertação foi averiguada a existência de trabalhos realizados na área de diagnóstico e terapêutica e definidos objetivos para a conceção de um resultado, que se espera ser positivo, tendo em conta a questão de investigação.
3. Desenho e conceção: um modelo de um artefacto de pesquisa pode ser considerado como sendo um esquema que integre uma contribuição para uma investigação. Nesta fase o artefacto foi desenvolvido, onde foram manipulados modelos, métodos, instâncias, entre outros. O conhecimento sobre a teoria empregue numa solução foi o meio necessário para transmitir os objetivos para a conceção. Na etapa seguinte, o artefacto, que incluiu os modelos de DM aplicados, foi desenvolvido.
4. Demonstração: nesta fase o objetivo compreende-se pela demonstração da aplicação do artefacto para a resolução de uma ou mais instâncias do problema. O conhecimento efetivo de como utilizar o artefacto para resolução do problema foi evidente nos recursos essenciais para esta etapa. Esta fase do projeto de dissertação expôs os resultados atingidos ao nível dos modelos aplicados.
5. Avaliação: nesta fase foi fundamental perceber o nível de sustentação, relativo ao artefacto, perante a solução do problema. A comparação entre os objetivos de uma solução, face aos resultados reais observados, foi efetuada, com base na aplicação do artefacto na demonstração, que necessita de métricas relevantes e técnicas de análise. Esta fase pode adquirir diversos formatos consoante a natureza do problema e do artefacto. No final deste processo, os responsáveis podem optar por voltar à fase 3 – Desenho e Conceção – para melhorar a eficácia do artefacto; ou avançar para a fase 6 – Comunicação – e depor melhorias para projetos futuros. A natureza da pesquisa ditou se

a iteração foi possível ou não. No final da etapa de demonstração, do presente projeto de dissertação, foi efetuada uma avaliação, de forma a perceber se os resultados foram de encontro aos objetivos estipulados. A conclusão da avaliação ocorreu na tomada de decisão entre a opção de voltar a desenhar, e conceber o desenvolvimento do projeto, ou, se os objetivos foram alcançados, permitindo a continuação do projeto para a última fase.

6. Comunicação: o objetivo desta fase final consiste na comunicação do problema e a sua importância, do artefacto, da sua utilidade, do rigor na conceção e da sua eficácia para os responsáveis e público-alvo como, por exemplo, profissionais na área do diagnóstico e terapêutica. Em publicações de pesquisa académica, os investigadores podem utilizar a estrutura deste processo para organizar o artigo, bem como a estrutura nominal de um processo de pesquisa empírica. A comunicação solicita o conhecimento da cultura disciplinar. Tendo em conta o presente projeto de dissertação, esta etapa pode ser visível aquando do desenvolvimento de relatórios, ou artigos científicos, providenciados à comunidade científica e, noutro cenário, numa apresentação final, pública, do projeto de dissertação.

A contribuição para o conhecimento da solução de resposta à questão de investigação, segundo Gregor e Hevner (2013), pode inserir-se em quatro momentos, como descrito na Figura 4.

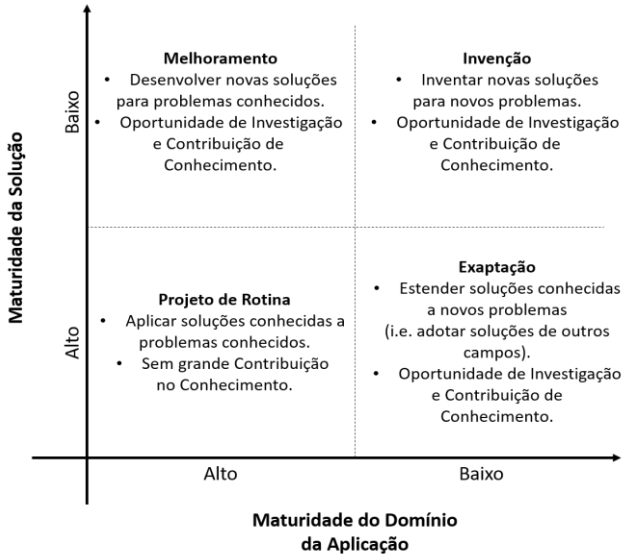


Figura 4 - Contribuição do conhecimento (adaptado de Gregor & Hevner, 2013*)*

Caso a maturidade da solução e a maturidade do domínio de aplicação sejam ambas altas, não existe grande contribuição no conhecimento, pelo que o artefacto a produzido foi

denominado de projeto de rotina. Se a maturidade da solução for baixa, mas a maturidade do domínio de aplicação for alta, então existe uma oportunidade de investigação e de contribuição para o conhecimento. Neste caso foram desenvolvidas novas soluções para problemas já existentes e conhecidos e o artefacto assume um papel de melhoramento, a nível do conhecimento. Se a maturidade da solução for alta e a maturidade do domínio da aplicação for baixa, então existiu a oportunidade de investigação e contribuição para o conhecimento. Neste caso as soluções existentes noutros campos foram aplicadas a novos problemas. Estes artefactos assumem o papel de exaptação que se entende pela utilização de algo numa função diferente daquela para que foi criado. Caso ambas as maturidades sejam baixas, então o nível de oportunidade de investigação é elevado, bem como o nível de contribuição para o conhecimento. Ou seja, a necessidade de criar novas soluções para novos problemas foi elevada. Assim sendo, os artefactos que se insiram neste momento são designados de invenções. Por maturidade da solução entende-se o nível de desenvolvimento e exploração que determinada resposta, a um problema, possui. Por maturidade do domínio da aplicação entende-se o nível de exploração do meio em que determinada solução se insere.

Nesta dissertação, com base no estudo efetuado, o artefacto a efetuado insere-se no momento de Exaptação, onde o objetivo foi aplicar modelos de DM a dados do tipo imagem. No capítulo 3 Materiais, Métodos e Ferramentas foram exploradas metodologias não científicas, como o CRISP-DM, que possibilitou um seguimento de lógica no desenvolvimento do projeto.

1.4. Organização do Documento

Ao longo desta dissertação serão abordados diversos conceitos que se relacionam com o tema. O presente documento encontra-se dividido em sete capítulos, para proporcionar a estruturação coerente e organizada do estudo efetuado:

- Capítulo 1 – Introdução: capítulo onde o enquadramento e a motivação, os objetivos e resultados, bem como algumas metodologias, são abordados. O objetivo deste capítulo é proporcionar a compreensão do tema do presente projeto de dissertação.
- Capítulo 2 – Estado de Arte: este capítulo apresenta o estudo efetuado sobre os diversos conceitos essenciais à compreensão do tema.
- Capítulo 3 – Materiais, Métodos e Ferramentas: capítulo onde é apresentada a metodologia a aplicar no decorrer do projeto, bem como uma apresentação da ferramenta PDME e um resumo da ferramenta R.

- Capítulo 4 – Trabalho Realizado: capítulo onde é apresentada a parte prática com base na metodologia do CRISP-DM.
- Capítulo 5 – Análise e Discussão de Resultados: capítulo onde são analisados os resultados obtidos e o porquê.
- Capítulo 6: capítulo onde é efetuada uma síntese do projeto, uma análise dos riscos verificados, as limitações e o trabalho futuro.
- Referências Bibliográficas: apresentação das referências utilizadas para execução do documento.

2. Estado de Arte

A enorme quantidade de dados, obtida em diversas áreas, tem sido uma constante preocupação ao longo dos anos. Isto porque, com o aumento exponencial da mesma, os processos de recolha, tratamento e armazenamento dos dados evoluíram com o passar do tempo. A automatização destes processos veio simplificar muito trabalho envolvido na análise dos dados.

Nesta dissertação, o tipo de dados a analisar são as imagens médicas. De acordo com Jiawei Han et al. (2012), este tipo de dados insere-se nos dados multimédia, ou dados de imagens de triagem. Mas, segundo Barbora Zahradnikova et al. (2015), existe outro conceito – *Image Mining* – que, ao contrário de outras técnicas de processamento de imagem, não pretende detetar um padrão específico nas imagens. O *Image Mining* permite identificar, e encontrar, padrões na imagem e obter conhecimento das imagens através de um *dataset* baseado na informação de baixo nível de uma imagem – informação ao nível das características dos píxeis: cor, textura, entre outras. Este conceito será abordado, de forma mais aprofundada, no presente capítulo. São várias as áreas de atuação relativamente à análise de imagem, como, por exemplo, a criminologia forense, onde se podem identificar impressões digitais relativas a algum crime, ou o reconhecimento das feições de um criminoso; a automação industrial e robótica, proporcionando a visão robótica de uma realidade virtual; a meteorologia e geografia, relativamente a análises de imagens via satélite; a educação, na visualização de informação com o suporte de tecnologia; ou na medicina, com a análise e deteção de anomalias via interpretação de raio X (Zahradnikova et al., 2015).

2.1. Estratégia de Pesquisa Bibliográfica

De forma a que o estado de arte seja efetuado corretamente, um método de pesquisa da documentação necessária foi idealizado. Assim sendo, no início da pesquisa, alguns conceitos foram considerados fulcrais para a realização desta dissertação. A recolha de documentação foi baseada nas palavras-chave e conceitos relacionados. Do tipo de documentação explorada fizeram parte livros das áreas em foco, artigos de conferência e revistas científicas. Alguns documentos foram fornecidos pelo orientador que foi considerada literatura muito importante. Os motores de pesquisa para procura de leitura foram o *Science Direct*, *Library Genesis*, *Google Schoolar*, IEEE *Xplore Digital Library*, RepositóriUM, entre outros. Para a seleção de documentação foram estabelecidos alguns critérios de forma a sistematizar a leitura por ordem de significância. Os três

níveis de significância da leitura podem ser compreendidos na seguinte ordenação: muito significante, que diz respeito à leitura indispensável e à análise cuidada do documento; significante, que corresponde à leitura superficial para captação de ideias relativas aos conceitos e ao tema da dissertação; e pouco significante, que coincide com leitura dispensável do documento, relativamente à escrita da dissertação, e por isso pode ser ignorado.

Na Figura 5 é possível visualizar o esquema da escolha relativa à documentação muito significante, significante e pouco significante. Cada documento é analisado e segue o fluxo do diagrama conforme o nível de significância.

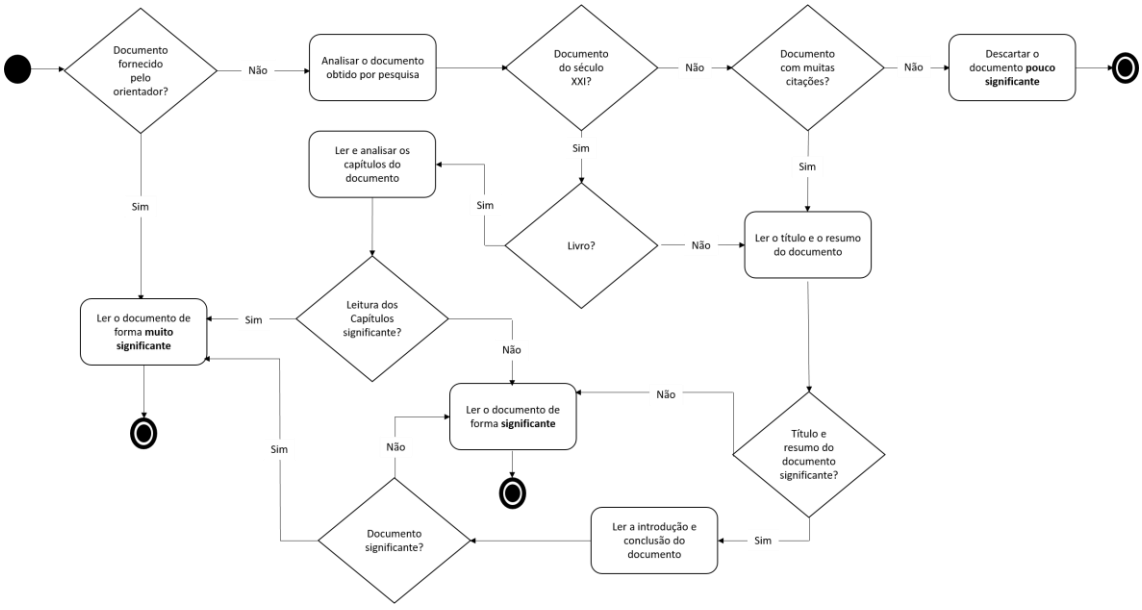


Figura 5 - Diagrama para o processo de seleção de documentos para leitura

2.2. Enquadramento Conceptual

Com a evolução das bases de dados e o crescimento da quantidade de dados, a necessidade de os compreender e de os utilizar de forma proveitosa veio suscitar novos desafios. Assim sendo, no final do século XX, surgiram as áreas de análise avançada de dados, que têm vindo a evoluir até ao momento, consoante as necessidades. O aparecimento do conceito de *data warehousing* veio providenciar uma arquitetura de armazenamento de dados, tal como o nome indica. Associada à tecnologia OLAP, a tecnologia *data warehousing* permite um repositório de fontes heterogéneas de dados, organizadas segundo um esquema local único, que facilita a tomada de decisão (Han et al., 2012). Mas esta tecnologia, por si só, não permite reconhecer padrões nos dados, nem perceber, de forma automática, qual a informação útil a retirar de todos os dados existentes. Como tal, ferramentas de *Data Mining*, bem como o conceito de descoberta de conhecimento, surgiram no final dos anos 80 como pode ser visualizado na Figura 6. Estas

ferramentas permitiram a classificação dos dados, a análise de *outliers* e deteção de anomalias nos dados, o agrupamento de dados com características semelhantes – *clustering* -, entre outras.

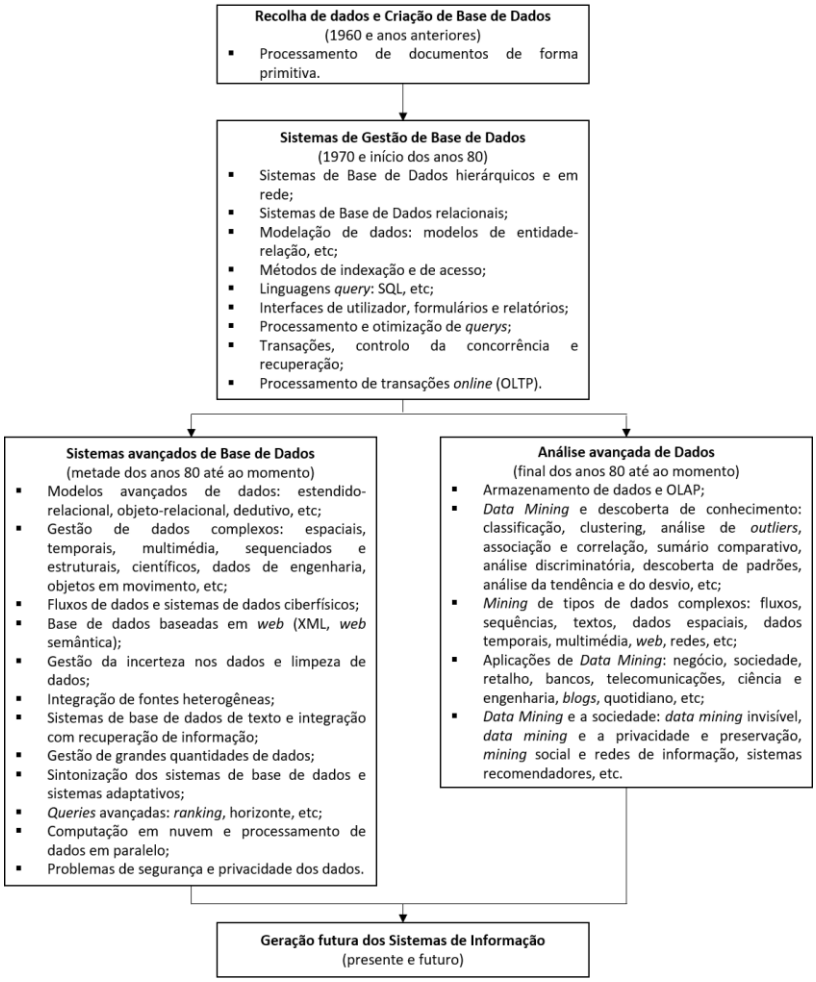


Figura 6 - A evolução da tecnologia do sistema de Base de Dados (adaptado de Han, Kamber, & Pei, 2012*)*

A capacidade de análise humana dos dados diminui, tendo em conta o aumento da quantidade dos mesmos. Se a capacidade de análise humana for considerada uma variável e a quantidade de dados outra, estas possuem uma correlação negativa. Isto é, uma variável tem tendência para diminuir, quando outra variável aumenta. Como a primeira variável – capacidade de análise humana – diminui, a capacidade de tomada de decisão também. Assim sendo, a informação útil precisa de ser filtrada, para que se obtenha o conhecimento fundamental.

A variedade de dados existentes permite múltiplas formas de análise. De acordo com Jiawei Han et al. (2012) existem os dados de bases de dados, os dados de *data warehouses*, os dados transacionais e outros tipos de dados. Os outros tipos de dados são particularmente diferentes às três categorias anteriores na sua forma, semântica e estrutura. Desta forma, estes tipos de dados podem ser reconhecidos em múltiplas aplicações: dados históricos, dados de fluxo,

dados espaciais, dados de multimédia e hipertexto, dados em rede, dados *web* ou dados de projeto de engenharia.

Segundo Ian Witten et al. (2011) os sistemas de aprendizagem, que permitem obter conhecimento através dos dados, podem ser aplicados em diversas áreas: *Web Mining*, decisões que envolvem o julgamento, imagens de triagem, previsão de carga, diagnóstico, *marketing* e vendas e outras aplicações.

A capacidade de analisar os dados de forma a obter informação que proporcione o conhecimento relativo a uma determinada área, tem sido um processo que tem despertado curiosidade ao longo dos anos. Isto porque, as organizações, por exemplo, conseguem otimizar os seus negócios com as oportunidades que advêm da exploração e avaliação dos dados existentes – oportunidades *data-driven*.

Segundo a definição sugerida por Enda Ridge (2015), a análise de dados é toda a atividade que envolva aplicar um processo analítico aos dados, para que se obtenha a introspeção dos mesmos. De acordo com a definição existente no livro de Tom Davenport "Competindo em Análises", a análise de dados entende-se por ser a utilização extensiva dos dados; a análise quantitativa e estatística; os modelos preditivos e exploratórios; e a gestão baseada em factos, que proporciona a tomada de decisões e ações, suportando as mesmas ou automatizando-as na sua totalidade (citado por Ridge, 2015, p.4). A análise de dados pode ser compreendida de várias formas e possuir variadas nomenclaturas, mas o importante é que tem sido a causa para a propagação de habilidades e produção de novas ferramentas e tecnologias.

2.3. Sistemas de Apoio à Decisão

Desde o século XVIII que a tecnologia, e sistemas integrantes nos procedimentos de manipulação de dados, progrediu. A documentação em papel é o processo mais rudimentar ou primitivo, comparativamente às tecnologias de armazenamento digital. O Homem chegou à conclusão que automatizar o processo permitiria aumentar a capacidade de armazenamento, bem como facilitar todo o seu progresso. Ao invés da existência de laboração meramente humana, o computador executaria o processo de forma mais ágil e com menor margem para erros. Nos finais do século XIX e início do século XX, os cartões perfurados eram a tecnologia principal da IBM para armazenamento de dados. Na atualidade, conceitos como *cloud storage* ou *Internet of Things* (IoT) são a novidade para o tema *Big Data* e armazenamento de grandes volumes de dados.

Numa análise cronológica é possível perceber a evolução do mercado de negócios. Isto é, à medida que a tecnologia vai evoluindo, as organizações tendem a evoluir os seus produtos ou serviços, de forma a compensar estes avanços e mudanças. Para tal, e segundo Efraim Turban et al. (2011), as organizações têm de ser ágeis e adotar, frequentemente, decisões estratégicas, táticas e operacionais, de forma rápida. Tais decisões têm de ser apoiadas por dados relevantes, que proporcionam a devida informação para obtenção do conhecimento necessário. Este conhecimento permite, então, a tomada de decisão acertada. O processo de aquisição do conhecimento essencial requer, muitas das vezes, suporte computacional.

Na Figura 7 é possível compreender o surgimento do modelo de suporte à decisão, como uma consequência da necessidade de resposta organizacional às alterações ambientais do negócio.

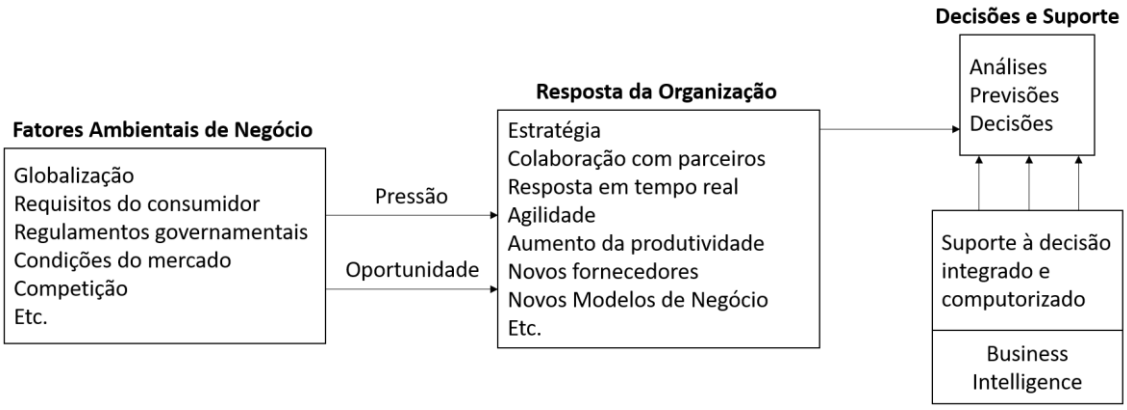


Figura 7 - As pressões dos Negócios, Respostas e Modelo de Suporte (adaptado de Turban, 2011*)*

O conceito de sistemas de apoio à decisão apareceu, primeiramente, no início dos anos 70. Uma das primeiras definições foi obtida por Scott-Morton (citado por Turban, 2011, p.16) que diz que um sistema de apoio à decisão é um sistema interativo baseado na tecnologia, que permite, a quem toma decisões, utilizar os dados e modelos para a resolução de problemas não estruturados. No final dos anos 70 o mesmo autor Scott-Morton, juntamente com Ken, promoveram a seguinte definição: os sistemas de apoio à decisão permitem a conjunção dos recursos intelectuais dos indivíduos, com as capacidades tecnológicas, para melhorar a qualidade das decisões, suportando a gestão de problemas semiestruturados (citado por Turban, 2011, p.16).

Na tomada de decisões é típico que os gestores ou responsáveis optem por seguir um processo de quatro passos (Turban, 2011):

1. Definir o problema, ou seja, uma situação de decisão que possa lidar com uma dificuldade ou uma oportunidade;
2. Construir um modelo que descreve um problema do mundo real;
3. Identificar soluções possíveis à resolução do problema identificado e avaliar as mesmas;
4. Verificar, selecionar e recomendar a solução provável para o problema.

Consoante a evolução da tecnologia, os gestores adquiriram novos tipos de capacidades. Num ambiente cada mais vez mais desenvolvido foi possível efetuar decisões de forma mais rápida e inteligente. Na metade dos anos 90 os conceitos de *Business Intelligence* e *Business Analytics* começaram a aparecer.

De acordo com Efraim Turban et al. (2011), o conceito de sistemas de *Business Intelligence* surgiu após o conceito de Sistemas de Informação Executiva (EIS – *Executive Information Systems*), anteriormente conhecidos por sistemas de apoio à decisão. Neste novo conceito, a arquitetura dos sistemas de BI é caracterizada por quatro componentes: um *data warehouse*, que armazena os dados; *business analytics*, que se caracteriza por ser um conjunto de ferramentas que possibilita a manipulação e análise dos dados do *data warehouse*; *business performance managemente* (BPM), para análise e monitorização da performance; e, por último, uma *user interface*, que permite a visualização agradável da informação proveniente dos dados - através de *dashboards*, por exemplo – (Turban, 2011). Para Diana Gonçalves, Maribel Yasmina Santos e Jorge Cruz (2011), os sistemas de BI são sistemas tecnologicamente preparados para suportar os decisores no processo de tomada de decisão. Este processo permite transformar os dados das organizações em informação útil, que proporcionará conhecimento aos *stakeholders*. Para o sucesso deste processo, estes sistemas são apoiados por tecnologia como: sistemas de *data warehousing*, que são repositórios onde ficam armazenados os dados históricos de cariz operacional e transacional extraídos dos sistemas operacionais; processos de extração, transformação e carregamento (ETL), para a seleção, transformação, limpeza e carregamento dos dados para o *data warehouse*; aplicações de análise que integram as tecnologias OLAP e DM (citado de Gonçalves et al., 2011, p.1 e p.2). Os sistemas OLAP baseiam-se nos modelos multidimensionais dos dados do *data warehouse* e possibilitam a análise da informação sob variadas perspetivas. As tecnologias de *Data Mining* possibilitam a descoberta de padrões nos dados, através da aplicação de algoritmos de análise exploratória.

Assim sendo, segundo Cody et al. em 2002 e Negash e Gray em 2003 (citado por Gonçalves et al., 2011, p.3), os sistemas BI combinam dados com ferramentas analíticas para

obtenção de informação importante ao processo de tomada de decisão, melhorando a disponibilidade e qualidade da mesma.

Em 2002, Stephen Wong et al. (2002) mencionaram os sistemas de armazenamento e comunicação de imagens – PACS – como simplificadores da gestão de imagens radiológicas e como causa da extinção de radiografias em papel. O principal objetivo destes *data warehouses* era proporcionar informação aos utilizadores, suportando as decisões ou suportando hipóteses de diagnóstico. As características que permitem distinguir os armazéns de imagens médicas de outros sistemas *data warehousing* podem ser compreendidas pela possibilidade de grandes volumes de imagens, e os seus devidos relatórios, serem adquiridos e arquivados centralmente, reduzindo a complexidade da preparação de dados e aquisição; a capacidade de processamento, registo, extração e quantificação das imagens, que possibilita a obtenção de informação, qualitativa ou quantitativa; o armazenamento de imagens que foca a aquisição e preparação dos dados com protocolos predefinidos, do que uma análise retrospetiva; e, para além disto, possibilita um suporte a uma abordagem baseada na verificação, pois providencia o acesso a ferramentas analíticas e estatísticas.

O problema deste tipo de *data warehouses* é, de acordo com Jefferson Teixeira et al. (2015), apenas permitem efetuar análises OLAP de tendências simples, como "Qual é o incidente do cancro mamário em 2011 na região sudoeste dos Estados Unidos da América?"; ou comparativas, como "Qual é o incidente do cancro mamário, nos últimos 3 anos, na região sudoeste dos Estados Unidos da América?"; e, por último, tendências múltiplas, como "Qual é o incidente do cancro mamário, nos últimos 3 anos e na região sudoeste dos Estados Unidos da América, considerando diferentes faixas etárias?" (citado de Teixeira et al., 2015, p.191). Deste modo, a necessidade de comparar informação com outras imagens das bases de dados existentes foi um fator que levou, em 2015, à proposta de uma nova arquitetura para *data warehouses* de imagens médicas por Teixeira et al. (2015). Devido a diferenças entre os dados, bem como a forma como os dados são manipulados em função da obtenção de informação de suporte à tomada de decisões, os *data warehouses* têm de ser estruturados de forma diferente.

Um conceito introduzido recentemente, acerca deste tema, são os sistemas em nuvem. Estes sistemas permitem o armazenamento de um grande volume de dados de forma remota. Na medicina, este conceito veio permitir e simplificar o acesso a relatórios médicos de pacientes, suportando o diagnóstico por parte do profissional de saúde. A aplicação de técnicas de análise

inteligente às imagens médicas permitirá, segundo Pranav Rajpurkar et al. (2017), um diagnóstico mais preciso em casos de pneumonia.

De acordo com Philippe Lambin et al. (2017), os sistemas de apoio à decisão podem auxiliar na decisão de uma terapia ou diagnostico mais apropriado para determinado paciente. A decisão não considera as escolhas do paciente e este não questiona a decisão do profissional de saúde.

Segundo Giuseppe Polese (2014), um sistema de apoio à decisão é apresentado, para que os profissionais de saúde possam analisar dados clínicos - como o histórico médico, os diagnósticos e terapias de pacientes -, de forma a detetar padrões comuns e obter conhecimento útil ao processo de diagnóstico. Técnicas de *data warehousing*, como *queries* OLAP, foram aplicadas para permitir a análise do diagnóstico por parte do profissional de saúde. De forma a facilitar a integração de dados provenientes de fontes diversas, ferramentas e abordagens de integração de dados visual foram utilizadas.

Os sistemas clínicos de apoio à tomada de decisão, segundo Gilmer Valdes et al. (2017), são ferramentas em crescimento com potencial para influenciar os cuidados de saúde. Estas suportam os profissionais de saúde na identificação eficiente, tendo em conta um histórico de planos de tratamento médico, de terapias para um novo paciente.

