Dissertação

Mestrado em Engenharia Informática e Computação Móvel

Utilização de técnicas de text mining sobre registos clínicos de epilepsia em crianças, para auxílio ao diagnóstico e classificação

Luís Miguel Oliveira Pereira

Leiria, Outubro de 2013

Dissertação

Mestrado em Engenharia Informática e Computação Móvel

Utilização de técnicas de text mining sobre registos clínicos de epilepsia em crianças, para auxílio ao diagnóstico e classificação

Luís Miguel Oliveira Pereira

Dissertação de Mestrado realizada sob a orientação do Doutor Rui Rijo, Professor da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e coorientação da Doutora Catarina Silva, Professora da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria.

Leiria, Outubro de 2013

À Minha Família

Agradecimentos

Quero agradecer em primeiro lugar aos meus orientadores o Professor Doutor Rui Rijo e à Professora Doutora Catarina Silva, de toda a disponibilidade e o apoio no desenvolvimento desta investigação através de sugestões e incentivos.

Gostaria de agradecer ao Hospital Santo André de Leiria por ter fornecido os registos médicos reais e anónimos que foram cruciais para o desenvolvimento deste projeto.

Agradeço à Dra. Margarida Agostinho pela disponibilidade, sugestões e explicações fornecidas para melhor entender a área de epilepsia, o processo de diagnóstico e sua classificação realizado no Hospital Santo André.

Quero também agradecer à Dra. Cristina Aniceto do Hospital Santo André de Leiria pela disponibilidade e apoio na elaboração deste trabalho.

Agradeço ao do Instituto Politécnico de Leiria (IPL), Escola Superior de Gestão de Leiria (ESTG), Departamento de Engenharia Informática (DEI) e à coordenação do Mestrado de Engenharia Informática e Computação Móvel (MEI-CM) pelo apoio e disponibilidade demonstrados.

Gostaria de agradecer aos meus pais pelo apoio e compreensão durante a elaboração desta dissertação.

Gostaria de expressar o meu agradecimento a todas as pessoas que contribuíram para a realização deste projeto quer diretamente ou indiretamente.

A todos, os meus sinceros agradecimentos!

Nota Prévia

Esta dissertação foi realizada em associação com o Hospital Santo André de Leiria, fornecendo um suporte técnico na interpretação de diversos fatores relevantes médicos como, sintomas, causas, eventos, entre outros. Além disso, foi possivel ainda a transcriçao de registos médicos que foram fundamentais para a análise dos resultados do processo proposto.

No ambito deste trabalho foram submitidas diferentes publicações:

- Pereira, L., R. Rijo, C. Silva, and M. Agostinho, Using Text Mining to Diagnose and Classify Epilepsy in Children, in IEEE HealthCom2013, October 9th 2013, Lisbon;
- Pereira, L., R. Rijo, C. Silva, and M. Agostinho, ICD9-based Text Mining Approach to Children Epilepsy Classification in HCist 2013, October 24th 2103 Procedia Technology, 2013;
- Pereira, L., Rijo, R., Silva, C., "Text Mining Applied to electronic medical records: literature review". International Journal of E-Health and Medical Communications. DOI: 10.4018/IJEHMC, ISSN: 1947-315X, EISSN: 1947-3168. Submitted: 11 March 2013. Conditionally accepted. Status: under review. Scopus index..

Resumo

A informação médica tem aumentado continuamente ao longo do tempo, produzindo-se quantidades elevadíssimas de dados. A análise e a extração desses dados oferecem possibilidades de reduzir o esforço e o tempo na sugestão e classificação de um diagnóstico.

O processamento dos dados médicos representa um grande desafio, considerando que estes dados são geralmente apresentados em texto livre e com vocabulário técnico específico. Entre os dados mais ricos e relevantes encontram-se os registos clínicos. A análise de registos clínicos é complexa pois para a realização de um diagnóstico correto é necessário ter em conta várias características como sintomas, exames, historial do paciente, tratamentos, medicamentos, entre outros. Além disso, esta análise requer um domínio de diferentes áreas de conhecimento para a realização de um diagnóstico fiável, entre outras *data mining*, *text mining*, registos clínicos eletrónicos, e a área clínica. Estes diagnósticos devem ainda ser classificados segundo normalizações, para que o médico possa tomar procedimentos e prescrever tratamentos mais corretos segundo determinadas classificações.

O presente trabalho sugere uma abordagem que incide na área de epilepsia infantil, analisando e extraindo informação relevante de registos clínicos eletrónicos, para ajudar os médicos a tomar decisões, tais como identificar e classificar diagnósticos, ajudar na prescrição de tratamentos, medicamentos e na sugestão de procedimentos. A epilepsia infantil é complexa e não linear, uma vez que os médicos têm de analisar diferentes causas, entre outras, genéticas, estruturais, metabólicas, e um diagnóstico errado pode modificar a vida de uma criança.

Os registos clínicos reais e anónimos foram fornecidos e transcritos com a ajuda do serviço de pediatria do Hospital Santo André. Os resultados alcançados são promissores, estando no entanto ainda longe dos desejados para permitir uma sugestão e classificação de diagnósticos de forma precisa e segura.

Esta abordagem permite ainda uma classificação dos diagnósticos baseadas em normalizações, de forma a sugerir os melhores procedimentos, prognósticos e tratamentos dependendo da classificação encontrada. Desta forma, será possível ajudar a reduzir o erro médico na classificação de diagnósticos, o erro na prescrição, e aumentar a eficácia no processamento dos dados médicos, poupando tempo e dinheiro.

Palavras-chave: Sistemas de Suporte à Decisão, Epilepsia, Registos Clínicos, Códigos ICD-9, Text Mining, Data Mining.

Abstract

Medical information is increasing each day, generating massive amounts of electronic data. This data can be extracted and analyzed thus offering possibilities to reduce time and effort suggesting and classifying diagnoses.

Processing these data represents a medical challenge, considering that these data are usually presented as free text and with specific technical vocabulary. Analysis of these records is complex because it is necessary to take into account several features, such as, symptoms, tests, patient history, treatments, medications, among others, to achieve a correct diagnosis. Moreover, this work requires mastery of different knowledge fields, among others, data mining, text mining, electronic medical records, and clinical area. These diagnoses should be classified according to specific standard codes, so physicians can take procedures and prescribe treatments more accurate.

This dissertation suggests an approach that focuses on the field of pediatric epilepsy, analyzing and extracting relevant information from electronic medical records to help doctors making decisions, such as identifying and classifying diagnosis, helping prescribing treatments, medications and suggesting procedures. Epilepsy in children are more complex and not linear to analyze than in adults, considering physicians need to take into account several causes, such genetic, metabolic, structured, and a misdiagnosis can change dramatically a child's life.

The real and anonymous clinical records were provided and transcribed with the help of the pediatric service of the Hospital Santo André of Leiria, Portugal. Results obtained are promising, but still far from the desired to allow an accurately and safely suggestion and classification of diagnoses.

This approach also allows the classification of diagnoses based on standard codes, in order to suggest the best procedures, prognosis and treatments according to the classification found.

This way, it is possible to reduce medical errors in diagnosis classification, reducing prescription error, and increase efficiency in the processing of medical data, saving time and money.

Key-Words: *Decision Support Systems, Epilepsy, Eletronic medical records, Standard Codes, Text Mining, Data Mining.*

Índice de Figuras

Índice de Tabelas

Lista de Siglas, Abreviações e Acrónimos

ARFF	*Attribute-Relation File Format*
CART	*Classification And Regression Trees*
CAIRN-DAMM	*Computer Assisted Medical Information Resources Navigation & Diagnosis Aid Based On Data Marts & Data Mining*
CRISP-DM	*Cross Industry Standard Process For Data Mining*
EEG	*Eletroencefalogramas*
XML	*Extensible Markup Language*
ML-Flex	*Flexible Toolbox for Performing Classification Analyses*
F1	*F-measure*
FURIA	*Fuzzy Unordered Rule Induction Algorithm*
GATE	*General Architecture For Text Engineering*
HITEx	*Health Information Text Extraction*
HIPAA	*Health Insurance Portability And Accountability Act De 1996*
HTML	*Hypertext Markup Language*
IH	*Informação Hospitalar*
IMS	*Intercontinental Medical Statistics*
IBM	*International Business Machines*

ICD-9	*International Classification Of Diseases, Ninth Revision*
ICF	*International Classification Of Functioning, Disability And Health*
JAPE	*Java Annotation Pattern Engine*
LibSVM	*Library for Support Vector Machines*
K-NN	*K-Nearest Neighbor*
KDD	*Knowledge Discovery In Databases*
MedLEE	*Medical Language Extraction And Encoding System*
NLTK	*Natural Language Toolkit*
ANNIE	*Nearly-New Information Extraction System*
PMBOK	*Project Management Body Of Knowledge*
REX	*Regenstrief Extraction Tool*
RCE	*Registos Clinico Eletrónicos*
RSE	*Registos De Saúde Eletrónicos*
MR	*Ressonâncias Magnéticas*
SEMMA	*Sample, Explore, Modify, Model And Assess*
SVM	*Support Vector Machine*
SNOMED-CT	*Systematized Nomenclature Of Medicine Clinical Terms*
UMLS	*Unified Medical Language System*
UIMA	*Unstructured Information Management Architecture*

Índice

1. Introdução

A informação médica eletrónica gera enormes quantidades de informação. Por exemplo, a base de dados *Medical Literature Analysis and Retrieval System Online* apresenta cerca de 12,5 milhões de registos, aumentando em cerca de 500 000 citações por ano [1]. Esta informação oferece diversas oportunidades para reduzir tempo e esforço de sugestão e classificação de um correto diagnóstico para cada paciente.

Existem diferentes tipos de registos eletrónicos que são utilizados em hospitais, fornecendo diversa informação médica como por exemplo, dados demográficos de cada paciente, exames laboratoriais, notas e tratamentos. Esta informação pode ser utilizada em diversas áreas ou tarefas como, pedidos, gestão de resultados, calendarização de compromissos e faturação que podem ser importantes para ajudar os médicos a compreender e cuidar dos pacientes [2]. Com a utilização de diversas ferramentas que permitam analisar e extrair os contextos relevantes destes documentos, é possível reduzir o esforço e tempo necessário para realizar uma correta análise e por sua vez uma classificação de diagnóstico para cada paciente.

O desenvolvimento de um processo que classifique registos clínicos eletrónicos apresenta inúmeros desafios, entre outros: dominar várias áreas de conhecimento, a análise de texto livre com léxico e semântica muito específica. Inicialmente, este processo requere um esforço para explorar diversas áreas de conhecimento necessárias para o seu desenvolvimento, como por exemplo *data mining*, *text mining*, na área médica e de inteligência artificial. Adicionalmente os textos médicos são apresentados em formato livre, ou seja, estes registos nem sempre contêm uma forma estruturada, podendo dificultar o processo de identificação e extração da informação relevante. Geralmente, cada médico tem a sua própria abordagem para descrever eventos ou sintomas, dependendo da experiência ou práticas médicas adquiridas. Além disso, a área médica possuí uma linguagem específica, que muitas vezes exige ferramentas adicionais para interpretar os termos, sintomas e informação semântica relevante em registos médicos.

Neste capítulo é abordada uma descrição do problema referindo dificuldades, processos e soluções, para um suporte à decisão medica num menor esforço e tempo. Além disto, é também apresentada a estrutura desta dissertação explicando os pontos mais relevantes.

1.1 Descrição do Problema

A epilepsia afeta cerca de 50 milhões de pessoas em todo o mundo [3] e prevê-se que cerca de 1 em 10 pessoas sofram pelo menos uma crise em toda a sua vida [4]. De acordo com o *Intercontinental Medical Statistics*[1] (IMS), epilepsia é a segunda doença neurológica mais comum em Portugal, afetando cerca de 70 000 pessoas cada ano.

O processo de análise e classificação de epilepsia é bastante complexo, exigindo tempo e esforço considerável [5]. Os médicos têm de ter diversos aspetos em consideração como sintomas, procedimentos, eventos, histórico do paciente e exames, de forma a determinar o diagnóstico de epilepsia e por sua vez, classificar o tipo de epilepsia definindo os procedimentos e tratamentos mais eficazes. De facto, alguns tipos de epilepsia necessitam de ser processados rapidamente para que as convulsões possam ser controladas, de forma a evitar lesões neurológicas irreversíveis e para que as pessoas possam viver normalmente.

O diagnóstico e a classificação de epilepsia não são lineares, uma vez que as pessoas que sofrem de epilepsia podem ter comportamentos, sintomas ou crises diferenciadas. Além disso, existe uma maior dificuldade de observação e classificação de epilepsia em crianças, pois é necessário uma análise adicional de diversas causas, e.g. genéticas, estruturais ou metabólicas. A responsabilidade destes diagnósticos em crianças é enorme já que pode mudar drasticamente a vida de uma criança, dado que um diagnóstico se pode traduzir num tratamento impreciso ou incorreto, podendo ser mesmo fatal caso não seja identificado ou controlado de forma adequada [6]. Além disto, a epilepsia em crianças tem um maior impacto do que em adultos, pois as crianças estão a compreender o mundo à sua volta e é necessário uma rápida análise e tratamento para que possam viver normalmente.

Estes diagnósticos podem ainda ser classificados de diversas formas utilizando normalizações, como a *International Classification of Diseases, Ninth Revision* (ICD-9),

[1] http://www.imshealth.com/portal/site/imshealth

muito utilizado em Portugal. Contudo Portugal é um dos últimos países a utilizar estes códigos, pois poderia utilizar códigos mais recentes como ICD-10.

Os códigos ICD-9 são utilizados para descrever o diagnóstico de um determinado paciente, incluindo sintomas, doenças ou distúrbios. Estes códigos permitem que todos os profissionais médicos possam compreender um diagnóstico da mesma forma em qualquer parte do mundo. Podem permitir também, uma boa qualidade de atendimento ao paciente e uma redução dos erros médicos na prescrição e diagnóstico. Estes códigos são muitas vezes utilizados para o controlo de financiamento entre organizações de saúde e do estado, ou entre organizações de saúde e as companhias de seguro. Na maioria das vezes, esta classificação é um processo manual e demorado, por exemplo, no caso de epilepsia, exige realização de exames complementares caros, como eletroencefalogramas (EEG) ou ressonâncias magnéticas (RM).

Desta forma, justifica-se a necessidade de um processo que permita esta análise e classificação, reduzindo o esforço e tempo para chegar a um correto diagnóstico e sua classificação. Este processo deve também reduzir o erro médico na prescrição de tratamentos ou medicamentos e aumentar a eficácia do processamento médico. De forma a garantir a adoção de uma nova abordagem pela comunidade médica, é necessário que um processo desta natureza apresente as justificações para uma determinada classificação de um diagnóstico e que esta seja compreensível pelo médico.

Consequentemente é proposto um processo para ajudar os médicos pediatras a tomar decisões. Este processo é elaborado num contexto real possibilitando um desenvolvimento e resultados mais relevantes. Esta abordagem proposta utiliza registos clínicos eletrónicos de crianças até aos 17 anos e com a utilização de um conjunto de técnicas de processamento, extração e aprendizagem, foi possível classificar um provável diagnóstico ou até mesmo a classificação do tipo de epilepsia, procedimentos e escolha dos melhores tratamentos com a utilização de códigos *standard,* como ICD-9.

Esta dissertação foi desenvolvida com o apoio do Hospital Santo André de Leiria, que disponibilizou os registos clínicos reais e anónimos eletrónicos e não eletrónicos. Além disto, disponibilizou um suporte técnico para identificar diversos aspetos relevantes, como sintomas, causas, eventos, medicamentos, entre outros, que desempenhou um papel fundamental no decorrer deste trabalho.

1.2 Estrutura da Dissertação

Esta dissertação encontra-se dividida em sete capítulos. Sendo que no próximo capítulo, são apresentados conceitos relacionados com o processo proposto, tais como, *data mining*, *text mining*, registos clínicos, *text mining* aplicado a registos médicos, epilepsia em crianças e códigos *standard.*

No terceiro capítulo é apresentada uma revisão da literatura, referindo os principais projetos encontrados sobre sistemas de suporte à decisão médica, classificação de procedimentos, tratamentos e diagnósticos, e sistemas de gestão hospitalar. Além disto, são ainda identificadas algumas questões de investigação relevantes relacionadas ao âmbito deste projeto.

O quarto capítulo apresenta a metodologia de investigação, onde é descrito o processo de investigação referindo, a gestão de projeto, calendarização, canais de comunicação e âmbito para concluir os objetivos iniciais. Além disso, são identificadas as questões de investigação de acordo com o processo de sugestão e classificação de um diagnóstico segundo a normalização ICD-9. Consequentemente é ainda especificada a abordagem, os procedimentos utilizados, dificuldades e soluções encontradas para o desenvolvimento deste processo.

