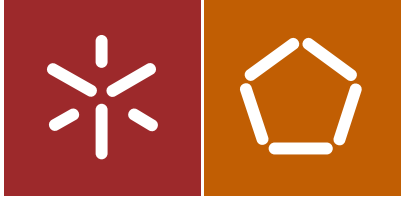
Universidade do Minho
Escola de Engenharia

Mário Sérgio Azevedo Ferreira

Sistema de Apoio à Decisão Clínica – Quality of Life System

outubro de 2015

Mário Sérgio Azevedo Ferreira Sistema de Apoio à Decisão Clínica – Quality of Life System
UMinho | 2015



Universidade do Minho
Escola de Engenharia

Mário Sérgio Azevedo Ferreira

Sistema de Apoio à Decisão Clínica – Quality of Life System

Dissertação de Mestrado
Ciclo de Estudos Integrados Conducentes ao Grau de
Mestre em Engenharia e Gestão de Sistemas de Informação

Trabalho efectuado sob a orientação de
Professor Doutor Luís Paulo Reis

e coorientação da
Professora Doutora Brígida Mónica Faria

outubro de 2015

DECLARAÇÃO

Nome: Mário Sérgio Azevedo Ferreira

Endereço eletrónico: a58754@alunos.uminho.pt **Telefone**: 914749782

Bilhete de Identidade/Cartão do Cidadão: 13716401

Título da dissertação: Sistema de Apoio à Decisão Clínica – Quality of Life System

Orientadores:

Professor Doutor Luís Paulo Reis

Professora Doutora Brígida Mónica Faria

Ano de conclusão: 2015

Mestrado Integrado em Engenharia e Gestão de Sistemas

Universidade do Minho, _____/_____/_____

Assinatura:

AGRADECIMENTOS

Ao Professor Doutor Luís Paulo Reis, meu orientador de mestrado, pela permanente disponibilidade e constante apoio durante todo o processo de desenvolvimento do presente projeto de dissertação, assim como pelas oportunidades de divulgação científica que proporcionou.

À Professora Doutora Brígida Mónica Faria, minha coorientadora de mestrado, pelos conselhos e comentários que contribuíram decisivamente para que fosse possível atingir os objetivos do projeto de dissertação.

A todos os elementos da Optimizer, mais concretamente ao Doutor Rui Lopes e ao Doutor Victor Carvalho, pela oportunidade de integrar um projeto tão importante como o QoLIS, e pelo apoio, disponibilidade e simpatia que demonstraram.

Aos elementos do LIACC que estiveram inseridos no projeto de dissertação, nomeadamente ao Doutor Joaquim Gonçalves, pela paciência nas explicações de todos os detalhes inerentes ao QoLIS, pelas opiniões e conselhos que contribuíram decisivamente para atingir os objetivos do projeto de dissertação.

Aos meus amigos pelo incentivo e pelos momentos de boa disposição que me ajudaram a ultrapassar momentos menos bons.

À Sara por acreditar sempre nas minhas capacidades, mas principalmente pelo amor, carinho e apoio incondicional ao longo desta e de outras etapas da minha vida.

Aos meus pais pelos valores e princípios que me transmitiram, e por me proporcionarem todas as condições para estar onde estou hoje.

RESUMO

A crescente relevância do conceito Qualidade de Vida no âmbito da Saúde, e a necessidade de aproveitar os benefícios da inovação das tecnologias e sistemas de informação são a sustentação para o desenvolvimento de um Sistema de Apoio à Decisão Clínica, que o presente projeto de dissertação pretende apoiar. Através de dados que retratam interações entre pacientes oncológicos e a instituição de saúde responsável pelo seu acompanhamento e tratamento, foi desenvolvido um processo de descoberta de conhecimento em bases de dados. Esse processo, desenvolvido de acordo com os princípios da metodologia CRISP-DM, identificou um conjunto de técnicas e modelos de Data Mining com capacidade para explorar, extrair e evidenciar padrões nos dados submetidos a tarefas de Clustering e Previsão. Os resultados obtidos através dos diferentes testes executados destacam modelos inerentes às Árvores de Decisão, Regras de Associação, Redes Neuronais, Vizinhos Mais Próximos e Classificadores de Bayes, avaliando os mesmos pela Sensibilidade, Especificidade, Erro quadrático médio, Erro médio e Tempo de aprendizagem apresentados. Nos testes realizados, através dos modelos de previsão que incidiram sobre as tarefas de Clustering, foram obtidas percentagens de acerto bastante elevadas, na ordem dos 80% aos 90%, enquanto nos testes de previsão de Qualidade de Vida as percentagens de acerto superaram, na grande maioria dos testes executados, o mínimo de percentagem estipulada de 70%.

O presente projeto de dissertação contribui assim para aumentar o número de estudos que fazem a aplicação prática do uso de Data Mining e Sistemas de Apoio à Decisão Clínica, procurando potenciar o processo de tomada de decisão por parte das equipas médicas especializadas, na procura de melhorar os tratamentos e consequentemente a qualidade de vida relacionada com a Saúde de doentes crónicos.

Palavras-Chave: Qualidade de Vida, Descoberta de Conhecimento em Bases de Dados, Data Mining, Clustering, Previsão

ABSTRACT

This dissertation project intends to support the development of a Clinical Decision Support System, motivated by the growing relevance of Quality of Life concept and the need to explore the benefits of innovation in technology and information systems, in health environments. Through a dataset which represents the different interactions between oncology patients and the health institution responsible for their monitoring and treatment, it has been developed a process of knowledge discovery in databases. This process, which followed the principles of CRISP-DM methodology, identified several Data Mining techniques with capabilities to explore, extract and highlight patterns in the dataset submitted to clustering and prevision tasks. The results obtained, through the different executed tests, highlight models as Decision Trees, Association Rules, Neuronal Networks, Nearest Neighbors and Bayes Classifiers. The evaluation of each model was based in metrics as Sensibility, Specificity, Mean square error, Mean error and learning time of the models. In the tests which were executed through the prevision models, with incidence in Clustering tasks, it were obtained high accuracy percentages between 80% and 90%, while in Quality of Life prevision tests the percentages exceeded the stipulated minimum percentage of 70%, in the most of the tests performed. This dissertation project contributes to increase the number of studies which makes the practical application of Data Mining and Clinical Decision Support Systems, demonstrating capabilities to enhance the decision-making process by the specialized medical teams, in order to improve treatments and consequently the Health related Quality of Life in persons with chronic diseases.

KEYWORDS: QUALITY OF LIFE, KNOWLEDGE DISCOVERY IN DATABASES, DATA MINING, CLUSTERING, PREVISION;

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

QdV – Qualidade de Vida

QdVRS – Qualidade de Vida Relacionada com a Saúde

DCBD – Descoberta de Conhecimento em Bases de Dados

CRISP-DM - Cross Industry Standard Process for Data Mining

EORTC – European Organization of Research and Treatment of Cancer

QoLIS – Quality of Life Information System

SADC – Sistema de Apoio à Decisão Clínica

DM – Data Mining

1. INTRODUÇÃO

O presente capítulo apresenta a motivação para a realização deste projeto de dissertação no âmbito da combinação das áreas de qualidade de vida e de Data Mining, intrinsecamente relacionadas com o Sistema de Apoio à Decisão Clínica (SADC) Quality of Life Information System (QoLIS). Em seguida são apresentados os objetivos a atingir, de acordo com a componente organizacional e académica inerentes ao projeto, sendo ainda apresentada a organização do presente documento.

1.1 Motivação

O sector da saúde acompanhou com especial atenção a crescente inovação tecnológica a nível das Tecnologias e Sistemas de Informação, com capacidade para controlar e armazenar os dados produzidos nas tarefas diárias de todos os seus intervenientes, com o intuito de melhorar a saúde pública e a assistência aos seus pacientes.

O ambiente hospitalar atual gera um grande e complexo volume de dados nas diversas interações com os pacientes, na atribuição dos recursos hospitalares, no diagnóstico de doenças, nos registos eletrónicos de saúde e nos aparelhos médicos, entre outros. Este fator é potenciador da utilização de ferramentas especializadas em armazenamento e capazes de transformar dados em informação que influencie, positivamente, a tomada de decisões na prática clínica. De acordo com essa expetativa e com a característica predominante do grande volume de dados, as técnicas e ferramentas do Processo de DCBD surgem como resposta, mais concretamente, através de Data Mining.

As técnicas de Data Mining têm alcançado resultados impressionantes em diversas indústrias. Como tal, o sector da saúde tem de aproveitar os avanços neste campo emocionante.

A organização Optimizer, caracterizada pela sua visão vanguardista no mercado dos Sistemas de Informação, e através de um projeto inovador, pretende desenvolver um Sistema de Apoio à Decisão Clínica, focado na avaliação da Qualidade de Vida de doentes oncológicos, denominado Quality of Life Information System. A organização pretende, através da avaliação dos índices de Qualidade de Vida adquiridos através de questionários dedicados ao efeito e devidamente inseridos no SADC, assim como através de informação adquirida através de todas as interações com os pacientes, fornecer às equipas médicas especializadas no tratamento de doenças oncológicas, informação fiável e rapidamente acessível, capaz de influenciar

positivamente a tomada de decisão na prescrição de terapias, melhorando assim a Qualidade de Vida individual dos doentes oncológicos. Para atingir os pronúncios enunciados, a organização pretende a utilização de técnicas de Data Mining, uma vez que reconhece a sua capacitação para responder a um projeto com as características do QoLIS. Esse fator é a principal motivação para o desenvolvimento deste estudo.

1.2 Objetivos

A Optimizer reconheceu na Universidade do Minho a oportunidade de desenvolver sinergias benéficas ao seu projeto, através de uma proposta de projeto de dissertação que sustentasse o desenvolvimento do Sistema de Apoio à Decisão Clínica que pretende implementar. Este projeto de dissertação pretende ainda construir contributos científicos relevantes na combinação da área da Saúde, nomeadamente nas intervenções com doentes oncológicos e através da medida dos seus índices de Qualidade de Vida, com a área de Descoberta de Conhecimento em Bases de Dados, mais concretamente, através da utilização de técnicas de Data Mining. Afirma-se a necessidade de classificar o presente projeto de dissertação, como um projeto de Data Mining, que de acordo com a ficha técnica referente ao QoLIS, (disponível em http://optimizer.pt/client/files/0000000001/1769.pdf), e de acordo com a componente académica, perfilam os seguintes objetivos:

- Executar um estudo exaustivo nas temáticas de QdVRS e Data Mining, com o objetivo de dominar os conceitos e conhecer as práticas que lhe são inerentes.
- Reunir um conjunto significativo de dados com conteúdo relacionado com QdVRS.
- Desenvolver um estudo exaustivo das técnicas de Data Mining, que garantam uma melhor performance do SADC.
- De acordo com os resultados do estudo realizado, aplicar as técnicas de Data Mining mais adequadas ao tipo de dados existentes.
- Realizar um conjunto de testes aos diferentes dados, utilizando as técnicas identificadas e analisar detalhadamente os resultados obtidos.
- Desenvolver uma integração da solução com *software* capaz de cumprir todos os requisitos do SADC.
- Analisar o impacto e a fiabilidade da solução desenvolvida, assim como a relevância dos resultados alcançados.

- Desenvolver conteúdo científico relevante.

1.3 Organização do documento

O presente documento está organizado pela seguinte estrutura de capítulos:

- **Introdução** – este capítulo pretende através de uma breve introdução às temáticas que compõem o projeto de dissertação, descrever alguns fatores que motivaram o desenvolvimento do estudo, assim como os objetivos de realização do mesmo, quer por parte do autor, quer por parte da organização Optimizer.
- **Abordagem Metodológica** – neste capítulo são explicitadas as metodologias utilizadas durante a componente prática do projeto de dissertação, bem como a estratégia de pesquisa bibliográfica desenvolvida para a realização dos dois capítulos de revisão de literatura.
- **Qualidade de Vida** – este capítulo pretende, através de uma revisão de literatura, elucidar as definições relacionadas com a temática Qualidade de Vida e Qualidade de Vida Relacionada com a Saúde, tal como escrutinar os instrumentos de medida que estão implementados no Sistema de Apoio à Decisão Clínica QoLIS.
- **Descoberta de Conhecimento em Bases de Dados** – este capítulo visa, através de um revisão de literatura, enquadrar as técnicas de Data Mining, explicitando o processo requerido para proceder à extração de conhecimento sobre determinados conjuntos de dados, assim como as medidas para perceber o seu resultado e contributo. Estão ainda destacadas as tarefas e áreas que utilizam as técnicas de Data Mining na área Médica, através de uma revisão sistemática de artigos.
- **Sistema de Apoio á Decisão Clínica para a Qualidade de Vida –** neste capítulo está refletido o desenvolvimento da componente prática do projeto de dissertação, através de uma adaptação da metodologia CRISP-DM. São neste capítulo descritos os dados extraídos do QoLIS, os tratamentos com vista à sua melhoria de qualidade e todos os resultados das técnicas e modelos de Data Mining
- **Conclusões** – neste capítulo é executada uma reflexão com base nas atividades e tarefas desenvolvidas durante todo o projeto de dissertação.

2. ABORDAGEM METODOLÓGICA

O presente capítulo apresenta a metodologia adotada para o desenvolvimento da componente prática do projeto de dissertação, bem como, a estratégia de pesquisa bibliográfica seguida no desenvolvimento dos capítulos de revisão de literatura. Adicionalmente são ainda abordadas algumas questões éticas inerentes ao presente projeto de dissertação.

2.1 Adoção de Metodologias

É perentório afirmar que o projeto em questão necessita de adotar uma metodologia de suporte a projetos de Data Mining, uma vez que é nesse sentido que o presente projeto de dissertação suportará o desenvolvimento do Sistema de Apoio à Decisão Clínica pretendido pela organização Optimizer. Como tal, a escolha recaiu pela metodologia CRISP-DM (Cross Industry Standard Process for Data Mining), apontada por diversos autores como sendo capaz de resolver qualquer tipo de projeto de Data Mining (Marbán et al., 2009). Segundo (Santos & Ramos, 2006) CRISP-DM trata-se de um modelo de referência que define as fases a seguir, as tarefas a executar e os resultados esperados pela realização de cada uma das mesmas. Foi desenvolvida por um grupo de empresas de referência (Teradata, SPSS, Daimler, Chrysler e OHRA), sendo caracterizada pela sua independência, face a funções industriais em que pode ser adotada, à ferramenta utilizada e ainda pela similaridade aos processos de descoberta de conhecimentos em Bases de Dados (Brandão et al., 2014;Marbán et al., 2009).

De acordo com a diversa literatura consultada, a metodologia CRISP-DM é descrita em seis fases:

1. **Compreensão do negócio** – Fase de compreensão dos objetivos do projeto e requisitos na perspetiva do negócio. Este levantamento conduzirá à conversão dos objetivos de negócio em objetivos do Data Mining, assim como o traçar de um plano para os atingir.
2. **Estudo dos Dados** – Nesta fase, através de diversas tarefas de exploração sobre um conjunto de dados iniciais, pretende-se fazer uma compreensão dos dados, identificação de problemas, a identificação de subconjuntos relevantes para posterior análise e identificação de conhecimento implícito.
3. **Tratamento dos Dados** – Fase de obtenção de um conjunto de dados para análise, através de algoritmos de Data Mining. Procede-se nesta fase a tarefas de preparação de

dados que incluem seleção de tabelas, atributos e registos, assim como a transformação e limpeza de dados com vista a posterior análise.

4. **Modelação** – Nesta fase são selecionadas as diversas técnicas de modelação a aplicar aos dados. As técnicas podem sofrer ajustes nos seus parâmetros no sentido de melhorar os resultados obtidos. Há ainda a preocupação com o facto de algumas técnicas não poderem ser aplicadas a determinados tipos de dados, o que pode implicar retroceder à fase de preparação de dados e aplicar transformações aos dados.
5. **Avaliação** – Fase de avaliação dos modelos obtidos na fase anterior e que apresentam qualidade de acordo com as medidas estabelecidas. A preocupação principal nesta fase é identificar se os modelos satisfazem, ou não, os objetivos de negócio, com o intuito de os utilizar, caso a satisfação se verifique.
6. **Desenvolvimento** – Geralmente, a identificação dos modelos não marca o fim do projeto de Data Mining. Sendo o propósito do modelo a obtenção de conhecimento através dos dados, esse conhecimento necessita de ser organizado e apresentado ao seu cliente, de maneira a que possa ser utilizado.

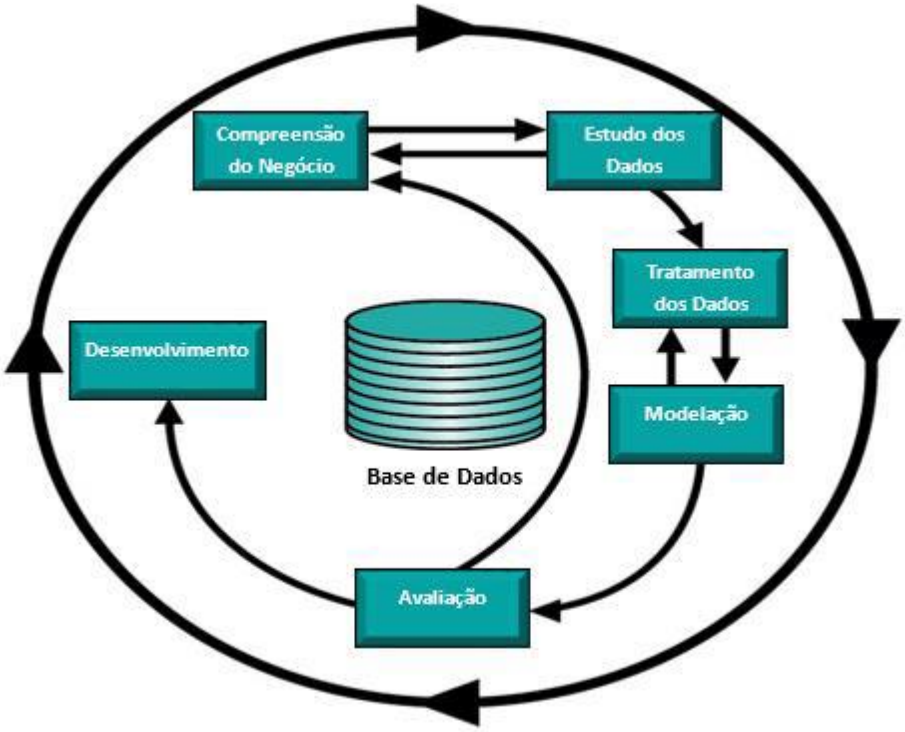


Figura 1- Metodologia CRISP-DM (adaptado de Chapman et., al, 2000)

As seis fases aqui descritas multiplicam-se em várias tarefas e atividades mais específicas, que na componente prática do presente projeto de dissertação foram devidamente adaptadas aos objetivos do mesmo.

2.2 Estratégia de Pesquisa Bibliográfica

A estratégia de pesquisa bibliográfica assentou sobre os conceitos de Qualidade de Vida e Data Mining. O relacionamento com os conceitos foi construído com base em estudos científicos, sugeridos pelo orientador e pelo percurso académico do autor. Esse relacionamento revelou-se útil para a definição de termos-chave, que viriam a influenciar a pesquisa bibliográfica positivamente. A pesquisa bibliográfica desenvolveu-se em inglês e português, através dos termos: *Quality of Life, Health-Related Quality of Life, Quality of Life Instruments and Measures, Cancer, Clinical Practice, Medicine, Knowledge Discover Database, Data Mining, Decision Support Systems, Clinical Decision Support Systems*. Os resultados da pesquisa bibliográfica foram alcançados através de cruzamentos entre os termos enunciados e utilizando portais de conteúdo científico, *Science Direct*, *SciELO, IEE Xplore*, *CiteSeerX, Wiley Online Library*, *Springer, Repositorium, Google Scholar*. É de frisar a importância do protocolo que a Universidade do Minho tem com a maioria das instituições referidas, facilitando o acesso a determinados documentos. O resultado da pesquisa bibliográfica não foi na sua totalidade aproveitado, uma vez que alguns dos termos utilizados são transversais a diversas áreas, resultando num número de documentos que poderiam não ir de encontro ao que se pretendia. A necessidade de execução de uma leitura ao resumo dos diferentes documentos reunidos, permitiu identificar quais os documentos com conteúdo que justificariam uma leitura integral. Esse processo naturalmente fomentou um maior conhecimento sobre autores relevantes nas temáticas em questão, levando à pesquisa individual dos mesmos. Foram ainda tidos em conta o ano e reputação dos artigos científicos. O resultado final deste processo de pesquisa bibliográfica pode ser consultado na secção de Bibliografia.

2.3 Questões Éticas

As principais preocupações éticas no decorrer deste projeto relacionam-se com a confidencialidade e proteção dos dados obtidos, não se afigurando qualquer distribuição ou utilização dos mesmos, que não os previstos originalmente junto da organização Optimizer. Salienta-se que o projeto em questão não desenvolverá qualquer tipo de atividade que ponha em risco a componente física dos pacientes.

3. QUALIDADE DE VIDA

O presente capítulo retrata a exploração de diversos contributos científicos, com o intuito de elucidar as definições associadas aos conceitos de Qualidade de Vida e Qualidade de Vida Relacionada com a Saúde, assim como, escrutinar os instrumentos de medida de QdVRS que integram o Sistema de Apoio à Decisão Clínica QoLIS que o presente projeto de dissertação pretende sustentar.