Nesta dissertação pretende-se, como previamente referido, desenvolver um artefacto que possibilite a classificação de imagens. Este artefacto será integrado, posteriormente, numa plataforma e as imagens a analisar serão imagens relativas ao diagnóstico e terapêutica na área da medicina. Com isto, um dos objetivos deste artefacto será o auxílio à tomada de decisões no âmbito de diagnósticos na saúde.

2.4. Data Mining

De acordo com os conteúdos já referidos anteriormente, os dados precisam ser analisados e tratados para que, com a existência de padrões nos mesmos, se possa extrair informação necessária ao conhecimento para a tomada de decisão acertada.

O *Data Mining*, de acordo com Hand et al. (2001), não é um processo que alguém efetua e termina, mas sim um processo contínuo de descoberta, interpretação e investigação de problemas ou oportunidades que surjam. A possibilidade de filtrar os dados e extrair informação de valor, aos interessados, é uma tarefa que resultou no conceito de DM. Este termo é considerado como um processo contínuo de análise de conjuntos de dados, particularmente extensos, de forma

a encontrar relações desconhecidas e a sumarizar os dados. Os resultados do processo têm de permitir o acesso a informação útil, de modo compreensível, ao utilizador. As relações desconhecidas, bem como a representação dos dados, são normalmente conhecidas como modelos ou padrões dos dados e incluem equações lineares, regras, *clusters*, gráficos, entre outros. O DM é um processo que tipicamente manipula dados previamente recolhidos. Assim sendo não assume qualquer relevância na estratégia de aquisição de dados. Isto é uma das razões pela qual o DM difere tanto da análise estatística, pois, ao contrário do DM, na análise estatística os dados são adquiridos utilizando estratégias eficientes para responder a questões específicas. Como consequência, o DM é conhecido como uma análise de dados secundária. O DM é geralmente definido num contexto mais amplo designado por descoberta de conhecimento em base de dados ou descoberta de conhecimento nos dados (KDD). O processo KDD é definido por um conjunto de passos como: a seleção dos dados *target*, o pré-processamento dos dados, a transformação dos dados se necessário, a aplicação do DM para extração de padrões ou relações nos dados e a interpretação e acesso das estruturas encontradas.

Apesar disto, os limites precisos da aplicação de DM no processo KDD não são fáceis de estabelecer. Para muitas pessoas, a transformação dos dados é uma parte intrínseca do DM. O processo de procura de relações, ou de procura de representações precisas e úteis, num conjunto de dados inclui os seguintes passos (Hand et al., 2001):

- determinação da natureza e estrutura da representação a utilizar;
- decisão de como quantificar e comparar o ajuste entre as diferentes representações e os dados;
- escolha de um processo algorítmico de otimização da função *score*;
- decisão dos princípios de gestão de dados necessários à implementação eficiente dos algoritmos.

A aplicação do processo de DM, segundo North (2012), permite localizar e interpretar padrões, que se entendem por ser indicadores de interesse, de forma a suportar a tomada de decisão. Este é um processo que, ao longo do tempo, tem sofrido várias alterações.

O termo DM, segundo Han et al. (2012), não representa todas as componentes existentes, ou seja, ao referir, por exemplo, a extração de ouro de pedras ou areia, o processo designa-se *Gold Mining* e não *Rock* ou *Sand Mining*. Quer isto dizer que o conceito DM deveria ser designado, na opinião do autor, por "*knowledge mining from data*". Como o conceito DM se foi tornando cada

vez mais popular, então os outros termos que representam o mesmo – como por exemplo, *knowledge mining from data*, *knowledge extraction*, *data/pattern analysis*, *data archaeology* e *data dredging* (citado de Han et al., 2012, p.6) – foram caindo em desuso. Algumas pessoas associaram o DM ao termo KDD, outras assumiram que este era apenas um passo no processo de descoberta de conhecimento. O processo de descoberta de conhecimento pode ser compreendido através dos seguintes passos, tal como pode ser visualizado na Figura 8:

1. Limpeza dos dados: etapa onde se removem os casos de ruído existentes nos dados, bem como os dados inconsistentes.
2. Integração dos dados: fase do processo onde várias fontes de dados se agrupam. Um dos pré-procedimentos utilizados pelas indústrias de informação é a combinação da limpeza dos dados com a integração dos dados e o resultado é armazenado num *data warehouse*.
3. Seleção dos dados: fase onde os dados relevantes à análise são obtidos da base de dados.
4. Transformação dos dados: etapa onde os dados são transformados e consolidados nas formas apropriadas para aplicação do *mining*, através da aplicação de operações de agregação. Algumas vezes a transformação dos dados e sua consolidação é efetuada antes do processo de seleção dos dados – um dos exemplos é o caso do *data warehousing*. A redução dos dados também pode ser aplicada para obtenção de representações mais pequenas, dos dados originais, sem afetar a integridade dos mesmos.
5. *Data Mining*: esta fase corresponde a um processo fundamental onde são aplicados métodos inteligentes, aos dados, para aquisição de padrões existentes nos mesmos.
6. Avaliação dos padrões: esta etapa possibilita a identificação de padrões realmente relevantes, que representem conhecimento baseado em medidas de interesse.
7. Apresentação de conhecimento: fase em que as técnicas de visualização e representação de conhecimento são utilizadas para apresentação do conhecimento aos *stakeholders*.

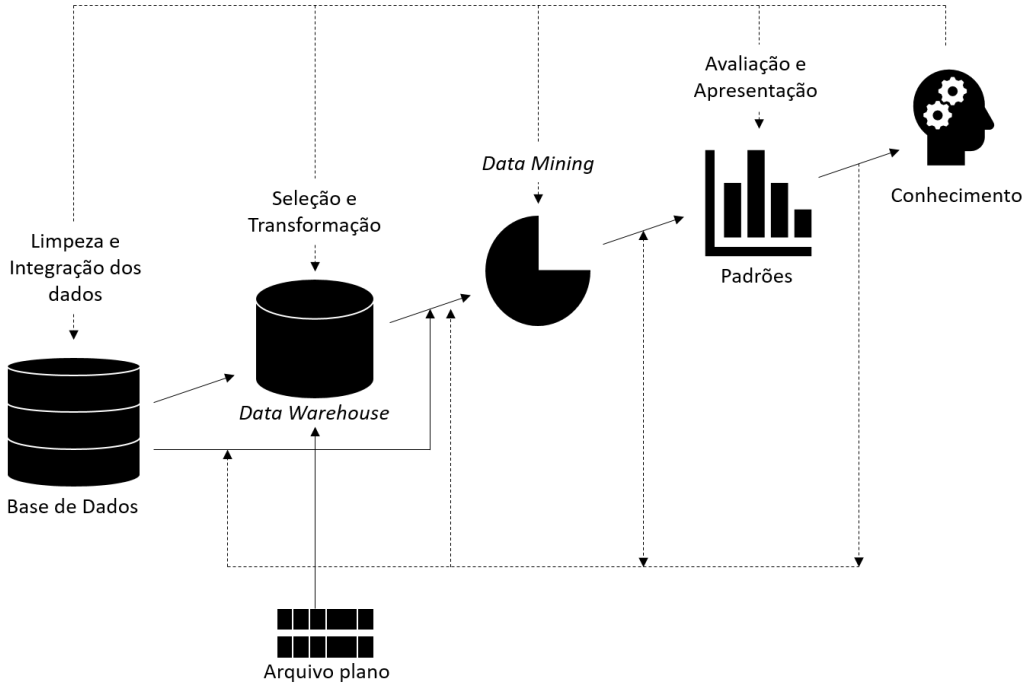


Figura 8 - Esquematização do processo, e respetivas fases, de Descoberta de Conhecimento (adaptado de Han et al., 2012*)*

Nos passos descritos acima, o DM equivale a uma fase do processo de descoberta de conhecimento. Isto porque possibilita o encontro de padrões escondidos para posterior avaliação. Contudo, na indústria da informação, nas investigações e noutras áreas, o conceito DM é utilizado para referir todo o processo de descoberta de conhecimento. Assim sendo, segundo Han et al. (2012), DM é o processo de descoberta de padrões e conhecimento interessantes de uma grande quantidade de dados. As fontes dos dados podem ser bases de dados, *data warehouses*, a *Web*, ou outros repositórios de informação, onde os dados são transferidos, de forma dinâmica, para o sistema. O DM adota técnicas de variadas áreas como: estatística, aprendizagem máquina, reconhecimento de padrões, visualização, algoritmos, computação de desempenho elevado, aplicações, recuperação de informação, *data warehouses* e sistemas de bases de dados.

O DM pode ser aplicado a variados tipos de dados, desde dados sequenciais, até dados na forma de texto, gráficos, vídeo e imagem. Com a evolução da tecnologia, também o DM evoluirá de forma a abranger novos tipos de dados que possam eventualmente surgir (Han et al., 2012).

A quantidade de dados está a aumentar de forma surpreendente. De acordo com Witten et al. (2011) estima-se que os dados armazenados, nas bases de dados mundiais, duplique a cada 20 meses. Com este aumento, a capacidade de compreensão humana, dos dados, tende a diminuir. Nestes dados pode encontrar-se informação potencialmente útil, que raramente é explícita, e da qual se podem obter variadas vantagens como a vantagem competitiva no mercado. Em DM, os dados são armazenados eletronicamente e a procura é automática ou suportada por

meios computacionais. Os dados podem ser solicitados automaticamente, identificados, validados e utilizados para a previsão por variados profissionais. O problema, como já foi referido, é o aumento dos dados que causa a necessidade de descobrir padrões, nos mesmos, que não estejam visivelmente destacados. Deste modo, a análise inteligente dos dados é algo fundamental neste processo. O DM compreende a resolução dos problemas, aquando da análise dos dados já existentes nas bases de dados. O processo de DM tem de ser automático, ou semiautomático, na descoberta de padrões. Os padrões permitem efetuar previsões, de forma adequada, nos novos dados. Existem dois tipos de padrões: os padrões de *black box*, cujas estruturas são efetivamente conhecidas, e os padrões de *transparent box*, cujas construções revelam as suas estruturas. Ambos são designados de estruturais, pois, tendo em conta a estrutura de decisão capturada, conseguem expor algo sobre os dados. Os profissionais utilizam o DM não só para a previsão, mas também para a aquisição de conhecimento.

Na perspetiva de Hand et al. (2001), as tarefas de DM podem ser separadas em análise exploratória dos dados, modelação descritiva, modelação preditiva, descoberta de padrões e regras e, por fim, recuperação por conteúdo. A finalidade da análise exploratória dos dados compreende-se pela análise dos dados, sem uma noção clara do que se procura. Algumas aplicações desta tarefa são os gráficos de setores, em que os ângulos das secções diferem. Ou seja, cada fatia do gráfico corresponde a uma variável e o ângulo da fatia é a medida de referência. A modelação descritiva permite obter caracterizações de todos os dados como, por exemplo, distribuições de probabilidades dos dados – estimação da densidade. Uma das aplicações desta tarefa é a análise de *clusters*. A análise de *clusters* proporciona a descoberta de conjuntos, onde os dados são agrupados tendo em conta características semelhantes entre si. A modelação preditiva é definida pela classificação e regressão de um conjunto de dados. Este tipo de tarefa de DM pretende obter um modelo que possibilite a previsão do valor de uma variável, com base no valor conhecido de outras variáveis. Na classificação, a variável a prever é categórica. Enquanto que na regressão, a variável a prever é quantitativa. A descoberta de padrões e regras é uma tarefa de DM que, ao contrário das tarefas anteriormente referidas, salienta a deteção de padrões e não a construção de modelos. Um dos exemplos apresentados pelo autor é o uso desta tarefa na astronomia. Com a identificação de estrelas ou galáxias invulgares é possível descobrir fenómenos que não são do conhecimento dos especialistas na área. As técnicas algorítmicas são sustentadas pelas regras associativas. A recuperação por conteúdo é a tarefa de DM que diz respeito à vontade de descobrir um determinado padrão no conjunto de dados, por parte do utilizador, que seja

semelhante ao padrão que este idealiza. Esta tarefa pode ser aplicada no caso das imagens, onde uma descrição de uma imagem pode ser utilizada para encontrar outras equivalentes, num grande conjunto destes objetos de estudo.

As tarefas de DM, segundo Han et al. (2012), podem ser classificadas em preditivas e descritivas. As tarefas preditivas efetuam indução nos dados atuais, de forma a obter previsões, que se entendem como o alcance de valores futuros de uma determinada variável. As tarefas descritivas determinam propriedades dos dados num conjunto de dados alvo. Das tarefas preditivas, consoante Gorunescu (2011), fazem parte a classificação, a regressão e a deteção de anomalias ou *outliers*. As tarefas descritivas incluem a descoberta de regras de associação, a descoberta de padrões sequenciais, o *clustering*, a sumarização e a visualização. Destas tarefas, as que serão aplicadas no presente projeto de dissertação são a classificação e o *clustering*.

Os componentes de algoritmos de DM para as tarefas anteriormente referidas são, de acordo com Hand et al. (2001), estruturas padrões ou modelos, funções *score*, métodos de otimização e pesquisa e estratégias de gestão dos dados. A primeira componente é constituída pela estrutura modelo – que apresenta um sumário global de todo o *dataset* –, ou pela estrutura padrão – que representa afirmações sobre regiões restritas do espaço abrangido pelas variáveis – , que indicam a estrutura fundamental ou as formas funcionais obtidas pelos dados. As funções *score* avaliam o ajuste de um determinado modelo ou estrutura parâmetro, relativamente a um conjunto de dados. A escolha da função *score* retrata a utilidade ou benefício espectável de um determinado modelo preditivo. Os métodos de otimização e pesquisa têm como objetivo determinar quais os valores de parâmetro e de estrutura que alcancem o mínimo ou máximo, tendo em conta o contexto, relativo ao valor da função *score*. As estratégias de gestão de dados, apesar de alguns algoritmos terem sido desenvolvidos sem focar esta componente, permitem compreender as maneiras de armazenar, indexar e aceder aos dados. Muitos dos algoritmos foram desenvolvidos na ideia de que os dados poderiam ser acedidos, de forma rápida e eficiente, na memória RAM. O problema está na imensidão de determinados conjuntos de dados e na ineficiência da aplicação desses algoritmos, nestes casos.

A aplicação do processo de DM pode criar problemas éticos, pelo que os profissionais têm de assumir responsabilidades aquando da manipulação dos dados. Um dos problemas éticos apresentados por Witten et al. (2011) é a discriminação de indivíduos. Um dos exemplos é o caso dos empréstimos bancários. Devido a questões raciais, escolhas sexuais, questões religiosas, ou outras, muitas pessoas não conseguem ter acesso a estes benefícios. Para além da falha ética

nestas situações, também é ilegal discriminar qualquer pessoa. A identificação de indivíduos com base nos dados existentes é outra preocupação ética. Muitos americanos podem ser identificados, segundo o autor, com base num código postal de cinco dígitos, numa data de nascimento e no género. O problema é que se os dados forem eliminados de determinados conjuntos, não existem bases para efetuar estudos e aplicar processos de DM. Outro problema prende-se com a questão da privacidade. Os responsáveis pela aplicação do processo de DM têm de saber quem pode aceder aos dados, qual o propósito da recolha dos mesmos e quais as conclusões que se pretendem obter. Os resultados do DM, juntamente com outro conhecimento, são apenas um meio para a tomada de decisão dos interessados.

2.4.1. Classificação

A tarefa preditiva de classificação insere-se na área de aprendizagem supervisionada, como definido por Han et al. (2012). A supervisão, em termos de aprendizagem, compreende-se pela previsão de uma variável, por exemplo, através dos exemplos identificados num conjunto de dados de treino.

O conceito de classificação surgiu com base no termo taxonomia (Gorunescu, 2011). A taxonomia apareceu como a ciência de classificar organismos vivos que, posteriormente, foi associada à classificação num contexto geral. O processo de classificação é o processo de inserir um determinado objeto ou conceito numa determinada categoria, baseado nas propriedades do objeto ou conceito. O procedimento de classificação consiste em quatro componentes essenciais: a classe, os preditores, o *dataset* de treino e o *dataset* de teste. A classe é a variável dependente, ou *target*, do modelo. Os preditores são as variáveis independentes do modelo e que suportam a obtenção da *target*. O conjunto de dados de treino contem os valores para as componentes anteriores e é utilizado na fase de treino do modelo. Ou seja, esta componente é utilizada na fase de reconhecimento da classe, com base nos preditores existentes. O conjunto de dados de teste contem os dados que serão classificados através do modelo classificador, previamente construído, e, assim, a precisão da classificação – ou desempenho do modelo – pode ser avaliada. O processo de classificação compreende a construção de um modelo classificador, a fase de treino do modelo e, por último, a avaliação do desempenho do modelo, com base na sua aplicação a conjuntos de teste.

Um exemplo da aplicação da previsão classificativa é a associação de imagens a categorias. Supondo que existem 3 categorias denominadas gato, cão e cavalo; o objetivo é direcionar as novas imagens, de acordo com as suas características, para a categoria adequada.

De forma a encontrar o modelo com o desempenho expectável é necessário comparar os comportamentos de vários modelos. Nesta análise, segundo Gorunescu (2011), é necessário ter em conta alguns conceitos relevantes, tais como a acuidade, a velocidade, a robustez, a escalabilidade, a interpretabilidade e a simplicidade. A acuidade, ou o nível de precisão de um modelo, é a capacidade de classificação acertada de um novo objeto. A velocidade refere-se à rapidez com que o modelo consegue processar os dados. A robustez demonstra a habilidade que o modelo possui para efetuar previsões precisas, mesmo na presença de ruído nos dados. A escalabilidade diz respeito à capacidade do modelo no processamento de um maior volume de dados. A interpretabilidade corresponde ao nível de facilidade na compreensão do modelo. A simplicidade corresponde à capacidade do modelo não ser complexo, apesar da sua eficiência.

Segundo Gorunescu (2011), após criação de alguns modelos de classificação, um passo importante prende-se com a avaliação dos seus desempenhos. O problema nesta avaliação é que não existem modelos que funcionem de forma perfeita. Ou seja, nem todos correspondem à expectativa total que se procura aquando da análise das medidas de avaliação. A acuidade classificativa é uma medida estatística que representa o nível de precisão com que o modelo classifica objetos. A avaliação do desempenho de um modelo de classificação assenta na contagem de objetos corretamente e incorretamente identificados. Assim sendo, com estes valores, é possível obter uma matriz de confusão. A matriz de confusão encontra-se estruturada numa tabela em que a classe prevista aparece na parte superior da mesma e, na parte lateral esquerda, encontra-se a classe observada. Desta forma, as células da tabela apresentam um valor correspondente os casos da classe observada que foram previstos pelo modelo.

Tabela 1 - Matriz de Confusão para um modelo de 2 classes (adaptado de Gorunescu, 2011*)*

Classificação	Classe Prevista		
Classe Observada		Classe = SIM	Classe = NÃO
	Classe = SIM	VP	FN
	Classe = NÃO	FP	VN

Com base na Tabela 1 é possível perceber qual é a célula que diz respeito ao:

- número de previsões corretas, para exemplos negativos (VN – Verdadeiro Negativo);

- número de previsões incorretas, para os exemplos positivos (FN – Falso Negativo);
- número de previsões incorretas, para os exemplos negativos (FP – Falso Positivo);
- número de previsões corretas para os exemplos positivos (VP – Verdadeiro Positivo).

Deste modo, a acuidade é calculada através da seguinte fórmula:

$$Acuidade = \frac{VP + VN}{VP + VN + FP + FN}.$$

A sensibilidade do modelo mede a quantidade de verdadeiros positivos que se encontram corretamente identificados, ou seja:

$$Sensibilidade = \frac{Número\ de\ VP}{Número\ de\ VP + Número\ de\ FN}.$$

A especificidade do modelo mede a quantidade de verdadeiros negativos que se encontram corretamente identificados, ou seja:

$$Especificidade = \frac{Número\ de\ VN}{Número\ de\ VN + Número\ de\ FP}.$$

O objetivo é que a classificação seja tão sensitiva quanto específica. O valor positivo previsto, ou VPP, diz respeito ao número de casos, com resultados de teste positivos, que foram corretamente identificados:

$$VPP = \frac{Número\ de\ VP}{Número\ de\ VP + Número\ de\ FP}.$$

O valor negativo previsto, ou VNP, diz respeito ao número de casos, com resultados de teste negativos, que foram corretamente identificados:

$$VNP = \frac{Número\ de\ VN}{Número\ de\ VN + Número\ de\ FN}.$$

Estas medidas estatísticas do desempenho de um modelo também podem ser compreendidas de outra forma (Gorunescu, 2011):

- A sensibilidade do modelo é também designada por taxa de verdadeiros positivos (taxa de VP) ou *recall*. Se a sensibilidade for 100% significa que o modelo classificador reconhece todos os casos positivos observados.
- Se a especificidade for 100%, isto significa que o modelo classificador reconhece todos os casos negativos observados.
- Teoricamente, um modelo pode atingir 100% na sensibilidade e na especificidade. Mas na prática isso é impossível.

A curva *Receiver Operating Characteristic* (ROC), de acordo com Gorunescu (2011), é utilizada para representar os resultados das previsões. Esta técnica, no caso da previsão classificativa, permite visualizar, organizar e selecionar os modelos classificadores, tendo em conta os seus desempenhos. Numa questão de classificação de duas classes, cada objeto é inserido ou nos positivos, ou nos negativos. Alguns modelos classificativos apenas permitem obter classificação de classes discretas, que apenas indicam a classe prevista do objeto. Outros permitem obter classificação contínua, onde diferentes limites podem ser aplicados para prever a associação de classes. A Figura 9 permite visualizar os dois tipos de curva ROC.

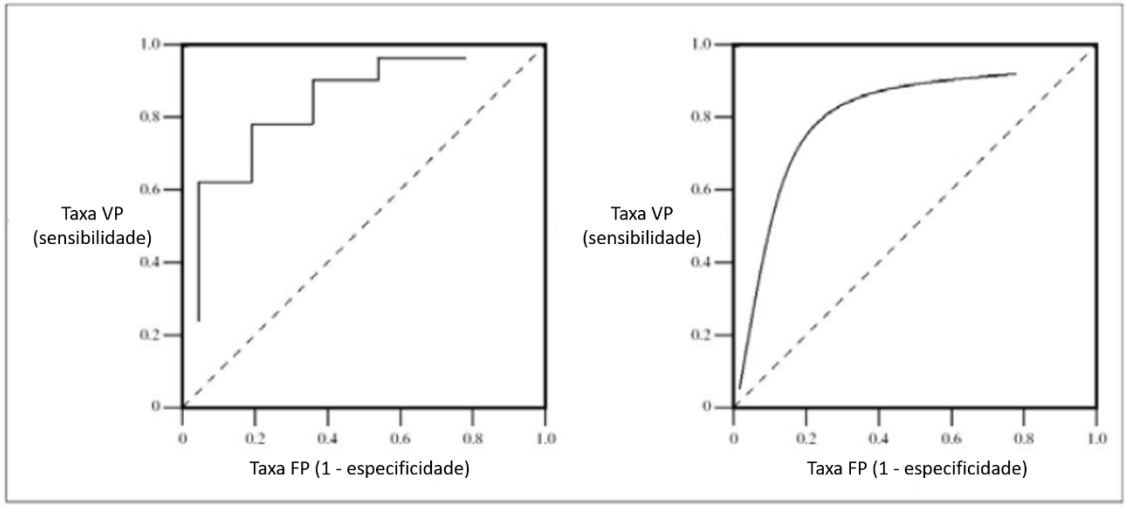


Figura 9 - Curva de ROC para a classificação discreta e contínua, respetivamente (adaptado de Gorunescu, 2011*)*

As curvas ROC são tipicamente apresentadas em gráficos de duas dimensões, em que os VP se encontram o eixo do y e os FP no eixo do x. De relevância são os seguintes pontos de um gráfico de uma curva ROC:

- O ponto (0,0) apresenta a estratégia da não emissão da classificação positiva. Isto é, não existem erros FP, mas também não existem VP.
- O ponto (1,1) apresenta a estratégia inversa do ponto anterior.
- O ponto (0,1) representa a classificação perfeita, ou seja, não existem FN nem FP.
- O ponto aleatório poderia situar-se na linha diagonal (ou linha da não discriminação) que vai desde o canto inferior esquerdo, até ao limite superior do canto direito. Esta linha divide o espaço ROC na parte superior à diagonal e na parte inferior à diagonal. A parte superior apresenta os valores de uma boa classificação. A parte inferior representa os resultados de uma má classificação.

- Um ponto no espaço ROC é melhor que outro, caso se encontre no noroeste do quadrado. Ou seja, a taxa de VP é maior e a taxa de FP é menor, ou ambos.

Para ser mais fácil a comparação de desempenhos de dois modelos classificadores, este parâmetro pode ser representado num único valor escalar. Um método é o cálculo da área abaixo da curva ROC (mais conhecido como AUC). Como este método representa uma parte da área do gráfico, o valor irá sempre variar entre 0,0 e 1,0. Quanto mais próximo o valor for de 1,0, melhor é a classificação obtida pelo modelo. A curva de ROC é mais sensitiva, quanto maior for o número de valores VP, e mais específica, quanto maior for o número de valores FP.

Alguns métodos de classificação, que também podem ser usados noutro contexto, são as árvores de decisão, os classificadores Naive Bayes, as redes neuronais, os algoritmos genéticos, os classificadores da vizinhança (*k-nearest neighbor*), entre outros (Gorunescu, 2011).

2.4.2. *Clustering*

O termo *clustering* significa agrupar objetos. Segundo Aggarwal e Reddy (2013), este termo possui diversas aplicações como a sumarização, a aprendizagem, a segmentação, entre outros. Este conceito pode ser considerado um modelo adequado, aquando da ausência de informação específica para a classificação de objetos. O problema do *clustering* passa pela divisão e associação de determinados dados a grupos, onde as características sejam o mais semelhantes possível.

O conceito de *clustering*, consoante Gorunescu (2011), pode ser considerado como um processo de classificação de objetos semelhantes em subconjuntos, dos quais fazem parte elementos com características comuns. Este processo classificativo nada tem a ver com a classificação preditiva. O problema surge na divisão de um conjunto de objetos, tendo em conta determinados atributos e com uma medida de similaridade, em grupos. Isto é, o objetivo é separar os objetos de modo a que, os que pertencem a um determinado aglomerado, sejam o mais similares possível e que, os que se encontram em diferentes *clusters*, sejam o menos similar possível a objetos que estejam noutros *clusters*. De forma simplificada, o processo de *clustering* será bem-sucedido se tanto a similaridade dentro do próprio *cluster*, como a dissimilitude entre clusters, for maximizada.

Este processo de agrupar objetos em diferentes *clusters* pode ser aplicado em diversas áreas. Algumas aplicações do *clustering* (Aggarwal & Reddy, 2013; Gorunescu, 2011):

- Segmentação de mercado: o objetivo é separar os consumidores em diferentes grupos, de acordo com as aquisições efetuadas. Aquando da criação desses *clusters*, os consumidores são considerados como alvo a atingir pelo *marketing* diverso de produtos e serviços disponíveis. Graficamente, a aplicação do *clustering* na segmentação de mercado – num exemplo de venda de automóveis, onde as características de um automóvel definem o grupo onde o objeto se insere – pode ser compreendida com a Figura 10.

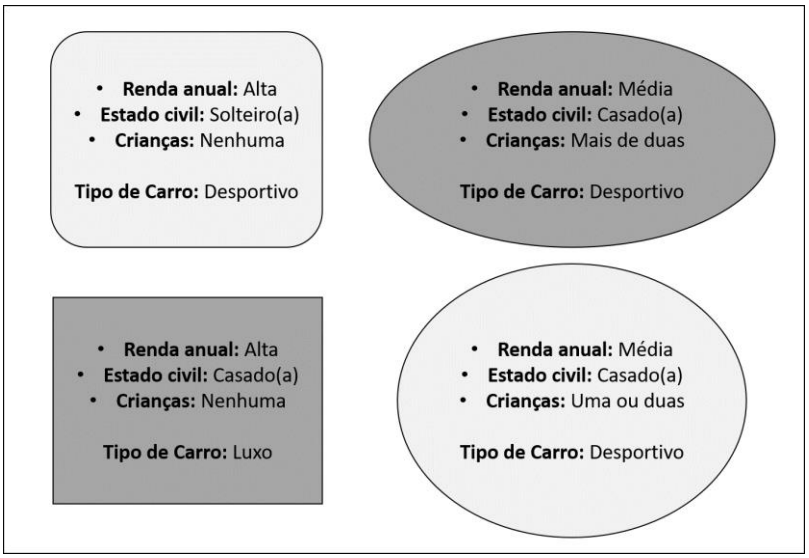


Figura 10 - Exemplo da aplicação do clustering na segmentação de mercado (adaptado de Gorunescu, 2011*)*

- Agrupamento de documentos: o propósito é encontrar grupos de documentação semelhante, de acordo com conceitos importantes, bem como o contexto onde se inserem, que estes contenham.
- Classificação de doenças: agrupação de sintomas e/ou tratamentos semelhantes.
- Deteção dinâmica de tendências: os dados são dinamicamente agrupados, numa transmissão de diversas vertentes, e podem ser identificados padrões ou alterações no meio onde se inserem. Um dos exemplos pode ser a análise de texto.
- Análise de dados multimédia: variados tipos de dados como áudio, vídeo ou imagem, são considerados dados multimédia. A associação de objetos semelhantes, como as fotografias, é um dos exemplos desta aplicação.