No quinto capítulo é especificada a implementação do processo apresentado nesta tese, bem como os procedimentos elaborados para a recolha do material necessário. Adicionalmente, neste capítulo são apresentadas as abordagens, bem como os algoritmos utilizados.

Os métodos de avaliação, resultados obtidos e a discussão desses resultados são apresentados no sexto capítulo. São também analisados e discutidos os resultados do processo de apoio ao diagnóstico e do processo de classificação dos códigos *standard*.

No sétimo capítulo, apresenta-se as principais conclusões, sendo resumidos os principais contributos mais relevantes deste trabalho e indicações para trabalho futuro.

2. Conceitos

Neste capítulo são apresentados ao leitor vários conceitos relevantes para uma melhor compreensão da abordagem proposta. O conceito de epilepsia em crianças é abordado, uma vez que um dos objetivos deste trabalho incide na sugestão de diagnósticos nesta área. Além disso, é apresentado o conceito de códigos *standard* para classificar corretamente um diagnóstico na área de epilepsia. Em seguida, são abordados os diferentes tipos de informação médica que permitem uma análise e extração dos aspetos para sugerir e classificar um diagnóstico. De forma a analisar e extrair estes aspetos relevantes de textos são necessárias técnicas de *text mining* que são também explicadas neste capítulo. Por fim, são apresentadas as técnicas chave para efetuar uma análise e extração de informação aplicada á área de registos médicos.

2.1 Epilepsia em Crianças

A epilepsia consiste na existência de convulsões (crises) recorrentes e imprevisíveis que podem ocorrer ao longo do tempo [7]. Uma convulsão é uma manifestação de descargas elétricas cerebrais que podem induzir sintomas de acordo com a sua localização específica no cérebro. Devido a estas descargas elétricas o cérebro não consegue realizar as tarefas normais, causando por exemplo, convulsões, distúrbios de linguagem, alucinações e perdas de consciência. Nem todas as convulsões poderão ser consideradas epiléticas, desta forma existem diferentes procedimentos que consideram as crises frequentes (pelo menos 2 vezes) e que não sejam provocadas por álcool, drogas, envenenamento, doenças ou outros eventos como crises epiléticas [7]. No entanto, é preciso ter em conta que cada hospital tem diferentes procedimentos para um mesmo registo médico.

A epilepsia poderá ser classificada de diferentes formas, dependendo entre outros, do motivo de ocorrência da primeira crise, da observação do paciente durante o episódio, do local de origem cerebral ou dos eventos que despoletaram a crise. Além disso, existem outros tipos de

classificação que podem ajudar no diagnóstico de epilepsia. Uma convulsão poderá ser classificada de diferentes formas, no entanto são geralmente classificadas como parciais, generalizadas ou desconhecidas [8]. Desta forma, as convulsões parciais consistem numa descarga elétrica com origem numa determinada localização do cérebro. Estas crises podem ainda surgir numa determinada localização, alastrando-se a outras partes do cérebro, dificultado a distinção entre crises generalizadas. Crises generalizadas são caracterizadas por uma instabilidade química em ambas partes do cérebro. Convulsões desconhecidas ou convulsões idiopáticas é uma outra classificação atribuída aos casos onde não foi possível identificar a causa da doença.

Exames como, EEG, ressonâncias magnéticas e exames físicos fazem parte do diagnóstico de epilepsia, contudo não existe diagnóstico sem EEG e sem a análise do médico. A história médica ou familiar pode também ser outro fator a ter em conta para identificar crises e ajudar no diagnóstico.

Existem várias diferenças entre epilepsia em adultos e crianças. Embora os tipos de convulsões possam ser as mesmas, as causas normalmente são diferentes. Geralmente, os episódios são mais frequentes em crianças podendo sofrer cerca de centenas por dia. Adicionalmente, as crianças tendem a responder de diferentes formas aos tratamentos do que os adultos, sofrendo diferentes efeitos adversos.

Estes tipos de crises podem ser classificadas de diferentes formas e por diferentes normalizações. Desta forma, na próxima secção apresenta o conceito de códigos *standard*.

2.2 Códigos Standard

A nosologia é a classificação sistemática de doenças [9]. No século XX, quando os programas de seguros médicos responsabilizaram os contribuintes em vez dos pacientes pela assistência médica, a nosologia tornou-se uma questão de grande interesse para os contribuintes públicos e privados [10]. As nosologias mais utilizadas incluem *International Classification of Diseases* 9 ou 10, (ICD-9 ou ICD-10) ou *Systematized Nomenclature Of Medicine Clinical Terms* (SNOMED-CT) [11]. Estas nosologias identificam unicamente um diagnóstico, descrições de sintomas e causas de morte dos seres humanos. A utilização destes códigos têm expandido desde a classificação da informação sobre morbidade e mortalidade para fins estatísticos de diversas aplicações administrativas, epidemiológicas ou de pesquisas em

serviços de saúde. Estes códigos possibilitam uma melhor consistência entre os médicos registando sintomas e atribuindo diagnósticos a cada paciente.

A área de epilepsia tem diversas classificações segundo a normalização ICD-9 situada em "outros distúrbios do sistema nervoso central", classificados com os códigos 340 a 349 [12]. Contudo foram apenas encontradas dez possíveis classificações de epilepsia que melhor se enquadrava a este projeto e são elas: 1) 345.0 "*Generalized nonconvulsive epilepsy"*; 2) 345.1 "*Generalized convulsive epilepsy"*; 3) 345.2 "*Petit mal status"*; 4) 345.3 "*Grand mal status*"; 5) 345.4 "*Localization-related (focal) (partial) epilepsy and epileptic syndromes with complex partial seizures"*; 6) 345.5 "*Localization-related (focal) (partial) epilepsy and epileptic syndromes with simple partial seizures*"; 7) 345.6 "*Infantile spasms"*; 8) 345.7 "*Epilepsia partialis continua*"; 9) 345.8 "*Other forms of epilepsy and recurrent seizures*"; 10) 345.9 "*Epilepsy, unspecified"*. Existem outros códigos de classificação que se deve de ter em conta neste processo, como por exemplo 779.0 "*Convulsions in newborn"*, 780.02 "*Transient alteration of awareness*", 780.2 "*Syncope and colapse"*, 780.31 "*Febrile convulsions*", e 780.39 "*Other convulsions and procedure codes*".

É ainda possível efetuar um mapeamento entre normas utilizando ferramentas como UMLS, fornecendo ainda um suporte para vocabulário médico, relações, sintaxe e sua morfologia, pois desta forma é possível classificar diagnósticos segundo as diferentes normas utilizadas.

A informação relevante para sugerir e classificar diagnósticos, como sintomas eventos, causas, entre outros, está presente na informação médica, como é apresentada na próxima secção.

2.3 Registos Clínicos

Existem diversas formas de informação médica como Informação Hospitalar (IH), Registos de Saúde Eletrónicos (RSE) e Registos Clinico Eletrónicos (RCE).

Estes sistemas de informação hospitalar estão relacionados entre si como está representado na Figura 1.

Os sistemas de informação hospitalar permitem gerir a informação médica, administrativa e financeira de um hospital. Podem conter, entre outros, um resumo da informação do histórico

de pacientes, testes, mecanismos de comunicação destes recursos para o exterior ou interior, bem como documentação de gestão administrativa ou calendarização dos funcionários.

Os RSE são um conjunto de registos médicos de um ou vários pacientes. Permitem seguir ou controlar os pacientes oferecendo mecanismos para aceder a esta informação médica em qualquer parte da instituição ou até mesmo em qualquer local [13]. Além de conter um resumo da informação relativa ao paciente, pode também conter outra informação como, notas de diferentes especialistas, conversas com a família do doente e informação de aplicações que possam fornecer um registo mais completo [14].

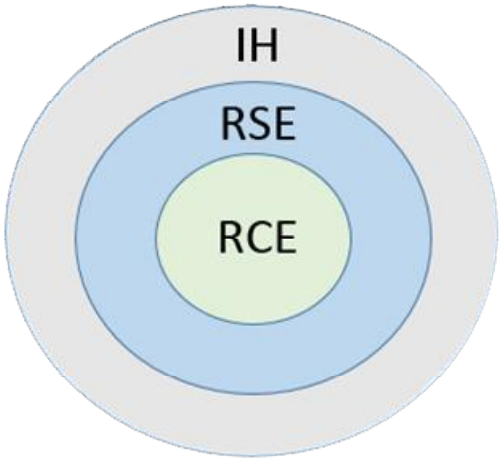


Figura 1 - Relação dos principais tipos de informação hospitalar, Informação Hospitalar (IH), Registos de Saúde Eletrónicos (RSE) e Registos Clínicos Eletrónicos (RCE)

Os RCE permitem guardar toda a informação do paciente em formato eletrónico. Esta informação pode conter sintomas, resultados de exames, anotações, observações feitas por um ou vários médicos, bem como a discussão com o paciente. Não só guarda esta informação, mas também fornece apoio à decisão do médico, organização, calendarização e comunicação. Esta informação poderá constar em artigos, pósteres ou então em relatórios escritos por médicos, que podem descrever as suas doenças, sintomas, historial médico e social, discrição de consultas, problemas, resultados de exames, etc. Estes textos clínicos muitas vezes não estão estruturados, são gramaticalmente incorretos, apresentam abreviações, termos culturais, entre outros, o que pode dificultar a sua interpretação e classificação.

Em 1907 a clinica Mayo, foi a pioneira na área de registos clínicos, fornecendo um processo centralizado de registos médicos dos pacientes, onde cada um teria a sua própria ficha. Mas

estes registos continuavam a estar desorganizados e só em 1960 é que Lawrence Weed começou a especificar uma normalização para estes registos de cada paciente. Estes recursos só poderiam ser acedidos por uma pessoa de cada vez, necessitavam de grande espaço para serem armazenados e teriam de estar organizados para um rápido acesso. Apenas em 1972 o instituto de Regenstrief desenvolveu um dos primeiros sistemas digitalizados de RCE [15].

Os registos clínicos podem apresentar-se em formato estruturado ou não estruturado. Porém, na maioria das vezes os dados encontram-se de uma forma não estruturada, com uma semântica específica dependendo de cada região e de cada escola de medicina, onde são utilizadas técnicas ou vocabulários adicionais, de forma a conseguir compreender e extrair conteúdo dos textos médicos. Isso faz com que a perceção do conteúdo seja mais difícil de entender, exigindo um maior esforço e tempo para extrair e classificar a informação.

Como é necessário proceder a análise destes registos médicos eletrónicos são necessárias técnicas de *text mining* para compreensão e extração de informação relevante em textos. Contudo, de forma a perceber esta técnica é primeiro necessário compreender o conceito de *data mining*, apresentado na próxima secção.

2.4 Data Mining

Data mining é um processo que permite compreender e descobrir padrões em grandes conjuntos de dados para adquirir conhecimento relevante [16].

É difícil saber quando este conceito realmente surgiu. Por exemplo, os algoritmos de Bayes são muito utilizados em *data mining*, sendo introduzidos no século XVIII [17]. Em 1950, foi elaborada uma análise de problemas em computadores, sendo também desenvolvidas as primeiras ferramentas de *software* relacionadas com análise estatística, já que os problemas começavam a ser relativamente complexos [17]. Contudo apenas em 1989-1991 foi introduzido o termo de *data mining* pelo investigador Gregory Piatetsky-Shapiro [18].

Existem diversas técnicas que podem ser usadas no processo de *data mining*, como por exemplo, Associação, Classificação, *Clustering*, e Predição [19]. A técnica associação permite identificar padrões de acordo com relações entre vários itens numa transação. Esta técnica possibilita, por exemplo a identificação de produtos que um consumidor costuma comprar, podendo verificar os produtos mais consumidos permitindo uma melhor campanha

de *marketing*. A classificação é uma técnica de aprendizagem, que é utilizada para classificar um item segundo um conjunto predefinido de classes ou grupos [20], como se pode observar pela Figura 2. São utilizadas técnicas de dedução como árvores de decisão, programação linear, redes neuronais, entre outras, de forma que seja possível aprender a classificar corretamente a informação. Serve por exemplo para classificar os tipos sanguíneos "A", "B", "AB" ou "O". Como se pode observar pela Figura 2, *clustering* é uma técnica que agrupa objetos com características similares [20]. Além de definir a classe para cada objeto agrupando-as. Assim, isto permite pesquisar mais rapidamente um objeto sempre que for necessário, já que é possível procurar pela sua classe em vez de pesquisar toda a informação. Pode-se aplicar em casos para agrupar textos que discutem o mesmo assunto. Predição é uma técnica semelhante a classificação que identifica relações entre os valores e permite prever futuros resultados [21]. Esta técnica pode ser usada para prever as receitas e efetuar saldos a um determinado produto. Para tal, tem de se ter em conta o historial das vendas e das receitas, de forma a construir uma curva de regressão para ajudar a analisar as futuras receitas.

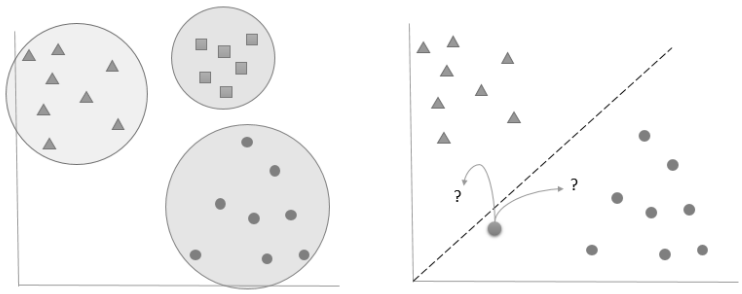


Figura 2 - Exemplo das técnicas de *Clustering* (na esquerda) e Classificação (na direita)

Data mining utiliza ainda métodos de *machine learning* que, de uma forma genérica, evoluíram da inteligência artificial. Estes algoritmos permitem analisar os padrões e aprender a partir dos dados, de modo a construir modelos que classificam a informação que foi ou não previamente conhecida. Existem diferentes estratégias de aprendizagem que podem ser utilizadas, como aprendizagem supervisionada, não supervisionada e semi-supervisionada [22]. A aprendizagem supervisionada é o processo que possibilita a construção de modelos com base em exemplos fornecidos e previamente classificados por uma entidade credível (supervisor). A aprendizagem não supervisionada pretende deduzir e classificar a informação sem conhecer o resultado. Por fim, a aprendizagem semi-supervisonada é um processo no qual apenas são conhecidos alguns resultados e a restante informação é deduzida, de forma a alcançar um resultado correto.

Existem várias metodologias de investigação que permitem definir a melhor forma de pôr em prática estes conceitos, como é o caso da metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), *Knowledge Discovery in Databases* (KDD) e *Sample, Explore, Modify, Model and Assess* (SEMMA).

Em 1996, Fayyad desenvolveu uma das primeiras metodologias de investigação, chamadas *Knowledge Discovery in Databases* (KDD) [23]. O KDD é o processo de utilização de métodos *de data mining* para extrair conhecimento de acordo com os objetivos especificados [24]. Existe cinco etapas para a realização deste processo são elas: Seleção, Pré-Processamento, Transformação, *Data mining* e Interpretação. A seleção permite selecionar informação a partir de um conjunto de variáveis ou conjuntos de informação. O pré-processamento consiste na limpeza da informação de forma a obter informação consistente. Na transformação é necessário aplicar técnicas e métodos relevantes ou reduzindo a dimensionalidade (número de variáveis). Na etapa de *data mining* é necessário procurar padrões de interesse na informação. Por fim, a interpretação ou avaliação dos padrões encontrados [25].