3.1 Perspetiva Histórica e Conceptual de Qualidade de Vida

Apesar da sua popularização e da exploração entusiástica por investigadores, clínicos, economistas, administradores e políticos (Campos & Neto, 2008; Guyatt et al., 1993; Farquhar, 1995), não existe uma conceptualização universal para o termo Qualidade de Vida. Esse facto surge como tópico de discussão em diversos estudos (Lin et al., 2013; Heutte et al., 2014) (Farquhar, 1995; Pais-Ribeiro, 2004; Guyatt et al., 1993, Gonçalves, 2012), considerando ainda o termo como uma combinação de perspetivas filosóficas, valores e princípios individuais, que segundo (Pais-Ribeiro, 2004) poderá influenciar a técnica de avaliação adotada. Em 1994, a Organização Mundial de Saúde definiu Qualidade de Vida como "a perceção que um indivíduo tem do seu lugar na vida, no contexto da cultura e do sistema de valores nos quais vive, em relação com os seus desejos, as suas normas e as suas inquietudes. É um conceito muito amplo, que pode ser influenciado de maneira complexa pela saúde física do indivíduo, pelo estado psicológico e pelo seu nível de independência, as suas relações sociais e as suas relações com os elementos essenciais do seu meio" (Group WHOQOL, 1994). (Farquhar, 1995) elucida algumas das contribuições relevantes para o estudo e evolução da temática, apontado a opinião de Hanestad em que, "a maioria das pessoas concordarão que Qualidade de Vida é um objetivo para um ou um conjunto de indivíduos" conotando essa afirmação com um pensamento apenas positivo relativo à medida da QdV. Frisa ainda algumas das primeiras preocupações na medida da QdV, como o facto de a Comissão do Presidente Eisenhower através do Relatório dos Objetivos Nacionais, em 1960, abordar a educação, a preocupação individual, o crescimento económico, a saúde na ótica do bem-estar, e a defesa de um mundo livre, bem como a preocupação retratada no discurso do Presidente Americano Lyndon Johnston em 1964, afirmando que " o progresso social não pode ser medido através do balanço dos bancos mas, através da Qualidade de Vida proporcionada às pessoas". Identifica ainda, na época de 1970,

um crescente uso do termo na investigação social, como, o crescente uso do termo QdV em experiências de intervenção clínica, principalmente no campo da Oncologia, Reumatologia e Psiquiatria.

3.2 Conceptualização de Qualidade de Vida Relacionada com a Saúde

Apesar de os termos Qualidade de vida e Qualidade de Vida Relacionada com a Saúde serem usados permutavelmente para abordar a mesma temática, existem pontos discrepantes que os distinguem (Pais-Ribeiro, 2004). Enquanto, genericamente descrito no ponto anterior, Qualidade de Vida é um conceito amplo influenciado por fatores psicológicos, mentais, físicos e sociais individuais, QdVRS foca-se nos efeitos provocados por algum tipo de doença e no impacto do seu tratamento. Mais concretamente, QdVRS reflete a forma como um indivíduo encara e reage ao seu estado de saúde, englobando aspetos não relacionados com o tratamento médico, como fatores físicos, emocionais, mentais e de bem-estar, assim como, desenvolver uma atividade profissional, a família, amigos e outras situações do quotidiano (Lin et al., 2013; Pais-Ribeiro, 2004). No estudo de (Lin et al., 2013) é caracterizado o conceito de QdVRS como sendo baseado no conceito de Saúde e de Qualidade de Vida, sendo influenciado pelas experiências, crenças, expetativas e perceções individuais, enunciando que um bom estado de saúde é mais que uma ausência de doença ou debilidades, mas também uma completa sensação de bem-estar físico, mental e social. QdVRS é também considerado como um conceito multifacetado que contem aspetos positivos e negativos de Saúde. Os aspetos negativos apontados visam doenças e disfunções, enquanto os positivos abordam o sentimento de bem-estar mental e físico, pleno funcionamento, aptidão física, ajuste e eficiência da mente e do corpo (Bowling, 2001). É ainda adicionada por (Bowling, 2001) a dinâmica do conceito resultante de experiências do passado, circunstâncias do presente e expetativas futuras. Sendo ainda explicitado que a perceção de QdVRS não está apenas dependente da capacidade física individual, mas está também relacionado com as preferências e prioridades da vida.

3.3 Instrumentos de medida de QdVRS em Oncologia

Para (Pimentel, 2003) a medida objetiva e precisa da QdVRS é imperiosa. Uma variedade de instrumentos tem sido desenvolvida com esse propósito, uma vez que um melhor conhecimento da QdVRS pode fornecer dados para uma tomada de decisão mais racional, quer para um individuo, quer para uma determinada população (Pimentel, 2003). A área oncológica

representa uma das áreas com maior interesse e utilidade na utilização de instrumentos de medida da QdVRS, uma vez que as doenças do foro oncológico e o seu tratamento têm um profundo impacto na vida dos pacientes (Gonçalves, 2012). A Organização Europeia para Pesquisa e Tratamento do Cancro enquadra-se como uma das organizações responsáveis pelo desenvolvimento de instrumentos de medida da QdVRS para a prática clínica, através de questionários como o EORTC QLQ-C30 ou do módulo específico para doentes oncológicos da cabeça e pescoço EORTC QLQ-H&N35 (Oliveira et al., 2011). Os benefícios potenciais apontados em (Oliveira et al, 2011), pelo uso de questionários para avaliação de QdVRS referidos refletem-se na maior facilidade do paciente para descrever as suas queixas relativamente à doença e/ou tratamento, na disponibilização de avaliações aos aspetos psicossociais dos pacientes às equipas médicas, reforçar a ideia no paciente que a sua QdV está a ser levada em conta pela equipa médica e que facilitará a comunicação entre as partes, assim como através do preenchimento do questionário será obtida uma avaliação concisa e validade da QdV permitindo a avaliação por itens do questionário que poderá ser usada no reconhecimento de possíveis problemas que podem ser reportados à equipa médica.

É transversal, quando são desenvolvidos estudos sobre a temática, a importância da medida da QdV, através de instrumentos objetivos (mais específicos) e subjetivos (mais genéricos). (Farquhar, 1995) aponta nos instrumentos mais objetivos a vantagem de não existirem tantos erros por parte do observador dos mesmos, mas refere o facto de não se preocuparem com os sentimentos dos visados. Já nos mais subjetivos, elucida o facto de se relacionarem com o julgamento das pessoas sobre a sua vida, destacando exemplos como a satisfação relativamente à atividade profissional, à saúde e moral. (Campos & Neto, 2008) apontam no seu estudo alguns dos instrumentos genéricos de medida mais utilizados no mundo: *SicknessImpact Profile (SIP), Nottingham Health Profile (NHP), McMaster Health Index Questionnaire (MHIQ), Rand Health Insurance Study (Rand HIS), The Medical Outcomes Study 36-Item Short Form Health Survey (SF-36),* Avaliação da Qualidade de Vida da Organização Mundial da Saúde *(WHOQOL-100),* focando ainda o facto de este tipo de medida ser obtida através de questionários de base populacional, mais direcionados a estudos epidemiológicos, ao planeamento e à avaliação do sistema de saúde. A nível dos instrumentos mais específicos classifica-os como mais capazes de avaliar, de forma individual e específica, determinados aspetos da QdV e atribui-lhes como principal característica o facto de medir alterações com maior sensibilidade, quer no historial, quer após determinada intervenção, e pelo facto de se adaptarem a uma determinada situação ou população. Num contexto de QdVRS (Guyatt et al.,

1993) desenvolvem um estudo sobre formas de estruturar e de melhor selecionar esses tipos de instrumentos para cada situação.

Uma vez que os questionários EORTC QLQ-C30 e EORTC QLQ-H&N35 estão já implementados no SADC que este estudo pretende sustentar, é oportuno descrever as suas características principais.

Questionário EORTC QLQ-C30

O questionário EORTC QLQ-C30 tem como objetivo principal a sua utilização em ensaios clínicos internacionais na luta contra o cancro (Koller et al., 2007). É composto por trinta itens, inseridos em cinco diferentes escalas funcionais – física, emocional, desempenho, cognitiva e social – três escalas de sintomas -fadiga, dor, náusea e vómito - seis itens para a avaliação de sintomas ou problemas adicionais (dispneia, perda de apetite, insónia, dificuldades financeiras, obstipação e diarreia) e uma escala global de QdV. Todas as escalas e itens variam numa pontuação dos 0 aos 100, sendo que à exceção das escalas funcionais e da escala global de QdV, em todas as outras escalas e itens simples, uma pontuação elevada indica pior QdV (Heutte et al., 2014; Oliveira et al., 2011; Pickard et al., 2009).

Questionário EORTC QLQ-H&N35

O questionário EORTC QLQ-H&N35 foi desenvolvido para o módulo específico de doentes oncológicos da cabeça e pescoço. Compreende trinta e cinco perguntas sobre sintomas e efeitos colaterais do tratamento, função social, imagem corporal e sexualidade. Incorpora sete escalas de sintomas (dor, deglutição, paladar e olfato, fala, alimentação em público, contacto social e sexualidade) e onze itens simples. Para todas as escalas e itens simples uma pontuação elevada significa pior QdV. Os dados obtidos correspondem ao estado do doente durante a última semana (Heutte et al, 2014; Koller et al., 2007).

3.4 Conclusões

Relativamente á temática Qualidade de Vida foi notório, durante o desenvolvimento da revisão de literatura, que a sua definição apesar de ambígua e subjetiva, aborda estados e expetativas de bem-estar físico, emocional, espiritual, psicológico e físico, num ponto de vista individual e que estendido ao termo Qualidade de Vida Relacionada com Saúde retrata ainda o modo como o individuo é afetado pela doença e pelo consequente tratamento. Foi ainda possível perceber

de que forma os instrumentos de medida, como o QLQ-C30 e o QLQ-H&N35, atuam e as escalas que adotam

.

4. DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O presente capítulo retrata a exploração de diversos contributos científicos, que contribuem para a compreensão do processo de extração de conhecimento necessário ao desenvolvimento da componente prática do presente projeto de dissertação e para a definição de Data Mining. São ainda descritas técnicas de Data Mining bem como as métricas de avaliação que permitem perceber o seu resultado e contributo É ainda apresentada uma revisão sistemática de artigos que evidencia a influência deste tipo de técnicas nas áreas e tarefas da Medicina.

4.1 Princípios

O volume de dados, gerados e armazenados no decurso de uma qualquer atividade, ultrapassa a capacidade de análise humana e torna impossível extração de conhecimento a partir desses mesmos dados, sem recorrer a um sistema que automatize esse processo (Fayyad et al., 1996). Este contexto justifica a existência da área de investigação de Descoberta de Conhecimento em Bases de Dados, genericamente definida por (Fayyad et al., 1996) como "o processo não trivial de identificação de padrões válidos e potencialmente úteis, percetíveis a partir dos dados", em que os princípios a si associados têm influência das áreas de Inteligência Artificial, Aprendizagem Automática, Reconhecimento de Padrões, Estatística, Base de Dados, Sistemas de Informação, entre outras. (Santos & Ramos, 2006; Ramos, 2014; Fayyad et al., 1996; Oded Maimon & Lior Rokach, 2010). Os algoritmos utilizados para procurar os padrões nos dados são denominados de Data Mining e são considerados o passo nuclear do processo de DCBD, que se desenvolve em várias fases. Esses padrões podem ou não representar conhecimento útil, sendo uma das fases que requer, normalmente, a participação do utilizador (Fayyad et al., 1996).

4.2 Processo de DCBD

Na definição de (Fayyad et al., 1996) relativamente à DCBD, os padrões referidos podem ser caracterizados por modelos, relações ou estruturas nos dados, que devem ser percetíveis após breve processamento. Os dados representam um conjunto de factos armazenados na caracterização de diversos padrões. O termo processo está associado à execução de diversos passos iterativos, iniciado com a seleção de dados a analisar e terminando com a interpretação dos resultados. Além de iterativo o processo de DCBD é interativo, uma vez que o seu

desenvolvimento requer a participação do utilizador sempre que é necessária a tomada de decisão (Santos & Ramos, 2006). A literatura revista (Oded Maimon & Lior Rokach, 2010; Fayyad et al., 1996) relativamente ao processo de DCBD descreve-a através de nove passos:

1. Desenvolver aprendizagem do domínio da aplicação e perceção de conhecimento relevante sobre o domínio, bem como a identificação dos objetivos a atingir no processo, na ótica do utilizador e do ambiente no qual se desenvolverá.
2. Seleção dos dados em que incidirá a descoberta, através dos algoritmos de Data Mining.
3. Pré-processamento e tratamento do conjunto de dados, através de um conjunto de estratégias de atuação, no que diz respeito ao aparecimento de dados incorretos ou omissos.
4. Transformação de dados, o que inclui a procura de configurações apropriadas para representar os dados, com o objetivo de diminuição do número de registos e/ou atributos em análise.
5. Seleção das técnicas de Data Mining (classificação, regressão, clustering, entre outras) que melhor se enquadram nos objetivos definidos.
6. Seleção dos algoritmos de Data Mining de acordo com a estratégia traçada para a obtenção de conhecimento sobre o conjunto de dados, previamente escolhido e tratado.
7. Implementação dos algoritmos de Data Mining, com intuito de descobrir padrões interessantes e úteis. Podendo este passo ser repetido até que os resultados pretendidos sejam úteis de acordo com os objetivos traçados.
8. Interpretação dos padrões descobertos, sendo possível a transformação dos mesmos em formatos percetíveis ao utilizador.
9. Utilização do conhecimento descoberto, com intenção de utilizá-lo num diferente sistema, ou apenas documentando o mesmo, para posterior utilização pelos interessados.

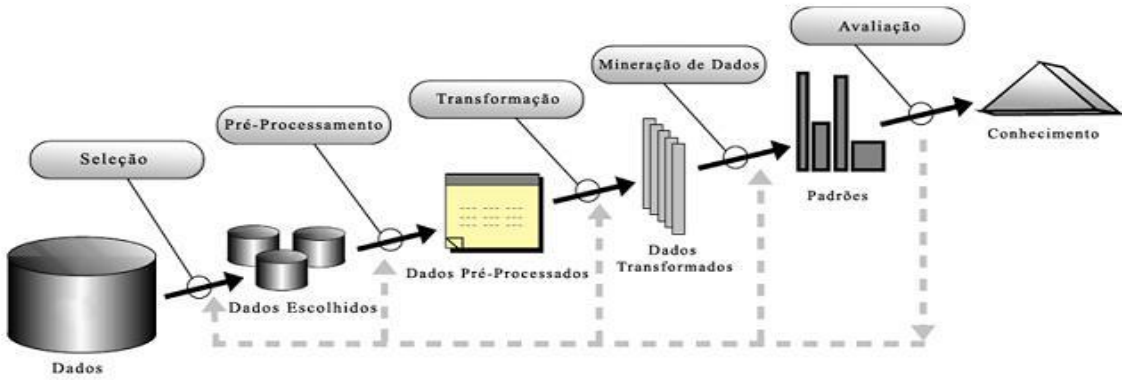


Figura 2- Processo de DCBD (adaptado de Usama Fayyad, 1996)

Em termos de trabalho, a fase de Data Mining, representa 20% no processo de DCBD, sendo melhor suportada por *software* do que as restantes. Todas as outras fases, desde a seleção de dados até à interpretação dos padrões encontrados, constituem mais uma questão de "arte" do que uma rotina que possa ser automatizada. (Andrienko & Andrienko, 1999; Santos et al., 1999).

4.3 Data Mining

O termo Data Mining é associado a análise de grandes repositórios de dados, geração de informação e descoberta de conhecimento. A sua função basilar enquadra-se na descoberta de padrões nos dados. É ainda associado à organização da informação de relações não facilmente percetíveis, regras de associação de estruturas, estimação de itens e valores desconhecidos para classificar objetos, compor agrupamentos (clustering) de objetos homogéneos, entre outros (Peña-Ayala, 2014). (Vlahos et al., 2004) define o termo DM, como um sistema de informação, direcionado para a pesquisa em repositórios de enorme dimensão, gerar informação e descobrir conhecimento. (Linoff & Michael, 2000), é mais simplista, definindo o termo DM como o processo de analisar grandes volumes de dados para descobrir padrões e regras.

Historicamente, a procura de padrões úteis em dados era associada a termos diferentes, que não Data Mining, tais como, Extração de Conhecimento, Descoberta de Informação, Colheita de Informação, Arqueologia de Dados e Processamento de padrões de dados, sendo maioritariamente usado por estatísticos, analistas de dados e gestores de sistemas de informação (Fayyad et al., 1996). (Oded Maimon & Lior Rokach, 2010), atribui como fator de desenvolvimento do Data Mining a crescente inovação e utilização de Sistemas Gestores de Bases de Dados. A área de Data Mining revela, atualmente, bastante importância na indústria de informação e na sociedade, devido à sua capacidade de transformar dados em informação e conhecimento, capacitando a sua utilização em diversas aplicações, desde análises de marketing, deteção de fraudes, gestão de relação com clientes, controlos de produção, exploração científica, saúde, medicina, entre outros.

4.4 Objetivos, Tarefas e Técnicas de Data Mining

Os objetivos de Data Mining recaem em dois tipos de abordagens, Verificação e Descoberta. Na Verificação são tidas em conta as hipóteses de análise do utilizador, enquanto na Descoberta, a procura de padrões é realizada de forma automática, sendo a mesma classificada

em duas diferentes categorias de tarefas: Descrição ou Previsão (Faria, 2013; Fayyad et al., 1996; Santos & Ramos, 2006). A Descrição permite identificar regras que caracterizam os dados analisados, enquanto a Previsão utiliza determinados atributos da base de dados para prever o valor de uma outra variável. Adicionalmente, na consideração dos modelos de Descrição, o melhor será o que apresenta uma precisão mais elevada, enquanto nos modelos de Previsão, a opção recairá pelo modelo que poderá não revelar resultados mais precisos, mas que permita adquirir conhecimento mais alargado sobre os dados analisados. (Santos & Ramos, 2006; Fayyad et al., 1996).

Assim, para alcançar os objetivos de previsão, podemos incluir as tarefas de Classificação e Regressão.

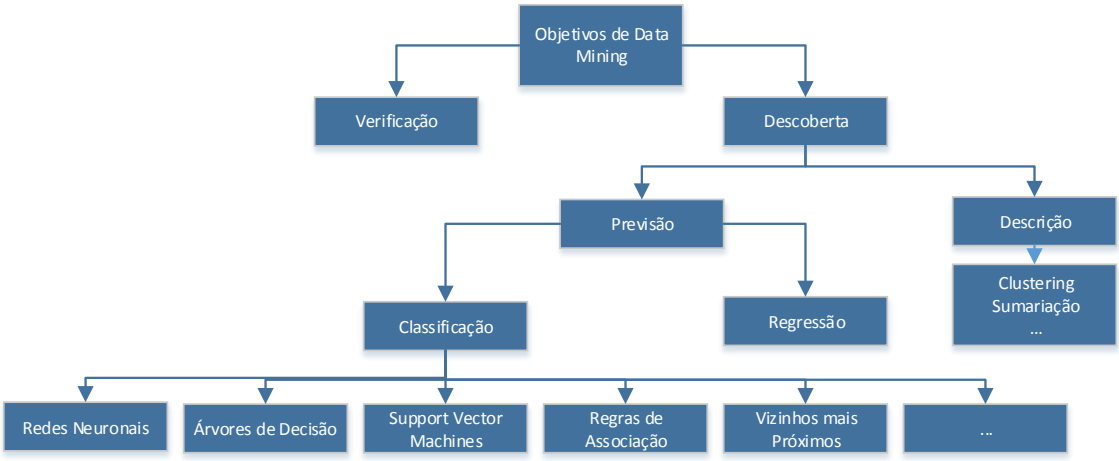


Figura 3 - Taxonomia de Data Mining (adaptado de Maimon & Rokach,2010)

A Classificação é considerada uma tarefa de aprendizagem supervisionada, que consiste na descoberta de uma função que vai associar um determinado caso a uma classe de entre as classes de classificação, com o propósito de classificar um novo objeto através do modelo de classificação. Este modelo desenvolve-se através da análise de registos num conjunto de dados de treino, sendo que os registos estão atribuídos a classes predefinidas, que constituem o conjunto de valores possíveis para o atributo de saída. O modelo obtido é usado para classificar o conjunto de dados de teste, permitindo verificar o seu desempenho na classificação de dados desconhecidos. Os resultados obtidos são analisados, no sentido de verificar o desempenho do modelo. A precisão do modelo é avaliada com base na quantidade de registos classificados corretamente, comparado com o valor armazenado no conjunto de dados de testes. Se a precisão do modelo for considerada aceitável, de acordo com o domínio da aplicação, este pode ser utilizado em tarefas de previsão para identificar a classe a que cada registo pertence (Santos & Ramos, 2006; Han & Kamber, 2006 Faria, 2013, Fayyad et al., 1996).

Na figura 4 (Santos & Ramos, 2006), através de um conjunto de 23 registos de clientes que solicitaram um crédito a uma entidade bancária, em que o símbolo '□' representa clientes que atrasaram ou falharam o cumprimento das prestações, e o símbolo 'Δ' representa os clientes cumpridores, é possível verificar que apenas dois registos foram mal classificados, pelo modelo em questão, refletindo-se em 91,3 de percentagem em termos de precisão. Se de acordo com o domínio da aplicação esses valores se constituírem como aceitáveis, esse modelo pode ser utilizado em tarefas de previsão para identificar as classes do registo.