De acordo com Aggarwal e Reddy (2013) os desenvolvimentos, na área do *clustering* de dados, assentam em diversas categorias como as categorias centradas na técnica, as categorias centradas no tipo de dados e nas categorias de compreensão adicional das variações do *clustering*. As categorias centradas na técnica são as categorias onde existe uma preocupação na escolha de métodos, como as técnicas probabilísticas, as técnicas baseadas na distância, entre outras, para

a realização do processo de *clustering*. Cada método terá as suas vantagens e as suas desvantagens e poderá ser adequado apenas em alguns cenários ou problemas. As categorias centradas no tipo de dados são as categorias onde, de acordo com o tipo de dados de estudo (ou seja: dados probabilísticos, dados em rede, entre outros), serão escolhidas as metodologias a utilizar no processo de *clustering*. As categorias centradas na compreensão adicional das variações no processo de *clustering* são as categorias que focam a análise visual ou, por exemplo, a análise supervisionada, que permitem compreender o desempenho do processo.

As abordagens à metodologia do processo de *clustering*, , segundo Gorunescu (2011), podem ser compreendidas por abordagens hierárquicas ou abordagens não-hierárquicas. As abordagens hierárquicas são as abordagens que representam uma sucessão de *clusters*. Ou seja, utilizam características semelhantes a *clusters* previamente definidos, produzindo um diagrama em árvore. Deste tipo de abordagem fazem parte os seguintes três tipos de *clustering*:

- Aglomerante (*bottom-up*): tipo de *clustering* onde os pares de objetos são sequencialmente conectados para obter um *cluster* maior. Este método equivale a inserir o objeto no *cluster* apropriado e juntar os *clusters* mais próximos até se obter um único relativo a um problema, ou terminar o processo.
- Divisivo (*top-down*): inicialmente, neste tipo de *clustering*, todos os objetos se encontram num *cluster* único. Consoante as características dos objetos, estes são separados e inseridos em clusters mais pequenos, até que a condição de paragem se alcance e o processo termine.
- Conceptual: este tipo de *clustering* consiste na associação de objetos que possuam uma propriedade comum ou que simbolizem um determinado conceito.

As abordagens não-hierárquicas consistem numa divisão inicial, dos objetos, em *clusters* não sobrepostos, de modo a que cada objeto pertença a apenas um conjunto. O processo de *clustering* consiste nos seguintes três passos principais: a definição de uma medida de similaridade, a definição de um critério para o processo de construção dos *clusters* e, por fim, o desenvolvimento de um algoritmo, para elaboração de *clusters*, tendo em conta o critério escolhido.

Na análise de clusters é ainda importante considerar a validação da estrutura do processo. Deste modo, segundo Gorunescu (2011), existem três tipos de validação: a validação externa, que equivale à comparação do processo com outras abordagens de segmentação ou classificação; a validação interna, que permite analisar os resultados do processo, com características existentes no conjunto de dados e sem dar importância a informação externa; e, por último, a validação

relativa, que diz respeito à comparação de dois modelos de *clustering* diferentes. Assim, os passos principais de um processo de análise de *clusters* são os seguintes:

- Preparação dos dados: recolher e organizar os dados para o processo.
- Escolha da medida de similaridade: selecionar a forma de calcular a distância entre os objetos.
- Conhecimento prévio: utilizar o conhecimento existente sobre a área, de forma a suportar a preparação dos dados e a escolha da medida de similaridade.
- Eficácia da estrutura dos *clusters*: otimizar a qualidade e o tempo dispensado no processo.

Segue-se uma síntese dos principais pontos a ter em consideração no processo de *clustering* (Gorunescu, 2011):

- Formulação do problema: selecionar os objetos para o processo.
- Escolha da medida de similaridade: selecionar a distância apropriada entre os objetos para o processo, de acordo com o critério proposto.
- Seleção do modelo para o processo de *clustering*.
- Seleção do número de *clusters*, ou da condição de paragem, para o processo.
- Ilustração gráfica e interpretação dos *clusters* existentes: retirar as devidas conclusões.
- Perceção da validade e robustez do modelo utilizando vários métodos como: repetir o processo recorrendo a outras medidas de similaridade, que correspondam com o contexto; repetir o processo recorrendo a ouras técnicas de *clustering* apropriadas; e, por último, repetir o processo várias vezes, mas ignorar, em cada iteração, um ou mais objetos.

Segundo Gorunescu (2011), uma medida de similaridade possui, para além de outros exemplo, as seguintes propriedades: as propriedades de continuidade, encontradas no reconhecimento de padrões, como a robustez na perturbação ou o ruído; e as propriedades de invariância que, no caso do reconhecimento de padrões, se compreendem como as medidas relativas a transformações paralelas que não variam. Alguns exemplos destas medidas de similaridade – medidas para analisar a divergência nas semelhanças entre dois objetos – consideram a distância como um vetor e podem ser: a distância de *Minkowski*, a distância ou medida de *Cosine*, a distância de *Tanimoto*, a distância ou índice de *Jaccard*, a distância de *Pearson'r*, a distância ou medida de *Mahalanobis* e as medidas clássicas de similaridade ou extensões *Fuzzy*.

Os modelos dos processos de *clustering* hierárquico, de acordo com Gorunescu (2011), são definidos por métodos de amalgamação, ou de ligação, como: ligação única, ou vizinho mais próximo; ligação completa, ou vizinho mais distante; ligação média, ou a média de grupo de pares ponderada e não ponderada; método centroide, ou centroide do grupo de pares ponderado (mediana) e não ponderado; e, por último, o método de *Ward's*. Já no caso do processo de *clustering* não-hierárquico, o modelo utilizado é conhecido como *clustering* de *k-means*. Ou seja, este método produz *k clusters* que dividem os objetos iniciais pelos grupos mais distintos possíveis. A técnica baseada no *cross-validation*, por exemplo, permite, automaticamente, determinar o número de *clusters* no conjunto de dados. Assim sendo, apesar do modelo iniciar com *k clusters* aleatórios, o objetivo é deslocar os objetos de forma a minimizar a variedade dentro dos *clusters* e a maximizar a variedade entre os *clusters*. O algoritmo para um modelo *k-means* possui os seguintes passos:

1. Selecionar os *k* pontos aleatórios para determinação do centro dos *clusters*.
2. Atribuir instâncias aos *clusters* mais próximos, de acordo com uma função de distância de similaridade.
3. Calcular o centroide, ou média, de todas as instâncias de cada *cluster*.
4. Agrupar os dados nos *k* c*lusters* onde *k* é um valor predefinido.
5. Voltar ao passo 3, até que os *clusters* possuam os mesmos pontos nas iterações consecutivas.

Outros algoritmos para a análise de *clusters*: *Expectation Maximization*, algoritmo utilizado no *clustering* de dados para aprendizagem máquina ou visão computacional; *Quality Threshold*, algoritmo alternativo ao *clustering* não hierárquico e não necessita de um valor predefinido para a quantidade de grupos; e, por último, *Fuzzy c-means*, algoritmo em que cada ponto possui um grau de associação aos *clusters*, ao invés de pertencer exclusivamente a um grupo.

As vantagens do *clustering* hierárquico são a compreensão intuitiva/fácil e a produção de *clusters* pequenos, que facilita a descoberta de conhecimento. As desvantagens são a dificuldade de lidar com grandes conjuntos de dados, bem como a sensibilidade a ruído e *outliers*, e a dificuldade de lidar com grupos de diversos tamanhos, ou grupos com formas convexas. Já as vantagens do *clustering* não-hierárquico são a facilidade de lidar com *clusters* circulares e com formas convexas; a facilidade e rapidez de computação para um maior número de variáveis, se o valor de *k* for pequeno; e a capacidade de produzir grupos mais firmes. As desvantagens são a dificuldade no processo se os dados possuírem *outliers*; o número ótimo de clusters ser afetado,

tendo em conta o valor de *k* definido; e a dificuldade no processamento de grupos muito diferentes (Gorunescu, 2011).

2.5. *Image Mining*

A versatilidade do tipo de dados disponíveis advém das diferenças existentes nas diversas áreas de atividade. As imagens são um objeto de foco em áreas como a medicina, pois suportam domínios como o diagnóstico e a terapêutica.

Segundo Wilhelm Burger e Mark Burge (2016), a capacidade de desenvolver um sistema ou programa que permita abordar a imagem e manipular os seus elementos, ou seja pixéis, é uma ideia fascinante e que atrai cada vez mais interessados. Apesar de tudo, obter uma solução robusta que permita processar uma imagem é algo ainda difícil. Isto porque o problema está na análise da imagem. A análise da imagem é um processo que pretende extrair informação útil dos objetos em estudo. O desafio está na capacidade de aplicar ferramentas e técnicas que permitam alcançar o sucesso do processo. O reconhecimento de padrões é fundamental para a perceção de informação em dados como as imagens. Existem vários tipos de imagens digitais que podem ser abordadas: fotografias digitais, documentos digitalizados, capturas de tela, entre outras.

A aquisição das imagens pode efetuar-se de diversas maneiras. Segundo Wilhelm Burger e Mark Burge (2016), existem 6 formas de obter imagens: o modelo da câmara de pinhole, a lente fina, forma digital, o tamanho e a resolução da imagem, sistemas de coordenadas da imagem e o valor do pixel. O modelo da câmara de pinhole é um modelo que consiste na obtenção de uma imagem através de uma câmara que não possui lente. Ou seja, a luz entra por uma entrada diminuta, formando uma imagem, invertida e pequena, na parede oposta à abertura. Ao contrário do modelo da câmara de pinhole, a lente fina é um modelo que permite obter uma imagem inversa, mas com melhor qualidade. Isto porque são utilizadas lentes finas e simétricas que proporcionam a entrada de luz, originando uma clareza na captação da imagem. A forma digital é caracterizada pela aquisição de uma imagem bidimensional (2D) que será posteriormente convertida em imagem digital para um computador. Este processo é determinado por três passos: a distribuição de luz tem de ser contínua e espacialmente verificada; a função resultante tem de ser demonstrada no tempo para permitir uma imagem fixa; e, por último, os valores resultantes têm de ser quantificados num alcance finito de inteiros, que permitam a sua representação em valores digitais. O tamanho e a resolução da imagem são componentes aplicadas a imagens retangulares, por exemplo, em que o tamanho diz respeito à largura *M* – número de colunas – e à altura *N* –

número de linhas – de uma imagem, em forma de matriz; e a resolução diz respeito a dimensões espaciais, como o número de ponto de uma polegada (*dpi – dots per inch*), de uma imagem. A resolução da imagem irá proporcionar a cálculo de distâncias entre objetos da imagem. No caso das imagens médicas, a dimensão de um tumor, num raio X, é um exemplo da aplicação da resolução da imagem. Os sistemas de coordenadas, das imagens, permitem a obtenção da localização de um objeto. No diagnóstico de um osso partido, estes sistemas permitirão localizar qual a secção, no raio X, onde se visualiza a ocorrência. Por fim tem-se o valor do pixel. As características do elemento de uma imagem dependem do tipo de imagem que representam. Os pixéis são, normalmente, considerados palavras binárias de tamanho *k*. Ou seja, um pixel pode representar quaisquer valores de 2^k. O valor exato dependerá do tipo de imagem a analisar: imagens em escalas de cinza, imagens binárias, imagens coloridas (RGB) ou imagens especiais.

Ao contrário dos dados tradicionalmente manipulados – dados numéricos, nominais, entre outros – as imagens são caracterizadas por algo mais complexo. Uma imagem é representada por uma matriz bidimensional *m* x *n* dos pixéis que a constituem, em que *m* e *n* determinam as dimensões da imagem e os pixéis assumem valores inteiros de acordo com o tipo da imagem (Teixeira et al., 2015). Ou seja, os pixéis podem assumir valores entre 0 e 1, se as imagens forem imagens binárias; valores entre 0 e 255, se as imagens forem caracterizadas por diferentes escalas de cinza e representadas por 8 bits; e, por outro lado, assumir três valores entre 0 e 255 cada um respeitante às três cores primárias do modelo RGB de imagens (Teixeira et al., 2015).

Um exemplo da aplicação do processo de IM, apresentado por Quellec, Charrière, Boudi, Cochener e Lamard (2017), é a deteção de retinopatia diabética através de um conjunto de 90000 fotografias provenientes do *2015 Kaggle Diabetic Retinopathy* e de um conjunto de dados privado de 110000 fotografias. A solução proposta permitia criar *heatmaps* dos pixéis que influenciavam a previsão ao nível da imagem.

As imagens, de acordo com Pedro Domingos (2015), possuem muito ruído e variam o seu conteúdo. Isto é, as imagens são constituídas por informação relevante e informação não relevante, sendo que o ruído é a parte não relevante da imagem. Relativamente à variação do conteúdo, que uma imagem apresenta, a informação de uma imagem não é a mesma informação que outra imagem possa conter. Desta forma, e tal como o nome indica, o *Image Mining* permite analisar imagens para obtenção de informação útil ao interessado. No caso desta dissertação, o objetivo é aplicar o IM a um *dataset* de imagens proveniente do projeto DEM.

O *Image Mining*, conforme Joanna Kazmierska e Julian Malicki (2008), citados por Balu e Devi (Balu & Devi, 2012), é uma extensão do *Data Mining* no domínio das imagens. Mas, segundo Barbora Zahradnikova et al. (2015), o IM é uma área que combina conhecimento e ferramentas provenientes do DM, de base de dados, de visão computacional, do processamento de imagem, da recuperação de imagem, da estatística, do reconhecimento, do *machine learning*, da inteligência artificial, entre outros. A aplicação do *Data Mining* não é suficiente para o processamento da imagem e, como tal, ao contrário de outras técnicas de processamento de imagem, o IM possibilita a identificação e obtenção de padrões, que proporcionam a aquisição de conhecimento. Este conhecimento é a resultante da informação baseada nos pixéis da imagem.

Após aquisição das imagens, o passo seguinte é a aplicação do processo de *Image Mining*. De acordo com Barbora Zahradnikova (2015), o processo de IM tipicamente aplicado pode ser visualizado na Figura 11.

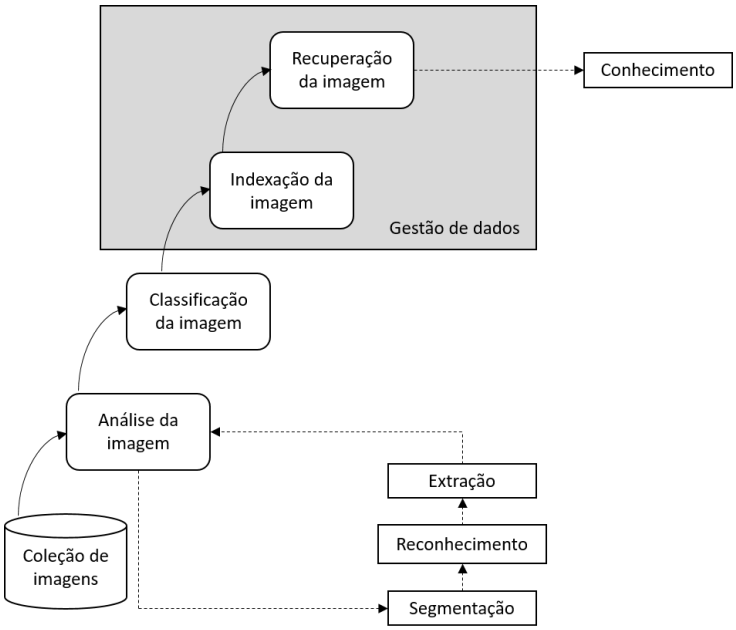


Figura 11 - Procedimento tradicional de Image Mining (adaptado de Zahradnikova et al., 2015*)*

Para obtenção de conhecimento, as imagens têm de passar primeiramente por uma análise. Nesta fase, o pixel da imagem e suas características são o foco de análise. Seguidamente, a imagem será classificada de acordo com a informação que vem da fase anterior. Por último, a imagem será indexada, e armazenada de acordo com a classificação que obteve, e essa informação poderá ser acedida para obtenção de conhecimento. As fases de indexação e recuperação de imagem fazem parte da gestão de dados que, neste caso, são imagens relativas à área de medicina.

Este subcapítulo aborda conceitos relevantes para o processo, tais como: *frameworks* de IM analisadas; abordagem à fase de análise de imagens; abordagem à fase de classificação de imagens; abordagem à gestão de imagens, que se compreende, de acordo com a Figura 11, pela indexação e recuperação de imagens; e, por fim, os problemas e limitações existentes para o *Image Mining*.

2.5.1. *Frameworks* de *Image Mining*

Uma *framework* permite perceber em que contexto será aplicado um determinado processo e como será desenvolvido. No caso do *Image Mining*, a existência de *frameworks* permite estruturar o processo ou caracterizar sistemas de IM.

Segundo Ji Zhang et al. (2001), uma base de dados de imagem que contenha dados de imagens sem tratamento, não pode ser acedida com propósitos de obter informação útil. Os dados têm de ser primeiramente processados, de forma a fornecerem informação válida, para aplicação posterior do IM. Um sistema de *Image Mining* é complicado, pois requer a aplicação de variadas técnicas desde a recuperação de imagem e os esquemas de indexação, até ao *Data Mining* e reconhecimento de padrões. Um bom sistema de IM tem de ser capaz de fornecer, ao utilizador, acesso eficaz ao repositório das imagens e ser capaz de gerar conhecimento e reconhecer padrões constituintes das imagens. Assim sendo, dois tipos de *frameworks* foram apresentados: *framework* orientada por função e *framework* orientada pela informação. A caracterização da *framework* orientada por função e da *framework* orientada pela informação, de acordo com Wynne Hsu et al. (2002), compreende-se pelo foco nas funcionalidades dos diferentes módulos componentes na organização do sistema de IM e pela estrutura hierárquica com ênfase na informação necessária aos vários níveis da hierarquia, respetivamente. A *framework* orientada por função tem o propósito de organizar e clarificar os diferentes papéis e tarefas a desempenhar no IM, mas falha na diferenciação de níveis de informação, existentes nos dados da imagem, antes do IM ser aplicado. Deste modo, Zhang et al. (2001) apresentam uma *framework* orientada pela informação que pretende salientar o papel da informação nos vários níveis. Estes níveis são distinguidos em quatro categorias:

- nível do pixel, que consiste na informação em estado natural como o pixel de uma imagem e as características primitivas como a cor, a textura e a forma;

- nível do objeto, que consiste na informação do objeto ou região baseada nas características do nível do pixel – algoritmos de *clustering*, associados ao conhecimento do domínio, permitem segmentar as imagens em regiões ou objetos significativos;
- nível do conceito semântico, que consiste na aplicação das regiões ou objetos identificados no nível anterior, no contexto das cenas representadas – raciocínio de alto nível e técnicas de descoberta de conhecimento são utilizadas, para gerar conceitos semânticos de alto nível e descobrir padrões interessantes;
- nível do padrão e conhecimento, que consiste na integração do domínio relativo a dados alfanuméricos e das relações semânticas obtidas através dos dados da imagem – os processos de *mining* são aplicados, para descobrir correlações úteis entre os dados alfanuméricos e os padrões das imagens.

A Figura 12 apresenta, de forma esquematizada, a estrutura da *framework* orientada pela informação, com os quatro níveis referidos anteriormente e as ligações entre eles. O utilizador pode ter acesso a informação proveniente dos quatro níveis, que contribuirão para a aquisição do conhecimento.

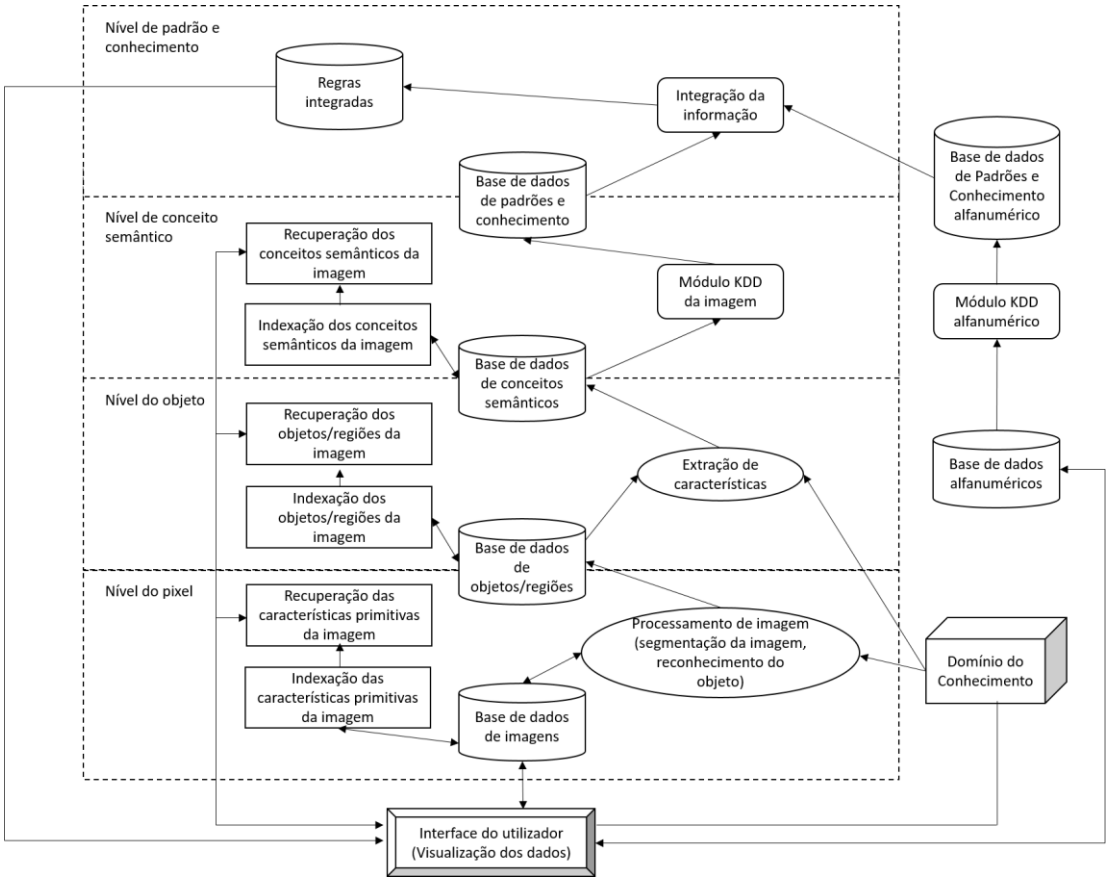


Figura 12 - Esquematização de uma Framework, de Image Mining, orientada pela informação (adaptado de Zhang et al., 2001*)*

As duas *frameworks* apresentadas são referidas por Gajjar e Chauhan (2012) e por Nilanjan Dey et al. (2015)e a *framework* que se adequa, e mais se assemelha ao procedimento tradicional referido na Figura 11, é a *framework* orientada pela informação. Esta relaciona a informação obtida pela análise do píxel da imagem, ao conhecimento que se pretende obter. Nesta dissertação, o pixel será o objeto de estudo para aquisição de conhecimento, no domínio do diagnóstico e terapêutica, na área da medicina.

2.5.2. Análise da Imagem

As imagens são caracterizadas por diversas componentes. Nas imagens em escalas de cinza, de acordo com Wilhelm Burguer e Mark Burge (2016), um único canal representa a intensidade, o brilho ou a densidade de uma imagem. Os valores são tipicamente positivos e os inteiros podem ir desde o valor 0 até ao valor $2^k - 1$. Numa imagem caracterizada tipicamente por escalas de cinza, os pixéis assumem $k = 8$ bits (1 byte) e os valores da intensidade vão de 0 a 255 – onde 0 corresponde ao brilho mínimo, ou seja, preto e 255 corresponde ao brilho máximo, ou seja, branco. Na área de medicina, 8 bits por pixel da imagem é insuficiente. Como tal, os pixéis podem assumir valores de 12, 14 ou até 16 bits neste tipo de imagens. Já nas imagens de cor (RGB), sabendo que cada uma profundidade de bit corresponde ao número de bits utilizados para representar um componente de cor, uma imagem de cor com profundidade de 8 bits equivale a um total de 24 bits por pixel. Nas imagens binárias, os pixéis apenas podem tomar dois valores – preto ou branco – utilizando um único bit por pixel (0 ou 1). Nas imagens de cor, como referido anteriormente, os pixéis assumem, tipicamente, 8 bits por componente de cor. Sabendo que estas imagens de cor são caracterizadas pelas três cores primárias – vermelho, verde e azul –, cada pixel assume $3 \times 8 = 24$ bits, com valores de intensidade de 0 a 255. As imagens de cor podem ter mais do que três componentes e, desta forma, cada pixel poderá tomar valores $4 \times 8 = 32$ bits, por exemplo. Este tipo de imagens designa-se imagens *pallete* ou imagens indexadas – imagens em que o número de componentes varia. As imagens especiais são as imagens que, tendo em conta as características do tipo de imagens anteriormente referidas, não possuem um formato definido para representar os valores da imagem. Ou seja, este tipo de imagem possui valores negativos, ou decimais, e as ferramentas de processamento de imagem podem não estar preparadas para efetuar o processo de análise. Este tipo de imagens pode ser encontrado na medicina, quando se utiliza a frequência de Nyquist para gerar imagens – obtêm-se valores negativos (Lima, 2008).

Relativamente à coloração das imagens, segundo Stan Zdonik e Jonathan Katz (2011), esta depende de quatro atributos característicos: intensidade, radiância, luminância e brilho. Tendo em conta a luz acromática, a intensidade é o único atributo relevante. Este é o cenário onde as escalas de cinza são inseridas, isto é, a intensidade varia de preto para branco, onde entre as duas cores se encontram as variações de tonalidades cinza. Por outro lado, na luz cromática, os restantes três atributos servem para medir a qualidade da fonte de luz. A radiância serve para medir a quantidade de energia emitida pela fonte de luz. A luminância serve para medir a quantidade de radiação percetível pelo observador. O brilho está associado à intensidade da luz e permite analisar a clareza do que se observa. Assim sendo, para além do modelo RGB, os autores apresentam o modelo de cores CMYK e o modelo de cores HSV. O modelo CMYK é conhecido como o modelo de quatro cores, que, ao contrário do modelo aditivo RGB, é um sistema de cores subtrativo. Ou seja, a cor ciano é a cor opositora da cor vermelha e atua como um filtro que a absorve a mesma. O modelo HSV permite a perceção humana das cores, pois permite a separação das três componentes de uma cor: tonalidade, saturação e valor.

Em termos práticos, a imagem é reconhecida como um conjunto bidimensional de valores dos pixéis constituintes que será processado por um determinado programa. Assim, de acordo com Wilhelm Burguer e Mark Burge (2016), primeiro a imagem tem de ser carregada para memória a partir de um ficheiro. Estes ficheiros podem assumir vários formatos e servem para armazenar, arquivar e partilhar dados relativos a imagens. Escolher o formato para o ficheiro é algo relevante, pois muitas plataformas apenas aceitam determinados formatos, ou a imagem pode possuir determinadas características que tenham de ser gravadas num formato específico. Os autores defendem que existem os seguintes formatos para imagens: imagem em formato raster ou vetor, formato de documento de imagem marcada (TIFF), formato de intercâmbio de gráficos (GIF), gráficos móveis em rede (PNG), formato a cargo do grupo de especialistas em fotografia conjunta (JPEG), o formato bitmap do Windows (BMP), o formato bitmap móvel (PBM), formatos de arquivo adicionais e, por fim, o formato em bits e bytes. O formato raster permite gravar as imagens que contenham valores de pixéis dispostos numa matriz regular, utilizando coordenadas discretas e, por outro lado, os gráficos vetoriais permitem representar objetos geométricos utilizando coordenadas contínuas. Caso os gráficos vetoriais tenham de ser apresentados num ecrã, as coordenadas contínuas são transformadas em coordenadas discretas, o que faz com que a imagem assuma o primeiro formato - raster. O formato TIFF suporta imagens caracterizadas por escalas de cinza, imagens indexadas ou imagens a cores e imagens especiais que contenham

valores inteiros ou decimais. Este tipo de documento permite o armazenamento de várias imagens com propriedades diferentes e, por isso, é um dos formatos de partilha mais utilizados para fotografia digital, por exemplo.