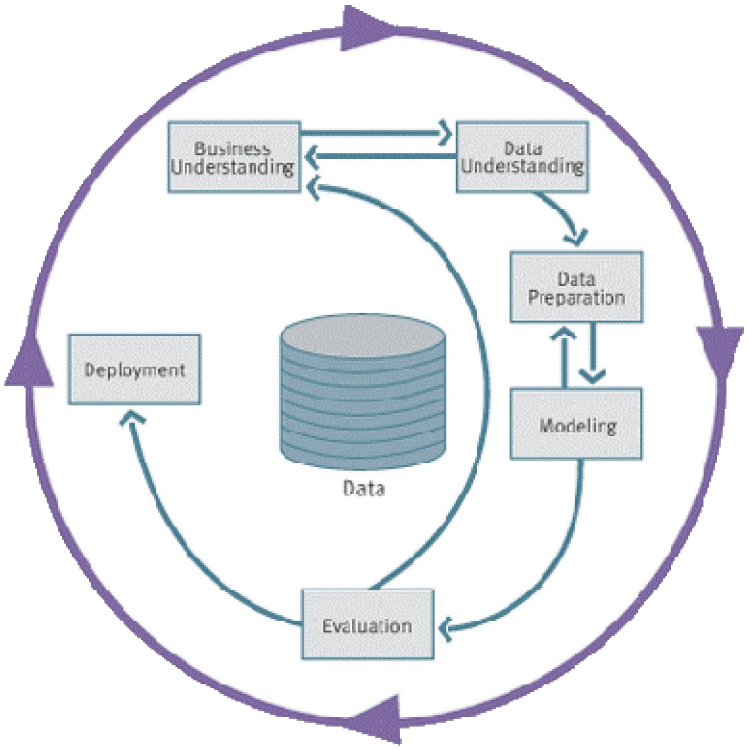


Figura 3 – Etapas do modelo CRISP-DM, figura adaptada [26]

Como se pode ver pela Figura 3, CRISP-DM [27] é constituída pelos seguintes passos *problem definition, data exploration, data preparation, modeling, evaluation* e *deployment.* No passo de *problem definition* é necessário clarificar o problema, os objetivos e requisitos. O passo de *data exploration* consiste em interpretar e explorar a informação. A *data preparation* consiste em extrair, limpar e formatar a informação necessária; em *modeling* são aplicadas funções, algoritmos ou redução de dimensões para obter a informação importante. Em *evaluation* determina-se o modelo que clarifica as expectativas ou objetivos, valida-se o modelo e decide-se como se vai utilizar os resultados do *data mining*. Em seguida observa-se os elementos similares entre itens através dos dados de treino e tenta-se otimizar o modelo caso seja possível. Por fim, *deployment*, em que se exporta os resultados obtidos, para caso de utilizar estes dados no futuro. [28, 29]

O *Sample, Explore, Modify, Model and Assess* (SEMMA) foi desenvolvido pelo instituto *Statistical Analysis System* e é outro exemplo de uma metodologia de *data mining*. Inicialmente este processo recolhe e determina a informação relevante. Esta informação deve ser completa, mas em pequena quantidade para ser eficiente. Em seguida tenta-se descobrir relações entre variáveis e anormalidades através de técnicas estatísticas, como *factor analysis, clustering,* entre outros, para uma melhor visualização da informação. Depois é necessário preparar a informação (selecionando, criando e modificando variáveis) de forma a construir um modelo que nos permita chegar ao nosso objetivo [25, 30].

Existem diversas ferramentas computacionais para realizar o processo de *data mining*, como por exemplo: Weka, *Unstructured Information Management Architecture* (UIMA), *Flexible Toolbox for Performing Classification Analyses* (*ML-Flex*)*, LingPipe* e S-EM. O weka é um conjunto popular de *software* de aprendizagem e análise, escrito em Java; UIMA desenvolvido pela *International Business Machines* (IBM) e é uma ferramenta para analisar conteúdo não estruturado como texto, vídeo e imagem; *ML-Flex* que permite a integração com terceiros, executando a análise de classificação em paralelo, produzindo relatórios *HyperText Markup Language* (HTML) e resultados de classificação; *LingPipe* é uma ferramenta para analisar a linguagem humana, escrito em java e S-EM ou *Spy-EM* que permite a análise consoante as técnicas de *text mining*. As ferramentas mais populares e *open source* são: *Rapid Miner* e R. *Rapid Miner* é uma ferramenta para classificação, descoberta de *clusters* e *outliers*, utilização de *association*, *text mining*, visualização da informação, análise sequencial, utilização de *prediction* de negócio; e R é uma ferramenta para classificação,

descoberta *clusters*, *association*, *text mining*, descoberta de *outliers*, visualização da informação, análise *web* e análise de redes sociais [31, 32].

2.5 Text Mining

Text mining é uma área de especialização de *data mining* aplicada à extração e análise de texto [33]. Estes textos podem estar em formato estruturado ou não estruturado, sendo que a informação é mais difícil de analisar e extrair em textos não estruturados.

Não é possível especificar com exatidão a data de quando este termo surgiu. No entanto em 1958, Luhn descreveu um sistema que permitia analisar automaticamente a informação de documentos de acordo com os interesses de uma entidade [34]. Apenas em 1999 Hearst refere a relação entre *text mining* e *data mining* [33].

Este processo é complexo porque requer, entre outros aspetos, estudos de frequência de palavras, classificação de palavras, a compreensão do significado de cada palavra, análise sintática e lexical consoante os objetivos que se pretende alcançar. Além disso, esta complexidade aumenta quando se procura resolver problemas de grande complexidade em que o número de características é elevado. Este é o caso de um diagnóstico médico em que, o número de fatores a ter em conta para efetuar um diagnóstico é enorme.

Diferentes ações de processamento podem ser utilizadas dependendo dos formatos estruturados ou não estruturados aplicados aos textos. Como muito destes textos poderão conter erros ortográficos, sintáticos, entre outros fatores, é necessário realizar um pré-processamento, utilizando uma verificação ortográfica, efetuando uma identificação da estrutura do documento, ou até mesmo remover *stopwords* se desejável, que são palavras que não fornecem um grande significado a uma frase ou expressão, e.g. “a”, “de”, “os” [35]. Em seguida, são utilizadas outras técnicas como t*okenization* [36] e *stemming* [37]. A técnica *tokenization* permite identificar e dividir o texto em palavras, frases, ou outros elementos como símbolos ou pontuações. O *stemming* consiste em identificar palavras com pequenas variações sintáticas, mas que referem significados semelhantes, e.g. “esperei”, “esperou”, “esperando”, etc.

É possível ainda utilizar outras técnicas, como por exemplo, *negation handling* que permite a deteção de negações e *entity recognition* ou reconhecimento de entidades [38], que são

utilizadas para classificar entidades analisando palavras, classes, terminologia similar e abreviações. A técnica *word sence disambiguation* pode também ser utilizada num pré-processamento para compreender o significado de cada termo de acordo com o seu contexto [36].

Existem diferentes métodos que podem ser utilizados em *text mining* adaptados do *data mining*, como *text summarization*, *information retrieval* e *clustering*. O *text sumarization* é uma técnica que permite analisar e extrair os pontos mais importantes de um texto, de forma a construir um resumo. A i*nformation retrieval* ou *document retrieval* permite localizar e extrair informação mais rapidamente através de consultas realizadas pelos utilizadores [39]. Existem outras técnicas que podem ser utilizadas para uma rápida extração como v*ector space model* [40]. Esta técnica representa documentos ou pesquisas através de vetores, de forma a tentar encontrar semelhanças entre eles. Estes vetores contêm as palavras-chave necessárias extraídas de cada documento. Desta forma, é ainda possível utilizar ontologias, para identificar e descrever palavras e suas relações. Estas ontologias estão organizadas de uma forma hierárquica, constituída por classes, subclasses, propriedades, atributos e instâncias [41].

Existem muitas ferramentas que ajudam no processo de *text mining*, como é o caso de *Alchemy* API, *Natural Language Toolkit* (NLTK), *Wandora*, *Protégé*, *General Architecture for Text Engineering*[2] (GATE), *Rapid Miner*[3], R[4] e *Textpresso*. *Alchemy* API[5] está na *cloud* e utiliza classificação e marcação semântica, identificação de linguagem, extração por *keywords*, categorização e muito mais, O NLTK[6] é um conjunto de bibliotecas de processamento de linguagem natural para classificar palavras, utilizando, *tokenization*, *stemming*, *tagging*, *parsing*, e análise semântica. O *Wandora*[7] é uma ferramenta que permite ajudar no processo de extração de informação, gestão, publicação e ainda ajuda a contruir ontologias. O *Protégé*[8] permite criar e editar ontologias, quer em plataforma web, quer plataforma Java. O GATE é uma ferramenta popular para extração de texto [36], *Rapid Miner* e R que também permite efetuar extração e análise de *text mining*. O *Textpresso* é um sistema

[2] http://gate.ac.uk/
[3] http://rapid-i.com/
[4] http://www.rdatamining.com
[5] http://www.alchemyapi.com/
[6] http://nltk.org/
[7] http://www.wandora.org/www/
[8] http://protege.stanford.edu/

baseado em ontologias para extrair informação clinica em textos médicos. Fornece um conjunto de *standards* categorizar entidades médicas e suas relações. Alem disso, processa estes documentos indexando as frases e utilizando termos ou padrões da ontologia em questão.

Como esta dissertação incide num processamento de informação médica com vista a classificar um correto diagnóstico, é apresentado na próxima secção os procedimentos, técnicas e ferramentas relevantes na área de *text mining* aplicada a registos médicos.

2.6 Text Mining Aplicado a Registos Clínicos

Como foi possível verificar, os registos médicos de cada paciente representam uma grande fonte de informação, difícil de analisar, utilizando muitas vezes um formato não estruturado com uma gramaticalidade e sintaxe complexa. Para tal é necessário recorrer a técnicas que permitam efetuar esta análise e extração mais rápida, reduzindo custos e recursos, de forma a proporcionar um melhor suporte para o médico, para o paciente e para a instituição.

Devido às dificuldades em extrair, classificar e analisar a informação médica foi desenvolvido, em 1986 [42], o *Unified Medical Language System* (UMLS). Este sistema fornece um grande conjunto de vocabulário médico, descrições, relações através de ontologias, sintaxes, morfologias (forma e estrutura) e possibilita uma verificação ortográfica. Permite ainda processamento de linguagem natural, ajudando ao desenvolvimento de sistemas de médicos baseados em *text mining* na língua inglesa.

Existem diferentes processos na extração e análise da informação do paciente, como a extração dos termos médicos, para deduzir o historial do paciente, encontrar relações entre esses diferentes termos, identificando entre outros, as diferentes partes do corpo e seus sinónimos.

Além disto, é importante manter a confiabilidade e segurança da informação de cada paciente [43]. Assim sendo, é necessário um processo que remova ou modifique as palavras que descrevem a informação pessoal de cada paciente. Esta confiabilidade e segurança poderão ser asseguradas utilizando, entre outros, a normalização *Health Insurance Portability and Accountability Act* de 1996 (HIPAA) [44], que permite proteger esta informação pessoal.

Várias abordagens podem ser utilizadas para extrair informação no campo da medicina. O *pattern-matching* é um exemplo muito utilizado para descobrir padrões em frases, expressões ou palavras. No entanto, este conceito não pode ser generalizado, ou seja, é difícil de estabelecer padrões relevantes para novos domínios ou até mesmo para diferentes linguagens. Outras possíveis abordagens aplicadas à área médica são *shallow* e *full syntactic parsing* para textos simples. *Shallow parsing* permitem identificar os vários componentes da frase segundo uma determinada sintaxe. Esta técnica é bastante útil, nomeadamente, para a realização de resumo dos documentos e para tradução de textos [45]. O *full syntactic parsing* é de uma forma geral semelhante, mas com uma estrutura sintática mais complexa [46]. A *Ontology-driven extraction* é outra abordagem que permite classificar e relacionar palavras, através de ontologias.

Como foi possível verificar na secção *Text Mining*, existem diversas técnicas para realizar um pré-processamento, tais como uma verificação ortográfica. Em seguida é geralmente utilizado um *tokenizer*, *negation handler* para detetar quando uma paciente não tem um certo sintoma e outras formas para analisar e interpretar a informação. Também pode ser utilizado *Word Sense Disambiguation* [36] que consiste em entender o sentido de cada palavra dependendo do contexto.

É importante também ter em conta a temporalidade que poderá ser utilizada para decifrar eventos ou termos médicos, uma vez que é importante conhecer quando um sintoma realmente surgiu, permitindo construir e analisar o histórico do paciente entre outros fatores.

Torna-se também difícil de analisar entidades, palavras ou classes de conceitos médicos, já que estes conceitos médicos contêm uma sintaxe e morfologia variável (por exemplo, sinónimos, semelhantes terminologias, etc). Em medicina existem diferentes sinónimos, abreviações e acrónimos que referem o mesmo conceito, sendo possível simplificar a análise destas diferentes palavras utilizando ontologias. Desta forma, é possível comparar a informação e suas relações entre as diferentes palavras simplificando a análise e extração de termos relevantes.

Um dos próximos passos consiste em determinar a relações entre os diferentes termos encontrados. Para tal, podem ser utilizadas várias técnicas como, *Graph-based Relation Extraction* onde se relaciona diferentes termos como o número da pressão arterial com a palavra pressão arterial, ou associação de doenças e sintomas, datas, partes do corpo de uma

forma semântica ou sintática. Pode também ser utilizado *Link Grammar Parser*, para poder analisar significados e ligações que possam existir entre frases ou palavras. Analisando apenas palavra a palavra, podendo ser necessário métodos para analisar *multi-words,* como "tem vindo a ficar".

Após este pré-processamento, poderá ser aplicada uma abordagem de *machine learning,* para aprender informação fornecida e construir modelos para classificar essa informação que não foi previamente deduzida. Existem diferentes algoritmos de *machine learning* para classificação, contudo deverá ser importante apresentar resultados de algoritmos de *white box*. Estes algoritmos permitem conhecer as características de que baseiam, reconhecendo possíveis sintomas ou fatores que levam a um determinado diagnóstico. A utilização destes algoritmos *white box* são importantes, pois aumenta a probabilidade de adoção de um sistema pela comunidade médica porque assim, é possível perceber a lógica e as características que levaram aos resultados obtidos.

Existem algumas ferramentas que podem ser utilizadas na área médica como GATE, *REgenstrief eXtraction tool* (REX) e *Health Information Text Extraction* (HITEx). GATE é uma ferramenta gratuita para *text mining* sendo muito utilizada na área médica de cancro e raios X [47]. REX é uma ferramenta que permite descobrir problemas que possam existir em por exemplo os raios X. HITEx permite ainda extrair, entre outros, a informação de diagnósticos, medicações e estados patológicos dos pacientes.

Nesta secção, foi possível entender o processo e técnicas utilizadas para classificar diagnósticos médicos eletrónicos. Contudo esta dissertação incide também na classificação de registos médicos na área de epilepsia para crianças. Desta forma, é então necessário perceber o conceito, os processos de classificação em epilepsia e as diferenças entre epilepsia em adultos e crianças.

2.7 Síntese

Foi apresentado neste capítulo uma contextualização ao leitor, para melhor compreender a abordagem proposta de análise de registos clínicos na área pediátrica de epilepsia para sugestão e classificação de um correto diagnóstico.

Foram descritos vários conceitos como por exemplo, epilepsia, códigos *standard*, registos clínicos, *data mining*, *text mining* e *text mining* aplicado a esta área médica. Foram abordadas as diferenças entre epilepsia infantil e epilepsia em adultos, tipos de classificação e informação a utilizar para analisar e extrair características como, sintomas, causas, eventos, entre outras. Apresentaram-se ainda os conceitos e técnicas para efetuar esta análise, de forma a sugerir e classificar um correto diagnóstico.

3. Revisão de literatura

Nesta secção são discutidos os principais trabalhos que analisam a informação médica utilizando técnicas de *text mining*, códigos *standard*, entre outros. Vários trabalhos foram analisados relativamente às áreas de suporte à decisão, classificação de procedimentos, tratamentos e diagnóstico, e gestão hospitalar. Algumas questões de investigação em aberto são também apresentadas para dar a conhecer ao leitor as áreas de maior relevância.

3.1 Projetos de Investigação Relacionados

Nesta secção são abordados os estudos relativamente a projetos de investigação encontrados nas áreas de sistemas de suporte à decisão, de classificação de procedimentos, tratamentos e diagnósticos, e na área de gestão hospitalar.

3.1.1 Sistemas de Suporte à Decisão

Os sistemas de suporte à decisão médica são desenvolvidos de forma a ajudar os médicos e outros profissionais a efetuar decisões mais informadas, classificando diagnósticos, examinando análises, entre outros. Além disso, estes sistemas podem controlar custos onde, é possível monitorizar os pedidos de medicamentos e gerir a complexidade clinica, isto é, acompanhar pedidos e realizar um atendimento preventivo. Estes sistemas podem também ajudar no apoio administrativo, classificando procedimentos e documentos, podendo reduzir o erro médico, erro na prescrição e evitar reações adversas no tratamento de um paciente, poupando tempo e dinheiro.

Existem diferentes exemplos de sistemas de suporte à decisão, como as ferramentas do *REgenstrief eXtraction*, que permitem descobrir padrões e utilizar regras baseadas em *text mining* para extrair informação de radiologias, notas de admissão e relatórios patológicos. Este sistema utiliza expressões regulares para detetar palavras-chave ou frases, de forma a

relacioná-las com um conceito específico, determinando o contexto de cada documento, para mais rapidamente consultar um documento sempre que necessário [48].

O *Medical Language Extraction and Encoding System* (MedLEE) é outro exemplo de sistema utilizado para diferentes tarefas. Tem sido utilizado para detetar características relacionadas com o cancro da mama e para processar radiologias. Utiliza também *machine learning* para detetar características anormais em relatórios de radiologia portuguesa. Além disso, foi utilizada uma estrutura de vigilância para identificar, eventos adversos relacionados com cateteres venosos e codificação da informação clinica [49]. O *CliniViewer* é um exemplo de muitas aplicações que utilizam *MedLEE* para resumir e navegar pelos textos clínicos [50].