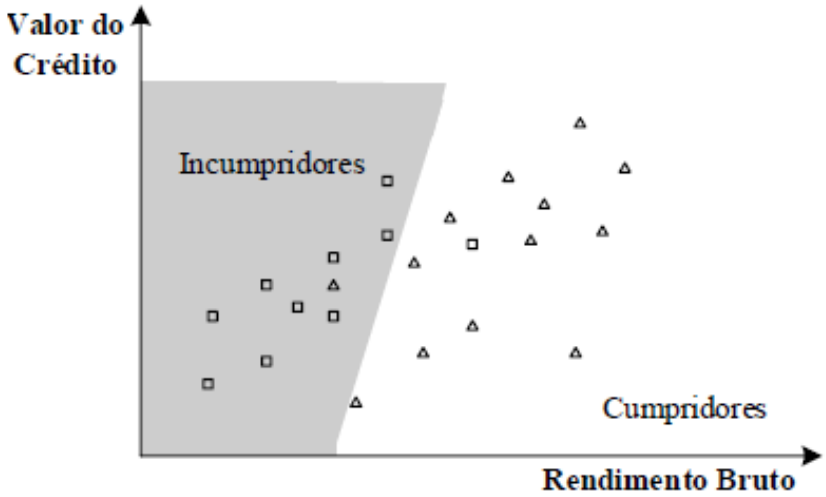


Figura 4 - Regressão Linear (Santos & Ramos, 2006)

Nas técnicas mais comuns, relativamente a tarefas de previsão, encontram-se as Árvores de Decisão, Regras de Associação, Análise Estatística, Redes Neuronais, classificadores de Bayes, entre outros.

Árvores de Decisão

A técnica de Árvores de Decisão desenvolve-se numa estrutura de árvore que representa um conjunto de decisões. Os algoritmos de indução deste tipo de técnica permitem gerar regras de classificação dos dados, baseados na informação armazenada na base de dados. É constituída por nós, ramos e folhas, sendo a sua representação facilmente interpretada pelos utilizadores. Nos nós encontram-se os atributos a classificar, nos ramos são descritos os valores possíveis para esses atributos e as folhas indicam as diversas classes em que o registo pode ser classificado. Na indução de uma árvore de decisão, que respeita o processo de desenvolvimento das tarefas de classificação, ou seja, utilizando o conjunto de dados de treino para identificar o modelo que classifica os dados atendendo à variável de saída, podemos obter árvores cujos

ramos podem refletir ruído ou *outliers* nos dados, referentes ao conjunto de dados de treino. Esse tipo de acontecimento poderá ser ultrapassado através de métodos de corte que permitem melhorar o desempenho da árvore na classificação de dados desconhecidos (Santos & Ramos, 2006; Han &Kamber, 2006;Faria, 2013, Fayyad et al., 1996, Oded Maimon & Lior Rokach, 2010).

Regras de Associação

As regras de associação são uma técnica de classificação que permite a identificação de relações entre os atributos existentes numa base de dados, representadas na forma de uma regra.

Se x Então y ou "X ⇒ Y"

Para medir objetivamente as regras encontradas são utilizados fatores de confiança, inferidos pelo número de transações efetuadas na base de dados que satisfazem X e Y e de suporte, através de um subconjunto de registos que satisfazem a união dos atributos que integram a parte antecedente (X) e consequente (Y). Através da confiança é possível inferir a força da regra, enquanto através do suporte é possível conhecer a significância estatística. Devem ser analisadas simultaneamente, uma vez que o suporte pode ser elevado (percentagem de registos que satisfazem a regra) e a regra possuir uma associação fraca, sendo diminuto o número de registos em que é possível prever Y, conhecendo X (Santos & Ramos, 2006; Han & Kamber, 2006; Faria, 2013, Oded Maimon & Lior Rokach, 2010).

A figura abaixo apresentada (Santos & Ramos, 2006), Figura 5, representa um conjunto de dados associados à compra de produtos numa superfície comercial. Os atributos disponíveis são Número e Produto, identificado o talão de compra e produto adquirido, respetivamente. Numa compra podem ser adquiridos um ou mais produtos.

No exemplo apresentado, a regra PÃO & MANTEIGA ⇒ LEITE (2:50%, 1) indica que os clientes que compram o produto Pão juntamente com o produto Manteiga, também compram Leite. Esta regra apresenta um suporte de 50%, o que significa que metade dos registos analisados verificam a referida regra. A confiança da regra apresenta o valor 1 (100%), uma vez que todos os registos em que foi verificada a ocorrência da compra Pão e Manteiga, também foi confirmada a aquisição de Leite.

Número	Produto
1101	Pão
1101	Manteiga
1101	Fiambre
2201	Leite
2201	Pão
2201	Queijo
3101	Leite
3101	Pão
3101	Queijo
4100	Margarina
4100	Pão
4100	Manteiga
4100	Leite

Produtos agrupados por compras

Número	Produto
1101	Pão, Manteiga, Fiambre, Leite
2201	Pão, Queijo, Leite
3101	Pão, Queijo, Margarina
4100	Pão, Manteiga, Leite

Exemplo de uma regra:

Pão & Manteiga ⇒ Leite (2:50%, 1)

Figura 5 - Regras de Associação (Santos & Ramos, 2006)

Vizinhos mais próximos

Esta técnica baseia-se no princípio de que registos semelhantes estão próximos uns dos outros, quando analisados numa perspetiva espacial. Os dados, representados por pontos no espaço, serão agrupados por semelhanças. Através deste agrupamento, serão identificadas regiões, denominadas classes ou segmentos, que apresentam características comuns para os registos que representam. A complexidade desta técnica aumenta à medida que cresce o número de registos a analisar, uma vez que cada registo é comparado com os restantes registos da amostra. (Santos & Ramos, 2006) Respeitando o princípio enunciado, surgem técnicas como o *k-means ou k-nearest-neighbour*. *K-means* é uma técnica frequentemente utilizada na segmentação (clustering) de dados. Inicia-se com a construção de k classes, em que k é um parâmetro de entrada. Segundo (Santos & Ramos, 2006) cada classe é representada pelo seu centro de gravidade, sendo que para determinar essa mesma classe, cada registo é transformado num ponto no espaço, apresentando tantas dimensões quanto os atributos em análise. O valor de cada atributo é interpretado como a distância da origem até à sua localização num dado eixo. K-*Nearest-Neighbour* é outra técnica utilizada na segmentação de dados, que se implementa pela identificação dos k elementos mais próximos de um dado registo (Santos & Ramos, 2006; Han & Kamber, 2006; Faria, 2013).

Redes Neuronais Artificiais

Uma Rede Neuronal Artificial (RNA) é um sistema de classificação modelado de acordo com os princípios do sistema nervoso humano. Desenvolve-se através de uma rede de ligações com pesos ajustáveis, contemplando as unidades de entrada encarregues de receber os dados a analisar, as unidades de saída que transmitem os sinais à saída da rede e as unidades intermédias, que contemplam um número ilimitado de níveis intermédios. As arquiteturas de RNA diferem no número de níveis intermédios permitidos. Nas redes *Perceptron*, não existe qualquer nível intermédio, o que torna o processo de aprendizagem mais simples mas limita a sua utilização a problemas aproximáveis através de funções lineares. As redes *Multi-perceptron* apresentam um ou mais níveis intermédios, permitindo aproximar qualquer função não linear. As diversas arquiteturas disponíveis podem variar na quantidade de nós de saída e na possibilidade de ajustamento dos pesos das ligações (Santos & Ramos, 2006).

A aprendizagem de uma rede inicia-se com a atribuição semelhante de pesos a todas as ligações da rede, sendo que através do conjunto de dados de treino e através de várias iterações sobre os mesmos são comparadas os valores de saída, até que se coadunem com o valor de classificação esperado. Uma desvantagem associada às RNA prende-se com o processo de aprendizagem não ser transmitido num formato percetível ao utilizador, uma vez que não é possível perceber como as decisões são tomadas.

Support Vector Machines

As *Support Vector Machines* são uma técnica de classificação de dados lineares e não lineares. Os algoritmos de SVM, tipicamente, fazem o mapeamento não linear para transformar os dados de treino originais em dimensões superiores. Nestas nova dimensão é traçado o hiperplano ótimo que executa a separação linear entre as classes. Através da figura 6, que retrata a compra ou não de um computador, é possível verificar o conjunto de dados linearmente separados por um número finito de hiperplanos. A escolha do melhor hiperplano recai pelo que tiver maior margem, o que significa uma maior separação entre as classes (Han & Kamber, 2006; Faria, 2013, Oded Maimon & Lior Rokach, 2010).

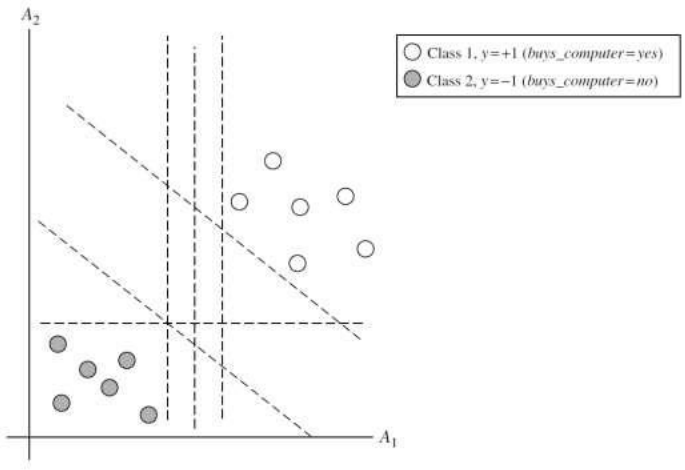


Figura 6 - Separação linear de duas classes (Han & Kamber, 2006)

Classificação Bayesiana

A classificação Bayesiana consiste na aplicação da teoria desenvolvida por Bayes, através da classificação de padrões, em termos probabilísticos, num conjunto de dados. Esta técnica de classificação assume que a presença de uma característica particular da classe não está relacionada com a presença de qualquer outra característica. Sendo a sua principal vantagem, o facto de apenas necessitar de um pequeno conjunto de dados de treino para estimar os parâmetros necessários para efetuar a classificação (Han & Kamber, 2006; Faria, 2013).

Para alcançar os objetivos de descrição podemos incluir as técnicas de Clustering e Sumariação.

Clustering

Clustering é uma tarefa de aprendizagem não supervisionada, que permite a identificação de um conjunto de classes ou segmentos, denominados por clusters. Os clusters surgem de agrupamentos detetados nos dados e que obedecem a métricas de similaridade, não tendo o utilizador qualquer influência na definição dos mesmos. A identificação de clusters nos dados poderão surgir pela utilização de diversos algoritmos, respeitando estratégias de divisão sucessiva de registos a segmentar, ou pelo agrupamento de registos em segmentos. (Santos & Ramos, 2006; Han & Kamber, 2006; Faria, 2013, Fayyad et al., 1996, Oded Maimon & Lior Rokach, 2010).

Sumariação

Para descrever resumidamente um determinado conjunto de dados é utilizada a tarefa de Sumariação. As referidas descrições são obtidas por generalização dos dados, sendo a determinação da moda ou do desvio padrão de uma amostra um exemplo disso mesmo.

Normalmente, tarefas de sumariação são utilizadas na exploração dos dados com intuito de formular hipóteses de análise futura, podendo ainda constituir um objetivo de Data Mining (Santos & Ramos, 2006).

4.5 Avaliação dos modelos de Data Mining

(Turban et al., 2011) elucida alguns dos fatores relevantes na avaliação dos modelos de classificação de Data Mining, como:

- **Acuidade de previsão**: capacidade do modelo prever corretamente a classe;
- **Velocidade**: custos computacionais envolvidos na geração e utilização do modelo;
- **Robustez**: capacidade de o modelo realizar previsões com uma acuidade razoável, independentemente do ruído ou presença de valores omissos nos dados;
- **Escalabilidade**: habilidade de construir um modelo de previsão eficiente face a um número considerável de dados;
- **Interpertabilidade**: capacidade de revelar conhecimento de uma forma percetível;

Nos estudos de (Lavrac, 1999; Santos & Azevedo, 2005) são descritos outro tipo de métodos de avaliação de eficiência dos modelos de classificação, como a Matriz de Confusão. Através desta avaliação é possível conhecer o número de classificações efetuadas corretamente pelo modelo, assim como as incorretas. A tabela seguinte retrata a explicação enunciada.

Tabela 1 - Matriz de confusão

Matriz de Confusão	**Previsão Classe 1**	**Previsão Classe 2**
Classe 1	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
Classe 2	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

Podem-se então resumir quatro indicadores:

- **Verdadeiros Positivos** (VP) – Número de elementos positivos classificados como tal;
- **Verdadeiros Negativos** (VN) – Número de elementos negativos classificados como tal;
- **Falsos Positivos** (FP) – Número de elementos positivos classificados como negativos;
- **Falsos Negativos** (FN) – Número de elementos negativos classificados como positivos;

Através da Matriz de Confusão podem ser calculadas diversas métricas, enunciadas na tabela seguinte:

Tabela 2- Métricas de avaliação

Métrica	Descrição	Representação
Acuidade	Taxa de capacidade de previsão do modelo;	$\frac{\mathbf{VP + VN}}{\mathbf{n}} * \mathbf{100}(\%)$
Sensibilidade	Taxa de Verdadeiros Positivos classificados;	$\frac{\mathbf{VP}}{\mathbf{VP + FN}} * \mathbf{100}(\%)$
Especificidade	Taxa de Verdadeiros Negativos classificados;	$\frac{\mathbf{VN}}{\mathbf{VN + FP}} * \mathbf{100}(\%)$
Curva ROC	Avaliação do desempenho de um modelo de classificação, pela relação entre a taxa de Sensibilidade e Especificidade;	Gráfica
AUC (Area Under Curve)	Avaliação do desempenho de um modelo de classificação pela área representada abaixo da Curva ROC	Gráfica

As métricas de Sensibilidade e Especificidade, segundo (Lavrac, 1999), em casos de aplicação médica são mais importantes do que a Acuidade. Esse facto prende-se com as características da área médica e a maior preocupação em tomar decisões baseadas em classificações corretas, do que num modelo que se revele mais ou menos capaz, dando a hipótese de erro.

4.6 Data Mining na Medicina

A medicina perfila-se como uma das áreas onde a aplicação de técnicas de Data Mining tem crescido exponencialmente. Esta realidade deriva da aposta em tecnologias em sistemas de informação, com capacidade para armazenar o grande e complexo volume de dados gerado no número considerável de interações e tarefas desenvolvidas nas instituições de saúde. O Data Mining surge assim associado à necessidade de extração de conhecimento das informações armazenadas nos registos eletrónicos de saúde, de forma a apoiar o processo de tomada de decisão das equipas médicas e a prática clínica.

O crescente interesse pela combinação destas duas áreas é sustentada pelo número de artigos publicados no portal científico da área médica, PUBMED, sobre o qual incidiu uma revisão sistemática de artigos, apresentada na tabela seguinte, que identificam uma clara utilização de técnicas de Data Mining sobre as tarefas de Diagnóstico, Prognóstico e Tratamento e sobre as áreas relacionadas com Farmácia, Doenças Cardiovasculares, Oncologia, Radiologias e Fraude.

Tabela 3- Revisão Sistemática da relação entre Data Mining e Medicina (Pubmed)

Título	**Autores**	**Sumário**	**Área de Intervenção**	**Tarefas de Data Mining**	**Técnicas e modelos de Data Mining**
Comparison analysis of data mining models applied to clinical research in traditional Chinese medicine	(Zhao Y, Xie Q, He L, Liu B, Li K, Zhang X, Bai W, Luo L, Jing X, Huo R, 2014)	Facilitar a seleção de modelos de Data Mining na prática clínica relacionada com a medicina tradicional chinesa	Diagnóstico e Tratamento	Não estão claramente identificadas	Não estão claramente identificadas
Data mining of solubility parameters for computational prediction of drug-excipient miscibility	(Alhalaweh A, Alzghoul A, Kaialy W, 2014)	Previsão da miscibilidade dos excipientes de um fármaco	Farmácia	Previsão e Descrição	K-means Clustering
Improving diagnostic accuracy using agent-based distributed data mining system	(Sridhar S., 2013)	Utilização de técnicas de Data mining para melhorar a precisão nas tarefas de diganóstico	Diagnóstico	Não estão claramente identificadas	Não estão claramente identificadas
Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes	(Austin PC, Tu JV, Ho JE, Levy D, Lee DS, 2013)	Utilização de técnicas de Data Mining para classificação e previsão de doenças	Insuficiência Cardíaca	Previsão e Classificação	Classification Trees, Random Forest, Support Vector Machines

HPVdb: a data mining system for knowledge discovery in human papillomavirus with applications in T cell immunology and vaccinology	(Zhang G, Riemer AB, Keskin DB, Chitkushev L, Reinherz EL, Brusic V, 2014)	Desenvolvimento de uma framework para extração de conhecimento em papiloma vírus humano com aplicação em imunologia de células T e vacinologia.	Diagnóstico e Prognóstico de Cancros	Não estão claramente identificadas	Não estão claramente identificadas
An overview of data mining algorithms in drug induced toxicity prediction	(Omer A, Singh P, Yadav NK, Singh RK, 2014)	Utilização de algoritmos de Data Mining para Previsão da toxicidade induzida por fármacos	Farmácia; Tratamento	Previsão e Descrição	Artificial Neural Networks, Support Vector Machine, K-mean clustering
Data mining in radiology.	(Kharat AT, Singh A, Kulkarni VM, Shah D, 2014)	Análise de dados relacionados com radiologia para facilitar processo de tomada de decisão.	Radiologia	Previsão e Descrição	Decision Trees
Comparsion analysis of data mining models applied to clinical research in traditional Chinese medicine.	(Zhao Y, Xie Q, He L, Liu B, Li K, Zhang X, Bai W, Luo L, Jing X, Huo R, 2014)	Seleção de modelos de Data Mining para resolver os problemas da Medicina Chinesa Tradicional	Diagnóstico e Tratamento	Não estão claramente identificadas	Não estão claramente identificadas
Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record.	(Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, Williams BA, deFilippi C, Ebadollahi S, Stewart WF, 2014)	Perceber a prevalência de sinais e sintomas de insuficiência cardíaca através de técnicas de data Mining	Insuficiência Cardíaca	Descrição	Text Mining

Noninvasive diagnosis of liver fibrosis: utility of data mining of both ultrasound elastography and serological findings to construct a decision tree	(Yada N, Kudo M, Kawada N, Sato S, Osaki Y, Ishikawa A, Miyoshi H, Sakamoto M, Kage M, Nakashima O, Tonomura A, 2014)	Estudo para melhorar o diagnóstico e avaliação da fibrose hepática através de árvores de decisão	Diagnóstico	Classificação	Decision Tree
The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome	(Penny KI, Smith GD, 2012)	Descoberta da influência de fatores sociodemográficas na avaliação de Qualidade de Vida Relacionada com Saúde	Síndrome de Intestino Irritável	Classificação	Artificial Neural Network, Classification Tree, Logistic Regression
Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges.	(Banaee H, Ahmed MU, Loutfi A, 2013)	Estudo com vista a fornecer uma visão geral de técnicas de Data Mining aplicadas a dados de sensores wearable no domínio da saúde	Sensores Wearable	Classificação.	Decision Tree, Neural Networks, Support Vector Machine, Association Rules, Bayesian Network, Logistic Regression
Exploring factors associated with pressure ulcers: a data mining approach.	(Raju D, Su X, Patrician PA, Loan LA, McCarthy MS, 2015)	Utilização de técnicas de Data mining sobre informações demográficas e medidas médicas para prever Úlceras de Pressão	Úlceras de Pressão	Previsão	Logistic regression, Decision trees, Random forests
Data mining in healthcare and biomedicine: a survey of the literature	(Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L, 2012)	Importância de Data Mining para a área da Saúde e a bio medicina	Estudo Geral na área médica	Previsão e Classificação	Não estão claramente identificadas

Drug safety data mining with a tree-based scan statistic	(Kulldorff M, Dashevsky I, Avery TR, Chan AK, Davis RL, Graham D, Platt R, Andrade SE, Boudreau D, Gunter MJ, Herrinton LJ, Pawloski PA, Raebel MA, Roblin D, Brown JS, 2013)	Detetar eventos adversos causados por fármacos, pela utilização da técnica de Data Mining de Árvores de Decisão	Tratamento	Previsão	Decision Tree
Prediction of survival in thyroid cancer using data mining technique.	(Jajroudi M, Baniasadi T, Kamkar L, Arbabi F, Sanei M, Ahmadzade M, 2014)	Previsão de sobrevivência relativamente ao Cancro da Tiróide, através da utilização	Doenças cardiovasculares	Previsão	Artificial Neural Network
Applying data mining techniques to improve diagnosis in neonatal jaundice	(Ferreira D, Oliveira A, Freitas A, 2012)	Aplicação de Técnicas de Data Mining para melhorar o diagnóstico da icterícia neonatal	Diagnóstico	Classificação	Decision trees, Neural networks
Using data mining to detect health care fraud and abuse: a review of literature.	(Joudaki H, Rashidian A, Minaei-Bidgoli B, Mahmoodi M, Geraili B, Nasiri M, Arab M, 2014)	Estudo da utilização de Data Mining para deteção de fraudes no sistema de saúde	Fraude	Não estão claramente identificadas	Não estão claramente identificadas

4.7 Conclusões

Na revisão de literatura desenvolvida sobre a área de Descoberta de Conhecimento em Bases de Dados foi reforçado o crédito a este tipo de abordagem, que desempenha com eficácia e eficiência a extração de conhecimentos de grandes repositórios de dados, através de técnicas que compõem o seu processo. Essas técnicas, resumidas pelo termo Data Mining, refletem a

componente mais importante deste projeto. Os contributos científicos consultados permitiram elaborar uma descrição dos objetivos, tarefas e técnicas que se pretendem implementar na componente prática do presente projeto de dissertação. Foi ainda possível concluir quais as tarefas e áreas da medicina que refletem uma maior utilização das técnicas de Data Mining, para apoiar a prática clínica e processo de tomada de decisão das equipas médicas.