O formato GIF, segundo Wilhelm Burguer e Mark Burge (2016), é um dos formatos mais utilizados para apresentação de imagens *online*. A sua capacidade de suporte de cores indexadas com variadas profundidades de bit, de compressão Lempel-Ziv-Welch, de carregamento de imagens entrelaçadas e habilidade de codificação de animações simples – imagens armazenadas num único documento – tornam o GIF um formato popular. O formato PNG suporta três tipos de imagem diferentes: imagens coloridas, em que cada pixel pode assumir o valor de 3×16 bits; imagens em escalas de cinza, em que cada pixel pode assumir o valor de 16 bits; e as imagens de cor indexada, que podem ter até 256 cores. Este formato deveria ser o formato de escolha para representação de imagens não comprimidas, sem perda e coloridas na *Web*. O formato JPEG define um método de compressão para escalas de cinza contínuas e imagens de cor como as fotografias. Atualmente é o formato mais utilizado para ficheiros de imagens (Burger & Burge, 2016). O formato BMP é um formato menos flexível que suporta imagens em escalas de cinza, indexadas e de cor, mas não suporta, de forma eficiente, imagens binárias, visto que cada pixel é armazenado num byte inteiro. O formato PBM consiste numa série de formatos de ficheiro simples que podem, opcionalmente, ser gravados num formato de texto, e lidos por um programa, ou editados num editor de texto. Os formatos de arquivo adicionais são o formato: RGB, RAS (*Sun Raster Format*), TGA (*Truevision Targa File Format*) e XBM/XPM (*X-Windows Bitcamp/Pixmap*). Concluindo, existe ainda o formato em bits e bytes que serve para reconhecimento da imagem ao nível do byte. Isto serve para, por exemplo, ler imagens em que se desconhece o seu formato.

Para além destes formatos de imagem, Stan Zdonik e Jonathan Katz (2011) apresentam os formatos PDF, PostScript e SVG. O PDF é um formato que permite apresentar documentos independentemente da aplicação onde foram desenvolvidos, ou do dispositivo onde estão a ser visualizados e da sua resolução. Estes documentos podem ser texto, gráficos ou, como foco desta dissertação, imagens. O formato PostScript é um formato composto por um conjunto de comandos que são compreendidos por um dispositivo de *output*. Como não representa os pixéis de forma direta, não pode ser acedido por programas convencionalmente utilizados para a manipulação da imagem. Desta forma, o PostScript consegue manipular textos e gráficos com melhor qualidade do que o formato de imagem raster, mas não permite armazenar imagens. O formato SVG é o

formato comummente utilizado pelos *browsers Web* modernos e suportam a maioria dos gráficos estáticos produzidos em R.

O processo de análise da imagem é um passo que, de acordo com Barbora Zahradnikova et al. (2015), necessita de ser efetuado em *Image Mining*. O objetivo compreende-se pela procura e extração de características que representam a imagem. Numa fase de pré-processamento da imagem, vários procedimentos são aplicados para diminuir o ruído e melhorar a resolução da mesma. Na diminuição do ruído, a média, a mediana e a filtragem de wiener são algumas das técnicas aplicadas. Já na otimização da resolução, técnicas como DWT baseadas na interpolação e a fusão de imagem em multi-resolução. O procedimento DWT permite decompor o sinal do input em conjuntos de funções denominados *wavelets*. A junção dos coeficientes destes conjuntos permite obter o sinal, como uma combinação linear, proporcionando a ponderação dos elementos (Alickovic, Kevric, & Subasi, 2018). O procedimento de fusão de imagens em multi-resolução permite combinar imagens com resoluções diferentes. O alinhamento dessas imagens possibilita a otimização da resolução das mesmas.

De acordo com Barbora Zahradnikova et al. (2015), o reconhecimento de um objeto provém da segmentação de uma imagem. O propósito é identificar objetos na imagem e dividir a imagem em regiões diferentes, tendo em conta o que se identifica. Para tal, modelos de aprendizagem supervisionada, que representam determinados padrões obtidos pela aplicação de um algoritmo de treino a um conjunto de dados de treino, têm de ser aplicados. Quando se consegue identificar objetos numa imagem, esta pode ser segmentada em diferentes áreas (Berlage, 2005):

- Segmentação baseada no marcador: os objetos são representados por uma área identificada com um marcador. o objeto a detetar é conseguido com a etiquetagem do espaço na imagem.
- Segmentação baseada no objeto: os objetos são segmentados sem determinação exata dos seus limites.
- Segmentação baseada no contorno: os contornos têm de ser baseados na compatibilidade precisa dos pixéis.

Os histogramas podem ser utilizados para facilitar, de forma visual, as estatísticas relativas a imagens. Segundo Wilhelm Burger e Mark Burge (2016), o histograma permite interpretar determinados problemas relativos à aparência, por exemplo, de uma imagem. Devido à

possibilidade de apresentarem distribuições de frequências, os histogramas permitem descrever as frequências dos valores de intensidades que ocorrem numa imagem.

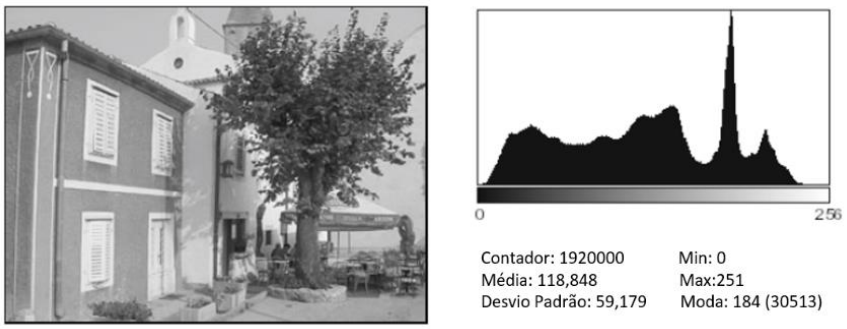


Figura 13 - Uma imagem de 8-bit, em escala de cinza, e um histograma que representa a distribuição de frequências dos 256 valores de intensidade (adaptado de Burger & Burge, 2016*)*

De acordo com o exemplo dado por Wilhelm Burger e Mark Burge (2016), e analisando a Figura 13, o histograma *h*, relativo a uma imagem *I* em escalas de cinza e com valores de intensidade num alcance de *I(u,v)*, apresenta *K* entradas onde $K = 2^8 = 256$ para uma imagem típica, em escalas de cinza, de 8-bit. Cada entrada do histograma é definida como

$$h(i) = \text{número de pixéis em } I \text{ com valor de intensidade } i,$$

para todos os $0 \leq i < K$ (citado de Burger & Burge, 2016). Deste modo, *h(0)* diz respeito aos pixéis com valor 0, *h(1)* os pixéis com valor 1 e assim por diante. Por fim, *h(255)* diz respeito ao número de pixéis de cor branca, ou seja, os pixéis com o valor máximo de intensidade $255 = K - 1$. O resultado da computação do histograma é um vetor *h*, de uma dimensão (1D), de tamanho *K*. Como o histograma não apresenta informação de onde provém cada entrada individual da imagem, este não fornece qualquer informação da organização espacial dos pixéis da mesma. O histograma permite analisar problemas provenientes no processo de aquisição de imagem, a nível de contraste ou alcance dinâmico, como a nível de resultados das fases do processamento aplicadas à imagem. O contraste e o alcance dinâmico, na fase de aquisição de imagem, dizem respeito, respetivamente, ao alcance de valores de intensidade utilizados numa imagem e ao número de valores distintos dos pixéis de uma imagem. Alguns defeitos da imagem também podem ser analisados com base nos histogramas. A saturação define-se pela receção inferior de luz muito clara ou luz muito escura, relativamente à receção de luz intermédia. Mas, caso a imagem possua luzes muito brilhantes, as partes escuras acabam por não ser bem capturadas e resultam num histograma de análise saturado numa das extremidades – neste caso, na extremidade da luz clara. Ao manipular uma imagem, defeitos como picos ou lacunas nas imagens podem ocorrer. Ao aumentar o contraste da imagem, as linhas do histograma separam-se,

causando lacunas. Ao diminuir o contraste, valores que antes eram distintos acabam por aparecer – suscitando maior número de entradas - originando picos no histograma. A compressão da imagem causa a redução do alcance dinâmico, ou seja, apenas alguns valores de intensidade podem ser visualizados. Isto proporciona a má qualidade da imagem.

Ainda sobre os histogramas, de acordo com Wilhelm Burger e Mark Burge (2016), é possível visualizar parâmetros estatísticos de uma imagem. O valor máximo e o valor mínimo da imagem *I* podem ser obtidos ao constatar o menor e maior valor do histograma, ou seja:

$$\min(I) = min\,\{i \mid h(i) > 0\},$$
$$\max(I) = \max\{i \mid h(i) > 0\}.$$

A média μ e a variância σ^2 de uma imagem I de tamanho $M \times N$, também podem ser calculadas. A média μ pode ser expressa da seguinte forma:

$$\mu = \frac{1}{MN} \times \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} I(u,v) = \frac{1}{MN} \times \sum_{i=0}^{K-1} h(i) \times i,$$

onde se obtém através dos valores dos pixéis de $I(u,v)$; ou do tamanho *K* do histograma *h*, onde $MN = \sum_i h(i)$ diz respeito ao total do número de pixéis. Com a média, também se pode obter a variância através da seguinte expressão:

$$\sigma^2 = \frac{1}{MN} \times \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} [I(u,v) - \mu]^2 = \frac{1}{MN} \times \sum_{i=0}^{K-1} (i-\mu)^2 \times h(i).$$

A mediana *m* de uma imagem é definida pelo menor valor do pixel que, por sua vez, é maior ou igual a uma metade de todos os valores dos pixéis. Ou seja, a mediana encontra-se no meio do total dos valores dos pixéis. O separador *i*, que destaca as duas metades do conjunto de valores dos pixéis, tem de ser identificado e tem de permitir a igualdade na soma dos valores de ambas as partes. Desta forma, a mediana pode ser obtida através da seguinte fórmula:

$$m = \min\,\{i \mid \sum_{j=0}^{i} h(j) \geq \frac{MN}{2}\}.$$

As alterações na intensidade, ou na cor da imagem, causam saliências na mesma como as bordas e os contornos, conforme Wilhelm Burger e Mark Burge (2016) referenciam. As bordas e os contornos são, ao olho humano, suficientes para realçar um objeto ou cenário. Desta forma, estes conceitos mostram-se relevantes para esta fase de processamento e análise da imagem. As bordas são caracterizadas por linhas chave que descrevem uma figura completa presente numa imagem.

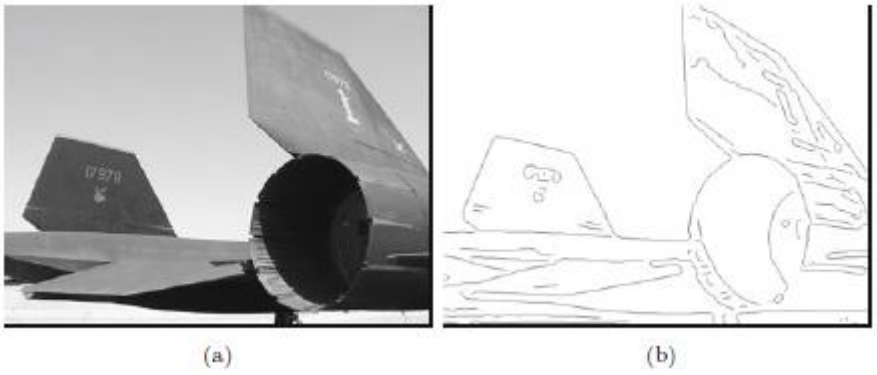


Figura 14 - Representação das bordas (b) da imagem original (a) (retirado de Burger & Burge, 2016*)*

Na Figura 14 pode ver-se que, numa imagem em escalas de cinza, é possível reconhecer a posição da imagem onde as intensidades variam distintivamente. Quanto maior for a variação da intensidade local, maior é a evidência da borda naquela posição da imagem. Os contornos compreendem-se pelas extensões opositoras, de um ponto existente numa borda, até que ambas se unam e formem um contorno fechado.

As características de uma imagem, segundo Barbora Zahradnikova et al. (2015), são normalmente representadas numericamente e providenciam uma representação matemática da mesma. A aquisição das características representa um processo de compressão da informação proveniente dos objetos distinguidos, num conjunto de atributos. A representação da imagem pode ser efetuada por descritores globais, mais fáceis de computar e que não tendem a possuir erros de segmentação; e descritores locais, que providenciam uma representação mais precisa e possibilitam a descoberta de padrões mais subtis (Martínez, Koenen, & Pereira, 2002).

Os descritores visuais, de acordo com José M. Martinez et al. (2002) e a norma ISO/IEC 15938 – mais conhecida por interface de descrição de conteúdo multimédia -, encontram-se divididos nas seguintes ferramentas de caracterização:

1. Elementos básicos: esboço em grelha, séries temporais, vista múltipla em 2D ou 3D, coordenadas espaciais 2D e interpolação temporal são exemplos de ferramentas utilizadas, posteriormente, por outros descritores.
2. Cor: o histograma de cores é a ferramenta de descrição comummente utilizada. Este providencia a fácil computação e permite a distribuição efetiva das características da cor de uma imagem. Os exemplos de descritores da cor de uma imagem podem ser compreendidos pelo espaço de cor, pela quantificação da cor, pela escalabilidade de cores, pelas cores dominantes, pela apresentação das cores, pela estrutura das cores,

pela agrupação de cor dos quadros e desenhos e, por último, pelos momentos de cor evidenciados numa imagem.

3. Textura: é uma ferramenta de descrição visual que caracteriza padrões visuais, de uma imagem, baseados na granularidade, direccionalidade e repetibilidade. Técnicas de filtragem de resolução múltipla e procedimentos DWT, bem como os conceitos apresentados na norma ISO/IEC 15938 (ex.: textura homogénea, textura não homogénea ou histograma de bordas e a procura da textura), são alguns dos descritores para esta característica.
4. Forma: descritores baseados na fronteira (ex.: formas retilíneas, aproximação poligonal, modelos de elementos finitos, entre outros), baseados na região (ex.: momentos estatísticos) e métodos de forma 3D são as técnicas mais utlizadas para a descrição desta característica.
5. Movimento (vídeos): descritores como a atividade do impulso, a captação de movimento, os parâmetros de movimento e a trajetória do movimento são os mais utilizados na caracterização deste conceito.
6. Localização: localizador de região e localizador espaço-tempo são exemplos de descritores desta característica.

Na análise da imagem será importante averiguar a existência de estruturas ou objetos numa dada localização. Como tal, Wilhelm Burger e Mark Burge (2016) referenciam os mapas de bordas binários como resultado do processo relativo à aplicação de uma operação limiar à robustez da borda. Estes mapas, considerados resultados preliminares, contêm pontos que podem não pertencer à verdadeira borda -falsos positivos – e, por outro lado, podem não conter determinados pontos, pontos que não são detetados, e por isso não constam no mapa – falsos negativos. Assim sendo, uma forma de localizar uma determinada estrutura, com ocupação significante na imagem, será selecionando um ponto aleatório da borda da mesma e ir analisando os pixéis da vizinhança. Caso estes pertençam à estrutura em foco, então os pixéis são adicionados a mapas de bordas contínuas – que se baseiam na força da borda e suas orientações -, ou a um simples mapa de bordas binário. Alguns dos problemas que se opõem a este processo de localização da estrutura são o facto das imagens possuírem ruído e a possibilidade da ocorrência de indecisões relativas à observação do contorno. Outra forma de localização de estruturas é a procura global de saliências que consistam em características de formas simples. Ou seja, apesar da imagem possuir ruído, é possível visualizar uma saliência relativa a uma circunferência, pois o nível cognitivo humano,

através do sistema visual, permite reconhecer formas triviais. Uma solução algorítmica apresentada por Wilhelm Burger e Mar Burge (2016) é a transformação de Hough.

O método de Paul Hough, segundo Illingworth et al. e Duda et al., é uma abordagem para localização de formas que podem ser definidas, parametricamente, de acordo com uma distribuição de pontos (citado por Burger & Burge, 2016, p.162). Este processo é utilizado em casos de deteção de linhas retas nos mapas de bordas. Numa imagem 2D, um segmento de linha pode ser especificado com o uso de dois parâmetros de valor real na função

$$y = k \times x + d,$$

onde k representa o declive da reta e d o ponto de interseção com o eixo vertical. O objetivo entende-se pela procura de valores k e d de modo a que exista o maior número de pontos, relativos à borda, na linha descrita por estes. Ou seja, obter linhas que possuam a maior quantidade de pontos relativos à borda de uma estrutura da imagem. O método de Hough averigua todas as possibilidades relativas aos segmentos de linha que passem por um único ponto da imagem. Toda a linha $L_j = (k_j, d_j)$ que passe por um determinado ponto $p_0 = (x_0, y_0)$ tem de satisfazer a seguinte condição:

$$L_j: y_0 = k_j \times x_0 + d_j$$

para valores apropriados k_j, d_j. Para um ponto $p_i = (x_i, y_i)$ a fórmula:

$$M_i: d = -x_i \times k + y_i$$

descreve a linha no parâmetro ou espaço de Hough. A relação entre o espaço de imagem (x,y) e o espaço de parâmetros (k,d) pode ser obtida com a observação da seguinte Tabela 2:

Tabela 2 - Relação entre o espaço de imagem (x,y) e o espaço de parâmetros (k,d) (adaptado de Burger & Burge,2016*)*

Espaço de imagem (x,y)		Espaço de parâmetros (k,d)
Ponto $p_i = (x_i, y_i)$	↔	$M_i: d = -x_i \times k + y_i$ Linha
Linha $L_j: y = k_j \times x + d_j$	↔	$q_j = (k_j, d_j)$ Ponto

Cada ponto p_i, e linha associada, corresponde exatamente uma linha M_i no espaço de parâmetros. O foco será analisar as linhas, do espaço de parâmetros, que se intersetam. Quantas mais linhas M_i se intersetarem num único ponto do espaço de parâmetros, mais pontos do espaço da imagem correspondem à borda da estrutura existente na própria imagem.

2.5.3. Classificação da Imagem

O objetivo desta fase do processo de IM é a categorização de objetos presentes numa imagem. Uns dos *inputs* recebidos, nesta etapa, equivale ao *output* criado na fase anterior relativa à análise de imagem. Assim, e de acordo com o reconhecimento de padrões existentes nos conjuntos de imagens é possível aplicar métodos de categorização de objetos, que permitirão classificar os novos objetos em categorias predefinidas, ou associar os novos objetos a grupos com características semelhantes entre si.

A. Classificação Supervisionada de Imagens

Com este tipo de tarefa preditiva de DM, previamente analisada no ponto 2.4.1 Classificação, pretende-se construir um modelo que permita classificar imagens nas categorias primeiramente definidas.

Alguns métodos referidos por Barbora Zahradnikova et al. (2015), de classificação de imagens são as árvores de decisão, a classificação baseada nas regras de associação, as máquinas de vetor de suporte, as redes neuronais, entre outros. Estes métodos são aplicados para a fase de treino do modelo. Desta forma, os classificadores obtidos permitiram, de acordo com as categorias definidas, etiquetar as novas imagens.

Um dos exemplos mais conhecidos, mencionado por Hand et al. (2001), é o sistema *Sky Image Cataloging and Analysis Tool* (SKICAT). Este sistema foi desenvolvido por Weir, Fayyad, Djorgovski e Roden (1992) e é um sistema integrado que permite classificar estrelas e galáxias de um vetor de características de 40 dimensões. Esta classificação é automática e possibilita a categorização de imagens digitais do céu. Os algoritmos de aprendizagem produzem árvores de decisão.

B. *Clustering* de Imagens

Com este tipo de tarefa descritiva de DM, previamente descrita no ponto 2.4.2 *Clustering*, é possível agrupar imagens com características semelhantes.

De acordo com Barbora Zahradnikova et al. (2015), o *clustering* representa um processo de categorização não-supervisionado das imagens. Ao contrário da classificação supervisionada, as imagens são agrupadas em clusters tendo em conta a sua similaridade e não as categorias predefinidas. O objetivo deste processo é a procura de propriedades comuns sem conhecer exatamente o tipo de dados. Com isto, os objetos serão separados em grupos, de objetos similares

entre si, que sejam diferentes dos restantes grupos existentes. A similaridade baseia-se nas características calculadas, da imagem, como a textura, a forma, entre outras.

Segundo Ankur Mahalle e Kuche (2015), este processo pode ser compreendido como o processo de organizar as imagens em grupos cujos membros sejam similares de alguma forma. Um *cluster* é uma coleção de objetos, similares entre si, e diferentes dos objetos dos restantes *clusters*. A noção de *cluster* equivale a grupos que possuem pequenas distâncias entre os seus elementos, possuem áreas densas de espaço de dados, intervalos ou distribuições estatísticas particulares. O processo de *clustering* é um processo iterativo de descoberta de conhecimento. Algumas técnicas de *clustering* de imagens podem ser os métodos baseados na limitação de histogramas, métodos baseados na deteção de bordas e métodos *Fuzzy c-means* com informação de bordas e localização.

2.5.4. Gestão das Imagens

As imagens possuem uma grande quantidade de informação. Como tal, a última fase do processo de IM tem de ser ponderada, para que seja possível a fácil procura e obtenção do devido conhecimento da base de dados da imagem.

O processo de IM, segundo Barbora Zahradnikova et al. (2015), não passa só pelas técnicas de DM aplicadas às imagens. Na gestão das imagens há que referir o armazenamento das imagens e a indexação e recuperação das mesmas. No armazenamento é preciso ter em conta a relatividade dos valores, a dependência na informação espacial e as interpretações múltiplas. A relatividade dos valores diz respeito ao contexto em que a representação numérica da imagem pode ser significante. A dependência na informação espacial diz respeito à interpretação do contexto de uma imagem, de acordo com a posição dos pixéis individuais na base de dados de imagens. As interpretações múltiplas dizem respeito às diferentes compreensões que se têm, aquando da existência de padrões das imagens. As imagens podem ser armazenadas de diversas maneiras, como referido no ponto 2.5.2 Análise da Imagem. O problema, muitas vezes, reside na dificuldade de aplicação do processo de IM nos dados. De forma a possibilitar a obtenção de imagens da base de dados, uma indexação adequada é necessária. Os métodos, baseados na similaridade, de indexação de imagens mais utilizados são a árvore *K-D-B*, a árvore *R*, a árvore *X*, entre outros.

As técnicas de obtenção das imagens de uma base de dados são as seguintes (Zahradnikova et al., 2015):

- *Query* para atributos associados: obtenção de imagens baseada nos atributos armazenados como metadados.
- *Query* para descrições: obtenção da descrição do contexto que é uma palavra chave atribuída às imagens.
- *Query* para conteúdo: organização das imagens através do conteúdo visual, por exemplo (como a textura, a cor, entre outros).

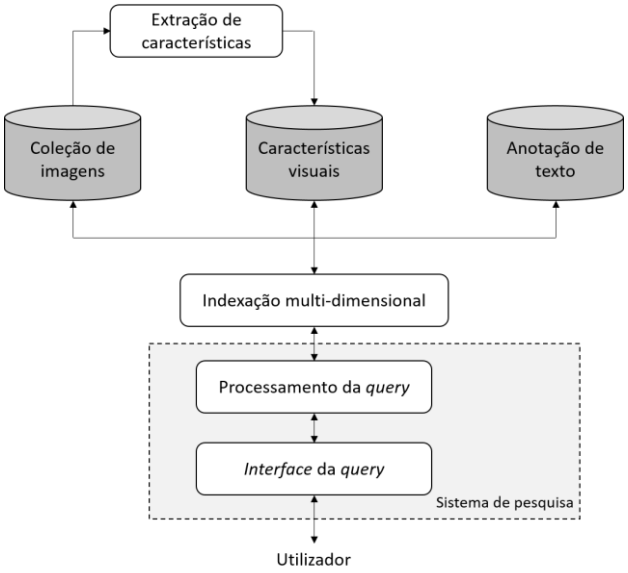


Figura 15 - Exemplo de um sistema multimodal de obtenção de imagens, no processo de IM (adaptado de Zahradnikova et al., 2015*)*

Com base na Figura 15 é possível perceber como funciona um sistema de obtenção de imagens, no processo de IM, e onde se encaixam as técnicas descritas anteriormente.

A extração da devida informação de uma grande quantidade de dados é obtida, segundo Barbora Zahradnikova et al. (2015), pela aplicação do classificador *Bayesian Naive*.

2.5.5. Problemas e Limitações

O processo de IM é um processo que tem vindo a ser explorado, em áreas como a medicina, para o suporte à tomada de decisão.

Porém, de acordo com Barbora Zahradnikova et al. (2015), ainda existem problemas a resolver, de forma a que a análise de imagens seja efetuada de forma eficiente e que seja obtido conhecimento. Alguns exemplos destes problemas são os seguintes:

- Avançar o nível do pixel de representação da imagem é um passo importante. Para um processo de IM bem-sucedido, o desenvolvimento de representações capazes de codificar a informação escondida, na imagem, é algo fundamental.

- Um passo importante no processo de IM é a classificação dos padrões obtidos. A obtenção automática de critérios de decisão adequados para o *clustering* é um obstáculo a superar.
- Um método de indexação que se ajuste aos requisitos também é um tópico relevante. A padronização de procedimentos de indexação e obtenção de conhecimento das imagens é um argumento importante no processo de IM.
- Uma linguagem de pesquisa para solicitação de padrões visuais e informações em forma de texto, que são metadados para relacionados com a imagem, necessita ser desenvolvida.
- A Web é a maior base de dados que contém grandes volumes de dados do tipo imagem. Obter conhecimento das imagens armazenadas *online* é ainda um desafio para o processo de IM.

2.6. Image Mining na Saúde

Neste ponto serão apresentados alguns casos de estudo de aplicação do processo de IM na saúde. Nalguns exemplos, o processo é designado por *Medical Image Mining*.

A dermoscopia é uma técnica importante, segundo Garcia Arroyo e Garcia-Zapirain (2017), na deteção prévia de melanomas. Assim sendo foi apresentado, neste exemplo, um método inovador de reconhecimento de padrões, para deteção de melanomas nas imagens de dermoscopia. O método consiste em duas fases. Na primeira fase, um processo de aprendizagem máquina supervisionado e após extração das características importantes como a textura e a cor das imagens, um modelo de classificação *fuzzy* de três classes ("*net*", "*hole*" e "*other*") foi aplicado. Na segunda fase, o padrão em rede de pigmentação é caracterizado, através de um processo de parametrização. A extração das diferentes características, das combinações relativas às máscaras de imagem obtidas pelas imagens de probabilidades, é efetuada, correspondendo aos cortes alfa adquiridos pelos conjuntos *fuzzy*. O método foi testado com uma base de dados de 875 imagens, que obteve, como resultados, 0.912 na métrica AUC e 88% de acuidade – com 90.71% de sensibilidade e 83.44% de especificidade. Estes resultados foram bastantes bons, o que possibilitou o desenvolvimento de um algoritmo inovador e que pode ser aplicado noutros problemas de reconhecimento de padrões.

Um exemplo já mencionado anteriormente é a aplicação de redes neuronais para a deteção pneumonia em raio X do peito. Este estudo foi apresentado por Rajpurkar et al. (2017) e consiste na aplicação de redes neuronais convencionais, de 121 camadas, a um conjunto de

dados de imagens raio X do peito, designado *ChestX-ray14*. Este conjunto de dados é de acesso público e contem mais de 100000 imagens de raio X frontais, relativas a 14 doenças. A comparação do desempenho do algoritmo foi efetuada com o desempenho de quatro radiologistas académicos, relativamente à precisão do diagnóstico. De acordo com os resultados, a métrica AUC do modelo obteve o valor 0.828 e os pontos relativos à sensibilidade e especificidade de cada radiologista, bem como a média das respostas dos profissionais, situaram-se abaixo da curva ROC. Isto significa que o algoritmo é tão, ou mais, capaz de detetar a pneumonia com precisão, comparativamente aos radiologistas.

O *Institute for Systems and Computer Engineering, Tecnology and Science* (INESC TEC), do qual faz parte a Universidade do Minho, desenvolveu um projeto denominado *Image Analysis and Machine Learning Plataform for Innovation in Diabetic Retinopathy Screening* SCREEN-DR (Oliveira, 2017). Esta plataforma permitirá avaliar as imagens recolhidas do fundo ocular dos utentes, detetando o grau de gravidade da patologia, de forma a auxiliar a tomada de decisão dos oftalmologistas.

Apesar de não ser um algoritmo específico para a classificação de imagens médicas, o DENSER, algoritmo desenvolvido por quatro investigadores da Universidade de Coimbra, permite reconhecer e classificar imagens melhor do que o Google *Brain* (Séneca, 2018). Este algoritmo utiliza redes neuronais (*Deep Evolutionary Network*) e é capaz de classificar 60 mil imagens em 100 categorias de 600 elementos mais rápido, e com menos *hardware*, do que os algoritmos que a Google possui. Mesmo que ainda não tenha sido divulgada mais informação sobre o algoritmo, os investigadores afirmam que este poderá ser aplicado em qualquer área que necessite ferramentas de reconhecimento de imagens.

Os exemplos acima descritos demonstram protótipos aplicados, ou possíveis modelos, na área de *Medical Image Mining*.

3. MATERIAIS, MÉTODOS E FERRAMENTAS

O presente capítulo da dissertação tem como objetivo apresentar, tal como o nome indica, os materiais, os métodos e as ferramentas explorados e utilizados.

Desta forma, a metodologia a aplicar no processo de *Image Mining* será o *Cross Industry Standard Process for Data Mining* (CRISP-DM). O artefacto a desenvolver será uma componente a integrar, num momento posterior, na plataforma *Pervasive Data Mining Engine* (PDME). A ferramenta R será a ferramenta de exploração dos modelos a implementar no contexto da dissertação.