O *SymTex* e *Mplus* utilizam análise semântica para inferir relações entre termos e o seu significado. Estas ferramentas podem ser utilizadas para analisar interpretações de exames pulmonares [51], para detetar pneumonias [52], classificar pacientes com traumas [53] e para analisar radiografias ao tórax [54].

3.1.2 Classificação de Procedimentos, Tratamentos e Diagnóstico

A classificação de diagnósticos, procedimentos e tratamentos tem sido uma abordagem bastante popular nos últimos anos. Em 2007, no *Computational Medicine Challenge,* no âmbito da execução de tarefas partilhadas em vários domínios médicos reutilizando registos médicos anónimos de radiologias, foi realizado um projeto para analisar e classificar estes registos de acordo com os códigos ICD-9 correspondentes [55].

Existem outros tipos de sistemas que permitem extrair códigos de diferentes contextos, como por exemplo *MedLEE*, que classifica a gravidade de uma pneumonia consoante os diferentes relatórios médicos [56].

O *Atigeo* é outro exemplo de um sistema que analisa registos clínicos eletrónicos e recomenda um código ICD-9 que represente esse diagnóstico. Além disso, permite também classificar procedimentos descritos em registos médicos. Esta abordagem foi proposta na *Text Retrieval Conference* em 2012, para promover a investigação e desenvolvimento de mecanismos de pesquisa em textos não estruturados, para poder identificar registos clínicos relevantes de acordo com determinadas consultas. Foi também utilizada a abordagem *Natural Language Pre-Processor* para reduzir a complexidade lexical e ambiguidade nos registos médicos e nas

consultas realizadas. Os *International Codes of Diseases* – 9th *Revision* (ICD-9) foram extraídos dos campos especificados de registos médicos para facilitar as pesquisas.

O *Computer Assisted Medical Information Resources Navigation & Diagnosis Aid Based on Data Marts & Data Mining* (CAIRN-DAMM) é um projeto para o hospital universitário de Areteion na Grécia, que consiste na gestão e consulta de documentos, classificação de diagnósticos com base em códigos ICD-9 e recolha de informação. Este projeto tem também o objetivo de armazenar informação médica, e.g. documentos, ficheiros multimédia, organizar e consultar documentos com base em *Natural Language Queries* [57]. O sistema interpreta linguagem humana permitindo consultas baseadas em termos, palavras-chave que aparecem nos respetivos documentos, consistindo por entidades que podem ser por exemplo diagnósticos, pessoas, organizações, entre outros. Além disso, é utilizada uma lista ordenada para apresentar uma classificação consoante o termo e suas relações, apresentando os documentos que mais se aproximam à pesquisa. Para cada documento é também guardado um diagnóstico baseado em ICD-9 de acordo com os termos apresentados em documentos.

Outro exemplo é o estudo que tem o objetivo de ajudar os profissionais a atribuir códigos ICD no Hospital Universitário de Geneva na Suíça. Este projeto utiliza códigos ICD-10 e um vocabulário francês para identificar e classificar palavras [58]. Adicionalmente, este método é bastante resistente a fenómenos de *overfitting* [59], que leva a que os resultados tenham uma menor percentagem de valores inconsistentes. Estes valores poderão ser causados pelo baixo número de resultados disponíveis ou pelo grande conjunto de atributos e seus possíveis valores. Os documentos são pré-processados removendo *stopwords*, *negation handling*, *stemming* e *spellchecking*, entre outros. Em seguida, é utilizada aprendizagem supervisionada para conhecer diagnósticos que não conseguiram ser previamente classificados através de regras.

Existem outros trabalhos que utilizam *text mining* em diferentes áreas de epilepsia, como é o caso do projeto de investigação do hospital psiquiátrico da Dinamarca [60], no qual consegue extrair informação através da recolha de descrições fenotípicas[9] de cada paciente dos registos médicos. Este projeto tem como objetivo a classificação com base em ontologias ICD-10, de

[9] Sistema de classificação de organismos baseado nas semelhanças ou diferenças consoante o número de características que podem ser observadas

forma a obter estatísticas da ocorrência de doenças e estatísticas estratificação de cada paciente.

Outro estudo foi realizado em várias organizações de saúde das clinicas *Kelset-Seybold* em Houston, para que fosse desenvolvido um algoritmo que permitisse a classificação de casos de epilepsia segundo a norma ICD-9. Esta classificação era feita através da extração da especificação de diagnósticos descritos pelos médicos nos campos preenchidos do registo médico, bem como a análise dos campos dos procedimentos e medicação utilizada [61]. Este estudo focou-se na construção de um algoritmo que poderia maximizar a sensibilidade e especificidade para aumentar a percentagem de valores corretamente classificados dos registos de pacientes adultos.

3.1.3 Sistemas de Gestão Hospitalar

Existem diferentes aplicações que permitem explorar e analisar informação médica para melhor identificar e acompanhar os pacientes de alto risco, projetar intervenções apropriadas e reduzir o número de intervenções e reclamações hospitalares, através de uma análise de custo benefício.

Estes sistemas podem também prevenir ataques terroristas [62], identificando surtos e projetando procedimentos, de forma a controlar estas epidemias.

O *Green, Amber, Red Delineation of Risk and Need* é um sistema que permite aumentar o custo-eficácia de prevenção e gestão de doenças [63]. O risco e a necessidade de prevenir doenças crónicas deve de ser avaliado, de forma a calcular o custo-eficácia da frequência e intensidade da intervenção que se deve realizar a pacientes de alto risco, ou pacientes com doenças cardíacas. Além disso, este sistema permite a construção de relatórios médicos na área de cardiologia automaticamente.

Existem outros sistemas que suportam a gestão hospitalar, identificando por exemplo a potencial falha de certos mecanismos e produtos, tais como máquinas de raio X e medicamentos. Com estes mecanismos é também possível uma avaliação e previsão da confiabilidade do produto através de condições específicas. Esta abordagem utiliza sensores para recolher informação e reconhecimento de padrões estatísticos para detetar mudanças na informação, isolar falhas e estimar a vida útil do produto. Assim sendo, permite a

identificação do desvio ou da degradação de um produto a partir de uma condição normal esperada, até uma previsão do estado e da sua fiabilidade [64].

3.2 Questões de Investigação em Aberto

Nesta secção são discutidas algumas questões em aberto sobre o *text mining* utilizado para processar registos clínicos eletrónicos em diversas áreas, nomeadamente epilepsia para sistemas de apoio à decisão, classificação de procedimentos, tratamentos e diagnósticos, sistemas de gestão hospitalar e formas de promover a aceitação destas tecnologias.

3.2.1 Sistemas de Suporte à Decisão

Os sistemas de suporte à decisão têm um enorme potencial para melhorar a qualidade dos cuidados médicos, orientando os médicos sobre os melhores procedimentos e tratamentos a utilizar.

Existem diversos campos com elevado potencial de investigação, tais como, a análise de relatórios de radiologia ou análise de exames de laboratório, e.g. exames ao sangue, tensão e ao coração [65]. Estes exames são bastante importantes na área da gestão médica, uma vez que os médicos precisam muitas vezes de rever e avaliar os resultados dos diferentes pacientes [66]. Além disto, as doenças cardíacas e cancro são outros exemplos de áreas de investigação promissoras. O cancro é principal causa de morte em muitos países e os tratamentos são excessivamente complexos e caros [67]. Existem estudos iniciais nesta área, classificando tipos de cancro através da verificação de genes *methylated* [68]. As doenças cardíacas são outra principal causa de morte, onde existem diversas doenças associadas com sintomas específicos [69] [70]. O *text mining* na área médica pode também ser utilizada para explorar e classificar imagens, como RM ou EEG [71] .

Além destes sistemas de enorme potencial no suporte à decisão, é ainda possível lembrar os médicos dos tratamentos de pacientes em alto risco, fornecendo ainda sugestões para alterar os procedimentos e tratamentos consoante a sua evolução e reação [65].

O *text mining* pode também ser aplicado para aumentar a eficácia de tratamentos comparando as causas, sintomas e analisando a evolução dos tratamentos. Desta forma, pode fornecer um papel importante na gestão médica no acompanhamento de pacientes de alto risco e na deteção de erros em prescrições inadequadas.

A epilepsia é uma área relevante de investigação onde a sugestão e classificação de diagnósticos, pode possibilitar uma redução do erro médico na prescrição, no diagnóstico e um aumento da eficiência na classificação. Assim, analisando todas as características relevantes de um registo médico é possível identificar os sintomas, causas, eventos, entre outros, de forma a reduzir o tempo e o esforço para a sugestão e classificação de um correto diagnóstico.

Nesta secção, foram apresentadas as áreas, nomeadamente epilepsia, de enorme potencial para o suporte à decisão médica. Contudo, os médicos necessitam de classificar diagnósticos, procedimentos e tratamentos consoante normas para que possam ser interpretados em todo o mundo. A próxima secção identifica também as áreas de enorme potencial de investigação para a classificação destes procedimentos, tratamentos e diagnósticos.

3.2.2 Classificação de Procedimentos, Tratamentos e Diagnóstico

A *World Health Organization* está a rever os códigos ICD, desenvolvendo a décima primeira revisão até 2015[10]. Esta revisão vai adicionar algumas funcionalidades para capturar o impacto de uma doença, entre outras modificações. Além disto, esta tarefa pretende dissolver as divisões de classificação entre adultos e crianças, e o reconhecimento de doenças entre diferentes culturas [72]. Desta forma, quer os médicos quer os investigadores podem contribuir para unificar esta revisão do código ICD com *a International Classification of Functioning, Disability and Health* (ICF) construindo uma normalização a nível mundial [73].

Os sistemas que usam códigos de classificação médica podem reduzir os custos, tempo e ajudar os médicos ou os pacientes a analisar diagnósticos, medicações ou até mesmo que procedimentos devem optar.

Existem diversas áreas onde é complexo e não é linear a classificação de um diagnóstico. Os médicos têm de ter em conta diversas características, como sintomas, causas, eventos, aspetos metabólicos, estruturais, genéticas, entre outras. Desta forma, é possível utilizar normalizações, para classificar diagnósticos consoante as características que um paciente possa ter.

[10] http://www.who.int/classifications/icd/revision/en

Nesta secção foram abordadas algumas áreas de relevante investigação para a classificação de procedimentos, tratamentos e diagnósticos. No entanto, existem outras questões para a administração hospitalar que podem ter um grande impacto para a investigação, como é apresentado na próxima secção.

3.2.3 Sistemas de Gestão Hospitalar

A redução de desperdício é uma das diversas questões em aberto na área de gestão hospitalar. Esta redução deste desperdício pode permitir a redução de milhões para os *Centers for Medicare & Medicaid Services*. Isto pode ser realizado através da adoção das melhores práticas quer pelos centros hospitalares, quer pelos médicos, reduzindo possíveis erros médicos, tratamento e atendimento ineficaz [74].

A previsão de surtos de doenças é outra área relevante para investigação, uma vez que a administração necessita de prever um surto de doença, de forma a tomar medidas mais eficazes consoante esse surto [75].

3.2.4 Promover a aceitação desses sistemas pela comunidade médica

Os médicos ainda estão apreensivos com este tipo de tecnologia, pois estes necessitam de soluções que permitam identificar as razões para uma classificação com uma precisão segura. Uma abordagem *white box* permite fornecer uma explicação de um determinado resultado, i.e. fornecendo um *feedback* na tomada de decisão, por exemplo é possível especificar os sintomas para um determinado resultado [76].

Segurança e controlo de acesso são outras funcionalidades que se deve ter em consideração para uma boa aceitação pela comunidade médica. A informação dos pacientes deve ser apenas acedida pelas pessoas autorizadas e devem ser tomadas medidas para ocultar ou substituir a informação privada de cada paciente.

3.3 Síntese

Foi possível observar nesta secção os principais projetos onde a partir de registos clínicos eletrónicos sugere-se um diagnóstico utilizando regras previamente definidas, utilizando de processos e técnicas de *text mining*. Verificou-se a utilização de algoritmos que pudessem detetar e classificar doenças a partir da descrição dos médicos em campos nos registo clínicos médicos.

Foi ainda possível verificar algumas áreas de investigação em aberto, como para sistemas de suporte à decisão médica, classificação de procedimentos, diagnóstico e tratamentos, sistemas de gestão hospitalar e procedimentos para aumentar a confiança da comunidade médica nestas tecnologias.

Foram apresentados alguns projetos relevantes na área de epilepsia, onde foi possível concluir que nenhum destes projetos permite uma identificação de epilepsia infantil ou de registos médicos portugueses. Além disso, muitos destes projetos apenas se limitam a identificar e extrair informações descritas pelos médicos em campos de registos médicos eletrónicos.

Uma solução que permita classificar e diagnosticar possíveis casos de epilepsia, classificando segundo as normalizações dos países respetivos seria uma mais-valia. Desta forma, no próximo capítulo será então abordada a solução proposta mediante as restrições dos projetos apresentados neste capítulo e as necessidades da comunidade médica.

4. Metodologia e Processo de Apoio ao Diagnóstico e Classificação ICD-9

Nesta secção é apresentada a metodologia de investigação escolhida, referindo e justificando como foi adaptada a este projeto. Em seguida, será apresentada a solução proposta tendo em conta os procedimentos e normalizações utilizadas pela comunidade médica.

4.1 Processo de Investigação

O processo de investigação permite através de uma pesquisa a recolha sistemática de informações, que obedecendo a um sistema de normas, é possível analisar e selecionar técnicas, processos e ferramentas necessárias para o processo proposto, permitindo explorar uma ideia de forma a resolver um ou mais problemas. Assim, é possível alcançar novas ideias, para identificar os problemas, dificuldades e possíveis soluções na área.

Existem várias metodologias de investigação como foi possível verificar na secção de *data mining*. Para este processo foi escolhida a metodologia CRISP-DM, já que é uma metodologia bastante utilizada e que melhor se adapta à elaboração deste projeto.

Desta forma, foi elaborado o processo baseado na metodologia de investigação CRISP-DM, apresentado na Figura 4, onde inicialmente se realizou uma revisão da literatura, identificando os principais procedimentos e ferramentas relativamente ao caso de estudo apresentado. Assim, é possível estabelecer as características necessárias para a realização de um processo na área médica aplicanda a *text mining*. Esta revisão permitiu saber que questões de investigação seria importante seguir, como metodologias de investigação, procedimentos e aspetos a ter em conta para elaborar este projeto. Em seguida, foi necessário escolher e testar as ferramentas que melhor se adaptavam à solução proposta. Foi então necessário identificar um processo para que fosse possível identificar e extrair a informação relevante a partir de registos médicos eletrónicos. Depois, foi identificada a melhor abordagem para efetuar um

diagnóstico médico, tendo em conta os procedimentos de uma instituição. Desta forma, foi necessário a ajuda de profissionais, de forma a entender as dificuldades, procedimentos e classificação utilizadas.

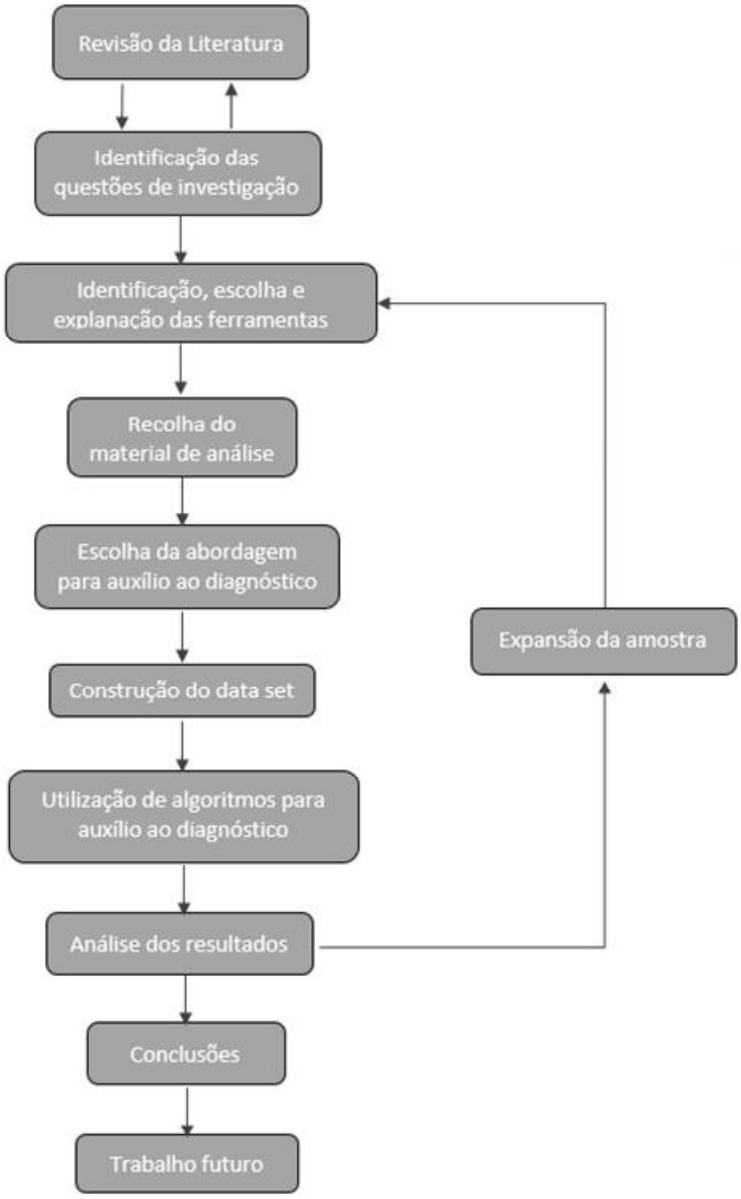


Figura 4 - Etapas do processo de investigação

Em seguida, foi efetuada uma construção do conjunto de dados, onde é necessário ter em conta as considerações sobre a informação que são importantes para a comunidade médica, como a segurança ou a confidencialidade.