5. SISTEMA DE APOIO À DECISÃO CLÍNICA PARA A QUALIDADE DE VIDA

O presente capítulo apresenta o desenvolvimento da componente prática do projeto de dissertação, onde o objetivo principal visa a dotar os módulos do Sistema de Apoio à Decisão Clínica QoLIS, de um conjunto de técnicas de Data Mining com capacidade para explorar, extrair e evidenciar padrões num conjunto de dados relacionados com patologias oncológicas. Considerando a implementação do processo de DCBD como necessário para alcançar esse objetivo, toda a componente prática foi desenvolvida segundo a metodologia CRISP-DM, através das suas seis diferentes fases: Compreensão do Negócio, Estudos dos Dados, Tratamento dos Dados, Modelação, Avaliação e Desenvolvimento. Contudo, as fases referidas estão devidamente adaptadas, de forma a refletir as atividades desenvolvidas no presente projeto de Dissertação. São ainda descritas as ferramentas tecnológicas utilizadas, nomeadamente a plataforma RapidMiner, responsável pela visualização, validação e otimização dos dados, e o *software* Weka, responsável pela implementação das técnicas e modelos de Data Mining necessárias ao processo de descoberta de conhecimento implementado.

5.1 Ferramentas Tecnológicas

5.1.1 Plataforma RapidMiner

A plataforma de *software* RapidMiner (RapidMiner, 2014) fornece um ambiente integrado e intuitivo, com capacidade para desenvolver *Machine Learning*, *Data Mining*, *Text Mining*, Análise Preditiva e Análise de Negócio, através de um conceito de operação modular. No desenvolvimento do presente projeto de dissertação, a sua utilização visou as capacidades que demonstra relativamente à visualização, validação e otimização de dados, assim como à facilidade demonstrada na conversão de formatos dos ficheiros. Saliente-se que para o nível académico o *software* está disponível sob uma licença *open source.*

5.1.2 Software Weka

O *software* Weka (Hall et. al, 2009) é reconhecido pelo vasto conjunto de algoritmos de *Machine Learning* que apresenta, demonstrando capacidades para o desenvolvimento de projetos de Data Mining. Os algoritmos podem ser aplicados diretamente nos datasets, em formato .arff, ou através da invocação das classes java, uma vez que é uma ferramenta *open source* sob licença da *GNU*. Assenta numa arquitetura modular e extensível, disponibilizando um interface gráfico multifacetado, que apresenta ferramentas com capacidade para executar diversas operações sobre dados, que serão utilizadas no presente projeto de dissertação, nomeadamente, pré-processamento de dados, classificação, regressão, desenvolvimento de clusters, regras de associação e visualização.

5.2 Plataforma QoLIS

O desenvolvimento de um Sistema de Apoio à Decisão Clínica, denominado QoLIS, pela organização Optimizer, enquadra-se como o fundamento principal para o desenvolvimento do presente projeto de dissertação. Este sistema, caracteriza-se como um sistema de informação inovador, destinado à área da saúde, focado no acompanhamento e tratamento de pacientes com patologias oncológicas. É pretendido que através deste sistema, sejam disponibilizadas às equipas médicas um conjunto de ferramentas para medir de forma sistemática a Qualidade de Vida Relacionada com a Saúde, assim como a possibilidade de executar uma previsão de comportamentos futuros de algumas patologias.

O sistema integra um conjunto de tecnologias recentes e será composto por diversos módulos (módulo de determinação de QdVRS, módulo de inferência e módulo de alertas), que através deste projeto de dissertação se pretende que sejam capazes de utilizar técnicas de inteligência artificial, como o Data Mining, para explorar, extrair e ajudar a evidenciar padrões sobre dados existentes nas instituições de saúde, transformando os mesmos em informações úteis e confiáveis, de rápido e fácil acesso, aumentando o conhecimento disponível, para auxiliar os médicos no processo de tomada de decisão, maximizando e potencializando a qualidade de vida dos pacientes.

Adicionalmente, o sistema permitirá conhecer melhor os pacientes visados, os seus padrões de comportamento, a sua histórica clínica, as motivações e as reações às terapias prescritas pelos médicos, assim como recolher informação sobre a perceção dos pacientes acerca do seu estado

de saúde, registar informação sociodemográfica, socioeconómica e o estado clínico em cada momento da consulta associado ao tratamento.

5.2.1 Antecedentes e Contexto Atual

O Sistema de Apoio à Decisão Clínica QoLIS assenta numa arquitetura de informação bem definida, orientada a serviços, com capacidade para integrar sistemas de registos existentes nas unidades de saúde. Atualmente o sistema contempla uma série de módulos (Inferência, Determinação de QdVRS, Alertas) que atuam sobre diferentes unidades, que visam o tratamento de patologias oncológicas. Importa salientar os módulos de Inferência e de Determinação da QdVRS, da Unidade Oncológica de Cabeça e Pescoço, onde é pretendido que o presente projeto de dissertação tenha influência.

A unidade referida dispõe de uma base de dados devidamente preenchida com informações recolhidas por meio de consulta no Instituto Português de Oncologia do Porto, relativas a pacientes com patologias oncológicas da cabeça e pescoço. O SADC QoLIS integra nessa Unidade os instrumentos de medida de QdVRS, EORTC QLQ-C30 e EORTC QLQ-H&N35, com o intuito de obter uma avaliação concisa da QdV dos pacientes, permitindo a avaliação por itens dos questionários, através de uma adaptação do modelo de Rasch (Gonçalves, 2012) e pelo modelo de cálculo da instituição responsável pela sua criação, a European Organisation for Research and Treatment of Cancer (EORTC).

Desta forma, e aliado ao estudo exaustivo previamente realizado sobre as temáticas Qualidade de Vida e Descoberta de Conhecimento em Bases de Dados, intrinsecamente relacionadas com o Sistema de Apoio à Decisão Clínica QoLIS, estão reunidas as condições para que nesta fase se inicie a componente prática do projeto de Dissertação, que visa dotar os referidos módulos, de um conjunto de técnicas de Data Mining de forma a explorar, extrair e ajudar a evidenciar padrões nos dados existentes sobre pacientes com patologias oncológicas.

5.2.2 Modelo de Rasch implementado

O modelo de Rasch implementado na plataforma QoLIS (Gonçalves, 2012) é comumente utilizado na Teoria de Resposta ao Item (Schumacker, 2010) como uma solução para o problema de mensuração de um traço latente contínuo a partir de um conjunto de indicadores dicotómicos. Na sua forma original o modelo responde dicotomicamente às questões de um instrumento de medida, sendo a probabilidade de obter uma resposta positiva a cada item dependente de dois parâmetros, a habilidade do sujeito e a dificuldade do item. É representado

pela expressão matemática que relaciona a probabilidade de sucesso (P) e a diferença entre a habilidade do sujeito θ e a dificuldade do item β, representada na seguinte equação:

$$P = \frac{e^{\theta-\beta}}{1+e^{\theta-\beta}}$$

Obtendo-se:

$$\ln\left(\frac{P}{1-P}\right) = \theta - \beta$$

Assim, este modelo atribui uma probabilidade de acerto à resposta para uma determinada questão em função de apenas dois parâmetros a serem estimados:

- Eficiência θ do sujeito "s"
- Dificuldade β do item "i"

No entanto, o modelo implementado na Plataforma QoLIS é uma extensão do mesmo para itens politómicos (Gonçalves, 2012; Quintão et. al, 2011), como os apresentados nos questionários que integram a plataforma. O modelo é calculado de acordo com o Modelo de Escalas de Classificação de Andrich (Andrich, 2006), que utiliza o método da máxima verosimilhança para estimar os parâmetros do modelo, no qual:

$$P(X_{ni} = x)^n = \frac{e^{x(\theta_{ni}-\beta_{i)}-\sum_{k=1}^{x}\tau_{ik}}}{Y_{ni}},\ x \in \{0,1,\dots,Y_i\}$$

Onde:

β_i representa a dificuldade do item i;

θ_n representa a habilidade do sujeit n;

m_i corresponde ao valor máximo da escala para o item i;

τ_{ik} é o valor de "$threshold$" da categoria k para o item i;

Y_{ni} é o factor de normalização dos parâmetros sendo determinado por:

$$Y_{ni} = \sum_{x=0}^{mi} e^{x(\theta_n-\beta_{i)}-\sum_{k=1}^{x}\tau_{ik}}$$

O "$threshold$" representa o local da variável latente (para um item em particular) no qual a probabilidade ser observado abaixo de uma categoria é a mesma de ser observada acima dela. Ou seja, τ_{ik} é o valor da variável latente em estudo para a qual no item i a probabilidade de ser observada abaixo da categoria k é igual à de ser observada acima da categoria k.

5.3 Objetivos da componente prática

Os objetivos para o desenvolvimento da componente prática do presente projeto de dissertação, podem ser descritos nos seguintes pontos:

- Utilização de informações sociodemográficas, clínicas e dos instrumentos de medida de QdVRS, na fase anterior ao tratamento, com intuito de detetar e identificar agrupamentos nos dados, de forma a categorizar os pacientes;

- Utilização das categorias de pacientes identificadas e das informações sociodemográficas, clínicas e de avaliação aferida pelos instrumentos de medida de QdVRS, em cada momento do tratamento, para executar uma previsão da Qualidade de Vida dos pacientes, na ordem dos 70%;

5.4 Estudo dos Dados

5.4.1 Extração de dados

Os dados iniciais foram prontamente cedidos pela Optimizer, após um processo de extração executado sobre a base de dados do QoLIS, com base em *queries* SQL. Foram extraídas e disponibilizadas quatro tabelas, em formato CSV, que agregam informações relativas a pacientes com patologias oncológicas da cabeça e pescoço. Foram ainda disponibilizados documentos que disponibilizam as parametrizações de todos os atributos que compõem as tabelas.

5.4.2 Descrição das tabelas

As parametrizações referidas permitem identificar os diferentes conteúdos retratados nas tabelas disponibilizadas. Na tabela Patients são apresentadas informações que permitem a caracterização sociodemográfica dos pacientes. Na tabela Appointments estão contempladas informações recolhidas por meio de consulta sobre os pacientes, que visam informações clínicas que podem retratar e influenciar o estado e os tratamentos das patologias oncológicas em questão. A tabela Clinic_Info agrega informações relativas aos tratamentos de quimioterapia, radioterapia e cirurgia executados sobre os pacientes. Na tabela Answer estão contidas as respostas aos instrumentos de medida da QdVRS, distinguindo um total de 30 itens de resposta

para o questionário EORTC QLQ-C30 e um total de 35 itens de resposta para o questionário EORTC QLQ-H&N35. Importa salientar que a adaptação do modelo de Rasch, referida anteriormente, tem nesta tabela alguma influência, uma vez que o cálculo de QdV presente é da sua responsabilidade.

A tabela seguinte apresenta as tabelas enunciadas de acordo com o conteúdo identificado, bem como o número de atributos e registos presentes individualmente.

Tabela 4- Tabelas extraídas da plataforma QoLIS

Tabela	**Conteúdo**	**Nº Atributos**	**Nº Registos**
Patients	Informações sociodemográficas	10	1211
Appointments	Informações Clínicas registas nas consultas	32	1853
Clinic_info	Informações Clínicas, relacionadas com os tratamentos	22	164
Answer	Informações relativas aos questionários EORTC QLQ-C30 e EORTC QLQ-H&N35	43	2640

5.4.3 Processo de criação dos datasets

Numa análise superficial é possível identificar que as tabelas disponibilizadas se encontram intrinsecamente relacionadas, uma vez que retratam informações de interações entre o paciente e a instituição de saúde onde decorre o tratamento das suas patologias oncológicas. Para aumentar o conhecimento e entendimento da interação refletida nas diferentes tabelas, assim como para possibilitar uma análise mais pormenorizada e abrangente, foi executado um processo através da plataforma RapidMiner, que permitiu a criação de um único dataset, convergindo todos os atributos presentes nas tabelas. Foi atribuído ao Número de Identificação da Consulta (Appoint_No) o papel de identificador das diferentes interações, visto que o Número de Identificação do Paciente (Pat_No), em alguns casos, estava presente em mais do que uma consulta. No entanto, no dataset resultante desse processo, foi identificada a replicação do Número de identificação da Consulta (Appoint_No), uma vez que um paciente poderia responder a dois diferentes instrumentos de medida de QdVRS, o EORTC QLQ-C30 e o EORTC QLQ-H&N35.

Tabela 5- Exemplo da replicação detetada

Appoint_No	***Pat_No***	**…**	***Quest_code***	**…**	***Answers***
2	*410*	*…*	*C30*	*…*	*P1: 1, P2: 2 … P30: 2*
2	*410*	*…*	*HN35*	*…*	*P1: 1, P2: 2 … P35: 2*

Para solucionar esta replicação foi executado um novo processo na plataforma RapidMiner, com a mesma base operacional, mas que executou a criação de dois datasets (QLQ-C30, QLQ-H&N35), que se distinguiam tendo em conta o atributo identificador do instrumento de medida de QdVRS (Quest_code) presente.

Esta solução, devidamente discutida com todos os intervenientes do projeto de dissertação, partiu do pressuposto que numa fase mais adiantada do projeto de desenvolvimento do Sistema de Apoio à Decisão Clínica surgiriam outros instrumentos de medidas de QdVRS, que atuariam sobre a mesma Unidade, e esta abordagem simplificaria a integração das tarefas de categorização e de previsão objetivadas.

5.4.4 Descrição de Dados

Para alcançar os objetivos pretendidos, para a componente prática do presente projeto de dissertação, torna-se essencial aprofundar o conhecimento sobre os dados que integram os datasets QLQ-C30 e QLQ-H&N35.

Desta forma foram novamente consultadas as parametrizações cedidas pela Optimizer, assim como executadas tarefas de visualização através da plataforma RapidMiner, que permitiram realizar um relatório de descrição de dados, apresentado na tabela seguinte, que visa a descrição dos atributos e a identificação do seu formato, bem como a amostra dos registos que os preenchem.

Tabela 6- Descrição dos dados

Atributo	**Descrição**	**Formato**	**Amostra**	
			Dataset QLQ-C30	Dataset QLQ-H&N35
Appoint_No	Número de Identificação da Consulta	Numérico	[2:1911]	[1:1911]
Appoint_Date	Data da Consulta	Data	[2010/03/19:2012/10/04]	[2010/03/19:2012/10/04]
Pat_No	Número de Identificação do Paciente	Numérico	[10392581:35821519]	[10392581:35821519]
Date_Birth	Data de Nascimento	Data	[1900/01/01:2196/09/23]	[1900/01/01:2196/09/23]
County_Code	Código de Distrito	Numérico	[2000:2283]	[2000:2283]
Gender_Code	Género	String	(M) Masculino = 1078; (F) Feminino = 183; (I) Indiferenciado = 2;	(M) Masculino = 1177; (F) Feminino = 203; (I) Indiferenciado = 4;
Edu_Deg_Code	Código de Habilitações	Numérico	[1:9]	[1:9]
Profession_Code	Código de Profissão	Numérico	[1000:1535]	[1000:1535]

Date_D	Data de óbito	Data	-	-
Civil_Status_Code	Código de estado civil	Numérico	[1:6]	[1:6]
Breed_Code	Código da Raça	Numérico	-	-
Nationality_Code	Código da Nacionalidade	Numérico	-	-
Doctor_Code	Código do Médico responsável	Numérico	-	-
Date_Px	Data da Próxima consulta	Data	[2014/12/04:2014/12/20]	[2014/12/04:2014/12/20]
Uni_Code	Código da Unidade	String	(UN1) Unidade Pescoço e Cabeça = 1263	(UN1) Unidade Pescoço e Cabeça = 1384
T_Code	Código do tamanho do tumor	String	TX=43; T0=160; T1=169; T2=182; T3=105;	TX=47; T0=181; T1=183; T2=204; T3=105;
N_Code	Código da Metastização local	String	NX=301; N0=165; N1=120; N2=26;	NX=332; N0=172; N1=129; N2=30;
M_Code	Código da Metastização regional	String	MX=523; M0=24; M1= 2;	MX=570; M0=27; M1= 2;
Ik	Índice de Karnofsky	Numérico	[0:1]	[0:1]
Pad_Code	Código do diagnóstico anatomopatológico	Numérico	[0:9]	[0:9]
Mo_Code	Código do momento da consulta	String	(1A) 1 Ano=71; (2A) 2 Anos=80; (3A) 3 Anos=36; (3M) 3 Meses=74; (4A) 4 Anos=15; (5A) 5 Anos=10; (6M) 6 Meses=59; (9M) 9 Meses=45; (C1) 1ª Consulta=249; (CG) Consulta de Grupo=119;	(1A) 1 Ano=78; (2A) 2 Anos=85; (3A) 3 Anos=39; (3M) 3 Meses=83; (4A) 4 Anos=16; (5A) 5 Anos=11; (6M) 6 Meses=65; (9M) 9 Meses=47; (C1) 1ª Consulta=266; (CG) Consulta de Grupo=121;
Hist_Code	Código do diagnóstico histológico	Numérico	-	-
Beh_Code	Código do Comportamento	Numérico	-	-
Deg_Code	Código do grau do tumor	Numérico	-	-
Topo_Code	Código da topografia	Numérico	-	-

N_Rec	Número da recidiva (default 0) ex: se for 1, é a 1ª recidiva	Binário	0=1263; 1=0;	0=1384; 1=0;
Smokes	Hábitos tabágicos	Numérico	(0) Fumador=298; (1) Ex-Fumador=606; (2) Não Fumador=208;	Fumador=327; Ex-Fumador=644; Não Fumador=231;
Years_Smk	Anos de Fumador	Numérico	(0) Não Fuma=56; (1) [1-10]=31; (2) [11-20]=61; (3) [21-30]=181; (4) [31-40]=231; (5) [40 +]=195;	(0) Não Fuma=68; (1) [1-10]=33; (2) [11-20]=67; (3) [21-30]=194; (4) [31-40]=257; (5) [40 +]=208;
Num_Cig	Número de Cigarros	Numérico	(0) [0]=52; (1) [1-10]=126; (2) [11-20]=319; (3) [21-40]=207; (4) [41-60]=61; (5) [61+]=9;	(0) [0]=64; (1) [1-10]=137; (2) [11-20]=341; (3) [21-40]=228; (4) [41-60]=68; (5) [61+]=9;
Years_Stop_Smk	Anos desde que parou de fumar	Numérico	(0) [0-1]=269; (1) [2-5]=152; (2) [6-10]=71; (3) [11-15=42; (4) [16+]=69;	(0) [0-1]=294; (1) [2-5]=166; (2) [6-10]=75; (3) [11-15]=43; (4) [16+]=74;
Drinks	Hábitos alcoólicos	Numérico	-	-
Years_Drink	Anos de consumidor de bebidas alcoólicas	Numérico	(0) [0]=262; (1) [1]=33; (2) [2-5]=8; (3) [6-10]=7; (4) [11-20]]=64; (5) [21+]=375;	(0) [0]=291; (1) [1]=34; (2) [2-5]=9; (3) [6-10]=9; (4) [11-20]=72; (5) [21+]=415;
Num_Lit_Beer	Número de litros de cerveja	Numérico	(0) [0]=296; (1) [1]=114; (2) [2]=54; (3) [3-5]=53; (4) [6-10]=16; (5) [11+] =12;	(0) [0]=327; (1) [1]=131; (2) [2]=61; (3) [3-5]=58; (4) [6-10]=18; (5) [11+]=14;
Num_Lit_Alc	Número de litros de bebidas brancas	Numérico	(0) [0]=359; (1) [1]=98; (2) [2]=32; (3) [3-5]=17; (4) [6-10]=2; (5) [11+]=6;	(0) [0]=397; (1) [1]=107; (2) [2]=38; (3) [3-5]=19; (4) [6-10]=3; (5) [11+]=6;
Num_Lit_Wine	Número de litros de vinho	Numérico	(0) [0]=163; (1) [1]=372;	(0) [0]=188; (1) [1]=396;

			(2) [2]=94; (3) [3-5]=71; (4) [6-10]=8; (5) [11+]=5;	(2) [2]=107; (3) [3-5]=82; (4) [6-10]=5; (5) [11+]=5;
Years_Stop_Drk	Anos desde que parou de beber	Numérico	-	-
Trach	Traqueotomia	Binário	(0) Sim=231; (1) Não=977;	(0) Sim=247; (1) Não=1074;
Alim	Alimentação	Numérico	(0) PER/OS=1095; (1) PEG=95; (2) SNG=20;	(0) PER/OS=1202; (1) PEG=101; (2) SNG=21;
Voice_Prot	Prótese fonatória	Binário	(0) Sim=27; (1) Não=620;	(0) Sim=27; (1) Não=658;
Icd_Code	Código do ICD	Numérico	[31,1:141,0]	[0:161,9]
Info_Appoint_No	Número da consulta que se determinou a informação	Numérico	[123:1752]	[0:1752]
Chemo_Exists	Tratamento de quimioterapia	Binário	(0) Não=1243; (1) Sim=20;	(0) Não=1365; (1) Sim=19;
Chemo_Date_I	Data inicial	Data	[2006/05/15:2012/03/23]	[2006/05/15:2012/03/23]
Chemo_Date_E	Data final	Data	[2007/01/30:2012/05/30]	[2007/01/30:2012/05/30]
Chemo_Cp_Code	Código do Protocolo Quimioterapia	Numérico	[1:11]	[1:11]
Chemo_Bp	Bomba (0 - não;1-sim)	Binário	(0) Não=4; (1) Sim=17;	(0) Não=4; (1) Sim=16;
Chemo_Int_Type_Code	Código do tipo (adjuvante, neoadjuvante, paliativo)	Numérico	-	-
Radio_Exists	Tratamento de Radioterapia	Binário	(1) Sim=13; (0) N=1250;	(1) Sim=12; (0) Não=1372;
Radio_Date_I	Data inicial	Data	[2003/01/29:2011/11/17]	[2003/01/29:2011/11/17]
Radio_Date_E	Data final	Data	[2003/03/11:2012/06/10]	[2003/03/11:2012/06/10]
Radio_Descr	Descrição da radioterapia	String	-	-
Radio_Num_Field	Número de campos de irradiação	Numérico	[0:999]	[0:999]
Radio_Int_Type_Code	Código do tipo (adjuvante, neoadjuvante, radical/curativa, paliativo, remissiva, profilática, abelativa)	Numérico	-	-
Radio_Intens	Intensidade da radiação	Numérico	[0:999]	[0:999]

Radio_Freq	Frequência dos tratamentos	Numérico	[0:999]	[0:999]
Radio_Margin	1 - positiva, 0 – negativa	Numérico	-	-
Radio_Info_Appoint_No	Nr. Consulta em que se determinou a informação	Numérico	[261:1752]	[261:1752]
Surg_Exists	Cirurgia	Binário	0=1256;1=7;	0=1377;1=7;
Surg_Icd_Code	Código do icd	Numérico	[27,5:30,4]	[27,5:30,4]
Surg_Surgery_Date	Data da cirurgia	Data	[2007/03/02:2012/02/15]	[2009/03/02:2012/02/15]
Surg_Info_Appoint_No	Nr. Consulta em que se determinou a informação	Numérico	[261:1354]	[261:1354]
Answer_Date	Data da resposta (Data Questionário)	Data	[2010/03/19:2012/10/04]	[2010/03/19:2012/10/04]
Quest_Code	Código do Questionário	String	C30=1263;	C30=1384;
QdV	Valor da QdV Global Questionário (cálculo rasch)	Numérico	[22:100]	[26:75]
Outfit	(cálculo do rasch)	Numérico	[0:11,3]	[0:9,7]
Infit	(cálculo do rasch)	Numérico	[0:9,8]	[0:9,9]
Error	(cálculo do rasch)	Numérico	[0,2:1,2]	[0,2:0,5]
P1:30	Resposta por item ao questionário	Numérico	[0:4]	
P1:P35	Resposta por item ao questionário	Numérico		[0:4]

5.5 Tratamento dos dados

5.5.1 Qualidade dos dados

Durante a atividade de descrição de dados foram utilizadas ferramentas e técnicas de visualização, que permitiram identificar alguns problemas e incoerências relativamente aos atributos e registos presentes nos datasets QLQ-C30 e QLQ-H&N35. Os problemas mais comuns visam o facto de existirem atributos sem qualquer preenchimento ou com uma percentagem significativa de valores omissos, assim como incoerências relativamente ao formato de preenchimento dos registos. Todos os problemas e incoerências detetados encontram-se descritos no Anexo I - Relatório de Qualidade de Dados.