3.1. Metodologia de Data Mining

Para a elaboração e desenvolvimento do processo de IM foi escolhida a metodologia CRISP-DM. Esta metodologia surgiu do trabalho de formalização e estandardização de uma abordagem, ao DM, por parte de algumas organizações como a Daimler-Benz (fabricante de automóveis), a provedora de seguros OHRA, os fabricantes de *hardware* e *software* NCR *Corp.* e o fabricante de *software* estatístico SPSS, *Inc.* (North, 2012). O CRISP-DM não foi desenvolvido para uma ferramenta específica e pode ser aplicado a qualquer tipo de dados.

O modelo de processo de DM, segundo Chapman et al. (2000), providencia uma visão geral do ciclo de vida de um projeto de DM. Este modelo contém as fases do projeto, as tarefas respetivas e as relações entre estas tarefas.

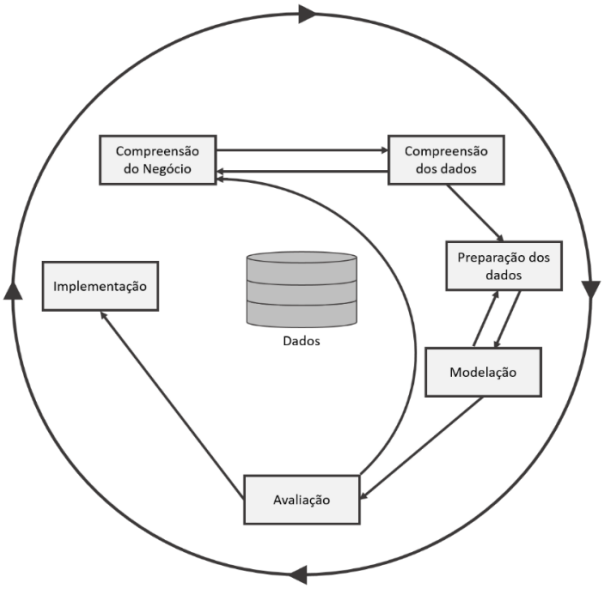


Figura 16 - Modelo referencial e fases do CRISP-DM (adaptado de Chapman et al., 2000*)*

Com base na Figura 16 é possível visualizar as seis fases que determinam o processo de DM: a compreensão do negócio, a compreensão dos dados, a preparação dos dados, a modelação, a avaliação e a implementação. Os passos podem repetir-se várias vezes e cada output de uma fase concede informação para a fase seguinte. O processo de DM não termina na implementação. Os resultados desta fase podem sugerir a criação de novos processos, de forma a melhorar a obtenção de conhecimento dos dados.

O presente documento segue as fases da metodologia e cada fase encontra-se descrita em secções distintas. Desta forma, os passos que constituem o processo de DM, segundo a metodologia de CRISP-DM, são os seguintes (Chapman et al., 2000):

- Compreensão do negócio: o objetivo desta fase inicial é compreender os requisitos através de uma perspetiva de negócio e transformar este conhecimento numa definição de um problema de DM. Desta forma é possível obter um plano inicial que procure alcançar os devidos propósitos.
- Compreensão dos dados: a aquisição dos dados é a etapa inicial deste passo. Posteriormente são efetuadas atividades que permitem a perceção dos dados. Estas atividades possibilitam: a identificação de problemas de qualidade nos dados, a descoberta de características elementares dos dados e a eliminação, caso necessário, de subconjuntos para formulação de hipóteses de informações não conhecidas.
- Preparação dos dados: nesta fase, todas as atividades essenciais à construção de um conjunto de dados final, para aplicação do modelo, serão aplicadas. As atividades são compreendidas como a seleção de tabelas, registos e atributos, ou a transformação e limpeza dos dados, através de ferramentas de modelação. Estas atividades serão aplicadas várias vezes e nunca numa ordem específica.
- Modelação: nesta fase, várias técnicas para o modelo são selecionadas e aplicadas, bem como os parâmetros são ajustados para valores ótimos. Algumas técnicas possuem requisitos específicos de acordo com o tipo dos dados. Assim sendo, voltar à fase anterior é algo, muitas vezes, necessário.
- Avaliação: nesta fase do projeto, os modelos que parecem ter qualidade elevada, numa perspetiva de análise de dados, são desenvolvidos. Mas estes têm de ser avaliados, bem como os passos efetuados durante o desenvolvimento, para que haja certezas de que o modelo satisfaz os objetivos de negócio. Um objetivo relevante é a determinação de

problemas de negócio importantes que não tenham sido considerados. No final, tem de ser efetuada uma decisão sobre o uso dos resultados de DM.

- Implementação: a criação de um modelo não é, geralmente, o fim do projeto. Mesmo que o objetivo do processo seja a aquisição de conhecimento sobre os dados, o conhecimento terá de ser organizado para que o utilizador o possa utilizar. Normalmente são utilizados modelos em tempo-real no processo de tomada de decisão das organizações. De acordo com os requisitos, a fase de implementação pode ser tão simples quanto a produção de relatórios, ou tão complexo quanto a implementação de um processo de DM repetível na organização. Em muitos dos casos é o utilizador, e não o analista de dados, que efetua os passos de implementação. Mesmo que o analista efetue todos os esforços da fase, o utilizador terá de compreender quais as ações a efetuar para que sejam realmente utilizados os modelos criados.

Cada fase do processo possui tarefas associadas. Para ser mais intuitiva a compreensão dos elementos de cada fase segue a seguinte Figura 17:

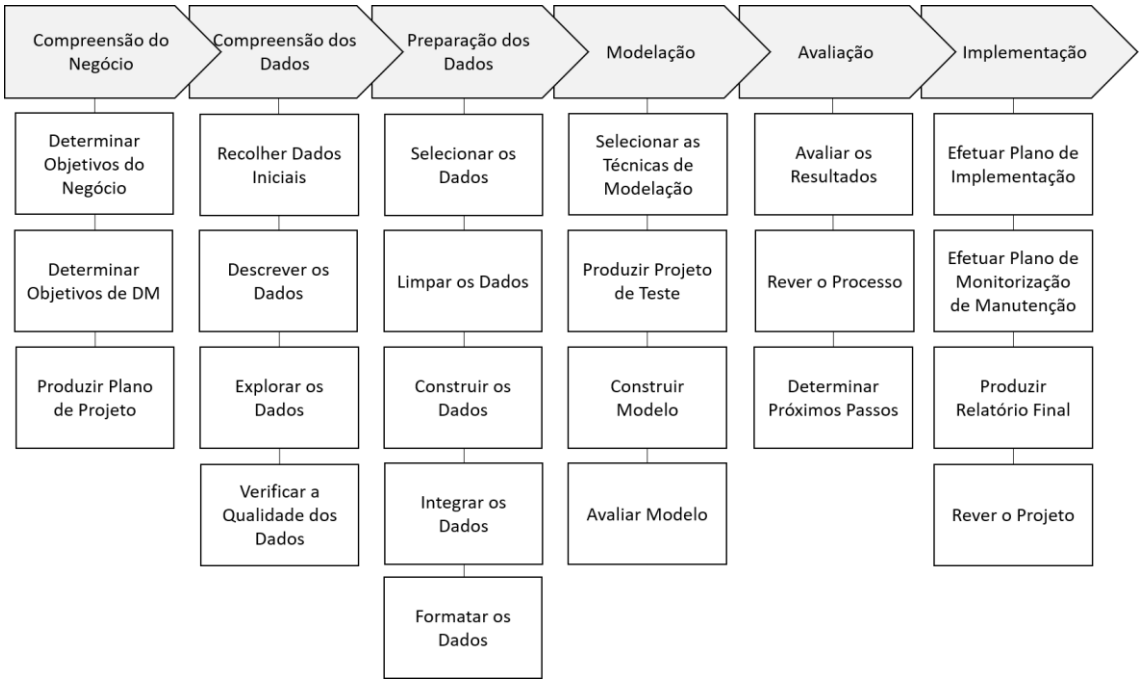


Figura 17 - Tarefas do modelo de referência CRISP-DM (adaptado de Chapman et al., 2000*)*

Na análise da Figura 17, o cabeçalho do esquema corresponde a cada fase do modelo de referência CRISP-DM. Cada fase possui as tarefas e, como exemplo, na primeira fase as tarefas são as seguintes: determinar objetivos de negócio, avaliar a situação, determinar objetivos de DM e produzir plano de projeto. O plano de projeto apresentará as tarefas de todo o projeto de dissertação, bem como os prazos associados. Tal como referido anteriormente, cada fase produz

outputs que serão necessários nas fases, onde se inserem, e nas fases posteriores. Os objetivos do negócio, por exemplo, são elaborados na fase da compreensão do negócio e são necessários para a elaboração do plano de projeto (*output* esperado na mesma fase), bem como para a fase de implementação. Isto porque, na fase de implementação, há que perceber se os requisitos foram cumpridos.

De forma a compreender a interação entre as metodologias *Cross Industry Standard Process for Data Mining* e *Design Science Research*, foi elaborada a **Erro! A origem da referência n ão foi encontrada.** de mapeamento entre as fases das duas metodologias.

A metodologia DSRM pode ser visualizada no ponto 1.3 Metodologia de Investigação e a metodologia CRISP-DM encontra-se detalhada acima. Desta forma poderão ser revistas as descrições de cada fase presente na Tabela 3.

Tabela 3 - Mapeamento entre as fases da metodologia DSRM e a metodologia CRISP-DM, para o projeto de dissertação

		Metodologia CRISP-DM					
		1. Compreensão do Negócio	2. Compreensão dos Dados	3. Preparação dos Dados	4. Modelação	5. Avaliação	6. Implementação
Mestodologia DSRM	1. Identificar o Problema e Motivação	X	X				
	2. Definir Objetivos de uma Solução	X	X				
	3. Desenho e Conceção			X	X		
	4. Demonstração					X	X
	5. Avaliação					X	X
	6. Comunicação						

Na Tabela 3, o mapeamento entre as fases da metodologia do CRISP-DM e da metodologia *Design Science Research Metodology* (DSRM) é efetuado. Corresponde à etapa de Compreensão do Negócio e à etapa de Compreensão dos Dados do CRISP-DM, as fases de Identificação do Problema e Motivação e de Definição de objetivos de uma Solução da Metodologia DSRM. Nestas fases é contextualizado o problema a solucionar que, no caso do projeto, corresponde à aplicação de modelos de DM para classificação e *clustering* de melanomas. Às fases de Preparação dos Dados e Modelação do CRISP-DM, corresponde a fase do Desenho e Conceção do DSRM. Nestas etapas, os dados são analisados e preparados de forma a aplicar os modelos de DM. A fase de Avaliação e Implementação do CRISP-DM, correspondem às fases de Demonstração e Avaliação do DSRM. Nestas fases, os principais objetivos são avaliar o desempenho dos modelos desenvolvidos.

A fase de Comunicação da DSRM corresponde à exposição dos resultados via publicação de artigos científicos, por exemplo. A fase de Implementação do CRISP-DM, no presente projeto de dissertação, coincide com as fases de Demonstração e Avaliação, de resultados, da DSRM. Assim sendo, relativamente ao mapeamento entre as duas metodologias, a fase de Comunicação da metodologia DSRM não possui ligação com nenhuma fase do CRISP-DM.

3.2. Ferramenta *Pervasive Data Mining Engine*

A PDME é uma ferramenta de DM que permite efetuar modelos, em tempo-real, de classificação e regressão de DM. Segundo Peixoto, Portela e Santos (2016), este protótipo foi desenvolvido para possibilitar a utilização e aplicação de modelos de DM. Além disso, o PDME possui características *pervasive*, como a invisibilidade e a ubiquidade, que melhoram a experiência do utilizador e proporcionam processos de DM inteligente e autónomos.

De acordo com Brian Carneiro (2017), a ferramenta PDME ajusta-se às necessidades e conhecimentos dos utilizadores. Ou seja, esta permite diferentes níveis de otimização como a configuração da execução do processo de DM em modo automático, manual ou a combinação de ambos. A simplificação do processo de DM ocorre na automatização do processo de carregamento, transformação, modelação, validação e apresentação dos resultados. A PDME é constituída por quatro componentes principais:

- Base de dados: local de armazenamento de todo o processo de base de dados, incluindo atividades intermédias e decisões efetuadas durante o processo. A base de dados serve de suporte às restantes componentes.

- Processamento: esta componente é responsável pela execução das tarefas de DM. Esta componente adota uma arquitetura multi-servidor, ou seja, cada servidor executa apenas uma tarefa. Desta forma a tarefa é executada a 100%, visto que é a única a ser processada num determinado momento.
- Controlo: componente responsável pelo processo de DM e possui a sua própria estratégia de escalabilidade.
- Interface: componente que fornece os serviços de *Data Mining Engine* ao utilizador.

De forma a ser compreendida a interação entre as diferentes componentes segue a seguinte Figura 18 com a arquitetura da ferramenta PDME:

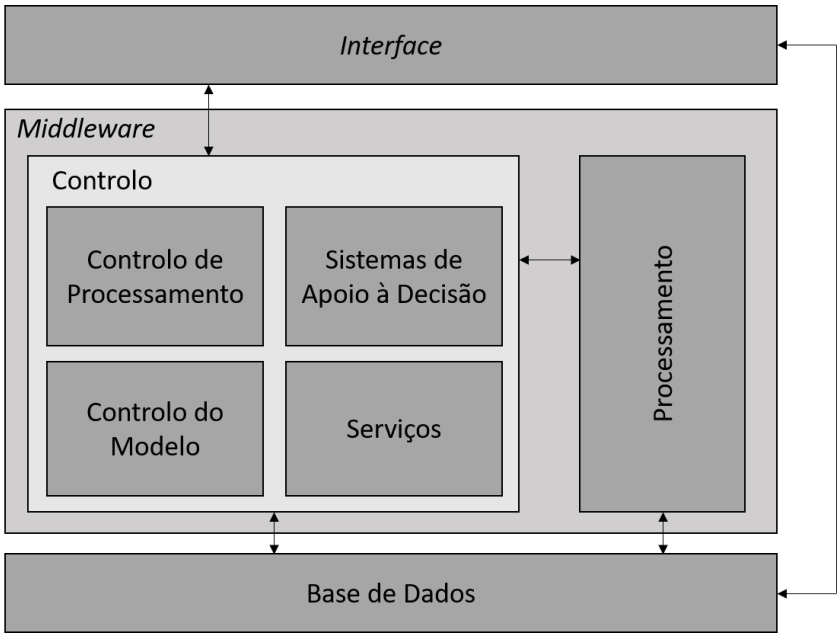


Figura 18 - Arquitetura da ferramenta PDME (adaptado de Carneiro, 2017*)*

Esta ferramenta permite a realização automática de processos de DM , construção de modelos de DM em paralelo, o registo de todas as instâncias do processo e a possibilidade de comparação entre eles (Carneiro, 2017).

Segundo Vitor Ribeiro (2017), a ferramenta PDME é capaz de efetuar processos completos de classificação de forma rápida e eficaz. Relativamente à interface, a ferramenta possui os seguintes constituintes:

- Registo ou *login*: constituinte onde o utilizador se pode registar na plataforma ou entrar na aplicação, caso já se tenha registado.
- Carregamento de dados: constituinte que permite, ao utilizador, inserir o conjunto de dados que pretende analisar, dando inicio ao processo de DM.

- Seleção do processo: constituinte que permite, ao utilizador, selecionar o processo de uma lista, para configuração.
- Configuração do processo: constituinte que permite ao utilizador determinar as tarefas do processo de DM que pretende executar de forma autónoma ou manual.
- Informação do processo: constituinte que permite a visualização do estado atual do processo, por parte do utilizador. Assim este sabe se é necessário efetuar alguma configuração.
- Seleção da *target*: constituinte que permite escolher a variável alvo, de todos os campos do conjunto de dados. Cada campo possui o nome e o tipo de dados.
- Avaliação da *target*: constituinte onde são definidos os objetivos do processo de DM, ou seja, as métricas e os valores desejados.
- Seleção do modelo: constituinte onde o utilizador escolhe quais os algoritmos a utilizar.
- Configuração do cenário: constituinte onde o utilizador escolhe quais as colunas que pretende utilizar para a aplicação do modelo. Ou seja, escolha do cenário.
- Configuração do modelo: constituinte onde o utilizador escolhe um dos cenários existentes e configura um modelo de DM.
- Avaliação: constituinte onde é visualizada a avaliação dos modelos de classificação.
- *Evaluation Regression*: constituinte onde é visualizada a avaliação dos modelos de regressão.
- *Scoring*: constituinte onde o utilizador pode aceder aos resultados.

A ferramenta PDME tem por base a linguagem R e por isso, a linguagem, bem como a ferramenta R, foi escolhida para o desenvolvimento da parte prática deste projeto.

3.3. Ferramenta R

A ferramenta R[1] será a ferramenta utilizada, neste projeto de dissertação, para a exploração de modelos de DM aplicados a dados do tipo imagem.

De acordo com Venables e Smith (2017), a ferramenta R é um *software* de pacotes integrados que facilita a manipulação, o cálculo e a apresentação gráfica de dados. As vantagens da ferramenta podem ser compreendidas pela eficácia do manuseamento e a facilidade de armazenamento de dados; pela existência de pacotes com operadores para cálculo de *arrays*, ou

[1] https://cran.r-project.org/

matrizes; pela existência de uma coleção de ferramentas para a análise de dados; pela existência de componentes, de análise de dados, gráficas de exibição no computador ou em impressão; e pela linguagem simples, efetiva e bem desenvolvida, da ferramenta. A ferramenta R é um meio para o desenvolvimento de métodos de análise interativa de dados e tem vindo a evoluir rapidamente, devido à extensão da sua coleção de *packages*. Alguns dos programas em R apenas são desenvolvidos para uma análise específica, pelo que é possível incrementar novos packages, de análise, à coleção.

4. Trabalho Realizado

Neste capítulo encontra-se apresentada, de forma estruturada e segundo a metodologia CRISP-DM, a parte prática do presente projeto de dissertação.

Numa fase inicial é efetuada uma contextualização que permite perceber o desenrolar do trabalho prático. Posteriormente, e para proporcionar um melhor detalhe do que foi desenvolvido, as fases do CRISP-DM são apresentadas.

4.1. Contextualização

Os dados fornecidos, no âmbito do projeto *Deux ex Machina* (DEM), dizem respeito a imagens de melanomas. No âmbito do projeto, o objetivo foi explorar e desenvolver componentes que permitissem a classificação e *clustering* de imagens. Estas componentes seriam integradas na plataforma *Pervasive Data Mining Engine* (PDME), numa fase posterior ao presente projeto de dissertação, de forma a proporcionarem um processo, em tempo-real, de imagens. Com base na metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), e as fases que a caracterizam, será possível a gestão do trabalho a executar ao longo do projeto.

Neste capítulo são retratadas todas as etapas da componente prática do projeto. Faz parte da componente prática, para além de outros aspetos, a análise de dados e a compreensão do objetivo a alcançar. As seis subsecções relativas ao CRISP-DM refletem isso mesmo. Na primeira subsecção da metodologia, tal como o nome indica, são compreendidos os objetivos do negócio, os objetivos do *Data Mining* (DM) e a ferramenta a utilizar no processo. Na fase de compreensão dos dados, os objetos de foco – imagens de melanomas - são analisados e é registada a forma como são manuseados. Ou seja, a origem destes é descrita e a exploração efetuada, de onde advêm perceções dos problemas e de qualidades existentes. Na terceira etapa, de preparação dos dados, é explicado o processo de seleção, limpeza e construção dos dados. Na modelação são descritas as abordagens, recorrendo a excertos de código produzido em R, efetuadas para a resolução do problema – classificação e *clustering* de imagens. A etapa de avaliação encontra-se subdividida em dois âmbitos: o âmbito de avaliação de testes e o âmbito da avaliação dos modelos. No primeiro foco, são analisadas as criações dos modelos, com base nos parâmetros a inserir. No segundo foco, é avaliado o melhor modelo - com base nos cenários aplicados aos dados de avaliação - de acordo com o desempenho da aplicação nos dados de teste. Por último, a etapa de implementação não demonstra a aplicação dos modelos ao nível de integração na ferramenta

Pervasive Data Mining Engine, visto não ser um objetivo do projeto. Nesta fase é discutida a contribuição dos modelos criados, num momento futuro.

4.2. Metodologia

O CRISP- DM, já apresentado e referido ao longo deste relatório, foi a metodologia escolhida para estruturação do processo de aplicação de modelos de DM a um conjunto de imagens.

4.2.1. Compreensão do Negócio

Tal como referido ao longo deste projeto, as imagens são definidas por um conjunto de peculiaridades como a cor, textura, entre outros. Essas peculiaridades são, muitas vezes, fatores a ter em conta aquando da análise de melanomas.

Os melanomas surgem com base num desenvolvimento anormal de células. Assim sendo, e de um ponto de vista clínico, estas lesões na pele podem ser classificadas em dois grandes grupos: melanocíticas e não-melanocíticas. Na Figura 19 podem ser visualizados alguns exemplos de lesões da pele.

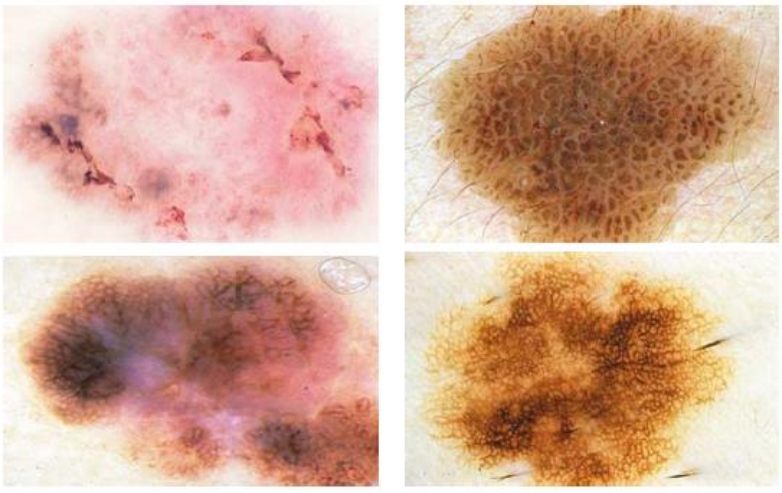


Figura 19 - Lesões melanócitas e Lesões não melanócitas (retirado de Barata & Celebi, 2017*)*

Segundo Barata e Celebi (2017), as lesões melanocíticas provêm de células melanócitas (mais conhecidas por melanócitos), que são as células responsáveis pela produção do pigmento protéico designado melanina. Por sua vez, as lesões não-melanocíticas derivam de outro tipo de células da pele, como as células basais ou as células escamosas. Esta divisão ocorre devido às diferenças acentuadas de certas características, como, por exemplo, a coloração. Cada tipo

encontra-se subdividido em tipos malignos e tipos benignos, como se pode ver no esquema da Figura 20. Um exemplo de lesões típicas são as marcas de nascimento. Ao contrário destas, as lesões irregulares apresentam, para além de dimensão maior, uma forma e borda desproporcional.

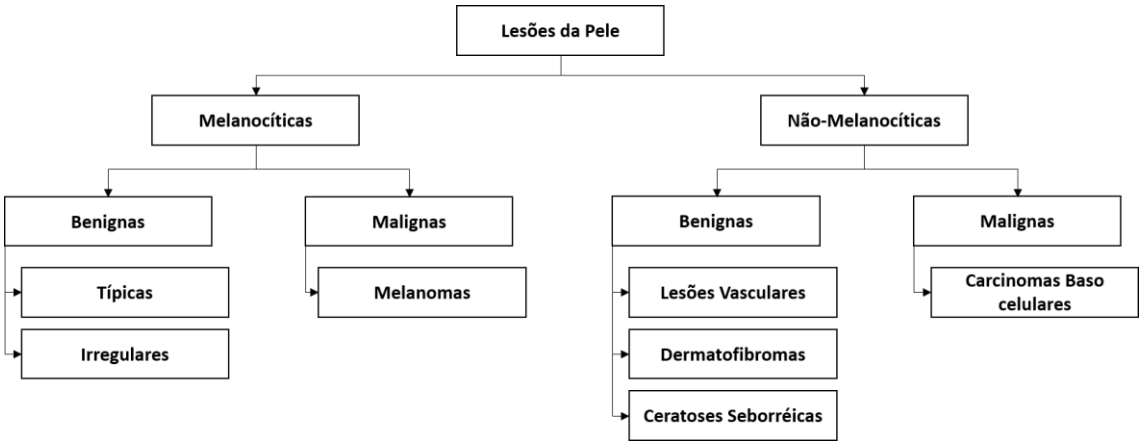


Figura 20 - Diferentes tipos de lesões da pele (adaptado de Barata & Celebi, 2017*)*

Os melanomas correspondem às lesões malignas do grupo das lesões melanocíticas. Comparativamente às lesões malignas do grupo das lesões não-melanocíticas, os melanomas apresentam um desenvolvimento mais acelerado. Ou seja, os melanomas evoluem mais rápido, em menos tempo, e possuem maior capacidade de propagação por outros tecidos – podendo originar a metástase a restantes órgãos do corpo. Uma lesão maligna pode ser facilmente confundida com uma lesão benigna. Nestes casos, os profissionais de saúde decidem efetuar um exame histológico – único exame que garante certezas na identificação de melanomas. Porém, este tipo de exames possui um custo elevado e não pode ser efetuado num paciente com uma quantidade significante de lesões. Caso seja efetuado, o paciente ficará com uma cicatriz no lugar da lesão. Assim sendo, a necessidade de deteção precoce é uma constante.

A aplicação de modelos de DM às imagens fornecidas terá como principal objetivo a classificação e *clustering* de melanomas. A métrica relativa à acuidade proporcionará a análise do desempenho de cada modelo na fase de validação e de testes. Quanto maior o valor da métrica, melhor o desempenho. Mas, devido à inexperiência na análise de objetos do tipo imagem, a probabilidade do valor não ser o expectável existe. Isto porque os resultados necessitam de uma validação por parte de um especialista que saiba perceber se a associação do resultado corresponde com o diagnóstico correto. A métrica *Loss* corresponde à probabilidade do modelo errar nas previsões. Ou seja, corresponde ao erro do modelo num caso em que prevê ser

melanoma, mas não é melanoma. O total de modelos a apresentar, dependerá de fatores como o seu desempenho e a capacidade de execução.

A ferramenta R possui variados *packages* que possibilitam a elaboração de modelos classificação e *clustering* de imagens. Porém, pode não ser suficiente. No caso da análise de certos modelos, ferramentas como o *Anaconda Navigator* poderão auxiliar funcionalidades de determinados packages como *tensorflow* ou *keras*.

Assim sendo, o objetivo principal do negócio foi a exploração e o desenvolvimento de modelos de classificação e *clustering* de melanomas. O objetivo de *Data Mining* compreende-se pela aquisição de acuidades acima de 75% e de erros menores que 35%, no desempenho dos modelos criados, no caso dos modelos preditivos. No caso dos modelos descritivos, o objetivo compreende-se pela aglomeração das imagens em *clusters* distintos.

4.2.2. Compreensão dos dados

Os dados fornecidos, no âmbito do projeto, são do tipo imagem. Correspondem a imagens de lesões da pele obtidas na Clínica da Pele, no Instituto Português de Oncologia do Porto - IPO. As imagens foram obtidas durante 6 sessões a 36 utentes – $\sim 44{,}4\%$ homens e $\sim 55{,}6\%$ mulheres. A média de idades corresponde a 46 anos, cuja idade mínima equivale a 6 anos e a idade máxima a 76 anos. Os dados foram adquiridos em sessões entre 2009 e 2013, com os utentes, por profissionais de saúde.

Estes objetos de estudo encontram-se divididos em duas pastas principais: imagens de dispositivos móveis – *smartphones* - (um HTC One e um Samsung S4, ambos *androids*); e imagens de um dermoscópio móvel (DermLite DL3), adaptável, com o auxílio do Samsung S4.

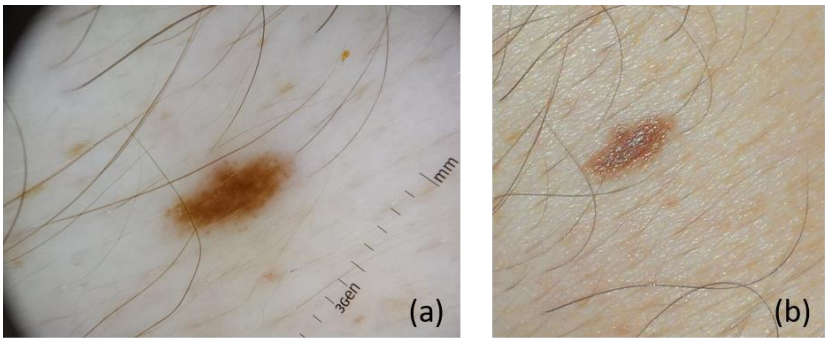


Figura 21 - Imagem captada pelo dermoscópio adaptável (a) e imagem captada por smartphone (b)

No total, e em teoria, deveriam existir 106 diferentes objetos na pasta de melanomas capturados pelo dermoscópio e 212 objetos na pasta de melanomas capturados pelos *smartphones*. Porém, na segunda pasta, apenas existem 184 objetos. Após análise, a variação ocorre pela a ausência de registos, tanto do HTC One, como do Samsung S4, de determinados melanomas.