Foi também necessário uma investigação de algoritmos ou métodos uteis para classificar diagnósticos médicos ou tipos de doenças, procedimentos ou tratamentos, com uma precisão e justificação do caminho tomado por esse algoritmo, de forma a ser aceite pela comunidade médica. Em seguida, os resultados foram analisados, identificando as possíveis restrições, problemas e características a serem modificadas, chegando a possíveis conclusões.

Como o processo de recolha de informação é complexo e demorado, foi necessário efetuar várias iterações, de modo a conseguir uma quantidade de informação suficiente para analisar e tirar conclusões dos resultados obtidos.

4.2 Gestão de Projeto de Investigação

As próximas secções descrevem o plano de trabalho delimitando a calendarização das atividades para este projeto. Além disso é apresentada a gestão de comunicações que permitiu um desenvolvimento mais eficaz. Estes planos são desenvolvidos e apresentados segundo as práticas para uma gestão de projeto de acordo com o *Project Management Body of Knowledge*[11] (PMBOK).

4.2.1 Plano de Trabalho

Foi então delineado uma calendarização das atividades para este projeto, que iam sendo modificados consoante as dificuldades identificadas ao longo deste processo, como podemos ver na Figura 5.

A revisão da literatura é uma tarefa que foi sendo realizada ao longo deste projeto, de forma a obter conhecimentos sobre técnicas e procedimentos. O relatório final foi outra tarefa iterativa, onde era sempre documentado os procedimentos, dificuldades e soluções encontradas. Inicialmente, foi feita uma pesquisa dos conceitos, como surgiu, técnicas e ferramentas utilizadas em *data mining*, *text mining*, em relatórios médicos, bem como uma investigação sobre a área de epilepsia.

Foi também planeada a construção de um pequeno protótipo para verificar se todas estas ferramentas conseguiam atingir os objetivos. No decorrer deste projeto foi ainda delineada a publicação de artigos em conferências ou jornais, sobre a aplicação, nomeadamente o processo de classificação de uma provável epilepsia e a arquitetura da classificação consoante

[11] http://www.pmi.org/PMBOK-Guide-and-Standards.aspx

os códigos *standard*, e sobre a revisão de literatura. Verificou-se ainda que era necessário algum tempo para a construção de um conjunto de dados, pois a recolha de registos médicos eletrónicos na área de epilepsia seria um processo árduo e demorado.

Posteriormente foi necessário a construção de ontologias e regras, para que fosse possível identificar e extrair características relevantes, tais como sintomas, eventos, causas, para que fosse possível a sugestão e classificação de epilepsia.

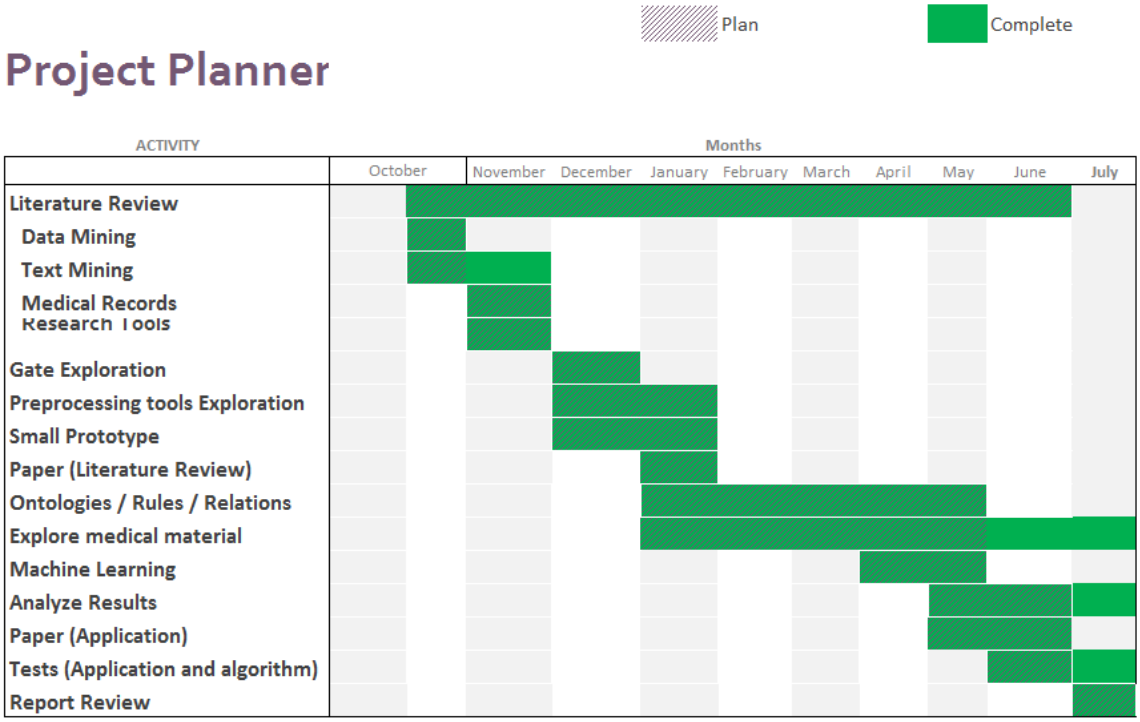


Figura 5 – Gráfico de Gantt para delinear o planeamento de projeto

Por fim, planeou-se a exploração do recurso de *machine learning* bem como os algoritmos que podem ser utilizados, de forma a escolher aquele que permite chegar aos objetivos com melhor precisão.

No decorrer deste projeto, foram também delineadas várias aplicações que permitissem a gestão e o planeamento, como por exemplo a aplicação Trello[12] que consiste na especificação da organização e da calendarização das atividades, e o GitHub[13] para controlar as versões do projeto.

[12] https://trello.com/
[13] https://github.com/

4.2.2 Gestão de Comunicações

Esta secção define como e quando a comunicação deve ser realizada, identificando os requisitos, funções e gestão da informação. Pela figura Tabela 1 é possível verificar os intervenientes e as suas responsabilidades ao longo deste projeto.

Tabela 1 - Identificação dos intervenientes e suas responsabilidades

Responsabilidade	Nome
Orientador	Rui Rijo
Coorientadora	Catarina Silva
Investigação	Luís Pereira
Consultora clínica	Margarida Agostinho

Tabela 2 - Planeamento das comunicações

Tipo de Comunicação	Frequência	Forma	Objetivo da Comunicação	Interveniente	Responsável	Entregas
Reunião Semanal	Semanal	Face a face Email	Será realizado um resumo das tarefas efetuadas, esclarecimento de dúvidas	Orientador Orientadora Investigação	Orientador	Ata da reunião
Reunião do estado do projeto	Quinzenal	Face a face	Explicado o trabalho realizado, apresentação de problemas e dúvidas, e próximos objetivos a realizar	Orientador Orientadora Investigação	Orientador	Documentação com os objetivos e abordagens tomadas, problemas, e soluções encontradas. Ata da reunião
Reunião de orientação técnica	Quinzenal	Face a face	Recolha de registos médicos, esclarecimento dúvidas, discussão de abordagens e técnicas	Orientador Orientadora Investigação Consultora clínica	Investigação	Ata da reunião
Apresentações	Esporadicamente	Face a face	Apresentação do trabalho efetuado	Orientador Orientadora Investigação Consultora clínica	Investigação	Ata da apresentação

Foi definido o mapa de comunicações explicando o objetivos, meio de comunicação, periocidade, intervenientes, o interveniente responsável pela comunicação, as entregas, identificando como foi gerida a informação, como se pode ver pela

Tabela 2.

4.3 Considerações Iniciais à Abordagem Proposta

Neste projeto utilizaram-se e transcreveram-se registos clínicos eletrónicos e registos não eletrónicos, já que estes contêm informação necessária para classificar diagnósticos, e.g., sintomas, resultados de exames, tipos de doenças, tratamentos. Tendo estes registos sido disponibilizados pelo Hospital Santo André de Leiria.

Como o desenvolvimento de uma aplicação que permitisse analisar e extrair as diferentes doenças, sintomas, procedimentos e tratamentos para toda a área da saúde, seria um projeto demasiado ambicioso, foi necessário escolher uma área médica relevante. Desta forma, a área de défice de atenção e hiperatividade foi considerada. Esta área de défice de atenção e hiperatividade constitui um dos problemas clínicos mais frequentes na infância, com um grande impacto a nível escolar [77]. Contudo, devido à inexistência de registos em formato eletrónico, verificou-se que não era possível a recolha de registos clínicos reais, que são essenciais para a realização deste projeto.

Devido à possibilidade da recolha de registos médicos eletrónicos reais de epilepsia infantil, bem como a validação e discussão dos aspetos técnicos pelo Hospital Santo André de Leiria facilitou a escolha da área de epilepsia infantil. Esta área é de grande relevância para o suporte e classificação de diagnósticos, tratamentos e procedimentos. De acordo com o *Intercontinental Medical Statistics*, epilepsia é a segunda doença neurológica mais comum em Portugal, afetando cerca de 70 000 pessoas cada ano, oferecendo ainda uma considerável complexidade no diagnóstico e na classificação do diagnóstico segundo os códigos ICD-9.

Sendo assim, foi necessário um apoio especializado na área por parte do serviço de pediatria do Hospital de Santo André, de forma a perceber que sintomas, medicação, procedimentos são utilizados na área de epilepsia infantil. Mesmo assim, esta é uma área complexa e não é linear, ou seja, um conjunto de sintomas não é sempre relacionado com um tipo de epilepsia. Além disso, duas pessoas com o mesmo tipo de crise podem sentir sintomas diferentes. Estes

sintomas podem conter diversos sinónimos, expressões ou tipos de qualificações, quantidades, negações, entre outros.

Testes preliminares foram realizados para traduzir os textos clínicos portugueses para inglês, uma vez que não foram encontradas ferramentas que efetuassem todo este processo de extração de informação em português. Desta forma, é possível verificar uma correta extração de informação relevante. No entanto, estes textos contêm termos difíceis de analisar e a sua tradução não era feita da melhor forma por diferentes ferramentas, impossibilitando uma boa classificação.

4.4 Identificação das Questões de Investigação

Como foi possível observar no capítulo de revisão de literatura, nenhum dos projetos apresentados relativamente a esta área, propõe uma identificação de epilepsia infantil ou de registos médicos portugueses. Além disso, muitos destes projetos apenas se limitam a identificar e extrair informações descritas pelos médicos em campos de registos médicos eletrónicos. Este projeto permite a identificação e extração de informação relevante de registos clínicos eletrónicos, de forma a recolher caraterísticas que permitam identificar a presença de epilepsia ou não. Além disso, permite ainda a classificação de diagnósticos segundo códigos ICD-9, identificando os procedimentos e tratamentos mais eficazes, reduzindo o erro médico o esforço, o tempo e aumentando a eficácia no processo de diagnóstico.

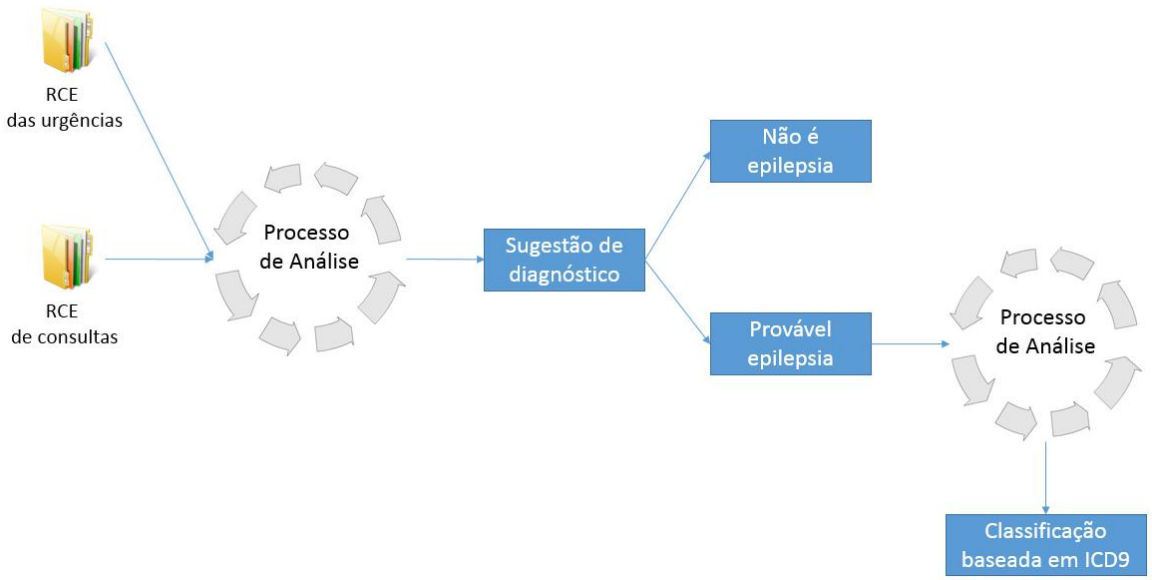


Figura 6 – Desafios do presente trabalho de investigação: diagnóstico e classificação ICD9

Verificou-se que o processo de identificação e extração de registos clínicos, e o processo realizado pelos profissionais médicos, apresenta diferentes desafios que se encontram na base deste trabalho. Estes desafios foram esquematizados, como se pode ver pela Figura 6, para que fosse possível o suporte à decisão médica, realizando diagnósticos e a suas classificações segundo os códigos ICD-9, bem como a sugestão de procedimentos e tratamentos mais eficazes.

Diferentes fontes de informação foram consideradas, como registos clínicos de consultas e urgências, para que fosse possível elaborar um processo de análise onde iria identificar e extrair a informação relevante, identificando sintomas e outros fatores importantes para chegar a uma sugestão de diagnóstico. Esta sugestão de diagnóstico poderia ser interpretada como sendo ou não uma provável epilepsia. Caso se tratasse de uma provável epilepsia seria necessário uma segunda análise para classificar um diagnóstico segundo a normalização ICD-9, sugerindo procedimentos e tratamentos mais eficazes.

As questões de investigação que estão na base deste trabalho são:

- Será possível a identificação de um processo que permita a extração de informação relevante de registos clínicos eletrónicos, de forma a recolher caraterísticas que permitam identificar a presença de epilepsia?

- Será possível a classificação de diagnósticos com base em códigos ICD-9, sugerindo procedimentos e tratamentos mais eficazes consoante o tipo de classificação?

Desta forma, foi analisado um processo geral que utilizaria técnicas de *Text Mining* para extrair informação, de forma a identificar se o paciente poderá ter ou não provável epilepsia. Além disso, esta abordagem iria também classificar o tipo de epilepsia segundo os códigos ICD-9, através de características presentes nos registos médicos. Assim, seria possível ajudar o médico a tomar decisões, reduzindo o esforço, tempo, e contribuindo para a redução do erro médico no diagnóstico, tratamentos ou procedimentos.

A partir da análise do processo e métodos de diagnóstico e na análise dos procedimentos efetuados pelo Hospital Santo André, verificou-se que os registos médicos provenientes de urgências ou consultas com as anotações do médico são processados de forma a reconhecer se o paciente poderá ter ou não uma epilepsia. Caso exista alguma probabilidade de epilepsia o médico usualmente realiza testes complementares, como por exemplo um EEG. Em seguida, os resultados destes testes irão confirmar ou não a análise realizada pelo médico, efetuando uma classificação de modo a seguir os procedimentos e tratamentos mais eficazes para cada paciente e tipo de epilepsia correspondente.