5.5.2 Inclusão/Exclusão de Dados

Tendo em conta os problemas identificados justifica-se o desenvolvimento da atividade de seleção de atributos, apresentada na tabela seguinte, que procederá à inclusão ou exclusão de atributos de acordo com critérios diretamente relacionados com a qualidade de dados apresentada, ou pela identificação de qualquer tipo de limitação relativamente à seleção e implementação dos modelos de Data Mining, presentes no *software* Weka, não descurando a exclusão de atributos que não se adequem aos objetivos pretendidos.

Atributo	**Inclusão**	**Razão para Inclusão/Exclusão**
Appoint_No	Não	Apenas identifica a consulta, não acrescentado informação relevante de acordo com os objetivos pretendidos
Appoint_Date	Não	Apenas identifica a data de realização da consulta, não acrescentado informação relevante de acordo com os objetivos pretendidos
Pat_No	Sim	Atributo necessário para identificação dos registos sujeitos a tarefas de categorização e classificação objetivadas
Date_Birth	Sim	Atributo caracterizador do paciente
County_Code	Não	Apenas identifica distrito de residência do paciente, não acrescentado informação relevante de acordo com os objetivos pretendidos
Gender_Code	Sim	Atributo caracterizador do paciente
Edu_Deg_Code	Sim	Atributo caracterizador do paciente
Profession_Code	Não	Excessiva variância do atributo
Date_D	Não	Não apresenta qualquer valor
Civil_Status_Code	Sim	Atributo caracterizador do paciente
Breed_Code	Não	Não apresenta qualquer valor
Nationality_Code	Não	Não apresenta qualquer valor
Doctor_Code	Não	Não apresenta qualquer valor
Date_Px	Não	Não apresenta qualquer valor
Uni_Code	Não	Todos os registos pertencem ao mesmo módulo (UN1), não acrescentado informação de acordo com os objetivos pretendidos

T_Code	Sim	Atributo caracterizador do tumor do paciente
N_Code	Sim	Atributo caracterizador do tumor do paciente
M_Code	Sim	Atributo caracterizador do tumor do paciente
Ik	Não	Apresenta > 90 % de valores omissos e incongruentes
Pad_Code	Sim	Atributo com informação relevante para a patologia oncológica de cabeça e pescoço
Mo_Code	Sim	Identifica o momento da consulta
Hist_Code	Não	Não apresenta quaisquer valores
Beh_Code	Não	Não apresenta quaisquer valores
Deg_Code	Não	Não apresenta quaisquer valores
Topo_Code	Não	Não apresenta quaisquer valores
N_Rec	Não	Todos os registos apresentam o mesmo valor – default, não acrescentado informação de acordo com os objetivos pretendidos
Smokes	Sim	Atributo com informação relevante para a patologia oncológica de cabeça e pescoço
Years_Smk	Sim	Atributo com informação relevante para a patologia oncológica de cabeça e pescoço
Num_Cig	Sim	Atributo com informação relevante para a patologia oncológica de cabeça e pescoço
Years_Stop_Smk	Sim	Atributo com informação relevante para a patologia oncológica de cabeça e pescoço
Drinks	Não	Não apresenta qualquer valor
Years_Drink	Não	Dependente do atributo “Drinks”
Num_Lit_Beer	Não	Dependente do atributo “Drinks”
Num_Lit_Alc	Não	Dependente do atributo “Drinks”
Num_Lit_Wine	Não	Dependente do atributo “Drinks”
Years_Stop_Drk	Não	Dependente do atributo “Drinks”, não apresentado qualquer valor
Trach	Sim	Atributo com informação clínica relevante para a patologia oncológica de cabeça e pescoço

Alim	Sim	Atributo com informação clínica relevante para a patologia oncológica de cabeça e pescoço
Voice_Prot	Sim	Atributo com informação clínica relevante para a patologia oncológica de cabeça e pescoço
Icd_Code	Sim	Atributo com informação relevante para a patologia oncológica de cabeça e pescoço
Info_Appoint_No	Não	Apenas identifica em que consulta foi registada a informação clínica, não acrescentado informação de acordo com os objetivos pretendidos
Chemo_Exists	Sim	Atributo que especifica se existiu a realização de tratamento
Chemo_Date_I	Não	Apenas identifica a data em que se iniciou o tratamento de quimioterapia, não acrescentando informação relevante de acordo com os objetivos pretendidos
Chemo_Date_E	Não	Apenas identifica a data em que terminou o tratamento de quimioterapia, não acrescentando informação relevante de acordo com os objetivos pretendidos
Chemo_Cp_Code	Sim	Atributo com informação relativa ao tratamento efetuado
Chemo_Bp	Não	Apresenta > 90% de valores omissos
Chemo_Int_Type_Code	Não	Não apresenta qualquer valor
Radio_Exists	Sim	Atributo que especifica se existiu a realização de tratamento
Radio_Date_I	Não	Apenas identifica a data em que se iniciou o tratamento de radioterapia, não acrescentando informação relevante de acordo com os objetivos pretendidos
Radio_Date_E	Não	Apenas identifica a data em que terminou o tratamento de radioterapia, não acrescentando informação relevante de acordo com os objetivos pretendidos

Radio_Descr	Não	Não apresenta qualquer valor
Radio_Num_Field	Não	Apresenta > 90% de valores omissos e incongruentes
Radio_Int_Type_Code	Não	Apresenta > 90% de valores omissos e incongruentes
Radio_Intens	Não	Apresenta > 90% de valores omissos e incongruentes
Radio_Freq	Não	Apresenta > 90% de valores omissos e incongruentes
Radio_Margin	Não	Não apresenta qualquer valor
Radio_Info_Appoint_No	Não	Apenas identifica em que consulta foi registada a informação relativa ao tratamento de radioterapia, não acrescentando informação relevante de acordo com os objetivos pretendidos
Surg_Exists	Sim	Atributo que especifica se existiu a realização de tratamento
Surg_Icd_Code	Sim	Atributo com informação relativa ao tratamento efetuado
Surg_Surgery_Date	Não	Apenas identifica a data em que realizou a cirurgia, não acrescentando informação relevante de acordo com os objetivos pretendidos
Surg_Info_Appoint_No	Não	Apenas identifica em que consulta foi registada a informação relativa à cirurgia, não acrescentando informação relevante de acordo com os objetivos pretendidos
Answer_Date	Não	Apenas identifica a data de resposta aos instrumentos de medida de QdVRS, não acrescentando informação relevante de acordo com os objetivos pretendidos
Quest_Code	Não	Os datasets já se encontram distinguidos
QdV	Sim	Apresenta relevância para o objetivo de previsão pretendido

Outfit	Não	Apenas regista medidas estatística no cálculo do modelo de Rasch, não acrescentando informação relevante de acordo com os objetivos pretendidos
Infit	Não	Apenas regista medidas estatística no cálculo do modelo de Rasch, não acrescentando informação relevante de acordo com os objetivos pretendidos
Error	Não	Apenas regista medidas estatística no cálculo do modelo de Rasch, não acrescentando informação relevante de acordo com os objetivos pretendidos
P1:30	Sim	Apresenta relevância para os objetivos de categorização e previsão pretendidos
P1:P35	Sim	Apresenta relevância para os objetivos de categorização e previsão pretendidos

5.5.3 Limpeza de dados

Após a clara identificação dos atributos necessários para alcançar as tarefas de categorização e previsão proposta, impõe-se a necessidade de proceder a operações de limpeza de dados, com o intuito de melhorar a qualidade de dados apresentada nos datasets QLQ-C30 e QLQ-H&N35. Desta forma são apresentados na tabela seguinte os tratamentos efetuados aos atributos e registos, tendo em conta os problemas e incoerências refletidas no relatório de qualidade dos dados.

Atributo	**Valores Omissos (%)**		**Qualidade dos Dados**	**Tratamento efetuado**
	Dataset QLQ-C30	Dataset QLQ-HN35		
Pat_No	0%	0%	Nenhum problema detetado	
Date_Birth	0%	0%	Datas em formato textual	Converter para formato data aaaa/mm/dd
Gender_Code	0%	0%	Presença do valor “Indiferenciado”	De acordo com a moda, transformar o valor em "Masculino”

Edu_Deg_Code	2%	2%	Presença de Valores Omissos	Atribuir o código "9999" parametrizando-o como "Desconhecido"
Civil_Status_Code	2%	2%	Presença de Valores Omissos	Eliminar linhas onde se verifica a presença de valores omissos
T_Code	48%	48%	Presença de valores omissos	Eliminar linhas onde se verifique a presença de valores omissos
N_Code	52%	52%	Presença de valores omissos	Eliminar linhas onde se verifique a presença de valores omissos
M_Code	57%	57%	Presença de valores omissos	Eliminar linhas onde se verifique a presença de valores omissos
Pad_Code	21%	21%	Presença de Valores Omissos	Atribuir o código "10" parametrizando-o como "Desconhecido"
Mo_Code	40%	41%	Presença de Valores Omissos	Atribuir o código "Desc" parametrizando-o como "Desconhecido"
Smokes	12%	13%	Presença de Valores Omissos e incongruentes	Consultar conjunto de decisões 1
Years_Smk	40%	40%	Presença de Valores Omissos e incongruentes	
Num_Cig	39%	39%	Presença de Valores Omissos e incongruentes	

Years_Stop_Smk	52%	53%	Presença de Valores Omissos e incongruentes	
Trach	4%	5%	Presença de Valores Omissos	Transformar Valores Omissos no Código “0”
Alim	4%	4%	Presença de Valores Omissos	Transformar Valores Omissos no Código “0”
Voice_Prot	49%	51%	Presença de Valores Omissos	Consultar conjunto de decisões 2
Icd_Code	96%	97%	Presença de Valores Omissos	Atribuir o código “0” parametrizando-o como “Desconhecido”
Chemo_Exists	98%	99%	Valores correspondentes a “0” aparecem em branco	Transformar valores em branco no valor “0”
Chemo_Cp_Code	98%	99%	Presença de Valores Omissos	Atribuir o código “0” parametrizando-o como “Desconhecido”
Radio_Exists	99%	99%	Valores correspondentes a “0” aparecem em branco	Transformar valores em branco no valor “0”
Surg_Exists	99%	99%	Valores correspondentes a “0” aparecem em branco	Transformar valores em branco no valor “0”
Surg_Icd_Code	99%	99%	Presença de Valores Omissos	Atribuir o código “0” parametrizando-o como “Desconhecido”
QdV	1%	1%	Presença de Valores Omissos	Eliminar linhas onde se verifique a presença de valores omissos

P1:30	2,8%	0%	Presença de valores incongruentes	Eliminar linhas onde se verifique a presença de valores incongruentes
P1:P35		3%	Presença de valores Incongruentes	Eliminar linhas onde se verifique a presença de valores incongruentes

Conjunto de Decisões 1:

- No atributo "**SMOKES**" os valores omissos "-1" transformar em "1" quando no atributo "**YEARS_STOP_SMK**" o seu valor for diferente de "-1";
- No atributo "**SMOKES**", os valores omissos "-1" transformar em "0", quando nos atributos "**YEARS_SMK**" e "**NUM_CIG**" o seu valor for maior que "0" e "**YEARS_STOP_SMK**" for "-1";
- No atributo "**SMOKES**" os valores omissos "-1" transformar em "2" quando o atributo "**Num_Cig**" tiver valores iguais a "0" e "-1";
- No atributo "**SMOKES**" os valores omissos "-1" transformar em "0" quando o atributo "**NUM_CIG**" tiver valores iguais a "3" e "2";
- Se no atributo "**SMOKES**" o valor for igual a "2", ou seja, Não Fumador, transformar os valores de "**YEARS_STOP_SMK**" em "-1";
- No atributo "**SMOKES**", se o valor for igual a "0", ou seja, Fumador, mas o atributo **"YEARS_STOP_SMK"** tiver valores iguais ou superiores a "1", transformar o valor de "**SMOKES**" em "1", ou seja, Ex-Fumador;
- No atributo "**SMOKES**" se o valor for igual a "0", ou seja, Fumador, transformar os valores omissos do atributo "**YEARS_SMK"** em "4", uma vez que é a moda, e os valores omissos do atributo "**NUM_CIG**" em "3", uma vez que também é o valor mais verificado;
- No atributo "**SMOKES**" se o valor for igual a "0", ou seja, Fumador, mas o atributo "**YEARS_STOP_SMK**" tiver valores iguais ou superiores a "0", transformar esses valores de **"SMOKES"** em "1", ou seja, Ex Fumador;
- No atributo "**SMOKES"** se o valor for igual a "1", ou seja, Ex Fumador, transformar os valores omissos do atributo "**YEARS_STOP_SMK**" em "0" que é a moda;

- No atributo "**SMOKES**" se o valor for igual a "1", ou seja, Ex Fumador, transformar os valores omissos e iguais a "0" em "4", uma vez que é a moda, assim como os valores omissos e iguais a "0" em "3" no atributo "**NUM_CIG**" pelo mesmo motivo;

Conjunto de Decisões 2

Os valores omissos do atributo "**Voice_Prot**" são influenciados pelo atributo "**Trach**", como tal, quando no atributo "**TRACH**" se verificar que o valor é igual a "0" o valor omisso do atributo "**Voice_Prot**" será "0". Para os restantes valores omissos do atributo de "**Voice_Prot**" o valor será "1".

5.5.4 Construção e Transformação de Dados

A tarefa de construção e transformação de dados pretende adequar as informações presentes nos Dataset QLQ-C30 e QLQ-H&N35 aos objetivos de Data Mining, delineados para o presente projeto de dissertação, tendo ainda em conta as limitações que existam relativamente à utilização de algumas técnicas e modelos de Data Mining, que se pretendem implementar.

Para atingir o objetivo que visa a utilização de informações sociodemográficas, clínicas e dos instrumentos de medida de QdVRS na fase anterior ao tratamento, com intuito de detetar e identificar agrupamentos nos dados, de forma a categorizar os pacientes, foi necessário proceder à criação e transformação de alguns dados, utilizando dados já presentes nos Datasets QLQ-C30 E QLQ-H&N35.

A nível da informação sociodemográfica foram tidas em conta informações que pudessem refletir a faixa etária, o género, o nível de educação e o estado civil dos pacientes. A nível das informações relativas aos hábitos tabágicos apenas se precisa identificar se o paciente é considerado fumador ou não fumador. A nível da informação relativa aos instrumentos de medida de QdVRS, a opção recaiu pelas dimensões, sintomas e avaliação da QdV, aferidos através de um procedimento de cálculo específico aos questionários EORTC QLQ-C30 e EORTC QLQ-H&N35.

Para extrair informação relativa às faixas etárias foi necessário proceder à construção de um novo atributo, em formato String, denominado **Age_Group**. Numa primeira fase foram calculadas as idades dos pacientes, através das datas de nascimento contempladas no atributo **Date_Birth**. Seguidamente as idades foram agrupados nas seguintes faixas etárias:

- [18-25] – Entre os 18 e 25 anos;

- [26-35] – Entre os 26 e os 35 anos;
- [36-45] – Entre os 36 e os 45 anos;
- [46-55] – Entre os 46 e os 55 anos;
- [56-65] – Entre os 56 e os 65 anos;
- [66-75] – Entre os 66 e 75 anos;
- [76-85] – Entre os 76 e 85 anos;
- [86-95] – Entre os 86 e 95 anos;
- [95+] – Superior a 95 anos;

O atributo **Edu_Deg_Code**, que apresenta informações relativas às habilitações académicas dos pacientes, sofreu uma transformação no seu formato, de numérico para String, de acordo com as seguintes parametrizações:

- Desconhecido – quando apresenta o valor 9999;
- Analfabeto – quando apresenta o valor 1;
- 4Ano – quando apresenta os valores 2 e 3;
- 9Ano – quando apresenta os valores 4 e 5;
- 12Ano - quando apresenta os valores 6 e 7;
- Universidade – quando apresenta os valores 8 e 9;

Relativamente ao atributo **Civil_Status_Code**, que apresenta informações relativas ao estado civil do paciente, existiu a transformação no seu formato, de numérico para string, de acordo com as seguintes parametrizações:

- Acompanhado – para os valores 2, 3, 7 e 8;
- Sozinho – para os valores 1,4,5 e 6;

Sobre o atributo **Gender_Code,** não foi necessário proceder a nenhuma transformação.

As informações relativas aos hábitos tabágicos, representados nos atributos **Smokes**, **Years_Smk**, **Num_Cig** e **Years_Stop_Smk**, deram origem a um novo atributo, de formato binário, designado **Smokes_YorN**, com o intuito de perceber se o paciente é considerado fumador (Y) ou Não Fumador (N),seguindo as seguintes parametrizações:

- **N** – para o valor 2 (Não Fumador) no atributo **Smokes**, independentemente dos valores dos restantes atributos;
- **Y** – para o valor 1 (Ex-Fumador) no atributo **Smokes**, combinado com os valores iguais e superiores a 2 no atributo **Years_Smk**, idendependentemente dos valores do atributo **Num_Cig**;
- **N** – para o valor 1 (Ex-Fumador) no atributo **Smokes**, combinado com os valores iguais e inferiores a 1 no atributo **Years_Smk**, independentemente dos valores do atributo **Num_Cig**;
- **Y** – para o valor 0 (Fumador) no atributo **Smokes**, independentemente dos valores dos restantes atributos;

A nível das dimensões, sintomas e avaliação da QdV, a transformação realizou-se tendo por base o Manual de Avaliação da Organização Europeia para Pesquisa e Tratamento do Cancro (EORTC, 2001), que contempla os procedimentos para o cálculo das escalas e itens dos questionários EORTC QLQ-C30 e EORTC QLQ-H&N35. Todas as escalas e itens simples são classificados no intervalo de 0 a 100. No cálculo das escalas funcionais, um valor de classificação mais alto representa um nível mais saudável de funcionamento, assim como um nível mais alto de classificação da escala de QdV representa uma maior Qualidade de Vida. Contrariamente, um nível mais alto nas escalas e itens simples dos sintomas representam um maior nível de problemas relacionados com os sintomas especificados.

O cálculo presente no Manual de Avaliação da EORTC desenvolve-se através dos seguintes procedimentos:

Cálculo da Pontuação Bruta

Quando os itens de resposta $P1, P2, \dots, Pn$ estão incluídos nas escalas:

$$Pontuação\ Bruta = PB = (P1 + P2 + \cdots + Pn)/n$$

Transformação Linear

Para obter o Resultado (R) das escalas funcionais, de 0 a 100:

$$R = \left\{1 - \frac{(PB - 1)}{intervalo}\right\} * 100$$

Para obter o Resultado (R) dos Sintomas e da avaliação da QdV, de 0 a 100:

$$R = \{(PB - 1)/intervalo\} * 100$$

O $intervalo$ é a diferença entre o valor máximo e mínimo possível da Pontuação Bruta (PB).