Para além das imagens, existe também uma folha de cálculo com informação de características dos 106 melanomas. Existe uma regra, designada por regra ABCD (*Asymmetry*, *Border*, *Color* e *Diameter*), que serve de auxílio à análise e deteção de melanomas. Segundo Messadi, Cherifi e Besaid (2014), esta regra foca a análise de quatro características visuais: a assimetria, a borda, a cor e o diâmetro. Deste modo, o ficheiro encontra-se estruturado de maneira a que cada linha corresponda a uma lesão e as colunas às características anotadas. Para além das peculiaridades registadas, o risco é uma das colunas do ficheiro. Esta coluna apresenta uma escala de 1 a 3 do risco provável de melanoma. O nível 1 corresponde ao risco baixo, o nível 2 ao risco médio e o 3 ao alto risco. Para ser mais fácil a compreensão do ficheiro, segue abaixo a Tabela 4. Esta Tabela 4 é constituída pelo nome do atributo e sua descrição, bem como os valores que o atributo pode assumir no ficheiro.

Tabela 4 - Apresentação da informação relativa aos dados fornecidos

Atributo	Descrição	Valores possíveis
ID da imagem	Esta coluna corresponde à imagem da lesão capturada.	1 a 106
Smartphone HTC	Caso esteja preenchida, então a imagem foi capturada pelo *smartphone* HTC.	1
Smartphone S4	Caso esteja preenchida, então a imagem foi capturada pelo *smartphone* S4.	1
Dermoscópio + *Smartphone* S4	Caso esteja preenchida, então a imagem foi capturada pelo dermoscópio adaptado no *smartphone* S4.	1

Atributo	Descrição	Valores possíveis
Assimetria (Eixo Maior e Eixo Menor)	Devido à forma das lesões (geralmente formas eclípticas), existem dois eixos principais: eixo maior – diâmetro mais longo – e eixo menor – diâmetro mais curto.	0 – eixo simétrico; 1 – eixo assimétrico.
Borda	Corresponde ao nível de irregularidade do contorno de uma lesão.	0 a 8: 0 – nível com menor irregularidade; 8 – nível com maior irregularidade.
Cor (Branco, Vermelho, Castanho Claro, Castanho Escuro, Preto, Azul Acinzentado)	Corresponde a tonalidade que uma determinada lesão apresenta. Esta pode assumir apenas uma cor, ou possuir uma junção das cores apresentadas.	0 – ausência de cor; 1 – presença de cor.
Risco de Melanoma	Corresponde ao risco, associado a uma determinada lesão, de ser melanoma.	1 – baixo risco; 2 – médio risco; 3 – alto risco.

O identificador da imagem permite perceber qual a imagem que está a ser tratada, numa escala de 1 a 106 lesões. Caso a imagem tenha sido capturada pelo *Smartphone* HTC, a coluna é preenchida com o valor de 1, caso contrário não assume nenhum valor. O mesmo acontece se for capturada pelo *Smartphone* S4, ou pelo dermoscópio adaptável. Se a lesão for assimétrica tendo em conta o eixo maior, a coluna é preenchida com 1 (tal como acontece para casos cuja assimetria derive no eixo menor da lesão. Se a lesão não for assimétrica em nenhum dos eixos, a coluna é preenchida com o valor de 0. A coluna correspondente à borda assume um valor na escala de 0 a 8. Quanto mais alto for o valor, maior é a irregularidade da borda da lesão. Nas colunas respetivas às cores que a lesão pode assumir, o valor de 1 corresponde á presença da cor e o valor 0 à sua ausência. Por último é apresentada a coluna correspondente ao risco da lesão. A escala vai de 1 a 3, em que 1 corresponde ao nível de risco menor de ser melanoma e 3 ao nível de maior risco.

Para análise, e compreensão de um objeto de estudo, foi escolhida a imagem com identificador 1. Com recurso aos packages *EBImage* e *rmarkdown*, será apresentada, abaixo, uma breve exploração da mesma. Na Figura 22 é possível verificar a intensidade de cores da lesão, por pixéis, através do histograma apresentado em (c). Sendo uma imagem de cor, os pixéis assumem valores na escala do modelo RGB. A lesão (a) corresponde à imagem escolhida para análise. A mesma lesão é apresentada em (b), mas agora sem o ruído que a envolve. Ou seja, a imagem original foi cortada para que apresentasse o máximo de lesão, relativamente à área da imagem.

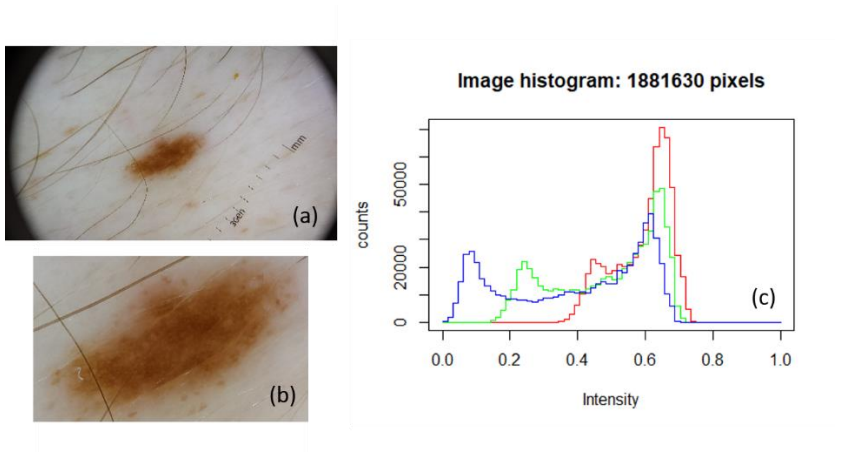


Figura 22 - Apresentação do histograma (c) relativo ao corte (b) efetuado à imagem da lesão (a)

A mesma imagem, cortada, foi convertida numa imagem em escalas de cinza. Desta forma, e através do histograma (b) na Figura 23, é possível perceber a intensidade assumida pelos pixéis da imagem. Desta forma é possível aplicar o método de segmentação de Otsu à imagem (a), apresentada na Figura 23.

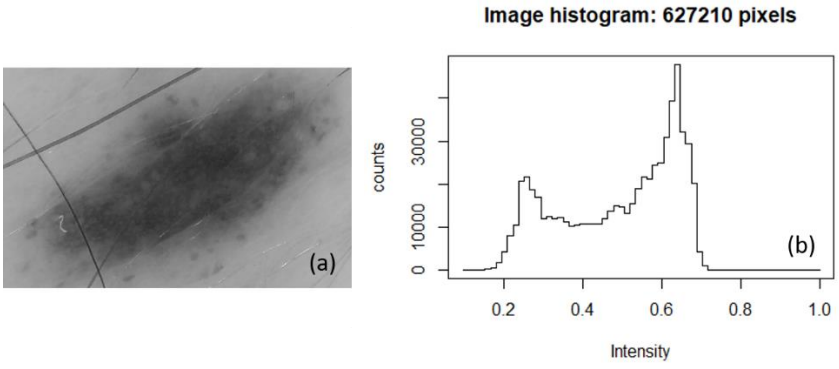


Figura 23 - Apresentação do histograma (b) da lesão em escalas de cinza (a)

O método passa pelo cálculo do valor de Otsu, com recurso à função *otsu()*. Este resultado é depois aplicado de forma a transformar o valor dos pixéis, que se encontrem abaixo desse

resultado, para que assumam o valor 0 (cor preta). Desta forma é possível obter uma nova imagem da lesão segmentada. Os valores de Otsu, aplicados às imagens, encontram-se no final do ponto 4.2.3 Preparação dos dados.

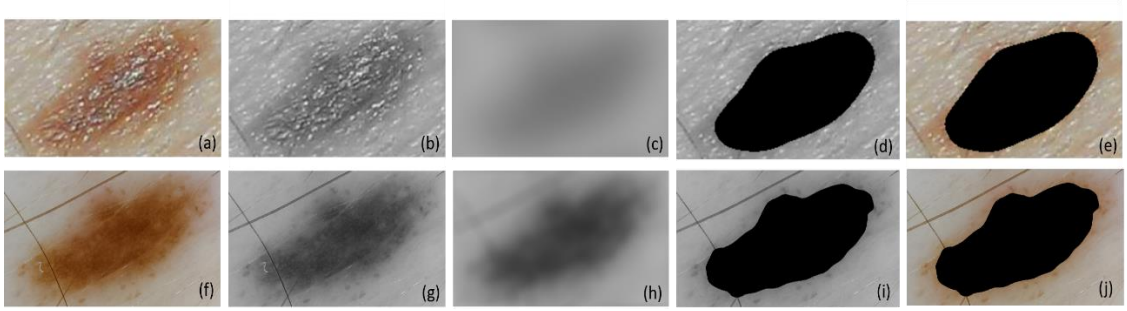


Figura 24 - Representação de fases de segmentação de uma lesão

Assim sendo, com base na Figura 24, é possível visualizar 5 diferentes formas de apresentar a mesma imagem: numa escala de cores (modelo RGB) - (a; f), em que cada pixel assume 3 valores entre 0 e 255; numa escala de cinzas – (b; g), cujo cada pixel ocupa 8 bits e pode assumir um valor entre 0 e 255; numa diminuição da nitidez – (c; h), de forma a eliminar ruído como pelos; e a segmentação da lesão na imagem – (d; e; i; j), cujo valor dos pixéis da lesão é igual a 0.

Um dos problemas, no processo de análise da imagem, é a qualidade que esta possui quando captada com certos dispositivos. Na Figura 24 é possível perceber a diferença de qualidade das imagens capturadas por *smartphones* e das imagens capturadas por dispositivos como o dermoscópio – imagem (a) e (f), respetivamente. Apesar de ser a mesma lesão, à primeira vista não parece. O reflexo captado na imagem (a), transparece a noção de que determinados pixéis assumem a cor clara, afetando o seu valor. Já na imagem (f), com base na localização do reflexo referido em (a), a nitidez da imagem - bem como a ausência de reflexos de cor - proporciona uma conclusão clara do que é visualizado. O mesmo ocorre com a segmentação da lesão. Nas segmentações (i) e (j) é possível visualizar os contornos e preenchimentos da lesão de forma mais precisa, em comparação com as segmentações (d) e (e).

4.2.3. Preparação dos dados

As imagens foram obtidas de duas formas, por captação com *smartphones* e captação com um dermoscópio adaptável. As imagens capturadas pelo dermoscópio apresentaram melhor qualidade. Assim sendo, estas foram escolhidas para o processo.

Os objetos, tal como visualizado anteriormente, possuem área de imagem que não é necessária e relevante ao estudo. Desta forma, as imagens foram cortadas individualmente e não recorrendo a um processo automático de área de corte. Isto porque, as lesões não se encontram na mesma região de imagem para imagem.

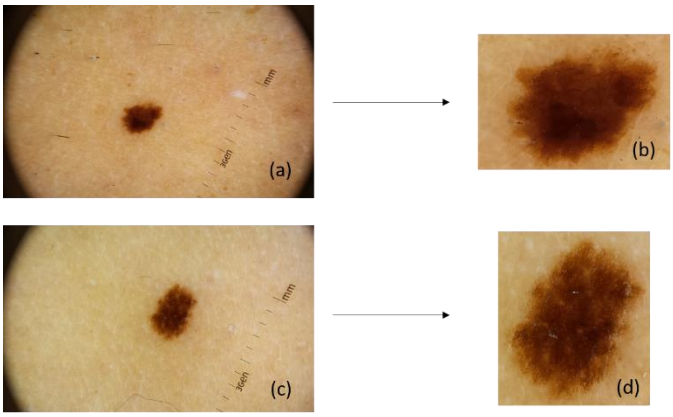


Figura 25 - Demonstração de diferentes pontos de corte em imagens de lesões distintas

Nos dois exemplos apresentados na Figura 25, a lesão em (b) foi obtida pelo corte aplicado à imagem (a) e a lesão (d) pelo corte aplicado à imagem (c). Imagine-se um ponto numa imagem e duas linhas sobre esse mesmo ponto, que percorram a largura e o comprimento da imagem. Ao focar um pixel, conseguimos ter uma coordenada (*x*,*y*) em que o valor de *x* corresponde a um ponto numa linha que percorra o comprimento de uma imagem; e *y* corresponde ao ponto numa linha que percorre a largura da mesma. Deste modo, cada lesão foi obtida com base na localização de dois pixéis: o pixel que limita o canto superior esquerdo e o pixel que limita o canto inferior direito. Em (b) o pixel correspondente ao canto superior esquerdo assume a coordenada (1710,1122) e o pixel correspondente ao canto inferior direito assume a coordenada (2352,1573). Já no caso da lesão (d), os valores da coordenada do primeiro pixel são (2039,2709) e do segundo são (692,1457). Por serem valores distintos, não foi possível aplicar um ciclo que possibilitasse o corte uniforme das lesões.

Devido a certas limitações encontradas, a serem explicadas em 6.2 Riscos Verificados e Limitações, a modelação terá por base o atributo relativo ao risco de ser melanoma. Os valores apresentados, correspondem a uma análise efetuada por profissionais de saúde, com base nas características das lesões.

Como analisado no ponto anterior, este atributo possui valores entre 1 e 3, que correspondem, respetivamente, ao nível mais baixo e mais alto de risco de melanoma. Por serem

relativamente poucas imagens para a quantidade de classes a analisar, então considerou-se que o nível 1 correspondia ao risco de não ser melanoma e os níveis 2 e 3, ao risco de ser melanoma. Desta forma existem duas classes para classificação e *clustering*: a classe de não ser melanoma (0) e a classe de ser melanoma (1).

Devido aos cortes aplicados - às imagens capturadas pelo dermoscópio adaptável -, de forma irregular de imagem para imagem, estas tiveram de sofrer um ajuste do seu tamanho. Estes planos de corte podem ser visualizados na Tabela 6. As dimensões que apresentavam diferiam no número de pixéis e por isso, recorrendo a um ciclo, todas as imagens foram ajustadas para que assumissem o comprimento, bem como a largura, igual a 100. Após aplicar a alteração, as mesmas duas imagens assumem as dimensões de 100x100. Para além disso, os valores dos pixéis também alteram ligeiramente. Isto acontece pela alteração da dimensão da imagem. O pixel assume um tom ligeiramente – não sendo nítida a diferença - mais claro que a sua versão anterior. Assim sendo, os cortes aplicados podem ser analisados na Tabela 6 que se encontra no final do presente ponto. Nesta tabela é especificada a imagem, o plano de corte e o valor de Otsu.

Para aplicação de alguns modelos, a estrutura dos dados é um fator relevante. Para o reconhecimento de imagem e posterior modelação, as imagens, que possuem uma forma não estruturada, têm de ser transformadas de forma a assumirem uma forma estruturada. Entende-se por forma não estruturada o tipo de dados apresentados como – "*Formal class 'Image'*". Desta forma, o *array* de imagens teve de sofrer alteração da sua estrutura.

Posteriormente, e depois de terem sido convertidos os valores associados ao atributo do risco de melanoma, estes foram declarados para emparelharem com o índice correspondente da lista de imagens. Devido à pequena quantidade de imagens, estas foram divididas em 70% para treino e 30% para teste. Não foi efetuada nenhum tipo específico de *cross-validation*. Isto porque, como já mencionado, a quantidade de dados era reduzida, então optou-se pela seleção das primeiras 70% imagens para a fase de treino e as restantes 30% para teste. Das imagens correspondentes à percentagem de 70% para treino, 20% representam a parte da validação. Assim sendo, os valores do risco foram convertidos para valores categóricos através da função *to_categorical()* do package *keras*.

Com base no resultado na Tabela 5, a primeira coluna corresponde à classe 0, ou classe dos não melanomas, e a segunda coluna à classe 1, ou classe dos melanomas. Cada linha corresponde a uma imagem. Quando uma posição se encontra preenchida a 1, significa que a

imagem possui uma lesão referente a um não melanoma ou a um melanoma, dependendo dos casos. Esta transformação está relacionada com o modelo a desenvolver e o tipo de dados, bem como a forma, que estes apresentam.

Tabela 5 - Exemplo de apresentação de dados na forma categórica

	[,1]	[,2]
[1,]	1	0
[2,]	0	1
[3,]	1	0
[4,]	1	0
[5,]	1	0

De forma perceber os cortes aplicados às imagens, segue a Tabela 6. Nesta tabela podem ser visualizados os identificadores das imagens (de 1 a 16), os planos de cortes e os valores de Otsu mencionados no ponto 4.2.2 Compreensão dos dados.

Tabela 6 - Plano de cortes, e valor de Otsu, das imagens capturadas pelo dermoscópio adaptável (da imagem 1 à imagem 16)

Imagem	Plano de Corte	Valor de Otsu
1	[1624:2658,1005:1646]	0,458984375
2	[1770:2177,772:1493]	0,404296925
3	[386:3772,452:2039]	0,504453125
4	[1923:2760,714:1282]	0,455078125
5	[2039:2709,692:1457]	0,388671875
6	[1617:2636,619:1260]	0,451171875
7	[1333:2956,415:1202]	0,501953125
8	[1347:2636,860:1311]	0,455078125
9	[1566:2905,918:1529]	0,494140625
10	[1937:2833,1049:1609]	0,408203125
11	[1668:2316,736:1493]	0,427734375
12	[1710:2352,1122:1573]	0,376953125
13	[1675:2913,612:1631]	0.537109375
14	[1595:2847,452:1748]	0,533203125
15	[1311:2483,357:1762]	0,498046875
16	[1733:2665,714:1464]	0,494140625

Cada plano de corte difere de imagem para imagem. Como previamente referido, as lesões não se encontram no mesmo local nas captações. Desta forma, existem planos cuja coordenada do pixel, que limita o canto superior da lesão, corresponde a (2011,1185) ou corresponde a (386,452), dependendo da região da lesão na imagem. Assim sendo, o plano de corte é definido pela abcissa do canto superior da lesão, abcissa do canto inferior da lesão, ordenada do canto superior da lesão e ordenada do canto inferior da lesão, respetivamente. Na Tabela 6 apenas se encontram os planos de corte das primeiras 16 imagens. Os restantes planos podem ser analisados no Anexo I.

No caso do valor de Otsu, como explicado, este permite compreender o limite da intensidade de cor, numa escala de cinzas. Cada pixel que possua valor abaixo desse limite assume a cor preta, ou seja, o valor do pixel igual a 0. Quanto maior for o valor, mais clara é a cor da lesão. Quanto menor for o valor de Otsu, mais escura é a cor da lesão.

4.2.4. Modelação

Para exploração de modelos classificativos de DM foram escolhidas as *Convolutional Neural Network* (CNN); através da aplicação de funcionalidades existentes nos packages *TensorFlow* e *Keras*. Como referido anteriormente, a instalação dos *packages* na ferramenta, por si só, não permitia o acesso a determinadas utilidades. Como tal, o *Anaconda Navigator* foi instalado. Isto porque os *packages* são muito utilizados em Python e, como o projeto foi desenvolvido em R, existem diferenças. Relativamente à exploração de modelos de *clustering*, um modelo previamente treinado para extração de características foi aplicado, bem como o método *k-means*.

A arquitetura de uma CNN corresponde às camadas que constituem o modelo e à forma como elas se encontram organizadas na estrutura. Existem camadas de *input*, camadas convolucionais, camadas de *maxpooling*, camadas *dropout*, camadas *flatten*, camadas *dense* e camadas de *output*. A camada de *input* é a camada que possui, como um dos parâmetros, a forma do *input* a receber pelo modelo (i.e., *input_shape*). As camadas convolucionais podem ser utilizadas como tipo de camadas *input*, ao definir o parâmetro mencionado anteriormente. Nestas camadas é criado um *kernel* convulocional, que, de acordo com o tamanho especificado, cobre certa região da imagem. Também conhecido como filtro, este possibilita a obtenção de um mapa de características da imagem. O filtro percorre a imagem e vai originando um *output* que servirá de entrada na camada seguinte. As camadas *maxpooling* permitem a redução da quantidade de

parâmetros a serem treinados na rede. Isto permite um controlo da redundância e excesso de parâmetros não significantes para a aprendizagem do modelo. As camadas *dropout*, tal como as camadas *maxpooling*, servem para prevenir o *overfitting*. Estas camadas transformam frações do *input* em 0, o que leva ao não processamento das mesmas. As camadas *flatten* permitem obter um vetor das características, permitindo a transformação do *input*. As camadas *dense* permitem adicionar uma camada densamente conectada ao *output*. Ou seja, permitem perceber a dimensão do *output* da camada. No caso do projeto, o *output* final está relacionado com as duas classes já mencionadas. Desta forma, a dimensão do output dos modelos será igual a 2.

A. Classificação

As CNN estão associadas, casualmente, a métodos de *Deep Learning* (DL). De acordo com Chollet e Allaire (2018), o DL é um subcampo do *Machine Learning* que enfatiza a aprendizagem por camadas sucessivas de representações com significado. Quantas mais camadas, maior a profundidade do modelo. Num primeiro modelo, os *packages TensorFlow* e *Keras* foram ambos aplicados.

Na Figura 26 encontra-se apresentada a arquitetura do primeiro modelo da CNN elaborada. O modelo possui uma primeira camada de input, que corresponde à junção da imagem de cor e da primeira camada convolucional. Como parâmetros, a camada de input recebe a imagem com dimensão 100x100 (largura x altura da imagem) e, por ser uma imagem de cor, o terceiro parâmetro de input é o valor 3 (possui valores na escala *Red, Green and Blue* - RGB). Já a camada convolucional é definida pelo número de filtros e o tamanho do *kernel*. De forma a perceber quantos parâmetros resultam desta fase, existe a seguinte fórmula de cálculo: $(2 \times 2 \times 3 \times 48) + 48$. A multiplicação corresponde ao tamanho do *kernel*,(2x2) pelo valor relativo às cores (3) que a imagem apresenta (escala RGB como já mencionado) e pelo número de filtros existentes (48). A esta multiplicação, acresce a soma da quantidade de filtros (48). O resultado desta fórmula corresponde a 624 parâmetros. Na segunda camada convolucional, o número de filtros foi reduzido para 32 e o tamanho do *kernel* manteve-se. Desta forma, o número de parâmetros da camada é calculado da seguinte forma: $(2 \times 2 \times 48 \times 32) + 32$. A fórmula corresponde à multiplicação do tamanho do *kernel*, pelo número de filtros da camada anterior e pelo número de filtros da atual camada; com a soma da quantidade de filtros da presente camada. O resultado corresponde a 6176 parâmetros nesta fase. A primeira camada de *maxpooling* e a primeira camada de *dropout* servem de camadas de controlo, desta forma não existem parâmetros

nestas fases. Na terceira camada convolucional o número de filtros aumentou para 64 e o tamanho do *kernel* manteve-se. A fórmula de cálculo dos parâmetros corresponde a: $(2 \times 2 \times 32 \times 64) + 64$. A multiplicação corresponde ao tamanho do *kernel* pelos filtros das camadas anteriores e os filtros da camada atual. A esta multiplicação acresce a quantidade de filtros existentes na camada e o resultado corresponde a 8256 parâmetros nesta fase. Na quarta camada convolucional o número de filtros e o tamanho do *kernel* mantêm-se. A fórmula para o cálculo dos parâmetros é a seguinte: $(2 \times 2 \times 64 \times 64) + 64$. Como resultado obtém-se 16448 parâmetros. Como anteriormente, a camada *maxpooling* e a camada *dropout* não apresentam parâmetros e servem de controlo da CNN. Tal como estas, a camada *flatten* não possui parâmetros, mas transforma parâmetros bidimensionais em parâmetros unidimensionais de forma a serem processados por camadas totalmente conectadas. A primeira camada *dense* possui 256 neurónios que processam o input fornecido pelas camadas anteriores e origina um output de: $(33856 \times 256) + 356 = 8667392$ parâmetros. A terceira camada de *dropout*, tal como as anteriores, existe para controlar a quantidade de parâmetros da rede. Na camada de output, que é uma camada *dense*, existem apenas 2 neurónios que correspondem às classes a prever pelo modelo. Desta forma, os parâmetros desta camada são obtidos com base na seguinte fórmula: $(2 \times 256) + 2 = 514$. Os valores resultado de parâmetros em cada camada pode ser confirmado no resultado da consola relativo ao processamento do primeiro modelo na ferramenta RStudio.

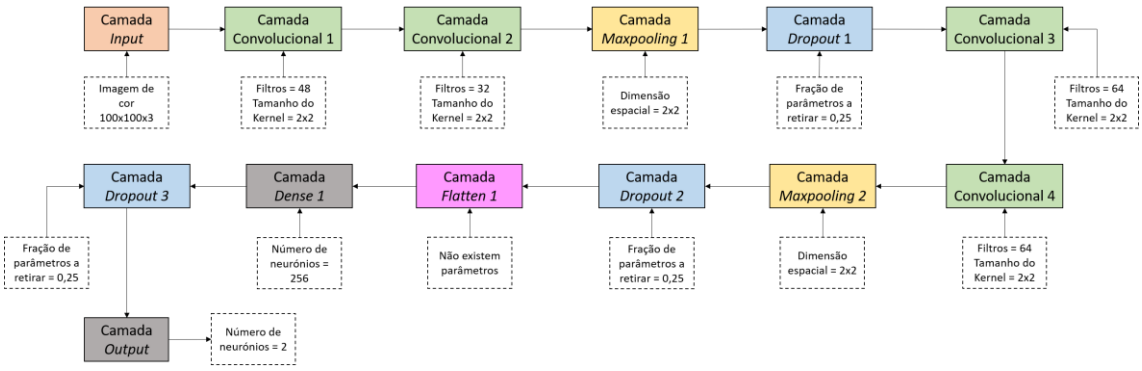


Figura 26 - Arquitetura do primeiro modelo elaborado

Como apresentado na Figura 26, e com base no resultado demonstrado Na Figura 27 é possível visualizar o resultado do processamento do modelo criado no RStudio As camadas e os tipos correspondentes, o *output* das camadas e a quantidade de parâmetros que cada camada possui podem ser analisados. De acordo com os cálculos efetuados para a arquitetura do modelo

na Figura 26, a quantidade de parâmetros processados foi de 8699410 (como se pode verificar na Figura 27).

```
#Modelo TensorFlow e Keras – Modelo 1
## _________________________________________________________________
## Layer (type)                 Output Shape              Param #
## =================================================================
## conv2d_1 (Conv2D)            (None, 100, 100, 48)      624
## _________________________________________________________________
## conv2d_2 (Conv2D)            (None, 99, 99, 32)        6176
## _________________________________________________________________
## max_pooling2d_1 (MaxPooling2D)   (None, 49, 49, 32)    0
## _________________________________________________________________
## dropout_1 (Dropout)          (None, 49, 49, 32)        0
## _________________________________________________________________
## conv2d_3 (Conv2D)            (None, 48, 48, 64)        8256
## _________________________________________________________________
## conv2d_4 (Conv2D)            (None, 47, 47, 64)        16448
## _________________________________________________________________
## max_pooling2d_2 (MaxPooling2D)   (None, 23, 23, 64)    0
## _________________________________________________________________
## dropout_2 (Dropout)          (None, 23, 23, 64)        0
## _________________________________________________________________
## flatten_1 (Flatten)          (None, 33856)             0
## _________________________________________________________________
## dense_1 (Dense)              (None, 256)               8667392
## _________________________________________________________________
## dropout_3 (Dropout)          (None, 256)               0
## _________________________________________________________________
## dense_2 (Dense)              (None, 2)                 514
## =================================================================
## Total params: 8,699,410
## Trainable params: 8,699,410
## Non-trainable params: 0
## _________________________________________________________________
```

Figura 27 - Modelo TensorFlow e Keras - Primeiro modelo criado

A CNN receberá como *input* o resultado de 100x100x3, que corresponde às dimensões das imagens e às cores que estas apresentam (RGB). De acordo com o sumário apresentado é possível perceber a arquitetura da CNN estudada. As camadas convolucionais são as camadas *core* que efetuam a maior parte computacional. Cada camada possui um conjunto de filtros que aprendem. Cada filtro percorre o *input* e produz um mapa de ativação, em duas dimensões, das posições que assume. Com isto, a rede vai aprendendo filtros que são ativados aquando da aparição de uma borda de determinado objeto na imagem. A camada de *pooling*, entre camadas convolucionais, e as de *dropout* permitem a eliminação do tamanho espacial e dos parâmetros, para que haja redução da computação na rede neuronal. A camada de *flatten* vai atenuando o *input* a receber. Ou seja, vai controlando a ordem das dimensões de *input*. A camada de *dense* permite obter o *output* do modelo.

A métrica escolhida para analisar o desempenho do modelo é a ACC, métrica que indica o valor da acuidade. Como função *fit* aplicada ao modelo é possível perceber qual a percentagem

de validação para os dados de treino, através do parâmetro *validation_split*; quantas vezes os dados de treino atravessaram a rede neural, através do parâmetro *epochs* (cada *epoch* corresponde a uma passagem de dados do início até ao fim, e do fim de volta ao início, da rede neuronal); e o número total de exemplos de treino presentes num único *batch*, através do parâmetro *batch_size*.