Devido à difícil aceitação da comunidade médica relativamente a estas ferramentas e técnicas, este processo deve sugerir e classificar um diagnóstico segundo os códigos ICD-9 explicando a razão dessa classificação, mencionando as características e sintomas que levaram a esse diagnóstico. Assim, o médico poderá ponderar a proposta de classificação, analisando essa solução e apresentando uma forma diferente de classificação, adicionando sintomas relevantes ou modificando o tipo de classificação, fornecendo detalhes para que a aplicação possa deduzir estas modificações no futuro.

Depois de uma investigação sobre as técnicas mais utilizadas, bem como os procedimentos e cuidados a ter em conta num processo de análise foi possível chegar a uma abordagem proposta, como se pode verificar na próxima secção.

4.5 Abordagem Proposta

O processo de extração de informação relevante de registos médicos eletrónicos é complexo [5]. Estes textos podem ser apresentados em formato não estruturado, contendo um

vocabulário complexo e composto de termos médicos, abreviações, acrónimos e termos de diferentes contextos geográficos e temporais que podem causar ambiguidade, levando à interpretação inconsistente de expressões. Existem outros problemas a ter em conta, salientando-se, entre outros, o facto de ter documentos que contenham frases ou expressões gramaticalmente incorretas para uma comunicação entre médicos mais rápida e clara, e podem ainda incluir erros ortográficos, nomeadamente em notas médicas.

Existem diferentes tipos de dados que poderiam ser usados para analisar e extrair informação relevante, e.g., entrevistas com os pacientes, análises laboratoriais, notas médicas, imagens, observações, que produzem grandes quantidades de informação. Esta informação fornece valores anormais (*outliers*) e diversas características relevantes ao diagnóstico, que podem ser um desafio na implementação ou no desempenho deste processo.

Devido às dificuldades apresentadas, é essencial efetuar uma limpeza prévia dos dados, simplificando e reduzindo os possíveis erros no processo de identificação e extração, como se pode observar pela abordagem proposta na Figura 7.

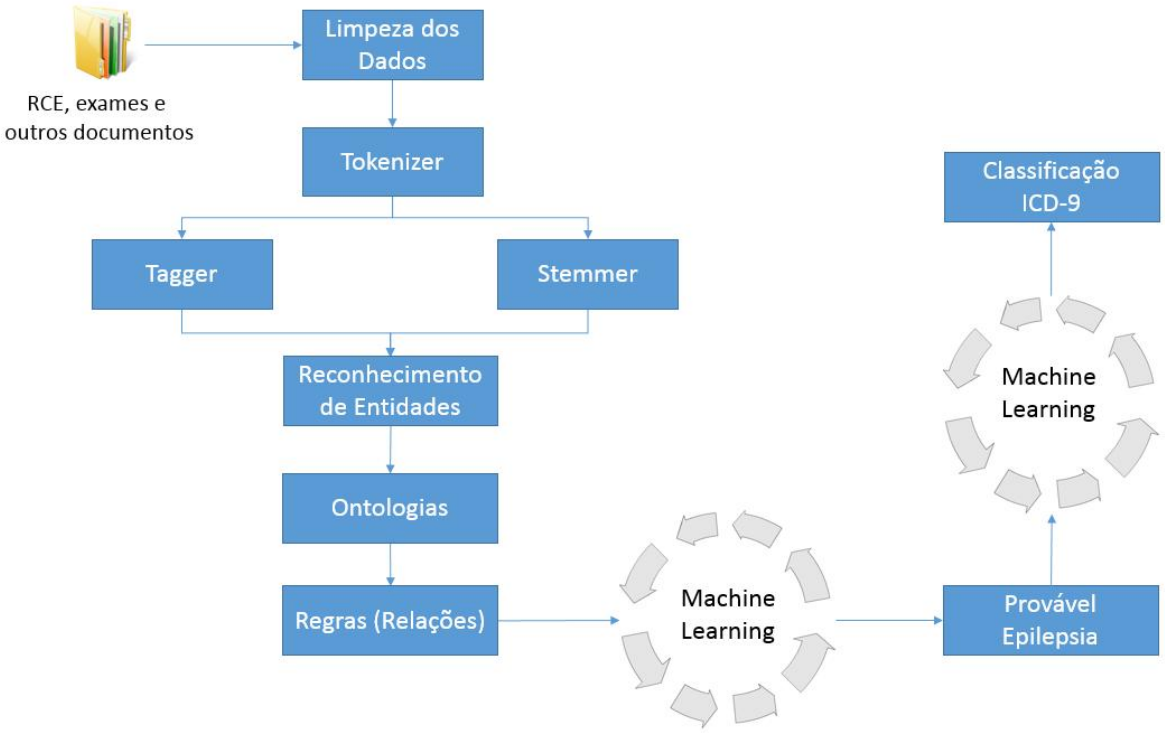


Figura 7 - Abordagem proposta para processar informação médica

Em seguida, será utilizada a técnica *tokenizer,* permitindo a identificação de frases, palavras, caracteres e pontuações. Um *tagger* com um dicionário Português deverá ser utilizado para classificar cada palavra gramaticalmente e *stemmer* (*stemming*) para identificar palavras com pequenas variações sintáticas, por exemplo a palavra "faço" seria classificada no infinitivo "fazer". O reconhecimento de entidades e as ontologias são técnicas utilizadas para atribuir categorias e relações a palavras, como por exemplo sintomas, ações e números. Além disto, estas ontologias podem facilmente ser utilizadas para a classificação de contexto em diferentes línguas. É ainda possível disponibiliza-las para que toda a comunidade médica as possa utilizar. Posteriormente, são utilizadas diferentes regras (relações) para identificar os vários sintomas de epilepsia.

É também necessário utilizar a técnica de *machine learning*, seguindo uma aprendizagem supervisionada, onde os dados fornecidos pelo hospital de Leiria com a sua determinada classificação são utilizados para desenvolver modelos que classifiquem o diagnóstico e sua classificação de registos futuros. Esta informação seria utilizada para identificar uma eventual epilepsia, bem como a sua respetiva classificação de acordo com o código ICD-9, permitindo sugerir um possível tratamento e procedimentos a considerar.

Verificou-se, através dos registos clínicos recolhidos, que era possível realizar uma classificação prévia segundo os códigos ICD-9. Isto é, embora o diagnóstico de epilepsia necessite das observações e dos resultados dos exames, como eletroencefalogramas, é possível obter uma decisão médica com base nas observações para sugerir procedimentos e tratamentos mais eficazes. Estas decisões médicas são necessárias em pacientes que necessitam de um rápido controle da doença efetuando um tratamento e medicação apropriado. Desta forma, é possível realizar tanto uma classificação com e sem o auxílio de exames, permitindo um suporte à decisão médica mais rápido e possibilitando um tratamento mais eficaz.

A identificação desta abordagem consistiu num processo iterativo, onde foram testados diversos procedimentos e técnicas até chegar a um processo que permitisse a identificação e a classificação de um correto diagnóstico segundo a normalização ICD-9. Assim sendo, foram identificadas e escolhidas diversas ferramentas para este projeto, como é possível verificar no próximo capítulo.

4.6 Síntese

Como foi mencionado, este projeto utiliza *text mining* em registos médicos eletrónicos portugueses no campo da epilepsia infantil, proporcionados pelo Hospital Santo André de Leiria. Tem como objetivo ajudar no processo de decisão de um médico, nomeadamente o diagnóstico do paciente, prescrição de medicamentos ou terapia e efetuando uma classificação segundo normalizações, como ICD9 (que é a norma de classificação adotada em Portugal). Permitindo assim diminuir o erro médico no diagnóstico e prescrição de terapia ou medicamentos. Além disso, possibilita o aumento da eficiência na análise e classificação, poupando tempo e tornando este processo mais fácil.

Foi apresentada a metodologia utilizada e como foi aplicada ao longo desta dissertação. Além disso, são identificadas as várias questões de investigação, apresentando o processo que os médicos efetuam e soluções para melhorar este processo.

5. Implementação do Processo de Sugestão e Classificação de Diagnósticos e Resultados

Neste capítulo são abordadas técnicas e procedimentos utilizados na construção e recolha de um conjunto de dados médicos, devido à sua dificuldade e importância para que fosse possível analisar os diferentes tipos de epilepsia.

Foi também necessário utilizar diversas ferramentas e técnicas *open source* que permitissem a sugestão e classificação de um diagnóstico correto. Neste capítulo são também abordados os algoritmos utilizados para a construção de um modelo, de forma a classificar futuros registos clínicos.

5.1 Recolha do Material de Análise

Foram utilizados e traduzidos alguns registos médicos de inglês para português que, embora escassos permitiram efetuar uma classificação, para verificar se era realmente possível uma correta classificação destes registos. Várias ferramentas de tradução de português para inglês foram encontradas, mas nenhuma destas permite ainda uma boa tradução. Como a classificação para ambos os casos não foi a mais desejada, e como este projeto envolve a área médica onde é importante obter uma classificação que verifique uma baixa taxa de erro, optou-se por efetuar este processamento manual, e com ajuda de uma pessoa experiente na área realizar a sua revisão e respetiva classificação. Deste modo optou-se pela recolha de registos reais e anónimos efetuando um protocolo com o Hospital Santo André de Leiria, uma vez que estes são considerados confidenciais, sendo necessária uma autorização para a sua utilização. Este processo de recolha de registos médicos foi transcrito presencialmente de modo a garantir a confidencialidade dos pacientes, traduzindo-se num processo lento e trabalhoso.

Quando a proposta de autorização da recolha de registos médicos foi aceite, observamos que este processo de recolha era complexo, lento e árduo. Nem toda a informação era apresentada num formato eletrónico, sendo necessária uma transcrição manual de algumas características, como resultado de exames, notas, diagnóstico secundário, diagnóstico final, entre outros. Além disto apenas era possível requisitar, no arquivo do hospital, cerca de 10 registos por reunião tornando o processo ainda mais lento.

Relativamente aos relatórios clínicos eletrónicos disponíveis, estes continham diversos erros gramaticais que teriam de ser corrigidos e informações privadas que teriam de ser removidas, como é possível observar pelo Anexo 1, um excerto de um relatório clinico.

Foi decidido utilizar também os textos traduzidos, uma vez que os registos recolhidos apenas continham os casos de epilepsia mais frequentes e seria uma mais valia utilizar casos onde o diagnóstico fosse menos comum.

5.2 Construção do Conjunto de Dados

Na construção de um conjunto de dados são necessários vários aspetos a ter em conta, nomeadamente, aspetos legais, étnicos e sociais que têm de ser considerados quando se gere informação médica, de forma a assegurar confidencialidade e segurança da informação pessoal de cada paciente. Assim sendo, foi necessário ocultar ou substituir esta informação confidencial de cada paciente.

Existem diferentes técnicas de ocultar esta informação como *anonymous data*, *anonymized data, de-identified data* e *identified data. A anonymous data* consiste em remover toda esta informação pessoal. *Anonymized data*, é utilizada para substituir a identificação do paciente por códigos que apenas são conhecidos por pessoas autorizadas e todo o resta da informação do paciente é removida. A técnica *de-identified data* que consiste em encriptar toda esta informação confidencial. Por fim, a *identified data* em que é utilizada quando existe um consentimento por parte do paciente.

Foi verificado que a identificação de um paciente poderia ser um fator importante para guardar uma possível evolução do paciente. Além disso, seriam necessários campos, como a idade e género, para melhor identificar sintomas. Desta forma, foram utilizadas as técnicas adaptadas de *anonymized data* e *anonymous data*. Adicionalmente, a identificação do paciente foi substituída por um código, de forma a ser desconhecida por pessoas não

autorizadas. Além disto, toda a informação pessoal do paciente foi removida exceto algumas características que são importantes para melhor diagnosticar fatores como, a idade, género e alguns aspetos sociais, e.g. pouco social.

Este processo foi realizado com registos clínicos reais fornecidos pelo hospital de Leiria. Estes registos contêm diagnósticos de epilepsia, suspeita de epilepsia e registos aleatórios de urgências. Para cada um destes registos fornecidos foram definidos diagnósticos finais, bem como os sintomas e características que levaram a tal diagnóstico.

Para tal, foram implementadas regras e ontologias, como foi mencionado na secção Escolha das Abordagens para o Auxílio do Diagnostico, de forma a identificar e classificar estas características. Estas características são identificadas de uma forma numérica para classificar se tem ou não uma provável epilepsia, e de uma forma nominal para classificar o diagnóstico segundo uma classificação ICD-9. Por exemplo, um diagnóstico foi classificado como "345.5", o que significava "*Localization-related (focal or partial) epilepsy and epileptic syndromes with simple partial seizures*".

Este processo de transcrição de registos na área médica foi complexa e demorada, não permitindo a transcrição de um grande número de registos ou uma grande diversidade de sintomas ou diagnósticos. Sendo assim, foram encontrados os tipos de diagnósticos de "*Complex focal seizure*", "*Simple focal seizure*" e "*Generalized convulsive epilepsy*". Sendo efetuado um pequeno teste com cerca de 19 registos clínicos, para verificar a possibilidade de uma correta classificação, como é apresentada na Tabela 3.

Tabela 3 - Frequência de tipos de epilepsia encontrados no conjunto de dados

		Frequência no conjunto de dados	Código ICD-9
Tipos de crises	Complex focal seizure	10	345.4
	Simple focal seizure	3	345.5
	Generalized convulsive epilepsy	6	345.1

Ao longo deste projeto verificou-se uma grande dificuldade ao recolher registos clínicos eletrónicos falsos negativos, ou seja registos que tivessem relevantes sintomas ou atributos de epilepsia, mas fossem classificados com outro tipo de diagnóstico ou registos onde se inicialmente pensou que o paciente sofresse de epilepsia, mas que realmente não tinha. Este

acontecimento verificou-se pois antes da consulta é efetuada uma análise pelos médicos nas urgências, identificando a provável doença associada, permitindo direcionar os pacientes para as respetivas especialidades. Como apenas foram recolhidos registos clínicos da área de pediatria sobre epilepsia, apenas foi possível utilizar os registos que inicialmente suspeitou-se de epilepsia. Esta abordagem permitia também assim identificar e classificar os casos de maior dúvida perante a comunidade médica.

5.3 Expansão do Conjunto de Dados

Devido ao reduzido número de registos clínicos e a distribuição irregular das diversas características, foi necessário utilizar uma técnica que permitisse a formação de novos registos, como é o caso de *crossover*. Esta é uma técnica que permite recombinação possibilitando novos registos médicos [78]. Para tal, é então necessário dividir cada relatório médico em partes iguais, que por sua vez vão servir de *input* à técnica de *crossover*, de modo a criar um novo registo médico eletrónico, construindo aleatoriamente por estas diferentes partes, como se pode ver pela Figura 8.

Como estes documentos podem não ter qualquer contexto significativo em termos médicos foi necessário que um profissional possa analisá-los e validá-los, de forma a serem utilizados neste processo de classificação de uma provável epilepsia.

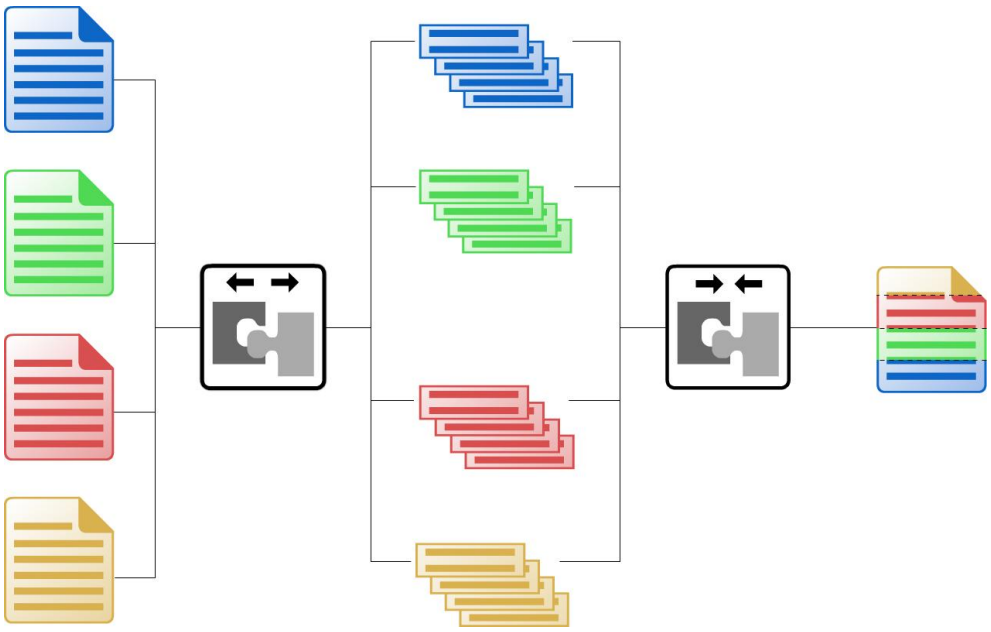


Figura 8 - Crossover em documentos de texto

Como é realizado um pré-processamento para criar uma lista de atributos ou sintomas associados ao diagnóstico de epilepsia, onde são classificados dependendo do seu aparecimento ou não no texto. Além disso, foi também possível utilizar esta técnica de *crossover* para criar novos conjuntos de valores. Assim sendo, divide-se estes atributos de cada registo médico, agrupando-os aleatoriamente formando novos casos. Para que os registos criados por esta técnica possam ser validados é necessário dar alguma informação adicional ao médico, uma vez que os valores para cada conjunto de atributos são anotados segundo uma classificação numérica, dificultando a sua perceção. Desta forma, foi fornecido um contexto textual a cada um desses atributos para ser mais fácil a sua análise e compreensão. Poderia ter sido feito ainda um *reverse engineering*, ou seja a partir de um conjunto de atributos tentar encontrar a parte textual que mais precisamente consiga retratar estes atributos, mas esta técnica pode não fornecer informação suficiente para que o médico possa analisar e compreender a situação.