No questionário EORTC QLQ-C30, as escalas são aferidas pelas seguintes parametrizações:

Tabela 7- Procedimento de cálculo para o questionário EORTC QLQ-C30 (adaptado de EORTC,2001)

Nome da Escala	**Número de itens (n)**	**Intervalo**	**Itens de Resposta (Versão 3.0)**
Qualidade de Vida			
QdV	2	6	P29, P30
Escalas Funcionais			
Física	5	3	P1 a P5
Funcional	2	3	P6, P7
Emocional	4	3	P21 a P24
Cognitiva	2	3	P20, P25
Social	2	3	P26,P27
Sintomas			
Fadiga	3	3	P10,P12,P18
Náusea	2	3	P14,P15
Dor	2	3	P9,P19
Dispneia	1	3	P8
Insónia	1	3	P11
Perda de Apetite	1	3	P13
Obstipação	1	3	P16
Diarreia	1	3	P17
Dificuldade Financeira	1	3	P28

No questionário EORTC QLQ-H&N35, as escalas são aferidas pelas seguintes parametrizações, seguindo os procedimentos de cálculo utilizados nos sintomas do questionário EORTC QLQ-C30:

Tabela 8 - Procedimento de cálculo para o questionário EORTC QLQ-H&N35 (adaptado de EORTC,2001)

Nome da Escala	Número de itens	Intervalo	Itens de Resposta (Versão 3.0)
Dor	4	3	P1 a P4
Deglutição	4	3	P5 a P8
Problemas nos Sentidos	2	3	P13,P14
Problemas na Fala	3	3	P16,P23,P24
Alimentação em público	4	3	P19 a P22
Função Social	5	3	P18, P25 a P28
Sexualidade	2	3	P29,P30
Dentes	1	3	P9
Abertura da boca	1	3	P10
Boca Seca	1	3	P11
Saliva Pegajosa	1	3	P12
Tosse	1	3	P15
Sentimento de Doença	1	3	P17
Medicamentos	1	1	P31
Suplementos nutricionais	1	1	P32
Tubo para alimentação	1	1	P33
Perda de Peso	1	1	P34
Ganho de Peso	1	1	P35

Para o objetivo que visa a utilização das categorias de pacientes e das informações sociodemográficas, clínicas e de avaliação aferidas pelos instrumentos de medida de QdVRS, em cada momento do tratamento, de forma a executar uma previsão de QdV dos pacientes, na ordem dos 70%, serão tidos em conta os atributos sociodemográficos utilizados no objetivo anterior, assim como os atributos que já integram os datasets QLQ-C30 e QLQ-H&N35, representativos das informações clínicas existentes.

As informações relativas aos hábitos tabágicos dos pacientes seguem as parametrizações executadas anteriormente, na construção do atributo **Smokes_YorN**.

As informações clínicas representadas pelos atributos **Trach** e **Voice_Prot**, representativas da realização de traqueotomia e da colocação de prótese fonatória, foram transformadas de acordo com as seguintes parametrizações:

- **Y** – para o valor 0;
- **N** – para o valor1;

As informações relativas à existência de tratamento de quimioterapia, radioterapia e cirurgia, representados pelos atributos **Chemo_Exists**, **Radio_Exists**, **Surg_Exists**, respestivamente, froram transformados de acordo com as seguintes parametrizações:

- **Y** – para o valor 1;
- **N -** para o valor 0;

O atributo **Alim**, representativo do tipo de alimentação, sofreu uma transformação a nível de formato, de numérico para string, de acordo com os seguintes parâmetros:

- **PER/OS** – para o valor 0;
- **PEG** – para o valor 1;
- **SNG** – para o valor 2;

Os atributos **T_Code**, **N_Code**, **M_Code**, **Pad_Code**, **Mo_Code**, **Icd_Code**, **Chemo_Cp_Code**, **Surg_Icd_Code** e **QdV**, representativos de informações clínicas relativas às patologias oncológicas e aos seus tratamentos, não necessitaram de qualquer intervenção nesta fase.

Na tabela seguinte são apresentadas as construções e transformações executadas sobre os Datasets QLQ-C30 e QLQ-H&N35, bem como os novos atributos que os compõem.

Tabela 9- Resumo das construções e transformações executadas aos datasets

Atributos	Formato	Construção	Transformação	Atributos	Formato
Dados comuns aos dois Datasets					
Pat_No	Numérico			Pat_No	Numérico
Date_Birth	Data	X	X	Age_Group	String
Gender_Code	String			Gender_Code	String
Edu_Deg_Code	Numérico		X	Edu_Deg_Code	String
Civil_Status_Code	Numérico		X	Civil_Status_Code	String
T_Code	String			T_Code	String
N_Code	String			N_Code	String
M_Code	String			M_Code	String
Pad_Code	Numérico			Pad_Code	Numérico
Mo_Code	String			Mo_Code	String
Smokes	Numérico	X		Smokes_YorN	Binário

Years_Smk	Numérico	X			
Num_Cig	Numérico	X			
Years_Stop_Smk	Numérico	X			
Trach	Binário		X	Trach	Binário
Alim	String		X	Alim	String
Voice_Prot	Binário		X	Voice_Prot	Binário
Icd_Code	Numérico			Icd_Code	Numérico
Chemo_Exists	Binário		X	Chemo_Exists	Binário
Chemo_Cp_Code	Numérico			Chemo_Cp_Code	Numérico
Radio_Exists	Binário		X	Radio_Exists	String
Surg_Exists	Binário		X	Surg_Exists	String
Surg_Icd_Code	Numérico			Surg_Icd_Code	Numérico
QdV	Numérico			QdV	Numérico
P1:30 **P1:35**	Numérico		X	P1:30 P1:35	Numérico
<u>Novos dados do Dataset-QLQ C30</u>					
P29, P30	Numérico	X	X	Escala_QdV	Numérico
P1 a P5	Numérico	X	X	Dimensão_Física	Numérico
P6, P7	Numérico	X	X	Dimensão_Funcional	Numérico
P21 a P24	Numérico	X	X	Dimensão_Emocional	Numérico
P20 a P25	Numérico	X	X	Dimensão_Cognitiva	Numérico
P26 a P27	Numérico	X	X	Dimensão_Social	Numérico
P10,P12,P18	Numérico	X	X	Fadiga	Numérico
P14,P15	Numérico	X	X	Náusea	Numérico
P9,P19	Numérico	X	X	Dor	Numérico
P8	Numérico	X	X	Dispneia	Numérico
P11	Numérico	X	X	Insónia	Numérico
P13	Numérico	X	X	Perda de Apetite	Numérico
P16	Numérico	X	X	Obstipação	Numérico
P17	Numérico	X	X	Diarreia	Numérico
P28	Numérico	X	X	Dificuldade Financeira	Numérico
<u>Novos dados do Dataset-QLQ H&N35</u>					
P1 a P4	Numérico	X	X	Dor	Numérico

P5 a P8	Numérico	X	X	Deglutição	Numérico
P13,P14	Numérico	X	X	Problemas nos Sentidos	Numérico
P16,P23,P24	Numérico	X	X	Problemas na Fala	Numérico
P19 a P22	Numérico	X	X	Alimentação em público	Numérico
P18, P25 a P28	Numérico	X	X	Função Social	Numérico
P29,P30	Numérico	X	X	Sexualidade	Numérico
P9	Numérico	X	X	Dentes	Numérico
P10	Numérico	X	X	Abertura da boca	Numérico
P11	Numérico	X	X	Boca Seca	Numérico
P12	Numérico	X	X	Saliva Pegajosa	Numérico
P15	Numérico	X	X	Tosse	Numérico
P17	Numérico	X	X	Sentimento de Doença	Numérico
P31	Numérico	X	X	Medicamentos	Numérico
P32	Numérico	X	X	Suplementos nutricionais	Numérico

5.6 Integração dos dados

A Optimizer acompanhou com especial atenção as atividades que visaram o tratamento de dados e a sua consequente melhoria de qualidade, refletindo-as na base de dados da plataforma QoLIS. Adicionalmente procedeu à extração de um novo dataset (C30Rasch) que apresenta uma estrutura idêntica à do dataset QLQ-C30, sendo que o cálculo das escalas, sintomas e avaliação da QdV refletem os cálculos efetuados pelo modelo matemático de Rasch presente na plataforma QoLIS. O dataset em questão foi devidamente integrado no presente projeto de dissertação, com o intuito de ser apenas sujeito ao objetivo de categorização, uma vez que apenas dispõe de informações sociodemográficas (Age_Group, Gender_Code, Edu_Deg_Code, Civil_Status_Code), clínicas (Smokes_YorN), e dos instrumentos de medida de QdVRS (Escalas, Sintomas e QdV, através do cálculo do modelo matemático de Rasch adaptado ao SADC QoLIS), referentes ao momento anterior ao tratamento. Importa salientar que não se prevê efetuar qualquer tipo de melhoria na qualidade de dados apresentada.

5.7 Modelação

5.7.1 Seleção de técnicas de Data Mining

A seleção de técnicas e modelos, com capacidade para atingir os objetivos definidos para o presente projeto de dissertação, caracteriza-se como uma das atividades mais importantes na descoberta de conhecimento pretendido. O sucesso do seu desenvolvimento está intrinsecamente dependente das características dos conjuntos de dados e das capacidades do software responsável pela implementação das técnicas e modelos selecionados. Desta forma, e analisando os objetivos de Data Mining, que visam a categorização e previsão de QdV dos pacientes oncológicos, impõe-se a necessidade de recorrer a técnicas de Clustering e Classificação.

O desenvolvimento de Clusters será executado através do modelo Simple k-Means (k-vizinhos-mais-próximos), enquanto a Classificação considerará modelos inerentes às Árvores de Decisão, Regras de Associação, Redes Neuronais, K-vizinhos-mais-próximo e aos classificadores de Bayes.

5.7.2 Desenvolver cenários de teste

Os cenários de teste a desenvolver devem distinguir os diferentes objetivos de Data Mining definidos. Desta forma, o primeiro cenário de teste visa o objetivo de categorização dos pacientes, enquanto o segundo cenário de teste visa a previsão de QdV dos pacientes.

Cenário 1 - Categorização dos Pacientes

A categorização dos pacientes será executada através do modelo Simple K-means, responsável pela deteção de agrupamentos nos dados presentes nos datasets QLQ-C30, QLQ-H&N35 e C30Rasch. Importa salientar que a categorização dos pacientes visa apenas a utilização de informação sociodemográfica, dos hábitos tabágicos e dos instrumentos de medida de QdVRS, na fase anterior ao tratamento. Desta forma impõe-se a necessidade de um critério de seleção relativo ao atributo Mo_Code, representativo do Momento da Consulta, com o valor igual a "CG" valor representativo da Consulta de Grupo efetuada antes dos tratamentos.

Na tabela seguinte são apresentados os atributos necessários ao desenvolvimento de Clusters.

Tabela 10 - Atributos necessários ao desenvolvimento de Clusters

Mo_Code = CG	Dataset QLQ-C30	Dataset QLQ-H&N35	Dataset C30Rasch
	Pat_No	Pat_No	Age_Group
	Age_Group	Age_Group	Gender_Code
	Gender_Code	Gender_Code	Edu_Deg_Code
	Edu_Deg_Code	Edu_Deg_Code	Civil_Status_Code
	Civil_Status_Code	Civil_Status_Code	Smokes_YorN
	Smokes_YorN	Smokes_YorN	Dimensão_Física
	Dimensão_Física	Dor	Dimensão_Funcional
	Dimensão_Funcional	Deglutição	Dimensão_Emocional
	Dimensão_Emocional	Problemas nos Sentidos	Dimensão_Cognitiva
	Dimensão_Cognitiva	Problemas na Fala	Dimensão_Social
	Dimensão_Social	Alimentação em público	Fadiga
	Fadiga	Função Social	Náusea
	Náusea	Sexualidade	Dor
	Dor	Dentes	Dispneia
	Dispneia	Abertura da boca	Insónia
	Insónia	Boca Seca	Perda de Apetite
	Perda de Apetite	Saliva Pegajosa	Obstipação
	Obstipação	Tosse	Diarreia
	Diarreia	Sentimento de Doença	Dificuldade Financeira
	Dificuldade Financeira	Medicamentos	Escala_QdV
	Escala_QdV	Suplementos nutricionais	

Cluster
Categorização

Na execução do modelo, Simple K-Means, serão testados agrupamentos de 3 a 6 clusters, através da medida de distância de Euclidean, utilizando os atributos representados na tabela.

Importa salientar que na avaliação do modelo Simple K-Means, apenas serão tidos em conta agrupamentos com percentagem igual ou superior a 10%. O modelo que apresentar melhores resultados, através de uma combinação entre a percentagem de distribuição dos clusters e o

menor erro quadrático médio, será sujeito a tarefas de classificação (Árvores de Decisão, Redes Neuronais, Regras de Associação), com intuito de perceber se a capacidade preditiva do modelo relativamente aos clusters permitirá que na presença de novos dados não seja necessária proceder ao desenvolvimento de novos clusters com vista à categorização dos pacientes, mas à sua previsão.

Cenário 2 – Previsão de QdV

O desenvolvimento do segundo cenário, que visa a previsão de QdV, está dependente da categorização dos pacientes contemplada no primeiro cenário, uma vez que essa categorização, juntamente com as informações sociodemográficas, clínicas e dos instrumentos de medida de QdVRS (Escalas Funcionais, Sintomas, Escala de QdV) é utilizada para alcançar uma previsão de QdV, com valores percentuais a rondar os 70%, em cada fase do tratamento, utilizando o atributo Mo_Code presente nos datasets QLQ-C30 e QLQ-H&N35 para as distinguir.

Devido ao número reduzido de registos presentes nos datasets, apenas será utilizado o processo de validação k-fold Cross-Validation, com k=5,onde k é o número de subconjuntos de dados e número de iterações necessários à execução da técnica de Classificação. Esse processo desenvolve-se através da divisão do conjunto de dados utilizado, de forma aleatória, para validação do algoritmo em K subconjuntos de dados igualmente distribuídos. Após essa divisão, iniciam-se os testes submetendo um dos subconjuntos de dados ao algoritmo pretendido, para que este os classifique (conjunto de dados de teste), e os restantes como conjunto de dados de treino. Desta forma todos os dados são utilizados para treino e para teste do algoritmo.

A execução dos modelos de classificação será executada de acordo com as configurações padrão contempladas no software Weka.

Na tabela seguinte são apresentados os atributos que integram os datasets submetidos às tarefas de classificação referidas.

Tabela 11- Atributos necessário à execução de tarefas de classificação

Dataset QLQ-C30	Dataset QLQ-H&N35
Categorização	Categorização
Pat_No	Pat_No
Age_Group	Age_Group
Gender_Code	Gender_Code

Edu_Deg_Code	Edu_Deg_Code
Civil_Status_Code	Civil_Status_Code
T_Code	T_Code
N_Code	N_Code
M_Code	M_Code
Pad_Code	Pad_Code
Mo_Code	Mo_Code
Description_Smokes	Description_Smokes
Trach	Trach
Alim	Alim
Voice_Prot	Voice_Prot
Icd_Code	Icd_Code
Chemo_Exists	Chemo_Exists
Chemo_Cp_Code	Chemo_Cp_Code
Radio_Exists	Radio_Exists
Surg_Exists	Surg_Exists
Surg_Icd_Code	Surg_Icd_Code
Dimensão_Física	Dor
Dimensão_Funcional	Deglutição
Dimensão_Emocional	Problemas nos Sentidos
Dimensão_Cognitiva	Problemas na Fala
Dimensão_Social	Alimentação em público
Fadiga	Função Social
Náusea	Sexualidade
Dor	Dentes
Dispneia	Abertura da boca
Insónia	Boca Seca
Perda de Apetite	Saliva Pegajosa
Obstipação	Tosse
Diarreia	Sentimento de Doença
Dificuldade Financeira	Medicamentos
Escala_QdV	Suplementos nutricionais
QdV	QdV

Considerando a necessidade de atribuição de uma classe aos registos pertences ao conjunto de dados de treino, o atributo **QdV**, utilizado nesta tarefa de classificação, sofrerá uma alteração no seu formato, de numérico para String, para que os modelos pretendidos possam ser executados. A alteração referida segue as seguintes parametrizações:

Tabela 12- Transformação do atributo QdV

Valores	**0 – 15**	**16 – 30**	**31 - 45**	**46 - 60**	**61 - 75**	**76 – 90**	**91 - 100**
Transformação	[0-15]	[16-30]	[31-45]	[46-60]	[61-75]	[76-90]	[91-100]

Saliente-se que o objetivo que visa a previsão de QdV está limitado pela obrigatoriedade de existir uma categorização dos pacientes (contemplada no cenário 1), que visa apenas o momento anterior aos tratamentos das patologias oncológicas. Desta forma torna-se impossível utilizar uma parte considerável das informações agregadas dos datasets QLQ-C30 e QLQ-H&N35, visto que apenas os registos que consideram o atributo **Mo_Code** com o valor "CG" foram categorizados, não existindo nos conjuntos de dados registos que retratem um acompanhamento de nenhum dos pacientes ao longo das diferentes fases de tratamento. Acrescenta-se ainda a limitação de executar o modelo de previsão distinguido a categorização atribuída, com o intuito de perceber de que forma é influenciada a previsão de QdV.

A avaliação dos modelos de classificação visará a percentagem de acerto do modelo e as métricas de Sensibilidade e Especificidade, aferidas pelos valores presentes na Matriz de Confusão, não descurando a velocidade de aprendizagem do modelo, e os valores de Erro Médio apresentados.

5.7.3 Construir e avaliar modelo

Cenário 1 – Categorização de pacientes

Tabela 13 - Resultados da execução do modelo Simple k-Means

	Dataset QLQ-C30	**EQM[1]**	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
Simple K-Means	K = 3	88	34%	47%	19%			
	K = 4	78	36%	43%	6%	15%		
	K = 5	75	30%	15%	4%	11%	40%	

[1] Erro Quadrático Médio

	K = 6	73	30%	13%	2%	6%	43%	6%
	Dataset QLQ-H&N35	**EQM**[1]	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
	K = 3	127	33%	18%	49%			
	K = 4	122	24%	16%	8%	51%		
	K = 5	116	20%	16%	6%	51%	6%	
	K = 6	111	20%	16%	4%	51%	6%	2%
	Dataset C30Rasch	**EQM**[1]	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
	K = 3	138	63%	14%	23%			
	K = 4	129	24%	13%	16%	47%		
	K = 5	120	23%	9%	16%	31%	21%	
	K = 6	112	23%	9%	16%	29%	17%	7%

Considerando os resultados obtidos na execução do modelo Simple k-Means sobre o dataset QLQ-C30, apenas o teste que visa a deteção de três agrupamentos (k=3) atingiu valores com a relevância estatística estipulada de 10%, apresentando um erro Quadrático Médio de 88 valores. Para o dataset QLQ-H&N35, verificou-se que apenas o teste que visa a três agrupamentos revelou valores com a relevância estatística estipulada de 10%, apresentando um erro Quadrático Médio de 127 Valores.

Os resultados obtidos na execução do modelo Simple k-Means, com vista à deteção de agrupamentos no dataset C30Rasch, apresentam para os três e quatro agrupamentos detetados a relevância estatística estipulada de 10%. Contudo o teste que visa os quatro agrupamentos verificou um menor valor de Erro Quadrático Médio.

Desta forma, os testes que visam os três agrupamentos detetados nos datasets QLQ-C30 e QLQ-H&N35, assim como o teste que detetou quatro agrupamentos no dataset C30Rasch, serão submetidos a tarefas de classificação. Impõe-se assim a necessidade de atribuir as categorizações obtidas aos conjuntos de dados, com intuito de perceber a percentagem de previsão do modelo.

Na tabela seguinte são apresentados os modelos de classificação executados, e as métricas de avaliação, através do processo de validação 5-fold Cross-Validation:

Tabela 14- Resultados dos modelos de previsão sobre os três clusters detetados no dataset QLQ-C30

	Modelo	Percentagens de Acerto	Erro Médio	Tempo (s)	Sensibilidade	Especificidade
Dataset QLQ-C30 K=3	*J48*	63%	0.2631	0.08	0,6875	0,8387
	MultilayerPerceptron	74%	0.1872	1.79	0,75	0,8064
	RandomForest	74%	0.3059	0.16	0,6875	0,8387
	PART	63%	0.2404	0.03	0,625	0,7419
	NNge	82%	0.1135	0.03	0,8125	0,9032
	IBk	57%	0.2955	0	0,5625	0,6774
	SMO	76%	0.279	0.26	0,8125	0,7419
	NaiveBayes	59%	0.2812	0.01	0,625	0,7096

Os resultados obtidos, em função da percentagem de acerto, para as tarefas de classificação executadas sobre os três agrupamentos de dados detetados no dataset QLQ-C30, destacam os modelos NNge, SMO, MultilayerPerceptron e RandomForest, com 82%,76%,74% e 74%, respetivamente. Analisando o Erro Médio verificado, o valor mais baixo é alcançado pelo modelo NNge. Numa análise global aos tempos de aprendizagem dos modelos, apenas o modelo MultilayerPerceptron demora mais que 1 segundo, sendo o modelo IBk o mais rápido nesse processo. A nível das métricas de avaliação aferidas pela Matriz Confusão, o modelo NNge volta a destacar-se dos anteriores, sendo o modelo que apresenta maior Sensibilidade e Especificidade.

Tabela 15- Resultados dos modelos de previsão sobre os três clusters detetados no dataset QLQ-H&N35

	Modelo	Percentagens de Acerto	Erro Médio	Tempo (s)	Sensibilidade	Especificidade
Dataset QLQ-H&N35 K = 3	*J48*	67%	0.251	0.03	0,5625	0,8181
	MultilayerPerceptron	89%	0.0869	2.26	0,9375	1
	RandomForest	81%	0.2554	0.17	0,9375	0,9090
	PART	75%	0.1882	0.01	0,75	0,8484
	NNge	81%	0.1224	0.05	0,875	0,9393
	IBk	73%	0.196	0	0,875	0,8787
	SMO	85%	0.2676	0.1	0,9375	0,9090
	NaiveBayes	67%	0.2274	0.02	0,8125	0,7575

Analisando os resultados obtidos na execução dos modelos de classificação, sobre os três agrupamentos detetados no dataset QLQ-H&N35, em função da percentagem de acerto, destacam-se os modelos MultilayerPerceptron, SMO e RandomForest. Sendo que o MultilayerPerceptron se verifica como o modelo que alcança um menor valor de Erro Médio, e classifica-se como o mais sensível e especifico. Contudo, o MultilayerPerceptron apresenta-se como o modelo mais lento a nível do tempo de aprendizagem.