Num segundo modelo, visível na Figura 28, utilizando apenas o *package* Keras, foi construída uma rede neuronal com apenas camadas *dense*.

```
#Modelo Keras – Modelo 2
## _________________________________________________________________
## Layer (type)                 Output Shape              Param #
## =================================================================
## dense_1 (Dense)              (None, 512)               15360512
## _________________________________________________________________
## dropout_1 (Dropout)          (None, 512)               0
## _________________________________________________________________
## dense_2 (Dense)              (None, 256)               131328
## _________________________________________________________________
## dropout_2 (Dropout)          (None, 256)               0
## _________________________________________________________________
## dense_3 (Dense)              (None, 128)               32896
## _________________________________________________________________
## dropout_3 (Dropout)          (None, 128)               0
## _________________________________________________________________
## dense_4 (Dense)              (None, 2)                 258
## =================================================================
## Total params: 15,524,994
## Trainable params: 15,524,994
## Non-trainable params: 0
## _________________________________________________________________
```

Figura 28 - Modelo Keras - Segundo modelo criado

Tal como explicado no modelo anterior, as camadas *dense* possibilitam a saída de output. A função *fit*, tal como no modelo anterior, possibilita perceber percentagens associadas a dados de validação, o número de "voltas" que os dados efetuam na rede, entre outros parâmetros. Basicamente corresponde à fase de treino do modelo apresentado na Figura 28.

De forma a ser possível a compreensão da arquitetura do segundo modelo criado, segue a Figura 29. Na arquitetura do segundo modelo, apenas foram utilizadas camadas dense e camadas *dropout*.

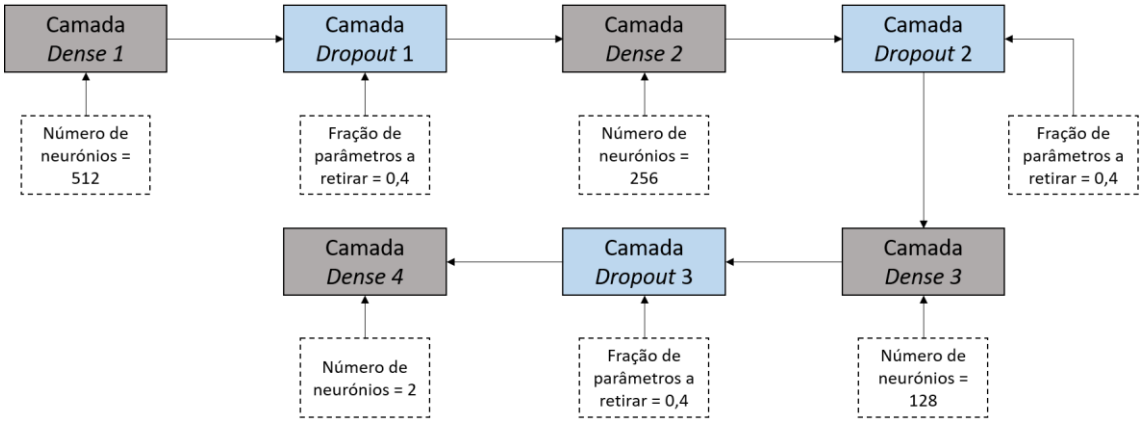


Figura 29 - Arquitetura do segundo modelo elaborado

As camadas *dense*, visíveis na arquitetura da Figura 29, possuem neurónios que processam a informação das imagens. Já as camadas *dropout* vão eliminando alguns parâmetros de forma a controlar o tempo de processamento da rede, por exemplo. Na primeira camada *dense* a fórmula que define o número de parâmetros é a seguinte: $(100 \times 100 \times 3 \times 512) + 512 = 15360512$. Nas camadas *dropout*, como já foi referido, não existem parâmetros a acrescentar, apenas a retirar de acordo com a fração indicada em cada camada. Na segunda camada *dense*, os parâmetros são dados de acordo com a seguinte equação: $(512 \times 256) + 256 = 131328$. Na terceira camada *dense*, os parâmetros desta fase são conhecidos com base na seguinte equação: $(256 \times 128) + 128 = 32896$. Na última camada *dense*, apenas existem dois neurónios – tal como no primeiro modelo criado – e a fórmula que define a camada é a seguinte: $(2 \times 128) + 2$. O resultado equivale a 258 parâmetros nesta fase. Comparativamente ao primeiro modelo criado, não existem camadas convolucionais (para além de outro tipo de camadas), que são camadas caracterizantes das CNN.

B. Clustering

Com base no trabalho efetuado por Simonyan e Zisserman (2014), existe um modelo previamente treinado num *dataset* – ImageNet - que possui 1,4 milhões de imagens identificadas e mil classes. Utilizar um modelo pré-treinado possibilita uma abordagem mais efetiva aquando posse de um número reduzido de dados, o que é o caso. Apesar do modelo ter sido treinado com

imagens de animais, nada impede o uso para classificação de diferentes classes como o ser ou não melanoma.

Na Figura 30 é possível analisar o modelo previamente treinado para o *clustering* das imagens. O modelo possui 14714688 parâmetros e é constituído por uma camada *input*, por treze camadas convolucionais e cinco camadas *maxpooling*.

```
#Modelo Clustering – Modelo 3

## Model
## _________________________________________________________________
## Layer (type)                 Output Shape              Param #
## =================================================================
## input_1 (InputLayer)         (None, None, None, 3)     0
## _________________________________________________________________
## block1_conv1 (Conv2D)        (None, None, None, 64)    1792
## _________________________________________________________________
## block1_conv2 (Conv2D)        (None, None, None, 64)    36928
## _________________________________________________________________
## block1_pool (MaxPooling2D)   (None, None, None, 64)    0
## _________________________________________________________________
## block2_conv1 (Conv2D)        (None, None, None, 128)   73856
## _________________________________________________________________
## block2_conv2 (Conv2D)        (None, None, None, 128)   147584
## _________________________________________________________________
## block2_pool (MaxPooling2D)   (None, None, None, 128)   0
## _________________________________________________________________
## block3_conv1 (Conv2D)        (None, None, None, 256)   295168
## _________________________________________________________________
## block3_conv2 (Conv2D)        (None, None, None, 256)   590080
## _________________________________________________________________
## block3_conv3 (Conv2D)        (None, None, None, 256)   590080
## _________________________________________________________________
## block3_pool (MaxPooling2D)   (None, None, None, 256)   0
## _________________________________________________________________
## block4_conv1 (Conv2D)        (None, None, None, 512)   1180160
## _________________________________________________________________
## block4_conv2 (Conv2D)        (None, None, None, 512)   2359808
## _________________________________________________________________
## block4_conv3 (Conv2D)        (None, None, None, 512)   2359808
## _________________________________________________________________
## block4_pool (MaxPooling2D)   (None, None, None, 512)   0
## _________________________________________________________________
## block5_conv1 (Conv2D)        (None, None, None, 512)   2359808
## _________________________________________________________________
## block5_conv2 (Conv2D)        (None, None, None, 512)   2359808
## _________________________________________________________________
## block5_conv3 (Conv2D)        (None, None, None, 512)   2359808
## _________________________________________________________________
## block5_pool (MaxPooling2D)   (None, None, None, 512)   0
## =================================================================
## Total params: 14,714,688
## Trainable params: 14,714,688
## Non-trainable params: 0
## _________________________________________________________________
```

Figura 30 - Modelo previamente treinado VGG16

O resultado da Figura 30, como mencionado, representa a arquitetura do modelo VGG16. Para cada imagem, a função *predict()* do *package Keras* será aplicada de forma a combinar o

modelo com as características extraídas das imagens. As características podem ser obtidas através da aplicação de uma função relativa à fase de pré-processamento do ImageNet – *imagenet_preprocess_input()*. Como resultado da aplicação da função *predict()* obtém-se um *dataframe* com informação das 12801 características associadas às 106 imagens. Esse resultado é guardado num ficheiro .RData e depois lido através da diretoria onde se encontra. Após ser feito o *load()*, os dados são carregados para um *plot*, de forma a perceber como se encontram aglomerados/distribuídos, tendo em conta as suas duas primeiras características.

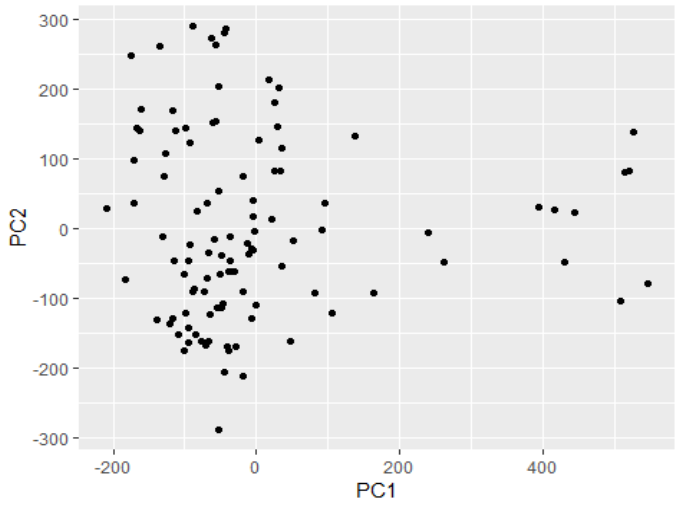


Figura 31 - Distribuição das imagens de acordo com as suas duas primeiras características

Com base na Figura 31, não há uma limitação que permita dizer, com certezas, a quantidade de *clusters* existentes para as duas primeiras características escolhidas. Apesar de ser visível a existência de, pelo menos, dois *clusters*.

4.2.5. Avaliação

Nesta etapa serão analisados os comportamentos dos modelos nas fases de treino e validação, bem como na fase de teste. Esta análise terá em conta o valor da métrica acuidade.

A. Classificação

No caso das CNN e na utilização do *TensorFlow* com o *Keras*, o resultado da aplicação da função *fit*, ou seja, o desempenho do modelo ao longo da fase de treino pode ser analisado na Figura 32.

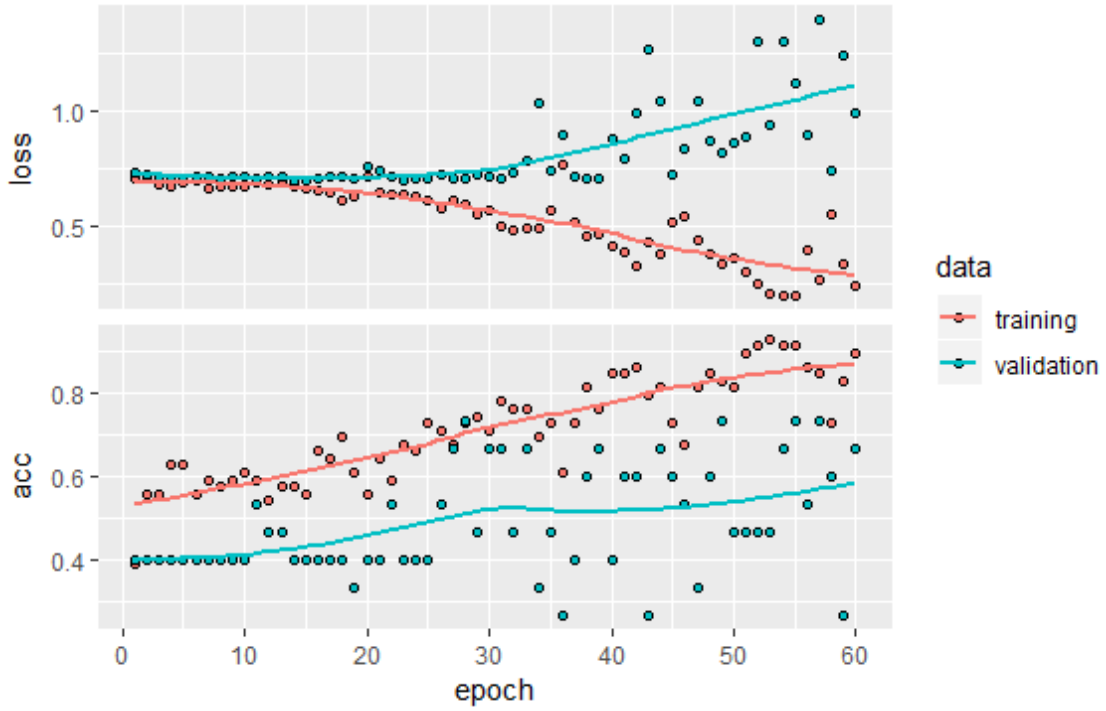


Figura 32 - Apresentação do desempenho do primeiro modelo na fase de treino

Conforme dito na legenda de cores da Figura 32, a linha vermelha corresponde aos resultados obtidos pela aplicação dos dados de treino ao modelo e a linha verde aos dados de validação. O primeiro gráfico corresponde à perda, ou seja, ao erro no não acerto da classe. Já o segundo gráfico corresponde à métrica escolhida, a acuidade. Esta métrica devolve o acerto do modelo ao classificar as imagens, neste caso. Ao longo dos *epochs* – parâmetros que correspondem às iterações do modelo – o valor da acuidade e da perda vai alterando. Quanto maior é a acuidade, menor é a perda e informação. Os valores da acuidade, na fase de treino, estiveram sempre acima dos 50% de acerto. Já no caso da acuidade. na fase de validação, os

valores não ultrapassaram os 80% de acerto. No caso da perda na fase de treino, os valores estiveram abaixo dos 80%. Já na fase de validação, houve ocorrência de valores acima de 100%.

De acordo com o desempenho do modelo foram elaboradas tabelas de apresentação de resultados. Na Tabela 7 pode ser visualizado o valor da métrica *Accuracy* (ACC) e o valor da métrica *Loss* do primeiro modelo criado.

Tabela 7 - Resultado da ACC para a fase de treino do primeiro modelo

ACC	Loss
~0.91	~0.36

Analisando a Tabela 7 pode dizer-se que a ACC do modelo, na fase de treino é de 91%. Já a métrica *Loss* obteve o valor de 36%.

Tabela 8 - Matriz de confusão dos resultados da fase de treino do primeiro modelo

		Atual	
		0	1
Previsto	0	34	6
	1	1	33

Com base na matriz de confusão apresentada na Tabela 8 é possível analisar que o modelo previu 40 imagens relativas a lesões como não melanoma e 34 imagens relativas a lesões de melanomas. Só que dessas 40 previsões de não melanomas errou em 6 (Falsos Negativos) e das 34 de melanoma errou em 1 (Falsos Positivos). Por causa disto, o modelo não atingiu 100% de ACC. Caso a acuidade fosse igual a 100%, a diagonal – onde os Verdadeiros Positivos e os Verdadeiros Negativos se encontram – seria a única a estar preenchida.

Na matriz de probabilidades de um modelo, cada linha corresponde a uma imagem com a lesão. A primeira coluna corresponde à probabilidade que o modelo tem de prever a classe 0 e a segunda coluna, à probabilidade do modelo prever classe 1. As duas últimas colunas correspondem à previsão efetuada. A forma de visualizar que o modelo falhou na previsão de uma classe é analisar se as duas últimas colunas, numa linha, apresentam valores diferentes. No caso de apresentarem, então o modelo falhou nessa previsão.

Com base na análise de sensibilidade e especificidade do primeiro modelo, os seguintes cálculos foram efetuados:

- Sensibilidade: $VP \div (VP + FN) \leftrightarrow 34 \div (34 + 6) = 0{,}85$ (85%);
- Especificidade: $VN \div (VN + FP) \leftrightarrow 33 \div (33 + 1) = 0{,}97$ (97%).

Um modelo diz-se sensível, no caso do presente projeto de dissertação, quanto maior for a sua capacidade de acertar na lesão não ser um melanoma. Um modelo diz-se específico, quanto maior for a sua capacidade para acertar na lesão ser melanoma. Com base nos cálculos efetuados, o modelo possui maior especificidade.

Na Tabela 9 podem ser visualizados os valores das métricas ACC e *Loss* para a fase de testes do primeiro modelo criado.

Tabela 9 - Resultado da ACC para a fase de teste do primeiro modelo

ACC	Loss
~ 0.44	~ 0.95

A ACC do modelo, na fase de testes, corresponde a 44%. Um valor bastante mais baixo que o obtido na fase de treino. O valor do *Loss* foi muito superior, comparativamente ao valor na fase de treino.

Tabela 10 - Matriz de confusão dos resultados da fase de teste do primeiro modelo

		Atual	
		0	1
Previsto	0	7	13
	1	5	7

Com base na Tabela 10, o modelo conseguiu prever 7 imagens com classe igual a 0, mas errou em 13 (Falsos Negativos). Conseguiu prever 7 imagens com classe igual a 1, mas errou em 5 (Falsos Positivos). Com a quantidade de acertos inferior à quantidade de erros em quase metade, a acuidade não podia assumir valores melhores.

Para a análise da sensibilidade e da especificidade seguem as seguintes equações:

- Sensibilidade: $VP \div (VP + FN) \leftrightarrow 7 \div (7 + 13) = 0{,}35$ (35%);
- Especificidade: $VN \div (VN + FP) \leftrightarrow 7 \div (7 + 5) = 0{,}58$ (58%).

Com base nos cálculos efetuados, o modelo é mais específico do que sensível. Tal como na fase de treino, as probabilidades do modelo prever classe 0, ou a classe 1, podem ser visualizadas com a função *predict_proba()*.

Relativamente ao modelo criado apenas recorrendo ao *package Keras* é possível perceber que a arquitetura difere em muito para o modelo criado anteriormente.

Tabela 11 - Resultado da ACC para a fase de treino do segundo modelo

ACC	Loss
~ 0.58	~ 0.69

A ACC obtida, na fase de treino do segundo modelo, pode ser visualizada na Tabela 11 e quer dizer que o acerto foi de 53%. Já a métrica *Loss* obteve 69% na fase de treino do segundo modelo.

Tabela 12 - Matriz de confusão dos resultados da fase de treino do segundo modelo

		Atual	
		0	1
Previsto	0	28	24
	1	7	15

De forma a perceber se o modelo é sensível ou específico foram efetuados os seguintes cálculos de acordo com a informação da Tabela 12:

- Sensibilidade: $VP \div (VP + FN) \leftrightarrow 28 \div (28 + 14) = 0{,}67$ (67%);
- Especificidade: $VN \div (VN + FP) \leftrightarrow 15 \div (15 + 7) = 0{,}68$ (68%).

De acordo com os resultados, o segundo modelo é mais específico do que sensível, na fase de treino.

Na Tabela 13 encontram-se apresentados os valores da métrica ACC e da métrica *Loss* para a fase de testes do segundo modelo criado.

Tabela 13 - Resultado da ACC para a fase de teste do segundo modelo

ACC	Loss
~ 0.53	~ 0.69

A ACC do modelo, na fase de testes, foi de 53%. A métrica *Loss* atingiu 69% na mesma fase de análise de desempenho relativo ao segundo modelo criado.

Tabela 14 - Matriz de confusão dos resultados da fase de teste do segundo modelo

		Atual	
		0	1
Previsto	0	12	15
	1	0	5

De forma a perceber a sensibilidade e a especificidade do modelo, na fase de testes, foram efetuados os seguintes cálculos com base na informação da Tabela 14:

- Sensibilidade: $VP \div (VP + FN) \leftrightarrow 12 \div (12 + 15) = 0{,}44$ (44%);
- Especificidade: $VN \div (VN + FP) \leftrightarrow 5 \div (5 + 0) = 1$ (100%).

Com base nos resultados obtidos, o segundo modelo possui uma percentagem de 100% relativa à sua especificidade.

B. Clustering

O objetivo do *clustering* é perceber, de acordo com as características de cada imagem, quais são semelhantes e quais diferem entre si. Desta forma foi criada a aglomeração presente na Figura 33. O *cluster* apresentado demonstra quais as imagens que correspondem aos k=4 centros definidos. Estes *clusters* correspondem à análise das 12801 características existentes nas imagens de melanomas. Cada imagem foi associada a um *cluster*, com base nas características verificadas, ou seja, o primeiro *cluster* – em caso de exemplo – correspondem ao conjunto de imagens que possuem X número de características. Assim sendo, fazem parte do primeiro grupo as imagens que verificam esse X número de características. Desta forma, cada imagem pode possuir a ocorrência de 12801 características ou apenas 1.

```
#Count
cluster_list %>%
  count(cluster_pca, class)

## # A tibble: 106 x 3
##    cluster_pca class     n
##          <int> <chr> <int>
##  1           1 101       1
##  2           1 102       1
##  3           1 103       1
##  4           1 104       1
##  5           1 105       1
##  6           1 106       1
##  7           1 11.       1
##  8           1 12.       1
##  9           1 13.       1
## 10           1 14.       1
## # ... with 96 more rows
```

Figura 33 - Contagem de classes do cluster criado com base nas características de cada imagem

Por sua vez, na Figura 34, é possível verificar em quantas imagens ocorre uma determinada característica. Ou seja, tendo-se uma característica A - como exemplo - e 106 imagens, é possível saber quantas estão associadas à característica exemplo. Assim sendo,as 12801 características foram centradas em k=4 *clusters*. Fazem parte do primeiro *cluster* – em questão de exemplo -, as imagens que verifiquem a ocorrência da característica Y.

```
#Count

cluster_list %>%
  count(cluster_feature, class)

## # A tibble: 106 x 3
##    cluster_feature class     n
##              <int> <chr> <int>
##  1               1 1.j       1
##  2               1 18.       1
##  3               1 2.j       1
##  4               1 21.       1
##  5               1 22.       1
##  6               1 25.       1
##  7               1 26.       1
##  8               1 27.       1
##  9               1 35.       1
## 10               1 36.       1
## # ... with 96 more rows
```

Figura 34 - Contagem de classes do cluster criado com base nas imagens que possuem uma determinada característica

Com base no *plot* gerado na Figura 35 é possível verificar mapeamento entre os clusters cujas imagens possuem determinadas características e entre os clusters em que, cada característica, ocorre em determinadas imagens diferentes. Ou seja, analisando a legenda correspondente ao *cluster_pca*: o círculo corresponde ao primeiro cluster, o triângulo ao segundo, o quadrado ao terceiro e a cruz ao quarto *cluster*. Assim sendo, em questão exemplo, fazem parte do primeiro grupo as imagens que possuem ocorrência de X (um número indeterminado) características. Cada conjunto de classes possui uma cor diferente de forma a ser possível mapear cada caso no gráfico. Apesar de não ser percetível, as colunas de cores diferem de tonalidade. Nenhuma classe possui a mesma cor.

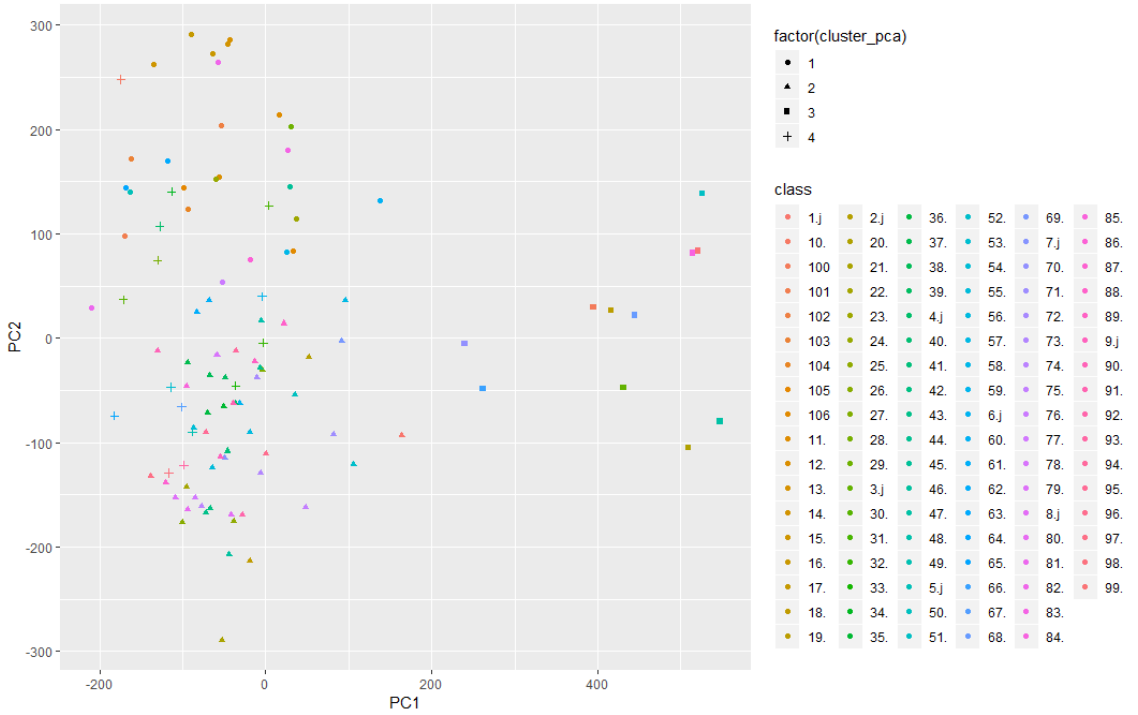


Figura 35 - Mapeamento de classes entre clusters

Para ser mais percetível o mapeamento foi elaborada a seguinte Tabela 15 com exemplos relativos às 16 primeiras imagens. Nesta é possível verificar, de forma mais precisa, a informação que se retira do mapeamento na Figura 35. Isto porque, cada linha da Tabela 15 corresponde a uma imagem, aos clusters a que esta pertence, bem como a classe a que diz respeito. Assim sendo, a primeira imagem corresponde a um triângulo de cor rosa. No caso da quarta imagem, esta corresponde ao quarto cluster (símbolo cruz) e possui uma tonalidade específica de cor verde.

Tabela 15 - Mapeamento dos clusters com as imagens e suas classes (da imagem 1 à imagem 16)

Identificador	*Cluster_pca*	*Cluster_feature*	Imagem	Classe
1	2	1	1.jpg	1.j
2	2	1	2.jpg	2.j
3	3	3	3.jpg	3.j
4	4	4	4.jpg	4.j
5	1	2	5.jpg	5.j
6	2	1	6.jpg	6.j
7	2	1	7.jpg	7.j
8	2	1	8.jpg	8.j
9	2	1	9.jpg	9.j

Identificador	*Cluster_pca*	*Cluster_feature*	Imagem	Classe
10	4	4	10.jpg	10.
11	1	2	11.jpg	11.
12	1	2	12.jpg	12.
13	1	2	13.jpg	13.
14	1	2	14.jpg	14.
15	1	2	15.jpg	15.
16	1	2	16.jpg	16.

Cada imagem corresponde a uma classe específica, como se pode visualizar com base na Figura 35 e na Tabela 15. O resto de informação, em tabela, do mapeamento entre *clusters* pode ser visualizado no Anexo II.

4.2.6. Implementação

Como referido desde o início do projeto, a exploração dos modelos de DM permitiria a criação de uma componente para uma integração futura ao nível da *Pervasive Data Mining Engine*. Isto porque a ferramenta não está preparada para analisar, processar e classificar imagens. Desta forma, não é possível apresentar um contexto de implementação apesar de este ser um passo da metodologia utilizada no projeto.

A exploração de modelos em R demonstra a capacidade de efetuar na análise de dados do tipo imagem da ferramenta. Apesar de ter sido necessária a complementação de funcionalidades com outras ferramentas, a ferramenta R foi o principal ambiente de desenvolvimento de toda a fase prática da dissertação. Desta forma, o mecanismo desenvolvido encontra-se apto para uma futura implementação na *Pervasive Data Mining Engine* (PDME).

No entanto, de um ponto de vista contínuo de aprendizagem, é uma oportunidade para serem explorados mais modelos de classificação e *clustering* de imagens. Sendo esta uma área em constante desenvolvimento e investigação, de certo que novas formas de aplicar modelos, bem como novas modelações, poderão surgir.

5. Análise e Discussão de Resultados

De forma a ser possível a classificação e o *clustering* de imagens de lesões da pele, três modelos de DM foram explorados. De um ponto de vista classificativo, as *Convolutional Neural Network* foram aplicadas com recurso aos *packages TensorFlow* e *Keras*. Num segundo modelo, apenas foram aplicadas funcionalidades disponíveis no pacote *Keras*. Por último, numa análise descritiva, foi utilizado o método *k-means* para o processo de *clustering*.

Relativamente aos dados disponíveis, a quantidade era reduzida. Isto é, 106 objetos não permitem obter um histórico de imagens com ocorrências idênticas ou distintas. Não havia muita diversidade que permitisse a aprendizagem fiável dos modelos. A transformação do atributo relativo ao risco das lesões serem, ou não, melanomas, também pode ter influenciado os resultados. Ou seja, os níveis originais de risco eram 3: risco baixo (45%), risco médio (40%) e risco alto (15%). O risco médio deixou de existir e assumiu-se que faria parte das lesões com probabilidade de serem melanomas. Desta forma apenas passaram a existir duas classes: risco de ser e risco de não ser melanoma. O problema é que se não houver grande discrepância nas características das lesões de risco médio, para as lesões de baixo risco - e se está a assumir que há -, a classificação pode não proporcionar os melhores resultados.