Por estes motivos, foi então decidido utilizar *crossover* nos diferentes registos médicos eletrónicos. Assim sendo, dividiu-se estes registos em partes iguais, uma vez que com esta ferramenta consegue-se saber o número de frases existentes em cada documento, de forma a repartir cada documento em partes iguais. Em seguida agrupa-se, de uma forma aleatória, cada duas ou mais partes destes documentos de forma a produzir um novo documento.

5.4 Implementação da Abordagem proposta

Foram exploradas diversas ferramentas que permitissem elaborar um pré-processamento como *Rapid Miner*, *R* e GATE. Optou-se pela ferramenta GATE, uma vez que é considerada uma das melhores ferramentas para o processamento e extração da informação em *Text Mining* [79]. Além desta aplicação ser bastante utilizada na área de medicina, permite ainda a utilização da maior parte das técnicas apresentadas na secção Abordagem Proposta, como ontologias, *tokenizer*, *machine learning* e reconhecimento de entidades.

No entanto, foi possível identificar algumas restrições desta ferramenta. Embora o GATE ofereça diversos *plugins*, muitos apenas conseguem classificar com grande precisão textos em inglês, apresentando grande complexidade na tradução de dicionários ou na modificação da aplicação.

Por exemplo, o *plugin* do GATE denominado por *A Nearly-New Information Extraction System* (ANNIE) permite extrair e classificar informação consoante um conjunto de regras,

utilizando, como por exemplo, um *tokenizer, stemmer*, para classificar textos em inglês. A modificação desta ferramenta para classificar textos em português implicaria uma tradução dos seus dicionários e regras. Além disso, também seria necessário treinar estas regras exigindo uma aprendizagem do processo. Outra possível solução seria utilizar um modelo adaptado a este *plugin* para classificar textos portugueses, como o OpenNLP. Estes modelos ainda são muito recentes e além de incompletos forneciam uma reduzida taxa de acerto, excluindo de imediato esta solução. Além disto, era essencial que esta técnica fosse adaptável às necessidades deste projeto. Ou seja, o processo de extração e análise na área médica é complexo, sendo necessárias algumas modificações para identificar, entre outros, resultados de exames laboratoriais, medicamentos e outras palavras que podem conter caracteres que muitas vezes assinalam uma quebra ou o final de uma frase, como "Ben-U-Ron", "145/90 mmHg", "10 a.".

Assim sendo, foi necessário utilizar outra ferramenta que permitisse a classificação de registos clínicos portugueses, como *Freeling*[14]. Esta ferramenta permite classificar e identificar palavras de acordo com a sua gramática, encontrando também entidades relevantes através das técnicas como, *tokenizer*, *tagger*, *stemmer* e reconhecimento de entidades.

Foi também verificado que seria necessário a utilização do *plugin* ANNIE, para analisar as palavras de cada ontologia desenvolvida, de forma a relacionar essas palavras com as dos textos clínicos. No entanto, a modificação deste *plugin* seria muito complexo de forma a classificar textos portugueses.

Além disso, era necessário utilizar o formato produzido pela GATE, para que fossem utilizadas outras técnicas como o *machine learning* do GATE.

Desta forma, foi necessário construir um mecanismo de integração para associar cada *output* das várias ferramentas utilizadas, como se pode verificar pela Figura 9.

Os registos clínicos são inicialmente processados através de um *tokenizer* disponibilizado pelo GATE, diferenciando palavras, frases e pontuações. Em seguida, os textos clínicos são processados pela ferramenta *Freeling,* que identifica cada palavra fornecendo um documento de texto com as respetivas classificações. Posto isto, a ferramenta de integração desenvolvida

[14] http://nlp.lsi.upc.edu/freeling

assimila cada resultado obtido das diversas ferramentas utilizadas, construindo uma estrutura *Extensible Markup Language* (XML), para ser interpretada pelo GATE, para que fosse utilizados outros *plugins*, como *Java Annotation Pattern Engine* (JAPE) e *machine learning*.

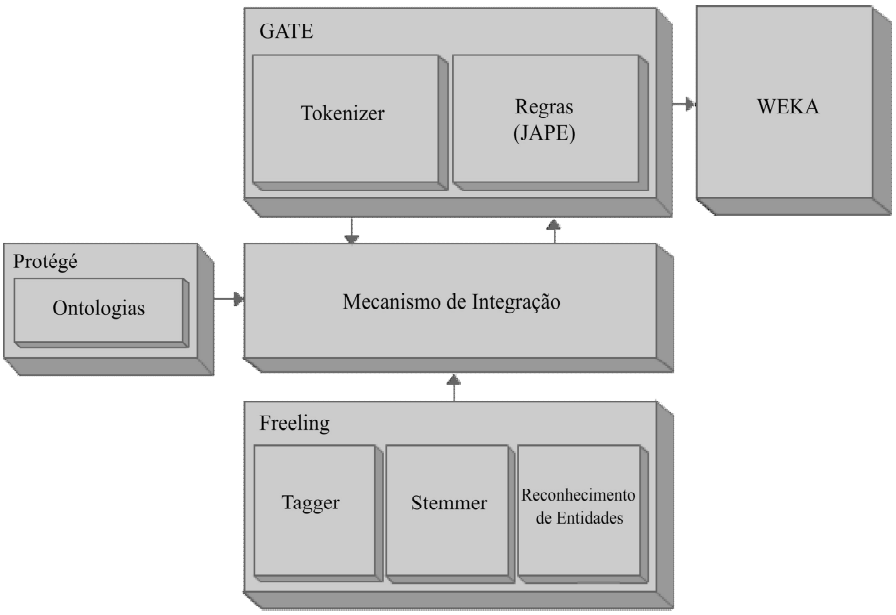


Figura 9 - Arquitetura da solução realizada

As ontologias foram desenvolvidas pela ferramenta *protégé*[15], bastante utilizada para o desenvolvimento de ontologias em diferentes trabalhos de investigação [80]. Esta ferramenta fornece um conjunto de características importantes para um desenvolvimento mais simples e rápido, como irá ser abordado na secção Ontologias. Posteriormente, são também utilizadas técnicas como o *tokenizer*, *tagger* e *stemmer,* classificando as ontologias relacionando-as com as palavras relevantes no texto. Estas classificações ou anotações são definidas por categorias, sob forma de hierarquia, permitindo um conteúdo necessário para identificar os sintomas relevantes, tais como a palavra "braço" que faz parte de regiões do corpo que por sua vez pertence à classe anatomia.

As palavras atribuídas a cada categoria das ontologias são inseridas na sua raiz, ou seja por exemplo a palavra "falou" é inserida na sua raiz "falar", de forma a ser mais fácil identificar e relacionar estas palavras com os textos clínicos processados, através da ferramenta de integração desenvolvida.

[15] http://protege.stanford.edu

Depois, foi necessário aplicar regras através da ferramenta JAPE, que permitem encontrar padrões em frases, palavras ou expressões que identifiquem sintomas, procedimentos, exames, entre outros, como será abordado na secção da Escolha das Abordagens para o Auxílio do Diagnóstico. Deste modo é possível fornecer diferentes características para que seja possível uma aprendizagem e classificação de uma provável epilepsia utilizando *machine learning*.

Outra restrição encontrada no decorrer deste projeto foi o facto de a ferramenta de *machine learning* do GATE não oferecer reconhecimento de características numéricas, ou seja, apenas reconhecia atributos nominais, como por exemplos sintomas [81]. Sendo assim, utilizou-se outra ferramenta, de forma a entender a informação para construir modelos que irão classificar o que não foi deduzido anteriormente. Assim sendo, decidiu-se pela utilização da ferramenta Weka[16].

Desenvolveu-se então um conjunto de regras que permitissem a construção de um ficheiro com o formato *Attribute-Relation File Format* (ARFF) onde os respetivos atributos encontrados na fase de pré-processamento são exportados para serem classificados no Weka.

5.5 Escolha das Abordagens para o Auxílio do Diagnóstico

Nesta secção são descritas as abordagens adotadas no desenvolvimento de regras e ontologias, para o auxílio ao diagnóstico e classificação de epilepsia.

5.5.1 Ontologias

Ontologia é um conjunto de informação sobre um determinado domínio, que poderá conter nomes de pessoas, locais, datas, preços, medicamentos, etc. Desta forma foi desenvolvido um conjunto de regras que permitem a identificação e combinação de diferentes palavras e frases.

Como na área de medicina são utilizados diferentes sinónimos, abreviações, acrónimos que referem um conceito, optou-se por utilizar ontologias de forma a proporcionar identificadores para descrever as palavras e as suas relações.

A Figura 10 apresenta a ontologia desenvolvida permitindo encontrar e relacionar palavras, como sintomas, eventos, causas, entre outros, para uma melhor interpretação do conteúdo de

[16] http://www.cs.waikato.ac.nz/ml/weka

um documento. Com a utilização de classes é possível analisar e classificar conjuntos de entidades, por exemplo classificar sintomas, negações, ações, quantidades, etc.

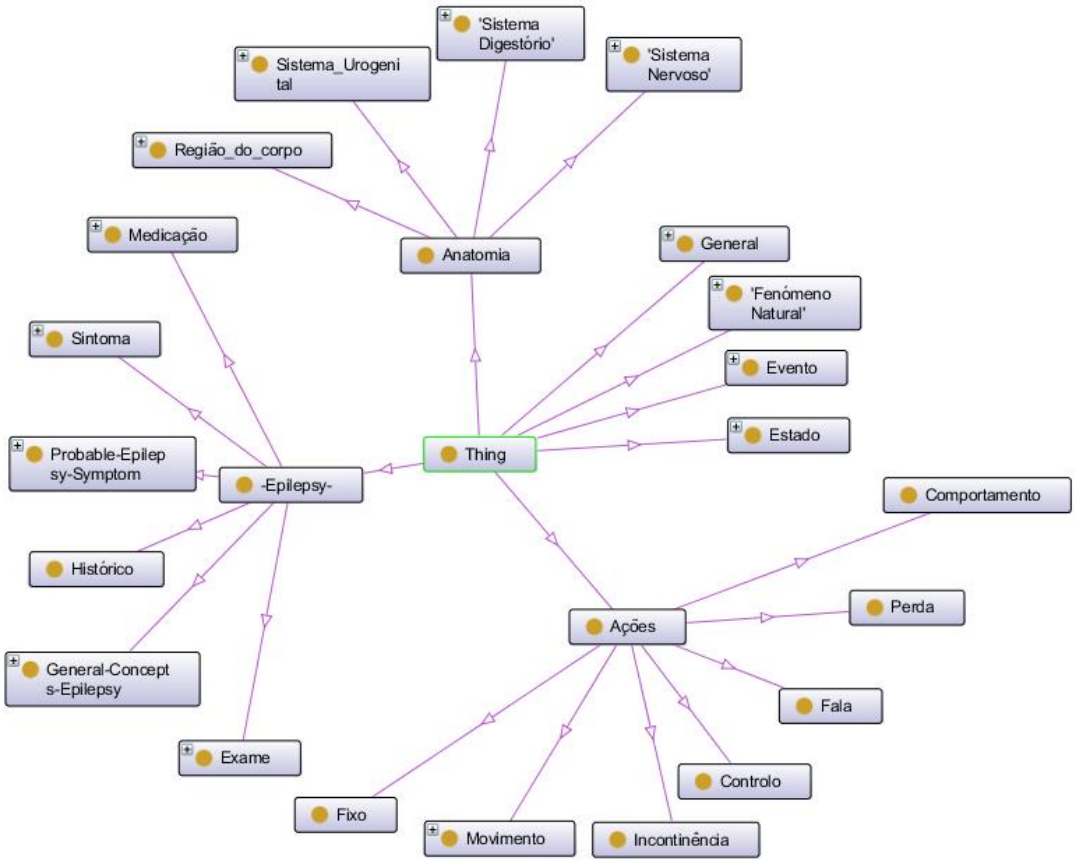


Figura 10 - Grafo da ontologia de suporte à análise da epilepsia

Estas ontologias foram desenvolvidas de uma forma gradual, onde à medida que os registos clínicos eletrónicos iriam sendo analisados seria possível adicionar palavras e classes a estas ontologias, como é possível verificar pelo Anexo 2 uma fase inicial das ontologias desenvolvidas.

Numa primeira fase, foram desenvolvidas ontologias para conseguir classificar sintomas, epilepsia, negações, quantidades, possibilidades e qualidades. Assim sendo, foi necessário construir uma classe para cada uma destas entidades com os devidos membros, por exemplo palavras que representassem sintomas quer de epilepsia ou de crises, como febre, emoções, aprendizagem. Em seguida foi necessário acrescentar uma entidade negação, de forma a conseguir identificar categorias para classificar os sintomas, tal como "não tem força".

Como na classificação de epilepsia é necessário distinguir crises generalizadas de localizadas sendo essencial uma classe de anatomia para identificar a origem do ataque, de forma a conseguir fornecer um diagnóstico e sugestão de medicação mais correta.

Estas ontologias foram construídas com base na *Unified Medical Language System* (UMLS) que é um conjunto de vocabulário estruturado de conceitos e suas relações na área médica. Alguns dos conceitos deste vocabulário estão traduzidos para diferentes linguagens, como o Português. Mas estas ontologias são complexas e difíceis de integrar num curto espaço de tempo. Por este motivo, apenas foi possível integrar algumas ontologias, nomeadamente, regiões do corpo que nos permitem deduzir a localização de um determinado tipo de epilepsia para melhor classificar um diagnóstico segundo a normalização ICD9.

Foi também encontrada uma ontologia portuguesa baseada em UMLS[17]. Esta ontologia apresenta classes com os tipos de epilepsia que existem, o que permitiu um melhor entendimento da classificação de epilepsia perante a normalização ICD9.

Nesta secção foi possível verificar o processo de identificação e anotação de palavras em textos clínicos. Contudo, para a sugestão de um correto diagnóstico é necessário a identificação e classificação de padrões que definam sintomas relevantes, como se pode verificar na próxima secção.

5.5.2 Regras

Depois da atribuição de categorias através de ontologias, foi necessário uma ferramenta que encontrasse padrões, para classificar palavras através de expressões regulares, de forma a produzir um maior contexto semântico, como foi introduzido na secção de Identificação, Recolha e Explanação das Ferramentas.

O JAPE é um *plugin* da ferramenta GATE que permite construir estas regras e é constituída por duas partes gramaticais, uma dessas partes permite identificar anotações especificadas por expressões regulares, a segunda parte descreve a ação a ser tomada sobre essas anotações. Anotação consiste na identificação de informação, similar a uma *tag*, de forma a especificar o conteúdo de uma imagem ou palavra, por exemplo identificar o nome de uma pessoa. Esta segunda parte que descreve a ação a ser tomada sobre os padrões encontrados, como estas

[17] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSH/

ações tomadas são descritas em java, além da possível elaboração de novas anotações é possível executar outro tipo de tarefas, como a elaboração de um ficheiro ARFF, dependendo dos atributos encontrados.

Para utilizar estas regras são então necessárias anotações, para descobrir padrões nos diferentes documentos e saber identificar os sintomas de uma crise epilética presentes no texto.

Estes sintomas permitem assim ajudar a verificar se o paciente tem uma provável epilepsia ou identificando o tipo de epilepsia através das crises que o paciente teve. Diversos sintomas foram analisados, e com a ajuda da equipa de pediatria do Hospital Santo André de Leiria foram extraídos e identificados os sintomas que teriam um maior importância para a classificação. Os sintomas identificados ao longo deste projeto são episódios paroxísticos, ou seja manifestações frequentes, de duração variável de movimentos distónicos[18], episódios de ausência, movimentos tónicos[19]/clónicos[20], movimentos involuntários ou espasmos, malformações cerebrais, controlo dos esfíncteres, fotossensibilidade, infeções do sistema nervoso, tumores cerebrais, desenvolvimento anormal, paralisia cerebral, antecedentes familiares de epilepsia, historial de episódios esporádicos ou episódios regulares, nascimento problemático, amnésia, confusão, hiperpnéia[21], perda de conhecimento, episódios similares e sonolência ou cansaço.