Tabela 16 - Resultados dos modelos de previsão sobre os quatro clusters detetados no dataset C30 Rasch

	Modelo	Percentagens de Acerto	Erro Médio	Tempo (s)	Sensibilidade	Especificidade
Dataset C30Rasch K = 4	*J48*	75%	0.1245	0.03	0,5882	0,9622
	MultilayerPerceptron	87%	0.0759	2.56	0,7647	0,9811
	RandomForest	85%	0.1647	0.17	0,7647	0,9811
	PART	71%	0.1505	0.02	0,6470	0,8679
	NNge	72%	0.1357	0.04	0,7058	0,9245
	IBk	78%	0.125	0	0,7058	0,9622
	SMO	81%	0.2702	0.31	0,6470	0,9811
	NaiveBayes	64%	0.1846	0.01	0,5882	0,8679

De acordo com os resultados obtidos, em função da percentagem de acerto para as tarefas de classificação executadas sobre o dataset C30Rasch, destacam-se os modelos MultilayerPerceptron, RandomForest e SMO, com valores percentuais de 87,85 e 81, respetivamente. Dos modelos destacados, o modelo MultilayerPercetron é o que apresenta o menor Erro Médio, mas apresenta-se como o mais lento a nível do tempo de aprendizagem. Relativamente às métricas aferidas pela Matriz de Confusão, os modelos MultilayerPercetron e RandomForest equiparam-se como os mais sensíveis e específicos.

Cenário 2 – Previsão de QdV

O presente cenário de previsão de QdV será desenvolvido de acordo com as categorizações identificadas no anterior cenário, de 3 Clusters para o dataset QLQ-C30 e QLQ-H&N35. Essas categorizações serão utilizadas como critérios de seleção, distinguindo as tarefas de classificação.

Tabela 17- Resultados obtidos na execução dos modelos de Data Mining sobre o dataset QLQ-C30

Dataset QLQ-C30	**Cluster 1**			**Cluster 2**			**Cluster 3**		
	%A[2]	EM[3]	T(s)[4]	%A[2]	EM[3]	T(s)[4]	%A[2]	EM[3]	T(s)[4]
J48	81	0.075	0.01	59	0.1569	0.01	44	0.2074	0
MultilayerPerceptron	75	0.0955	2.23	63	0.1209	2.8	44	0.1874	1.17
RandomForest	68	0.105	0.1	68	0.1464	0.07	44	0.196	0.09
PART	81	0.075	0.01	59	0.1479	0.02	44	0.2074	0
NNge	68	0.1042	0.03	68	0.1061	0.01	44	0.1852	0.02
IBk	68	0.1595	0	63	0.1612	0	33	0.2481	0
SMO	75	0.1574	0.03	68	0.2071	0.04	55	0.214	0.05
NaiveBayes	70	0.1045	0.01	72	0.085	0	67	0.1125	0

Numa análise global, os resultados obtidos na execução dos modelos de classificação sobre os três agrupamentos de dados detetados no dataset QLQ-C30, em função da percentagem de acerto, destacam o modelo NaiveBayes, uma vez que se apresenta como o único modelo com capacidade preditiva igual ou superior ao valor mínimo estipulado de 70%, em dois dos três agrupamentos presentes. No entanto os modelos J48, Part, MultilayerPerceptron, e SMO apresentam percentagens acima dos 70% estipulados na previsão efetuada sobre o Cluster1. Relativamente à minimização do Erro Médio, para o Cluster 1 destaca-se o modelo PART, enquanto nos Clusters 2 e 3 se destaca o modelo NaiveBayes.

A nível dos tempos de aprendizagem todos modelos destacados variam entre 0 a 2 segundos, sendo o modelo MultilayerPerceptron o mais lento em todas as previsões efetuadas. Importa salientar que as métricas de avaliação Sensibilidade e Especificidade não foram tidas em conta na tarefa de classificação presente, devido ao número de registos reduzido que cada cluster dispõe.

Tabela 18 - Resultados obtidos na execução dos modelos de Data Mining sobre o dataset QLQ-H&N35

Dataset QLQ-H&N35	**Cluster 1**			**Cluster 2**			**Cluster 3**		
	%A[2]	EM[3]	T (s) [4]	%A[2]	EM[3]	T (s) [4]	%A[2]	EM[3]	T (s) [4]
J48	87	0.0952	0.01	22	0.4127	0	75	0.1188	0.01
MultilayerPerceptron	75	0.1281	2.11	66	0.1878	1.17	83	0.0966	3.4
RandomForest	93	0.127	0.11	66	0.2588	0.06	79	0.1591	0.09

[2] Percentagens de Acerto
[3]Erro Médio
[4] Tempo de Aprendizagem do Modelo (segundos)

PART	87	0.0795	0	11	0.4339	0.01	62	0.1882	0.01
NNge	81	0.0938	0.03	77	0.1111	0.02	79	0.1042	0.05
IBk	75	0.1847	0	77	0.2058	0	54	0.2542	0
SMO	75	0.225	0.03	88	0.2722	0.04	83	0.2167	0.04
NaiveBayes	93	0.0318	0.01	66	0.1667	0	83	0.0739	0.01

Os resultados obtidos na execução dos modelos de classificação sobre os três agrupamentos de dados identificados no dataset QLQ-H&N35, em função da percentagem de acerto, destacam os modelos SMO e NNge, com percentagens de acerto superiores ao estipulado de 70%, para todos os clusters. Contudo, se a análise dos resultados for executada individualmente para cada Cluster, podem-se ainda destacar os modelos NaiveBayes e RandomForest para o Cluster 1, e os modelos NaiveBayes e MultilayerPerceptron para o Cluster 3, uma vez que apresentam elevadas percentagens de acerto. No caso do Cluster 2, apenas três modelos alcançam valores iguais ou superiores ao mínimo estipulado para o presente cenário, nomeadamente, os modelos anteriormente destacados, SMO e NNge, e o IBk. Relativamente à minimização do Erro Médio apresentado o modelo NNge apresenta valores de destaque em todos os Clusters classificados. A nível de tempo de aprendizagem os modelos comportam-se de forma semelhante, abaixo de 1 segundo, com exceção do modelo MultilayerPerceptron que varia entre 2 e 4 segundos. Importa salientar que as métricas de Sensibilidade e Especificidade não foram utilizadas para avaliação dos modelos devido ao número reduzido de registos presente em cada um dos clusters submetidos às tarefas de classificação.

5.8 Discussão de Resultados

Os resultados alcançados nos diferentes cenários contemplados permitem a sugestão de um conjunto de técnicas e modelos, com capacidade para atingir a objetivação geral e da componente prática do presente projeto de dissertação.

Na categorização dos pacientes, através das informações sociodemográficas, clínicas e dos instrumentos de medida da QdVRS, os resultados obtidos revelam que a homogeneização dos dados presentes visam um total de três agrupamentos para os datasets QLQ-C30 e QLQ-H&N35 e um total de quatro agrupamentos para o dataset C30Rasch. As categorizações identificadas, quando sujeitas a tarefas de classificação, revelaram percentagens de acerto bastante elevadas, destacando os modelos NNge, SMO, MultilayerPerceptron e RandomForest. No entanto, o modelo MultilayerPerceptron na generalidade dos diferentes testes executados

sobre os datasets categorizados enquadra-se como o mais assertivo, apresentado níveis de Sensibilidade e Especificidade relevantes. Negativamente os resultados apontam-no como sendo o modelo mais demorado a nível do tempo de aprendizagem.

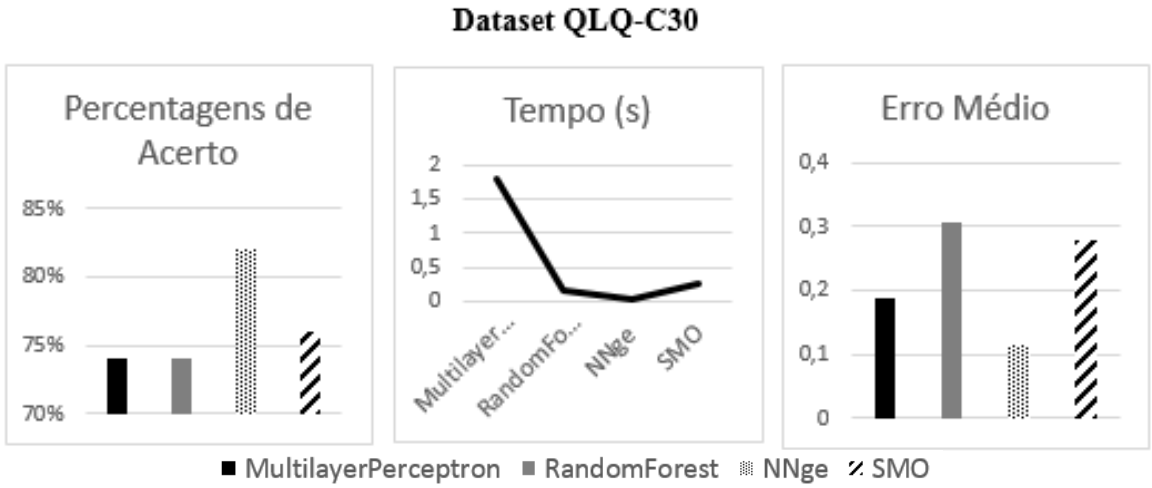


Figura 7 - Gráficos comparativos das métricas de avaliação na execução dos diferentes modelos destacados (Dataset QLQ-C30)

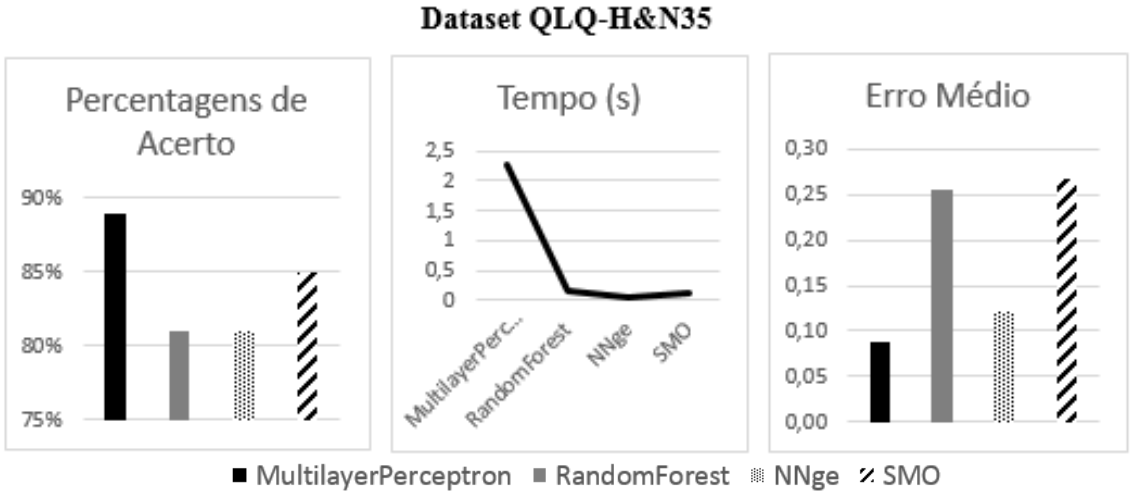


Figura 8 – Gráficos comparativos das métricas de avaliação na execução dos diferentes modelos destacados (Dataset QLQ-H&N35)

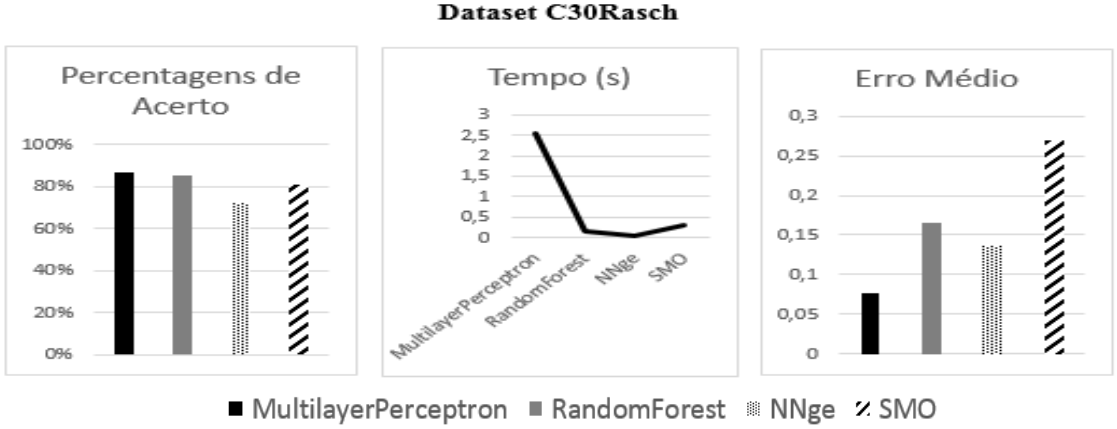


Figura 9 - Gráficos comparativos das métricas de avaliação na execução dos diferentes modelos destacados (Dataset C30Rasch)

Na previsão de QdV destacam-se os modelos NaiveBayes, NNge e SMO, com percentagens de acerto iguais ou superiores aos 70% estipulados. Contudo o número reduzido de registos presentes nos datasets QLQ-C30 e QLQ-H&N35 limitou o cálculo das métricas de avaliação relativas à Especificidade e Sensibilidade dos modelos.

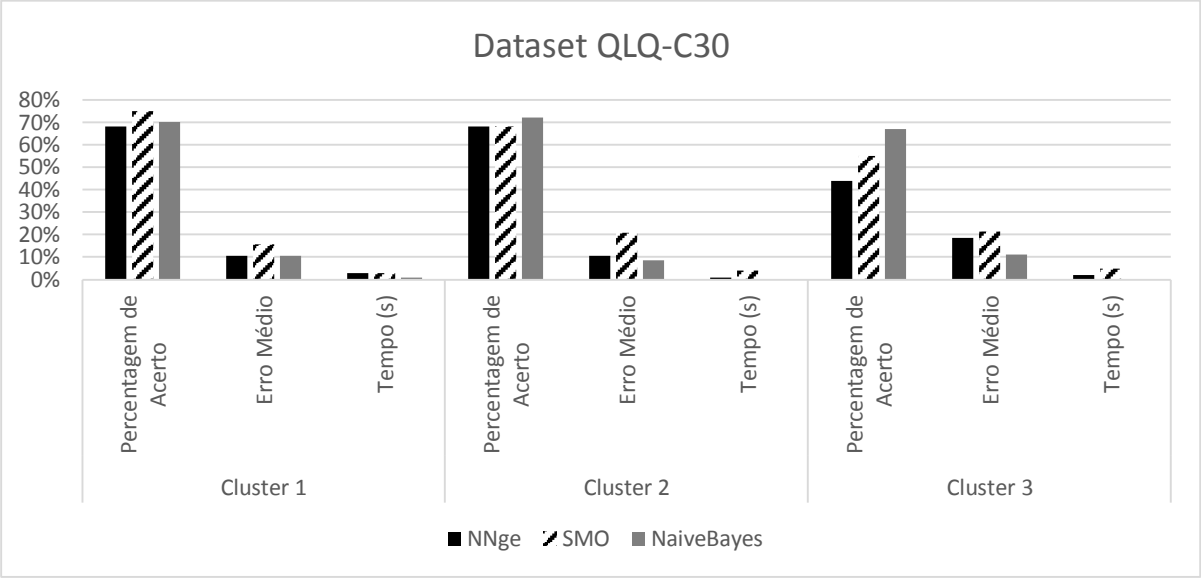


Figura 10 - Gráficas comparativos das diferentes métricas de avaliação relativos à previsão de QdV (Dataset QLQ-C30)

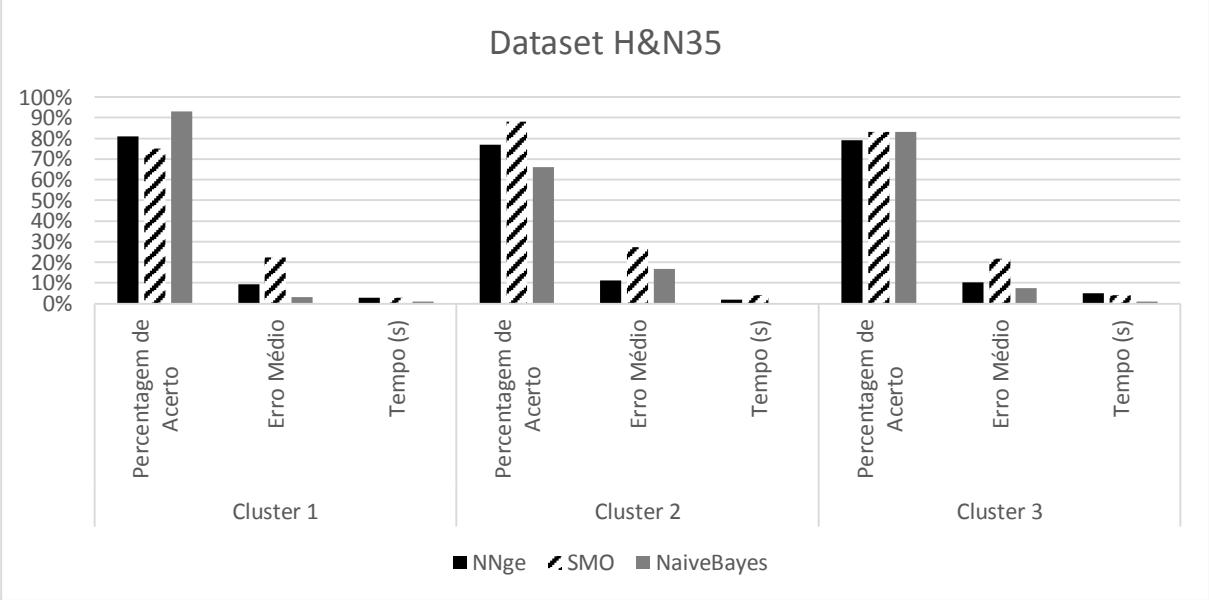


Figura 11- Gráficas comparativos das diferentes métricas de avaliação relativos à previsão de QdV (Dataset QLQ-H&N35)

5.9 Implementação na Plataforma QoLIS

O projeto de dissertação foi desenvolvido em estreita colaboração com a Optimizer, precipitando a que no momento atual todos as tarefas e atividades desenvolvidas na componente prática se encontrem já implementadas na plataforma QoLIS.

6. CONCLUSÕES E TRABALHO FUTURO

6.1 Conclusões

O projeto de dissertação desenvolve um processo de descoberta de conhecimento em bases de dados, que permitiu a identificação de técnicas de Data Mining, com capacidade para influenciar os módulos de inferência e de determinação de QdVRS do SADC QoLIS, em desenvolvimento pela Optimizer. O processo desenvolvido foi precedido por uma revisão de literatura sobre as áreas de Qualidade de Vida e Descoberta de Conhecimento de Bases de Dados, onde foram adquiridos conhecimentos essenciais para apoiar a realização do processo de descoberta de conhecimento em bases de dados, através do estudo de técnicas de Data Mining e ainda conhecimento relativamente aos conceitos e práticas de QdVRS e aos questionários avaliação da QdV, que se encontram integrados no SADC, nomeadamente o questionário EORTC QLQ-C30 e EORTC QLQ-H&N35.

O processo de descoberta de conhecimento foi desenvolvido na componente prática do presente projeto de dissertação, através da adaptação da metodologia CRISP-DM aos objetivos de Categorização dos pacientes e Previsão de QdV, de acordo com os dados extraídos da base de dados do QoLIS. Essa extração resultou num conjunto de dados, que refletia interações entre pacientes e a instituição de saúde responsável pelo acompanhamento e tratamento das suas patologias oncológicas, que através de um processo executado na plataforma RapidMiner foi dividido em dois datasets, o dataset QLQ-C30 e o dataset QLQ-H&N35.

O dataset QLQ-C30 apresenta atributos caracterizadores do paciente sociodemograficamente, informações clínicas obtidas por meio de consulta, informações clínicas sobre os tratamentos e informações, apenas respeitantes ao questionário EORTC QLQ-C30. O dataset QLQ-H&N35 apenas se distingue do anterior nas informações a nível do questionário, uma vez que apenas contempla informações respeitantes ao questionário EORTC QLQ-H&N35.

Estes datasets foram utilizados durante a realização do processo de descoberta de conhecimento, após uma série de tratamentos, transformações e construções de atributos, que resultaram numa melhoria substancial da qualidade de dados apresentada.

Foi ainda integrado o dataset C30Rasch, que resultou da aplicação das operações de tratamento, transformação e construções referidas na base de dados do SADC, com a peculiaridade de as

informações respeitantes aos questionários serem obtidas através de um cálculo do modelo de Rasch.

Foram então desenvolvidos cenários de teste, distinguidos pelos objetivos de Categorização e Previsão definidos.