No primeiro modelo, a acuidade obtida foi de, aproximadamente, 91%. Esta foi a percentagem de acerto que o modelo tinha na previsão das classes existentes na fase de treino e validação. Já a métrica de *Loss* atingiu os 36%. Uma das possíveis causas é a existência de ligeiras diferenças entre lesões. Outra opção é existirem mais lesões de uma determinada classe na parte de teste, comparativamente aos dados da parte de treino – problema do *overfitting*.

No segundo modelo o processo, a arquitetura do modelo não possui nenhuma camada convolucional. Neste modelo a *Accuracy* (ACC) obtida com este processo foi de aproximadamente 58%. Porém, o valor da métrica de *Loss* foi de 69%.

Para o *clustering*, tal como já referido, foi utilizado um modelo pré-treinado, VGG16. Devido à sua aplicação num conjunto de mais de 1 milhão de imagens e 1000 classes, existe vantagem na aplicação a conjuntos de dados mais pequenos. Ou seja, torna-se mais fácil extrair as características das imagens e foi possível a aglomeração dos objetos em 4 *clusters* com características semelhantes. Apesar de várias tentativas de reformulação do modelo, as classes continuam a ser iguais ao número de imagens a aglomerar. Assim sendo, cada imagem ficou

associada a uma classe própria e, por sua vez, cada classe ficou associada a um dos 4 *clusters* existentes.

De forma a serem mais percetíveis os resultados obtidos, segue a Tabela 16. Nesta Tabela 16 é possível verificador o identificador do modelo criado, os packages utilizados em cada modelo, o resultado da métrica ACC e da métrica *Loss* (para os modelos preditivos) e o número de *clusters* (para o modelo descritivo).

Tabela 16 - Tabela de comparação de resultados obtidos nos modelos

Identificador do Modelo	*Packages* aplicados	Descrição	Métrica ACC	Métrica *Loss*	Número de Clusters
1º	*TensorFlow* e *Keras*	*Convolutional Neural Network* com funcionalidades dos packages TensorFlow e *Keras*.	91%	36%	-
2º	*Keras*	Modelo tendo por base funcionalidades do *package Keras*.	58%	69%	-
3º	*K-Means*	Modelo criado com base num pré-treinado (VGG16) e no método *k-means*.	-	-	4 clusters 106 classes

De acordo com os objetivos de Data Mining definidos no ponto 4.2.1 Compreensão do Negócio, apenas um modelo preditivo conseguiu superar a ACC de 75%. Nenhum modelo preditivo conseguiu possuir valores de Loss abaixo dos 35%. No caso do modelo descritivo, pode verificar-se a aglomeração de imagens por 4 *clusters* distintos.

Uma das grandes dificuldades na elaboração do projeto foi a perceção de todo o processo de classificação e *clustering* de imagens. Os resultados obtidos transpõem a falta de experiência na execução de modelos com objetos do tipo imagem. Como tal, existe a necessidade de garantir a viabilidade do processo por um especialista. O principal do projeto era garantir a criação de um modelo passível de ser testado. Os resultados são alvo de análise de uma segunda fase externa a este projeto de dissertação.

6. CONCLUSÕES

A procura de informação, num conjunto de dados, para aquisição do devido conhecimento, tem sido um processo muito requisitado. Isto porque existem decisões que, se não forem ponderadas e suportadas, poderão não ser bem-sucedidas. Neste projeto o foco foi a criação de modelos que permitissem a exploração e a possibilidade de testar, em contexto real, as componentes criadas.

6.1. Síntese do Trabalho Efetuado

A elaboração do projeto foi dividida em duas fases: uma fase teórica e uma fase de teor prático. Na fase teórica foram dedicados meses a leituras para que fossem compreendidos os conceitos base relacionados com o tema. Já a fase prática, para além da continuidade de leituras sobre métodos e processos específicos, foi caracterizada pela exploração de técnicas de DM a aplicar em objetos do tipo imagem.

Na primeira fase foram identificados alguns modelos que, na parte prática, não conseguiram ser aplicados. Alguns por incompatibilidades, ou meramente por falta de técnica e experiência a trabalhar com imagens. As redes neuronais sempre foram a técnica mais lida e que aparecia como sendo a que proporcionava melhores resultados.

Aquando do início da elaboração e execução de modelos, foram analisados exemplos desenvolvidos por indivíduos mais ou menos experientes. Isto porque a imagem, ao contrário de dados numéricos, possui uma estrutura bem diferente. Essa estrutura não pode ser processada por modelos na sua forma original. Há todo um processo de transformação que era desconhecido e continua a mostrar-se ser um grande desafio.

Num período mais recente foram obtidos resultados com os *packages TensorFlow* e *Keras*, para uma abordagem classificativa do problema, e o método *k-means* para uma abordagem descritiva. Apesar dos resultados não serem os expectáveis, a parte prática é focada na exploração das imagens e das técnicas mencionadas.

Assim sendo, o primeiro modelo criado e apresentado obteve uma *Accuracy* (ACC) de 91% e *Loss* de 36%. O segundo modelo obteve uma ACC de 58% e *Loss* de 69%. Relativamente ao modelo descritivo, pode verificar-se a aglomeração das 106 imagens/classes pelos 4 *clusters* distintos.

De acordo com a metodologia CRISP-DM, o projeto seguiu, na fase prática, todas as suas etapas, à exceção da fase de implementação. A fase de implementação da metodologia corresponde à caracterização da integração das componentes desenvolvidas numa plataforma, por exemplo. No caso do projeto, o foco – que já foi mencionado ao longo do relatório – foi a exploração de modelos passíveis de serem testados em contextos reais. Numa fase alheia a esta dissertação, o objetivo passará por analisar os resultados obtidos e integrar as componentes desenvolvidas.

Tendo por base o processo de *Data Mining* (DM) é possível interligar o trabalho desenvolvido com a fase final do processo, a aquisição de conhecimento. O processo inicia com a aquisição dos dados que, no caso do projeto, foram fornecidos pelo orientador. Posteriormente existe uma etapa de seleção e transformação dos dados a explorar. Esta etapa corresponde ao início da parte prática do projeto, bem como às três primeiras fases da metodologia CRISP-DM: Compreensão do Negócio, Compreensão dos dados e Preparação dos dados. A fase de DM para extração de informação relevante corresponde à fase de Modelação da parte prática do projeto. Seguidamente os modelos são avaliados e caracterizados de acordo com o seu desempenho. Por fim, e de acordo com a evolução do trabalho efetuado, o conhecimento é adquirido. Os resultados do presente projeto de dissertação, para além da investigação efetuada, serão uma mais valia para uma posterior análise e exploração do tema.

De acordo com a elaboração do presente projeto de dissertação é então possível dar resposta a questão de investigação referida no ponto 1.2 Objetivos deste relatório: De que forma os modelos de *Data Mining* suportam a interpretação de imagens? Assim sendo, os modelos de DM já permitem uma exploração de imagens a um nível bastante satisfatório. Apear de não ser possível efetuar um processamento de imagem numa arquitetura de modelo, não é algo que impossibilite a exploração dos dados. A imagem é analisada, processada e segmentada para que seja possível ser compreendida pelos modelos a desenvolver. Ao acrescentar determinadas funcionalidades da linguagem *Python* e da ferramenta *Anaconda Navigator*, a ferramenta R ficou apta para um ambiente de desenvolvimento de *Convolutional Neural Networks* (CNN). Os resultados deste trabalho são uma boa base para a continuação de exploração do processo de interpretação de imagens com base em modelos de DM.

De acordo com a metodologia *Design Science Research Methodology* (DSRM) mencionada no ponto 1.3 Metodologia de Investigação, o artefacto elaborado insere-se no momento de

Exaptação. Isto porque, utilizando modelos de *Data Mining*, foi possível classificar e aglomerar imagens de melanomas, com base nas suas características. Desta forma é verificada uma oportunidade de investigação e contribuição de conhecimento.

6.2. Riscos Verificados e Limitações

Tal como todos os projetos, o presente projeto de dissertação esteve sujeito à ocorrência de riscos. De forma a ter conhecimento dos riscos associados a esta dissertação foi elaborada a Tabela 17. Esta tabela apresenta o identificador do risco, o risco e a ação aplicada para atenuar o mesmo.

Tabela 17 - Tabela de riscos

ID	Risco	Ação, ou ações, atenuante
1	Má compreensão do negócio	Voltou a analisar-se a questão do projeto, bem como os objetivos a atingir.
2	Complexidade do projeto	Houve uma procura e leitura adequadas, para que fosse possível alinhar os conceitos e facilitar a perceção do tema.
3	Falta de tempo	Compensar com desenvolvimentos mais sucintos e não tão complexos.
4	Má execução do planeamento do projeto	Remodelação de etapas a efetuar e das dependências entre cada uma delas.

Entende-se por limitação um obstáculo que surge e por vezes impede a execução de tarefas, por exemplo. Outras vezes a boa execução de uma tarefa em si. Não é algo que esteja sob controlo.

Desta forma, a principal limitação para a realização do projeto foi a falta de conhecimento do tema. Apesar do conceito de *Data Mining* não ser desconhecido, a aplicação de técnicas a imagens é um problema completamente novo. As imagens têm de ser tratadas para que os

modelos as consigam processar. Outra das limitações foi o facto do autor se encontrar a trabalhar e ter de conciliar horários para desenvolver a componente prática do projeto.

A máquina onde foram explorados os modelos também não ajudou. Com processador i7-4700MQ CPU @ 2.40GHz e sistema operativo de 64 bits, o problema foi a memória de apenas 4GB. Não era possível armazenar o carregamento das imagens do dermoscópio.

6.3. Trabalho Futuro

Como referido ao longo deste projeto, as imagens são um objeto de estudo muito particular. Para além dos avanços notórios na área do *Image Mining*, ainda é difícil superar certos obstáculos relacionados com a natureza do objeto.

Os modelos explorados são exemplos de técnicas que podem ser aplicadas a conjuntos de imagens. Mas há sempre espaço para mais exploração, mais modelação e novas formas de aplicar o conhecimento.

O trabalho efetuado vai de encontro às expectativas e objetivos definidos inicialmente. Como tal, uma meta de trabalho futuro seria otimizar o desempenho dos modelos elaborados. Depois explorar e aplicar uma maior quantidade de mais modelos, tanto preditivos, como descritivos. Posteriormente seria aumentar a quantidade de dados, de forma a elevar a variabilidade de formatos de lesões, tonalidades, entre outras características. Outro objetivo passaria por melhorar o desenvolvimento da componente, para que se esta se encontre de acordo com os requisitos da ferramenta *Pervasive Data Mining Engine* (PDME). A sua integração e a obtenção de resultados, através da ferramenta, é um passo externo à presente dissertação. Como tal, os seguintes passos dizem respeito ao projeto *Deux ex Machina* (DEM).

Para concluir, com base no desenvolvimento do projeto de dissertação, está a ser elaborado um artigo científico para publicação.

REFERÊNCIAS BIBLIOGRÁFICAS

Aggarwal, C. C., & Reddy, C. K. (2013). *DATA Custering Algorithms and Applications*.

Alickovic, E., Kevric, J., & Subasi, A. (2018). Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical Signal Processing and Control*, *39*, 94–102. https://doi.org/10.1016/j.bspc.2017.07.022

Balu, R., & Devi, T. (2012). Design and Development of Automatic Appendicitis Detection System using Sonographic Image Mining, *1*(3), 67–74.

Barata, A. C. F., & Celebi, S. D. J. dos S. S. M. C. D. M. E. (2017). Automatic Detection of Melanomas Using Dermoscopy Images. Retrieved from http://vislab.isr.ist.utl.pt/wp-content/uploads/2017/05/cbarata_phd_thesis.pdf

Berlage, T. (2005). Analyzing and mining image databases. *Drug Discovery Today*. https://doi.org/10.1016/S1359-6446(05)03462-8

Burger, W., & Burge, M. J. (2016). *Digital Image Processing*. *Electronics and Power* (Vol. 24). https://doi.org/10.1007/978-1-4471-6684-9

Carneiro, B. N. (2017). Clinical Intelligence - Definição de Processos de ETL e DW.

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

Chollet, F., & Allaire, J. J. (2018). *Deep learning with R*. Retrieved from https://www.manning.com/books/deep-learning-with-r

Dey, N., Karâa, W. B. A., Chakraborty, S., Banerjee, S., Salem, M. A. M., & Azar, A. T. (2015). Image mining framework and techniques: a review. *International Journal of Image Mining*, *1*(1), 45. https://doi.org/10.1504/IJIM.2015.070028

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. https://doi.org/unm Centennial Lower Level 2Q387 .D66 2015

Dresch, A., Lacerda, D. P., & Antunes, J. A. V. (2015). Design Science Research. *Design Science Research*, 67–102. https://doi.org/10.1007/978-3-319-07374-3_4

Gajjar, T. Y., & Chauhan, N. C. (2012). A Review on Image Mining Frameworks and Techniques. *International Journal of Computer Science and Information Technologies*, *3*(3), 4064–4066.

Garcia-Arroyo, J. L., & Garcia-Zapirain, B. (2017). Recognition of pigment network pattern in dermoscopy images based on fuzzy classification of pixels. *Computer Methods and Programs in Biomedicine*, *153*, 61–69. https://doi.org/10.1016/j.cmpb.2017.10.005

Gonçalves, D., Cruz, J., & Yasmina, M. (2011). Implementação de um Sistema de Business Intelligence para a análise da Doença Pulmonar Obstrutiva Crónica. Retrieved from http://repositorium.sdum.uminho.pt/bitstream/1822/13835/1/Capitulo_Livro_Final_BI_2010.pdf

Gorunescu, F. (2011). *Data mining: concepts and techniques*. *Chemistry &* https://doi.org/10.1007/978-3-642-19721-5

Gregor, S., & Hevner, A. R. (2013). P OSITIONING AND P RESENTING D ESIGN S CIENCE Types of Knowledge in Design Science Research. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.2753/MIS0742-1222240302

Gregor, S., & Jones, D. (2007). The Anatomy of a Design Theory. *Journal of the Association for Information Systems*, *8*(5), 312–335. https://doi.org/Article

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third Edition*. *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). *Principles of data mining*. *Drug safety : an international journal of medical toxicology and drug experience* (Vol. 30). https://doi.org/10.2165/00002018-200730070-00010

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92. https://doi.org/http://aisel.aisnet.org/sjis/vol19/iss2/4

Hsu, W. (2002). Image Mining : Trends and Developments, 1–26.

Kazmierska, J., & Malicki, J. (2008). Application of the Naïve Bayesian Classifier to optimize treatment decisions. *Radiotherapy and Oncology*, *86*(2), 211–216. https://doi.org/10.1016/j.radonc.2007.10.019

Lambin, P., Zindler, J., Vanneste, B. G. L., De Voorde, L. Van, Eekers, D., Compter, I., … Walsh, S. (2017). Decision support systems for personalized and participative radiation oncology. *Advanced Drug Delivery Reviews*, *109*, 131–153. https://doi.org/10.1016/j.addr.2016.01.006

Lima, J. J. P. de. (2008). *Física em Medicina Nuclear, Temas e aplicações*. Coimbra: Imprensa da Universidade de Coimbra. https://doi.org/http://dx.doi.org/10.14195/978-989-26-0387-2

Martínez, J. M., Koenen, R., & Pereira, F. (2002). MPEG-7: The generic multimedia content description standard, Part 1. *IEEE Multimedia*, *9*(2), 78–87.

https://doi.org/10.1109/93.998074

MESSADI, M., CHERIFI, H., & BESSAID, A. (2014). Segmentation and ABCD rule extraction for skin tumors classification. *Journal of Convergence Information Technology*, *9*(March), 21.

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/

Oliveira, E. (2017). INESC TEC desenvolve plataforma para rastreio da retinopatia diabética « Notícias UP. Retrieved February 13, 2018, from https://noticias.up.pt/inesc-tec-desenvolve-plataforma-para-rastreio-da-retinopatia-diabetica/

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee. (2008). Peffers et al. (2008) A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Peixoto, R., Portela, F., & Santos, M. F. (2016). Towards a Pervasive Data Mining Engine—Architecture Overview (pp. 557–566). Springer, Cham. https://doi.org/10.1007/978-3-319-31307-8_58

Polese, G. (2014). A decision support system for Evidence Based Medicine. *Journal of Visual Languages & Computing*, *25*(6), 858–867. https://doi.org/10.1016/j.jvlc.2014.09.013

Quellec, G., Charrière, K., Boudi, Y., Cochener, B., & Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, *39*, 178–193. https://doi.org/10.1016/j.media.2017.04.012

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. https://doi.org/1711.05225

Ribeiro, V. H. da S. (2017). Análise Estatística em R utilizando o Pervasive Data Mining Engine.

Ridge, E. (2015). *Guerrilla Analytics*. *Guerrilla Analytics*. https://doi.org/10.1016/B978-0-12-800218-6.00018-7

S.Mahalle, A., & Kuche, S. H. (2015). Image Mining and Clustering Based Image Segmentation. *International Journal of Advance Research in Computer Science and Management Studies*, *3*(3), 50–54. Retrieved from www.ijarcsms.com

Séneca, H. (2018). Exame Informática | DENSER: o algoritmo criado em Coimbra que superou o Google Brain. Retrieved February 14, 2018, from

http://exameinformatica.sapo.pt/noticias/software/2018-01-22-DENSER-o-algoritmo-criado-em-Coimbra-que-superou-o-Google-Brain

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved from http://arxiv.org/abs/1409.1556

Teixeira, J. W., Annibal, L. P., Felipe, J. C., Ciferri, R. R., & Ciferri, C. D. de A. (2015). A similarity-based data warehousing environment for medical images. *Computers in Biology and Medicine*, *66*, 190–208. https://doi.org/10.1016/j.compbiomed.2015.08.019

Turban, E. (2011). Decision Support and Business Intelligence Systems, 9/E. *Prentice Hall*, 696. https://doi.org/10.1017/CBO9781107415324.004

Vaishnavi, V. K., & William Kuechler, J. (2007). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*.

Valdes, G., Simone, C. B., Chen, J., Lin, A., Yom, S. S., Pattison, A. J., … Solberg, T. D. (2017). Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiotherapy and Oncology*, *125*(3), 392–397. https://doi.org/10.1016/j.radonc.2017.10.014

Van Aken, J. E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. *Journal of Management Studies*, *41*(2), 219–246. https://doi.org/10.1111/j.1467-6486.2004.00430.x

Venables, B., & Smith, D. M. (2017). *An Introduction to R An Introduction to R* (Vol. 3). https://doi.org/10.1016/B978-0-12-381308-4.00001-7

Weir, N., Fayyad, U. M., Djorgovski, S., Roden, J. C., & Rouquette, N. (1992). Skicat - a Cataloging and Analysis Tool for Wide-Field Imaging Surveys. *Astronomical Data Analysis Software and Systems Ii*, *52*, 39–44. Retrieved from https://ntrs.nasa.gov/search.jsp?R=19940018073

Witten, I. H., Eibe, F., & Hall, M. a. (2011). *Data Mining Practical Machine Learning Tools and Techniques*.

Wong, S. T. C., Hoo, K. S., Knowlton, R. C., Laxer, K. D., Cao, X., Hawkins, R. A., … Arenson, R. L. (2002). Design and Applications of a Multimodality Image Data Warehouse Framework Design and Applications of a Multimodality Image Data Warehouse Framework. *Journal of the American Medical Informatics Association*, *9*(3), 239–254. https://doi.org/10.1197/jamia.M0988

Zahradnikova, B., Duchovicova, S., & Schreiber, P. (2015). Image Mining : Review and New Challenges. *International Journal of Advanced Computer Sience and Applications*, *6*(7), 242–

246. https://doi.org/10.14569/IJACSA.2015.060732

Zdonik, S., & Katz, J. (2011). *SpringerBriefs in Computer Science Series Editors*. https://doi.org/10.1007/978-1-4471-2179-4

Zhang, J., Hsu, W., & Lee, M. L. (2001). Image mining: Issues, frameworks and techniques. *Proceedings of the 2nd ACM SIGKDD International Workshop on Multimedia Data Mining (MDM/KDD'01)*, 1–7. https://doi.org/10.1016/S0952-1976(02)00019-2

ANEXO I

Na Tabela 18 é possível visualizar os planos de corte correspondentes às restantes imagens que não foram mencionadas no ponto.4.2.3 Preparação dos dados

Tabela 18 - Plano de cortes, e valor de Otsu, das imagens capturadas pelo dermoscópio adaptável (da imagem 17 à imagem 106)

Imagem	Plano de Corte	Valor de Otsu
17	[1071:2374,408:1660]	0,486328125
18	[1602:3036,532:1515]	0,490234375
19	[1508:3029,707:1944]	0,546171875
20	[1675:2796,903:1704]	0,443359375
21	[2141:3022,881:1828]	0,458984375
22	[1580:2563,729:1580]	0,474609375
23	[1034:1828,787:1529]	0,462890625
24	[2061:2607,998:1668]	0,470703125
25	[1828:2578,954:1442]	0,447265625
26	[1762:2301,1078:1522]	0,376953125
27	[1515:2410,721:1821]	0,498046875
28	[1806:2396,1078:1770]	0,404296875
29	[2068:2745,1122:1784]	0,439453125
30	[1464:2578,517:1573]	0,431640625
31	[1573:2206,1136:1697]	0,345703125
32	[1835:2520,962:1595]	0,384765625
33	[1959:2570,1216:1872]	0,412109375
34	[1602:2090,823:1224]	0,376953125
35	[1959:2745,969:1770]	0,482421875
36	[1464:2905,780:2141]	0,560546875
37	[1056:2869,648:1901]	0,587890625
38	[1100:2323,816:1755]	0,513671875
39	[860:2367,30:1821]	0,521484375
40	[1500:2469,306:2141]	0,564453125
41	[1697:2833,714:1690]	0,470703125
42	[1697:2549,1173:2017]	0,419921875
43	[2170:2891,874:1449]	0,423828125
44	[1813:2738,510:1857]	0,490234375

Imagem	Plano de Corte	Valor de Otsu
45	[1296:2570,721:1551]	0,552734375
46	[1777:2520,823:1748]	0,419921875
47	[1427:2083,729:1449]	0,392578125
48	[1551:2396,685:1427]	0,427734375
49	[1260:2862,568:1675]	0,458984375
50	[1311:2177,605:1631]	0,443359375
51	[1784:2898,678:1624]	0,392578125
52	[1347:2694,714:1595]	0,427734375
53	[1409:2118,1399:2069]	0,408203125
54	[2089:2924,1176:2050]	0,443359375
55	[1953:2885,1312:2351]	0,419921875
56	[1564:2749,1137:2429]	0,529296875
57	[1263:2982,496:3060]	0,595703125
58	[1487:2526,1108:1788]	0,529296875
59	[1574:2293,1137:1817]	0,404296875
60	[1380:2429,1283:2205]	0,458984375
61	[1739:2584,1380:2244]	0,498046875
62	[1671:2710,1205:2215]	0,482421875
63	[1496:2468,1399:2448]	0,439453125
64	[1661:2769,1185:2536]	0,501953125
65	[1613:2788,1108:1788]	0,509765625
66	[1613:2400,982:1768]	0,412109375
67	[892:1220,449:940]	0,494140625
68	[1797:2963,1321:2710]	0,517578125
69	[1729:2827,1011:1924]	0,564453125
70	[2011:2944,1185:2273]	0,513671875
71	[1846:2448,807:1710]	0,419921875
72	[1856:2613,1516:2234]	0,474609375
73	[1555:3973,1040:2128]	0,547890625
74	[1885:3449,1156:2011]	0,494140625
75	[1661:2866,982:1477]	0,439453125
76	[1807:2905,1030:2128]	0,580078125
77	[1613:2429,1710:2584]	0,388671875

Imagem	Plano de Corte	Valor de Otsu
78	[1759:2973,933:1914]	0,447265625
79	[1778:2497,1079:1895]	0,404296875
80	[1885:2526,1011:1661]	0,408203125
81	[1535:2798,1195:2361]	0,533203125
82	[2108:2837,1545:2176]	0,482421875
83	[2215:2778,1545:2176]	0,451171875
84	[1137:2866,1195:1992]	0,439453125
85	[1496:2885,340:1564]	0,470703125
86	[1632:2351,1176:1875]	0,513671875
87	[1477:2458,991:1652]	0,419921875
88	[1691:2846,982:2021]	0,494140625
89	[1768:2769,1011:2215]	0,509765625
90	[1885:2788,991:2069]	0,474609375
91	[1564:2720,1127:2089]	0,470703125
92	[836:3526,1253:2098]	0,501953125
93	[1710:2720,1312:2098]	0,408203125
94	[1535:2497,1156:2205]	0,490234375
95	[1185:2914,1011:2662]	0,544921875
96	[1671:2574,1166:2060]	0,451171875
97	[1292:2536,739:2536]	0,494140625
98	[1584:2409,991:2293]	0,412109375
99	[1797:2973,1079:1982]	0,478515625
100	[1428:2361,807:2409]	0,498046875
101	[1312:3118,1166:2108]	0,431640625
102	[1759:2623,1331:2040]	0,439453125
103	[1710:2963,739:2001]	0,478515625
104	[1351:2953,1049:2215]	0,486528125
105	[1729:2458,1409:2189]	0,427734375
106	[1623:2448,632:2283]	0,517578125

ANEXO II

Neste anexo é possível verificar a continuidade do mapeamento, entre clusters, resultante do ponto 4.2.5 Avaliação. Desta forma segue a seguinte Tabela 19.

Tabela 19 - Mapeamento dos clusters com as imagens e suas classes (da imagem 17 à imagem 106)

Identificador	*Cluster_pca*	*Cluster_feature*	Imagem	Classe
17	1	2	17.jpg	17.
18	2	1	18.jpg	18.
19	3	3	19.jpg	19.
20	3	3	20.jpg	20.
21	2	1	21.jpg	21.
22	2	1	22.jpg	22.
23	1	2	23.jpg	23.
24	1	2	24.jpg	24.
25	2	1	25.jpg	25.
26	2	1	26.jpg	26.
27	2	1	27.jpg	27.
28	4	4	28.jpg	28.
29	1	2	29.jpg	29.
30	4	4	30.jpg	30.
31	4	4	31.jpg	31.
32	4	4	32.jpg	32.
33	4	4	33.jpg	33.
34	4	4	34.jpg	34.
35	2	1	35.jpg	35.
36	2	1	36.jpg	36.
37	2	1	37.jpg	37.
38	2	1	38.jpg	38.
39	2	1	39.jpg	39.
40	2	1	40.jpg	40.
41	2	1	41.jpg	41.
42	2	1	42.jpg	42.
43	2	1	43.jpg	43.
44	2	1	44.jpg	44.

Identificador	*Cluster_pca*	*Cluster_feature*	Imagem	Classe
45	1	2	45.jpg	45.
46	2	1	46.jpg	46.
47	3	3	47.jpg	47.
48	2	1	48.jpg	48.
49	4	4	49.jpg	49.
50	2	1	50.jpg	50.
51	3	3	51.jpg	51.
52	2	1	52.jpg	52.
53	4	4	53.jpg	53.
54	2	1	54.jpg	54.
55	2	1	55.jpg	55.
56	2	1	56.jpg	56.
57	2	1	57.jpg	57.
58	1	2	58.jpg	58.
59	4	4	59.jpg	59.
60	1	2	60.jpg	60.
61	2	2	61.jpg	61.
62	2	2	62.jpg	62.
63	1	2	63.jpg	63.
64	1	2	64.jpg	64.
65	4	4	65.jpg	65.
66	3	3	66.jpg	66.
67	4	4	67.jpg	67.
68	3	3	68.jpg	68.
69	2	1	69.jpg	69.
70	3	3	70.jpg	70.
71	2	1	71.jpg	71.
72	2	1	72.jpg	72.
73	2	1	73.jpg	73.
74	2	1	74.jpg	74.
75	2	1	75.jpg	75.
76	1	2	76.jpg	76.
77	2	1	77.jpg	77.

Identificador	*Cluster_pca*	*Cluster_feature*	Imagem	Classe
78	2	1	78.jpg	78.
79	2	1	79.jpg	79.
80	2	1	80.jpg	80.
81	1	2	81.jpg	81.
82	1	2	82.jpg	82.
83	1	2	83.jpg	83.
84	3	3	84.jpg	84.
85	2	1	85.jpg	85.
86	1	2	86.jpg	86.
87	2	1	87.jpg	87.
88	2	1	88.jpg	88.
89	2	1	89.jpg	89.
90	2	1	90.jpg	90.
91	2	1	91.jpg	91.
92	2	1	92.jpg	92.
93	4	4	93.jpg	93.
94	2	1	94.jpg	94.
95	2	1	95.jpg	95.
96	2	1	96.jpg	96.
97	2	1	97.jpg	97.
98	4	4	98.jpg	98.
99	3	3	99.jpg	99.
100	3	3	100.jpg	100
101	1	2	101.jpg	101
102	1	2	102.jpg	102
103	1	2	103.jpg	103
104	1	2	104.jpg	104
105	1	2	105.jpg	105
106	1	2	106.jpg	106