O primeiro cenário contemplou o objetivo de Categorização de Pacientes, através do modelo Simple k-Means, executado sobre informações sociodemográficas, clínicas e dos instrumentos de medida de QdVRS, na fase anterior ao tratamento, que integram os datasets QLQ-C30, QLQ-H&N35 e C30Rasch. Os resultados obtidos permitiram a categorização através dos três agrupamentos detetados nos datasets QLQ-C30 e QLQ-H&N35, e quatro agrupamentos no dataset C30Rasch. Esses agrupamentos foram atributidos aos dados como classes, permitindo a execução de modelos de classificação com vista à previsão da categorização efetuada. Os resultados nessa atividade permitiram destacar os modelos NNge, SMO, RandomForest e MultilayerPerceptron.com percentagens de acerto a bastante elevadas. Sendo que o último referido, se destaca nas métricas de avaliação de Sensibilidade e Especificidade, mas ao mesmo tempo se caracteriza como o mais lento nos tempos de aprendizagem do modelo.

O segundo cenário contemplou o objetivo de Previsão de QdV, através de modelos inerentes às Árvores de Decisão, Regras de Associação, Redes Neuronais, aos Vizinhos mais Próximos e aos Classificadores de Bayes. Os modelos foram executados sobre os datasets QLQ-C30 e QLQ-H&N35, integrando para a tarefa de classificação pretendida as informações sociodemográficas, clínicas e dos instrumentos de QdVRS, em todas as fases do tratamento categorizadas. Os resultados obtidos destacaram os modelos NaiveBayes, NNge e SMO, atingindo em metade dos testes executados percentagens de acerto iguais ou superiores ao mínimo estipulado de 70%. Contudo, estes resultados foram limitados pelo número reduzido de registos apresentados nos datasets e nos momentos de tratamento categorizados.

Pode-se assim concluir que os resultados revelaram modelos com capacidade para explorar, extrair e evidenciar padrões nas atividades desenvolvidas pelos módulos da Unidade de Pescoço e Cabeça que integram o Sistema de Apoio À Decisão Clínica. Porém, existe a necessidade de executar o processo de descoberta de conhecimento em bases de dados, novamente, quando o volume de dados aumentar, de maneira a que a previsão de QdV apresente resultados mais sustentados.

Importa ainda salientar que o presente projeto de dissertação preenche uma lacuna observada na comunidade científica, que visa a falta de estudos que relacionem a aplicação prática do uso de Data Mining e Sistemas de Apoio à Decisão Clínica na área de Qualidade de Vida.

6.2 Limitações e Trabalho Futuro

Os resultados dos modelos de previsão de Qualidade de Vida foram limitados pelo número reduzido de dados com informações respeitantes ao momento anterior ao tratamento, uma vez que o agrupamentos de dados que permitiu a categorização de pacientes estava dependente do momento referido. Este fator restringiu a obtenção de elementos da Matriz de Confusão com capacidade para proceder ao cálculo das métricas de Sensibilidade e Especificidade, impedindo uma análise sustentada dos resultados obtidos na tarefa objetivada.

Desta forma, os modelos executados foram apenas avaliados pela percentagem de acerto, pela minimização do erro médio e ainda pelo tempo de aprendizagem dos modelos. Impõe-se assim a necessidade de no futuro acompanhar as diferentes fases do tratamento, repercutindo essa informação nos dados, de forma que seja possível ultrapassar a limitação referida. Esta realidade pode precipitar a necessidade de executar novamente os testes aos modelos e consequentemente sua análise, existindo a hipótese de serem destacados diferentes modelos para integrar os módulos do SADC QoLIS apresentados no presente projeto de dissertação, para executar a previsão de Qualidade de Vida.

BIBLIOGRAFIA

A. Simon Pickard, James W. Shaw, Hsiang-Wen Lin, Peter C. Trask,Neil Aaronson,Todd A. Lee, David Cella. (2009). A Patient-Based Utility Measure of Health for Clinical Trials of Cancer Therapy Based on the European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire. *Value in Health, 12*.

Alexandra Oliveira, Pedro L. Ferreira, Bárbara Antunes, Francisco L. Pimentel. (2011). OnQol: Electronic device to capture QoL data in oncology: Difference between patients 65 years or older and patients younger than 65 years of age. *Journal of Geriatric Oncology*, 253-258.

Alhalaweh A, Alzghoul A, Kaialy W. (2014). Data mining of solubility parameters for computational prediction of drug-excipient miscibility. *Drug development and industrial pharmacy*, 904-909.

Andreia Brandão, E. P. (2014). Managing Voluntary Interruption of Pregnancy using Data Mining. *Procedia Technology* , 1297-1306.

Andrich, D. (2006). Item Discrimination and Rasch-Andrich Thresholds Revisited. *Rasch Measurement, 20*(2), 1055-1057.

Andrienko, G. L., N. V. Andrienko. (1999). Knowledge Extraction from Spatially Referenced Databases: a Project of an Integrated Environment. *Varenius Workshop on Status and Trends in Spatial Analysis.* Sta. Barbara: CA.

Austin PC, Tu JV, Ho JE, Levy D, Lee DS. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, 398-407.

Banaee H, Ahmed MU, Loutfi A. (2013). Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors, 13*(12), 17472-17500.

Bowling, A. (2001). *MEASURING DISEASE: A review of disease-specific Quality of Life Measurement Scales* (Second Edition ed.). Buckingham . Philadelphia: Open University Press.

Edlund M, T. L. (1985). Quality of life: an ideological critique. *Perspect Biol, 28*, 591-607.

Efraim Turban, R. E. (2011). *Decision Support and Business Intelligence Systems.* Oklahoma: Prentice Hall.

EORTC. (2001). *EORTC QLQ-C30 Scoring Manual.* Brussels: European organisation for Research and Treatment of Cancer.

Faria, B. M. (2013). *Classificação de pacientes para adaptação de cadeiras de rodas inteligente.* Aveiro: Departamento de Electrónica, Telecomunicações e Informática.

Farquhar, M. (1995). Elderly people's definitions of Quality of Life. *Pergamon, 41*(10), 1439-1446.

Ferreira D, Oliveira A, Freitas A. (2012). Applying data mining techniques to improve diagnosis in neonatal jaundice. *BMC medical informatics and decision making*, 143.

George E. Vlahos, Thomas W. Ferratt, and George Knoepfle. (2004). The use of computer-based information systems by German managers to support decision making. *information systems by German managers to support decision making, 41(6)*, 763-779.

Gloria Phillips-Wren, P. S. (2008). Mining lung cancer patient data to assess healthcare resource utilization. *Expert Systems with Applications, 35*, 1611-1619.

Gonçalves, J. J. (2012). *Plataforma para avaliação da Qualidade de Vida Relacionada com a Saúde em Oncologia.* Porto: Universidade Fernando Pessoa.

Gordon H. Guyatt, David H. Feeny, Donald L.Patrick. (1993). Measuring Health-related Quality of Life. *Annals of Internal Medicine, 118*, 622-62.

Gordon S.Linoff, Michael J. A. Berry. (2000). *Mastering Data Mining: The art and science of Customer Relationship Management.* New York: Paperback.

Group WHOQOL. (1994). Development of the WHOQOL: Rationale and current status. *International Journal of Mental Health, 23(3)*, 24-56.

Jajroudi M, Baniasadi T, Kamkar L, Arbabi F, Sanei M, Ahmadzade M. (2014). Prediction of survival in thyroid cancer using data mining technique. *Technology in cancer research and treatment, 13*(4), 353-359.

Joudaki H, Rashidian A, Minaei-Bidgoli B, Mahmoodi M, Geraili B, Nasiri M, Arab M. (2014). Using data mining to detect health care fraud and abuse: a review of literature. *Global journal of health science, 7*(1), 194-202.

Kamber, J. H. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann.

Kharat AT, Singh A, Kulkarni VM, Shah D. (2014). Data mining in radiology. *The indian journal of radiology & imaging, 24*, 97-102.

Kulldorff M, Dashevsky I, Avery TR, Chan AK, Davis RL, Graham D, Platt R, Andrade SE, Boudreau D, Gunter MJ, Herrinton LJ, Pawloski PA, Raebel MA, Roblin D, Brown JS. (2013). Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiology and drug safety, 22*(5), 517-523.

Lavrac, N. (1999). Selected techniques for Data Mining in Medicine. *Artificial Intelligence in Medicine*, 3-23.

Manuel Filipe Santos, Carla Sousa Azevedo. (2005). *Data Mining Descoberta de conhecimento em bases de dados.* FCA - Editora de Informática, Lda.

Mar Marcos, J. A.-S. (2013). Interoperability of clinical decision-support systems and electronic health records using archetypes: A case study in clinical trial eligibility. *Journal of Biomedical Informatics, 46*, 676-689.

Maribel Yasmina Santos, I. R. (2006). *Business Intelligence : tecnologias da informação na gestão de conhecimento.* Lisboa: FCA - Editora de Informática. Obtido de http://hdl.handle.net/1822/6198

Maribel Yasmina Santos, L. A. (1999). *A descoberta de conhecimento em bases de dados geográficas através da explicitação semântica.* DSI - Engenharia da Programação e dos Sistemas Informáticos.

Mário Ferreira, L. P. (2015). Data Mining e Sistemas de Apoio à Decisão em Aplicações Clínicas e Qualidade de Vida. *CISTI*, (pp. 271-275). Aveiro.

Mark A. Musen, Y. S. (2006). *Biomedical Informatics: Computer Applications in Health Care and Biomedicine.* Springer.

Mark Hall, E. F. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1).

Maryane Oliveira Campos & João Felício Rodrigues Neto. (2008). Qualidade de Vida: um instrumento para a promoção de saúde. *Revista Baiana de Sáude Pública*.

Means, G. &. (2000). *Meta-capitalism: The e-business revolution and the design of 21st century companies and markets.* New york: John Wiley & Sons Inc.

Michael Koller, Neil K. Aaronson, Jane Blazeby, Andrew Bottomley, Linda Dewolf, Peter Fayers, Colin Johnson,John Ramage, Neil Scott, Karen West. (2007). Translation procedures for standardised quality of life questionnaires: The European Organisation for Research and Treatment of Cancer (EORTC) approach. *European Journal Of Cancer, 43*, 1810-1820.

N. Heutte, L. Plisson, M.Lange, V.Prevost, E. Babin. (2014). Quality of Life tools in head and neck. *European Annals of Otorhinolaryngology, Head and Neck diseases, 131*, 33-47.

Nura Esfandiari, Mohammad Reza Babavalian, Amir-Masoud Eftekhari Moghadam, Vahid Kashani Tabar. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 4434-4463. doi:http://dx.doi.org/10.1016/j.eswa.2014.01.011

Oded Maimon & Lior Rokach. (2010). *Data Mining and Knowledge Discovery Handbook: Second Edition.* London: Springer.

Omer A, Singh P, Yadav NK, Singh RK. (2014). An overview of data mining algorithms in drug induced toxicity prediction. *Mini reviews in medical chemistry, 10*, 345-354.

Oscar Marbán, J. S.-B. (2009). Toward data mining engineering: A software engineering approach. *Information Systems*, 87-107.

Pais-Ribeiro, J. (2004). Quality of life is a primary end-point in clinical settings. *Clinical Nutrition*, 121-130. doi:10.1016/S0261-5614(03)00109-2

Paolo Fraccaro, D. O. (2015). Behind the screens: Clinical decision support methodologies - A review. *Health Policy and Techonology, 4*, 29-38.

Peña-Ayala, A. (2014). Educational Data Mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*, 1432-1462.

Penny KI, Smith GD. (2012). The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome. *Journal of clinical nursing, 21*, 2761-2671.

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR),Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler),Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler). (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS.

Pimentel, F. L. (2003). *Qualidade de Vida do Doente Oncológico.* Porto: De autor.

Raju D, Su X, Patrician PA, Loan LA, McCarthy MS. (2015). Exploring factors associated with pressure ulcers: a data mining approach. *International journal of nursing studies, 52*(1), 102-111.

Ramos, L. F. (Março de 2014). Deteção e Caracterização Geo-Espacial das Zonas de Acumulação de Acidentes Rodoviários. Guimarães, Braga, Portugal: Universidade do Minho.

RapidMiner. (2014). RapidMiner Studio. Obtido em 2015, de http://www.rapidminer.com

Riccardo Bellazi, B. Z. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International journal of medical informatics, 77*, 81-97.

S., S. (2013). Improving diagnostic accuracy using agent-based distributed data mining system. *Informatic for health & social care*, 182-195.

Santos, C. (2007). *Análise dos resultados do WHOQOL-100 utilizando Data Mining.* Paraná: Universidade Tecnológica Federal do Paraná.

Schumacker, R. E. (2010). Item Response Theory. *Applied Measurement Associate LLC*.

Sergio Cavalheiro, S. M. (2013). A Data Mining system for providing analytical information on brain tumors to public health decision makers. *Computer methods and programs in biomedicine, 109*, 269-282.

Shaker H. El-Sappagh, S. E.-M. (2014). A distributed clinical decision support system architecture. *Journal of King Saud University – Computer and Information Sciences, 26*, 69-78.

Shu-Hsien Liao, P.-H. C.-Y. (2012). Data Mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications, 39*, 11303-11311.

Sónia Quintão, A. R. (2011). Avaliação da escala de auto-estima de Rosenberg mediante o modelo de rasch. *Psicologia, 25*(2). Obtido de http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0874-20492011000200005&lng=pt&tlng=pt

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*.

Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, Williams BA, deFilippi C, Ebadollahi S, Stewart WF. (2014). Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of cardiac failure*, 459-464.

Xiao-Jun Lin, I-Mei Lin, Sheng-Yu Fan. (2013). Methodological issues in measuring health-related quality of life. *Tzu Chi Medical Journal, 25*, 8-12. doi:http://dx.doi.org/10.1016/j.tcmj.2012.09.002

XindongWu, V. K.-H. (2007). Top 10 algorithms in Data Mining. *IEEE International Conference on Data Mining (ICDM)* (pp. 1-37). Springer.

Yada N, Kudo M, Kawada N, Sato S, Osaki Y, Ishikawa A, Miyoshi H, Sakamoto M, Kage M, Nakashima O, Tonomura A. (2014). Noninvasive diagnosis of liver fibrosis: utility of data mining of both ultrasound elastography and serological findings to construct a decision tree. *Oncology*, 63-72.

Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems, 36*(4), 2431-2448.

Zhang G, Riemer AB, Keskin DB, Chitkushev L, Reinherz EL, Brusic V. (2014). HPVdb: a data mining system for knowledge discovery in human papillomavirus with applications in T cell immunology and vaccinology. *Database: the journal of biological databases and curation*.

Zhao Y, Xie Q, He L, Liu B, Li K, Zhang X, Bai W, Luo L, Jing X, Huo R. (2014). Comparsion analysis of data mining models applied to clinical research in traditional Chinese medicine. *Journal of Traditional Chinese Medicine*, 627-634.

ANEXO I – RELATÓRIO DE QUALIDADE DOS DADOS

Atributo	Formato	Valores Omissos (%)		Qualidade dos Dados
		Dataset QLQ-C30	Dataset QLQ-HN35	
Appoint_No	Numérico	0%	0%	Nenhum problema detetado
Appoint_Date	Data	0%	0%	Datas em formato textual "aaaammdd"
Pat_No	Numérico	0%	0%	Nenhum problema detetado
Date_Birth	Data	0%	0%	Datas em formato textual
County_Code	Numérico	1%	1%	Presença de Valores Omissos
Gender_Code	String	0%	0%	Presença do valor "Indiferenciado"
Edu_Deg_Code	Numérico	2%	2%	Presença de Valores Omissos
Profession_Code	Numérico	3%	3%	Presença de Valores Omissos
Date_D	Data	100%	100%	Não apresenta qualquer valor
Civil_Status_Code	Numérico	2%	2%	Presença de Valores Omissos
Breed_Code	Numérico	100%	100%	Não apresenta qualquer valor
Nationality_Code	Numérico	100%	100%	Não apresenta qualquer valor
Doctor_Code	Numérico	100%	100%	Não apresenta qualquer valor
Date_Px	Data	100%	100%	Não apresenta qualquer valor
Uni_Code	String	0%	0%	Nenhum prolema detetado
T_Code	String	48%	48%	Presença de valores omissos
N_Code	String	52%	52%	Presença de valores omissos
M_Code	String	57%	57%	Presença de valores omissos
Ik	Numérico	95%	94%	Presença de Valores Omissos e Incongruentes
Pad_Code	Numérico	21%	21%	Presença de Valores Omissos

Mo_Code	String	40%	41%	Presença de Valores Omissos
Hist_Code	Numérico	100%	100%	Não apresenta qualquer valor
Beh_Code	Numérico	100%	100%	Não apresenta qualquer valor
Deg_Code	Numérico	100%	100%	Não apresenta qualquer valor
Topo_Code	Numérico	100%	100%	Não apresenta qualquer valor
N_Rec	Binário	0%	0%	Todos os valores em Default
Smokes	Numérico	12%	13%	Presença de Valores Omissos
Years_Smk	Numérico	40%	40%	Presença de Valores Omissos
Num_Cig	Numérico	39%	39%	Presença de Valores Omissos
Years_Stop_Smk	Numérico	52%	53%	Presença de Valores Omissos
Drinks	Numérico	100%	100%	Não apresenta qualquer valor
Years_Drink	Numérico	41%	40%	Presença de Valores Omissos
Num_Lit_Beer	Numérico	57%	56%	Presença de Valores Omissos
Num_Lit_Alc	Numérico	59%	59%	Presença de Valores Omissos
Num_Lit_Wine	Numérico	44%	43%	Presença de Valores Omissos
Years_Stop_Drk	Numérico	100%	100%	Não apresenta qualquer valor
Trach	Binário	4%	5%	Presença de Valores Omissos
Alim	String	4%	4%	Presença de Valores Omissos
Voice_Prot	Binário	49%	51%	Presença de Valores Omissos
Icd_Code	Numérico	96%	97%	Presença de Valores Omissos
Info_Appoint_No	Numérico	96%	97%	Presença de Valores Omissos
Chemo_Exists	Binário	98%	99%	Valores correspondentes a “0” aparecem em branco
Chemo_Date_I	Data	98%	99%	Datas em formato textual “Aaaammdd”
Chemo_Date_E	Data	98%	99%	Datas em formato textual “Aaaammdd”
Chemo_Cp_Code	Numérico	98%	99%	Presença de Valores Omissos
Chemo_Bp	Binário	98%	99%	Presença de Valores Omissos
Chemo_Int_Type_Code	Numérico	100%	100%	Não apresenta qualquer valor
Radio_Exists	Binário	99%	99%	Valores correspondentes a “0” aparecem em branco
Radio_Date_I	Data	99%	99%	Datas em formato textual

				"Aaaammdd"
Radio_Date_E	Data	99%	99%	Datas em formato textual "Aaaammdd"
Radio_Descr	String	100%	100%	Não apresenta qualquer valor
Radio_Num_Field	Numérico	99%	99%	Presença de Valores Omissos e Incongruentes
Radio_Int_Type_Code	Numérico	100%	100%	Presença de Valores Omissos e Incongruentes
Radio_Intens	Numérico	99%	99%	Presença de Valores Omissos e Incongruentes
Radio_Freq	Numérico	99%	99%	Presença de Valores Omissos e Incongruentes
Radio_Margin	Numérico	100%	100%	Não apresenta qualquer valor
Radio_Info_Appoint_No	Numérico	99%	99%	Presença de Valores Omissos
Surg_Exists	Binário	99%	99%	Valores correspondentes a "0" aparecem em branco
Surg_Icd_Code	Numérico	99%	99%	Presença de Valores Omissos
Surg_Surgery_Date	Data	99%	99%	Datas em formato textual "Aaaammdd"
Surg_Info_Appoint_No	Numérico	99%	99%	Presença de Valores Omissos
Answer_Date	Data	0%	0%	Datas em formato textual "Aaaammdd"
Quest_Code	String	0%	0%	Nenhum problema detetado
QdV	Numérico	1%	1%	Presença de Valores Omissos
Outfit	Numérico	0%	0%	Nenhum problema detetado
Infit	Numérico	0%	0%	Nenhum problema detetado
Error	Numérico	0%	0%	Nenhum problema detetado
P1:30	Numérico	2,8%	0%	Presença de Dados incongruentes
P1:P35	Numérico		3%	Presença de Dados incongruentes

ANEXO II – CONTRIBUTO CIENTÍFICO PRODUZIDO

Data Mining e Sistemas de Apoio à Decisão em Aplicações Clínicas e Qualidade de Vida

Autores: Mário Ferreira, Luís Paulo Reis, Brigida Mónica Faria, Joaquim Gonçalves, Álvaro Rocha

Conferência: CISTI'2015 - 10ª Conferência Ibérica de Sistemas e Tecnologias de Informação

Ano:2015

Resumo: O desenvolvimento de novas tecnologias, sistemas de informação, sistemas de apoio à decisão e algoritmos de predição de parâmetros clínicos utilizando aprendizagem computacional e data mining, abre um conjunto de novas perspetivas em muitas áreas da saúde. Neste contexto, apresenta relevância o conceito de Qualidade de Vida (QdV) no âmbito da saúde e a possibilidade de desenvolver Sistemas de Apoio à Decisão Clínica (SADC) que o utilizem. Através da expetativa individual, de bemestar físico, psicológico, mental, emocional e espiritual dos pacientes, variáveis discutidas e medidas na área de investigação de Qualidade de vida, pretende-se fazer um estudo dos dados para estabelecer correlações com dados laboratoriais, farmacêuticos, socioeconómicos, entre outros, obtendo conhecimento a nível de padrões comportamentais de doentes crónicos, alcançando uma série de dados confiáveis e de fácil acesso, capazes de potenciar o processo de tomada de decisão por parte das equipas médicas especializadas, na procura de melhorar os tratamentos e consequentemente a qualidade de vida relacionada com a Saúde de doentes crónicos. Neste artigo são estudados e comparados estudos relacionados que desenvolvem sistemas de apoio à decisão e predição na área clínica, com ênfase para os estudos na área da qualidade de vida.

Palavras-Chave:Qualidade de Vida, Data Mining, Predição, Sistemas de Apoio à Decisão Clínica;