Posto isto é feita uma deteção de *split annotations*, isto é, são símbolos ou palavras que indiquem a quebra do sentido de uma frase, como conjunções ou pontuações. Estas anotações serão importantes para detetar sintomas numa frase, permitindo construir regras para que possam ser encontradas entre *split annotations*. Sendo assim os sintomas são identificados a partir de expressões regulares, através do conjunto de anotações relacionadas com um determinado sintoma até um *split annotation* numa determinada frase. Na Figura 11 é possível identificar uma simples regra para encontrar palavras relacionadas com amnésia num texto, onde é realizada uma expressão regular sobre todas as anotações "Amnesia" construídas, quer por outras regras, quer por ontologias até um *split annotation*.

[18] Consiste numa contração dos músculos
[19] Consiste numa contração súbita dos músculos
[20] Consiste em movimentos involuntários em ambas as partes do corpo
[21] Hiperventilação

Mas identificando apenas se um sintoma está pressente ou não num texto, não nos daria informação suficiente para conseguir extrair uma semântica satisfatória, de forma a saber se um paciente sofria ou não de epilepsia.

```
/* ------------------------------------------------------Amnesia ---------------------------------------*/
Phase: Amnesia
Input: Amnesia SplitAnnotationConjuction
Options: control = appelt debug = true


Rule:   amnesia
(
  {Amnesia}
  )
:amnesia
{SplitAnnotationConjuction}
-->
:amnesia.AmnesiaRule = {result="1", rule = "Amnesia", String = :amnesia@string}
/*-------------------------------------------------------------------------------------------------------*/
```

Figura 11 - Exemplo de uma regra JAPE

Foi introduzida a técnica de *Negation handling* que permite detetar negações e verificar a relação dessas anotações entre palavras ou expressões. Foi adicionada uma categoria "negação" à ontologia, identificando as negações mais utilizadas pelos médicos como "não", "nem", "nenhum", "jamais", "rejeição", entre outras, que foram observadas nos registos clínicos. Assim sendo, as regras foram elaboradas de modo a que se consiga encontrar um determinado sintoma no texto, verificando também se a negação desse sintoma é verificada no texto.

Foi também adicionada a técnica de reconhecimento de dúvida relativamente a sintomas, permitindo saber quando o médico teve dúvidas ao descrever sintomas. Além da eventual pontuação e das palavras que introduzem dúvida à frase, como por exemplo a utilização da palavra "talvez", foi também necessário classificar expressões identificadas pelos médicos, como é o caso da introdução de palavras entre parenteses, com ou sem pontuação, e.g. "(?)" ou "(dor de cabeça)".

Para todas estas características, foi também necessário identificar casos onde uma provável dúvida influenciasse outros sintomas dessa mesma frase, como por exemplo "ele teve amnésia e movimentos tónicos?", onde a utilização de conjunções, e.g. "e", "ou", "mas", entre outros, introduz dúvida na restante parte da frase.

Desta forma, foi utilizada uma classificação baseada em *rank* consoante a utilização das diferentes regras. Este *rank* pode ter diferentes valores como "-1", "1" e "2". O valor "-1" é atribuído quando é identificado um sintoma seguido de uma negação, referindo que o paciente

não sofre desse sintoma. O valor "1" é atribuído quando existe alguma dúvida se o paciente tem ou não um determinado sintoma e "2" se é um sintoma. Como o próximo passo seria exportar estas classificações para ser utilizado por um algoritmo de *machine learning*, para que fosse possível deduzir uma provável classificação, e como cada documento poderá conter inúmeras destas anotações, foi necessário efetuar um mecanismo que simplificasse esta interpretação. Assim sendo, foi necessário construir uma regra que permitisse a verificação da presença ou não de determinados sintomas em cada documento, onde a ausência de um sintoma é classificado com o valor "0".

Em seguida, foi implementado um mecanismo que permitisse construir o ficheiro ARFF, onde todos os sintomas ou atributos fossem identificados juntamente com o diagnóstico de provável epilepsia ou de classificação ICD, previamente atribuído na fase de treino. Sendo atribuído o *rank* respetivo a cada um desses sintomas.

Este sistema utiliza anotações para construir regras e encontrar informação relevante. Por este motivo é possível uma classificação para outras línguas, bastando para isso apenas inserir as palavras relacionadas a cada ação ou sintoma nas ontologias e escolher outra linguagem fornecido pelo *tagger* do *Freeling*.

Foi utilizada a técnica de reconhecimento de graus de intensidade permitindo uma melhor ponderação entre sintomas ou atributos, para que se possa obter um diagnóstico com o menor erro possível. Assim sendo, pretende-se com a utilização de graus de intensidade encontrar expressões realizando uma classificação numérica de acordo com as palavras relacionadas com o item de interesse. Por exemplo, a expressão "tem muita febre" tem um maior impacto de "tem febre" ou "não tem febre". Estas classificações foram elaboradas através de regras e ontologias que continham estas diferentes quantidades, como se pode ver pela secção de Ontologias.

Um grau de intensidade de uma expressão pode ter *rank* elevado quando se encontram palavras como, muito, demasiado; e baixo quando são extraídas expressões como, pouco e pequena. Sendo assim a expressão extraída como "O paciente tem muita febre" terá uma maior classificação do que "O paciente tem febre". Além disto, foi também utilizada uma classificação que dependeu da frequência de sintomas que um paciente pode ter nos diferentes episódios, ou nas várias idas às urgências. Por este meio, caso se verifique o mesmo sintoma

em diferentes crises, esse atributo é classificado como tendo uma frequência elevada, ou se esse atributo for apenas mencionada uma única vez a sua classificação será baixa.

Verificou-se que os médicos não utilizam a frequência nem a intensidade para sublinhar a importância de um sintoma e por isso, estas técnicas de identificação de graus de intensidade e frequência foram retiradas do processo. Apesar disto, é um contributo que será relevante para aplicações futuras do resultado deste trabalho a outras patologias.

5.6 Algoritmos utilizados no Auxílio ao Diagnóstico

A técnica de *machine learning* contém diferentes processos para deduzir modelos (funções) a partir de dados fornecidos, no qual podem ser utilizados para classificar novos dados.

Foi utilizada uma aprendizagem supervisionada, uma vez que era possível deduzir um modelo a partir de registos clínicos previamente classificados (dados de treino), permitindo construir modelos para que seja possível deduzir resultados de futuros registos clínicos.

Desta forma, foram escolhidos algoritmos que permitissem efetuar uma aprendizagem supervisionada, como *K-Nearest Neighbor* [82], que é um algoritmo fácil de implementar e de conhecer as características que foram utilizadas para chegar a um possível resultado (*whitebox*). Este algoritmo classifica cada registo baseando-se nas características mais próximas, ou seja cada característica ou atributo é classificada segundo a sua vizinhança.

Foram também utilizados outros algoritmos de aprendizagem supervisionada como árvores de decisão (*tree algorithms*) é um processo que permite expressar, sob forma de um grafo ou um modelo, um conjunto de condições que é necessário ocorrer, de forma a chegar a um resultado. Este processo é bastante popular na área médica para classificar padrões, uma vez que se torna mais fácil de analisar, é relativamente fácil de construir e permite obter uma boa precisão [83]. Por exemplo, é possível analisar que após uma pessoa ter eventualmente uma crise generalizada sente sonolência.

Classification And Regression Trees (CART) é um exemplo de um algoritmo de árvores de decisão desenvolvido por Breiman [84]. Este algoritmo permite manipular tanto variáveis categóricas como contínuas e consegue também analisar valores em falta, como o algoritmo C4.5. CART tenta construir regras baseadas no atributo que mais consegue diferenciar os

valores, em seguida quando esta regra é selecionada é dividida em dois recursivamente (apenas divide estes valores de uma forma binária, em 2), até quando o CART detetar que não é possível obter uma maior ganho.

É possível verificar através dos testes realizados pelo D. Lavanya [83] que o algoritmo CART (ou SimpleCART em Weka) produz melhores resultados, com uma melhor precisão quando a complexidade é elevada. Como a precisão é um fator importante no âmbito do diagnóstico médico utilizou-se este algoritmo.

Algoritmos difusos são outro exemplo bastante utilizado em medicina, que diferem um pouco dos algoritmos convencionais, ou seja os algoritmos convencionais constroem regras com limites e transições abruptas entre classes diferentes, mas os algoritmos difusos permitem que os intervalos sejam graduais e que se possam construir regras de uma forma mais percetível. Além disso, permitem que estados indeterminados possam ser tratados, e desta forma classificar conceitos não quantificáveis, como por exemplo temperatura e os seus estados como quente, médio ou frio, onde um conjunto de regras são construídas com base em dados de treino, de forma a referir que temperatura quente provavelmente é descrita como superior a 30ºC, média entre 15 e 30ºC e fria se inferior a 15ºC.

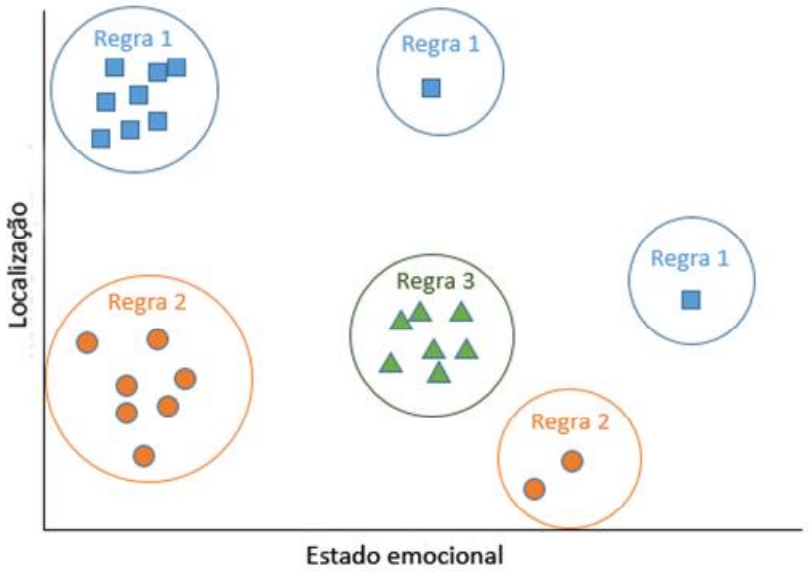


Figura 12 - Exemplo de regras difusas consoante o estado emocional das pessoas e a sua localização

Na Figura 12, pode-se verificar um exemplo de classificação difusa, no qual se construiu regras consoante o estado emocional das pessoas, como por exemplo, contente, indiferente, triste, considerando diferentes localizações.

A ferramenta *Weka* também permite a utilização deste tipo de classificação através do classificador *Fuzzy Unordered Rule Induction Algorithm* (FURIA).

Foi também decidido utilizar algoritmos *Case-based Reasoning* que permitem resolver problemas utilizando experiencias prévias. Consiste em guardar os resultados para os diferentes problemas decorridos, para que sejam eventualmente utilizados em problemas que possam vir a surgir.

A técnica de *Case-based Reasoning* é constituído por 4 etapas: *Retrieve*, *Reuse*, *Revise* e *Retain*. *Retrieve* consoante um determinado problema permite recuperar os casos relevantes que ajudem a resolver o problema atual, a sua solução e como se chegou a essa solução. *Reuse* consiste em mapear as soluções de casos anteriores para o problema objetivo. Assim sendo, é possível a adaptação de soluções anteriores para conseguir se chegar a uma solução para um novo problema. *Revise* consiste em testar a solução e se necessário rever, alterando para possíveis soluções onde o objetivo é o mais desejado. *Retain* consiste em guardar a informação mais importante como resultado de uma experiência. Esta informação consiste na solução e a justificação para se chegar a essa solução. Como nem todos estes passos são importantes para a realização deste processo apresentado, foi possível utilizar o classificador IBK fornecida pela ferramenta *Weka* que permite a utilização das etapas *Retrieve, Reuse* e *Retain*.

Como a análise médica não consiste apenas de um conjunto de regras mas também num conjunto de experiências, foi então utilizada esta técnica.

Por fim, foi utilizado um método de classificação *black box*, ou seja um algoritmo que fornecia resultados mas não era possível conhecer as razões de um determinado resultado. Sendo assim, *Support Vector Machine* (SVM) foi utilizado como forma de comparação entre as diferentes medidas de avaliação.

5.7 Sequência de Testes Realizados

Ao longo do processo iterativo apresentado nesta dissertação foram realizados diversos testes para alcançar resultados aceitavéis. Estes resultados permitiram uma evolução na identificação e extração de informação relevante, para que os algoritmos consiguissem de uma forma mais clara e precisa obter um diagnóstico e classifica-lo consoante a normalização ICD-9, como se pode ver pela Tabela 4.

Na tabela 4, são apresentados os diferentes testes mais relevantes referindo o objetivo de cada um, a amostra e os algoritmos utilizados, consoante a ordem realizada.

Tabela 4 - Sequencia de testes realizados

Ordem	Descrição	Objetivo	Amostra	Algoritmos
1	Teste inicial	Diagnósitco	19	IBK
2	Teste inicial	Classificação segundo a norma ICD-9	19	IBK
3	Teste com mais registos recolhidos	Diagnósitco	30	IBK, SVM
4	Teste inicial à técnica de crossover	Diagnósitco	53	IBK, SVM, CART,
5	Teste à técnica de crossover (registos adicionais)	Diagnósitco	70	IBK, SVM, CART, FURIA
6	Teste a sintomas adicionais relevantes	Diagnósitco	113	IBK, SVM, CART, FURIA
7	Teste à classificação do diagnóstico	Classificação segundo a norma ICD-9	51	IBK, SVM, CART, FURIA

5.8 Síntese

Como foi possível neste capítulo, uma recolha de material foi feita em colaboração com o Hospital Santo André de Leiria, essencial para a construção do conjunto de dados para a realização de vários testes, aumentando a precisão e segurança dos resultados.

Além disso, um mecanismo de integração foi desenvolvido e apresentado neste capítulo, para associar os resultados das diferentes ferramentas e técnicas utilizadas, para que fosse possível a identificação dos sintomas e características relevantes, e com a utilização da técnica de *machine learning* ajudar a sugerir e classificar um diagnóstico.

6. Análise e Discussão de Resultados

Neste capítulo são abordadas as medidas de avaliação utilizadas para analisar os resultados obtidos, de acordo com a área médica. Em seguida, são apresentados os resultados obtidos quer no apoio ao diagnóstico, quer na classificação do diagnóstico, justificando e analisando os problemas e possíveis soluções encontradas.

6.1 Medidas de Avaliação

Existem diferentes processos de classificação, como *multiclass*, *one-class* e *binary*. O processo *binary* permite classificar um atributo ou classe em dois grupos, com uma determinada propriedade ou sem essa propriedade. A *multiclass* consiste em classificar um atributo ou classe em mais de duas propriedades, e.g., o tempo pode ser quente, frio ou normal. A *one-class* tenta distinguir uma classe de todas as outras, identificando as prováveis classes que um objeto pode pertencer [85].

Foi utilizada uma abordagem de múltiplas classes *one-vs-all* (*multiclass*) de forma a avaliar a tarefa de decisão, onde foram definidos também diferentes resultados possíveis de classificação: verdadeiro positivo (VP), falso positivo (FP), falso negativo (FN) e verdadeiro negativo (VN), como se pode ver pela Tabela 5.

Foram também utilizadas diferentes medidas para classificar o desempenho de cada algoritmo e.g.: taxa de erro ((FN+FP)/(VP+FN+FP+VN)), recall (R=VP/(VP+FP)), precisão (P=VP/(VP+FN)) e *F-measure* (*F1*) onde se combina a *recall* com a precisão (F1=2*P*R/(P+R)). Foram também utilizadas medidas como especificidade (VN/(FP+VN)) e sensitividade (VP/(VP+FN)) que são medidas bastante utilizadas em medicina, uma vez que identificam a taxa de resultados positivos ou negativos num conjunto de dados.

Tabela 5 - Matriz de confusão

Resultados		Patologia	
		Presente	Ausência
Diagnóstico	Positivo	**Verdadeiro Positivo (VP)** (diagnostico positivo, patologia presente)	**Falso Positivo (FP)** (diagnostico positivo, ausência de doença)
	Negativo	**Falso Negativo (FN)** (diagnóstico negativo, patologia presente)	**Verdadeiro Negativo (VN)** (diagnostico negativo, patologia ausente)

6.2 Apoio ao Diagnóstico

Este processo de classificação foi iterativo, onde várias funcionalidades iam sendo adicionadas à medida que o processo de classificação era testado e os resultados obtidos analisados. Inicialmente foi sendo testado um simples algoritmo *K-Nearest Neighbor* (K-NN) identificado os sintomas ou atributos mais comuns entre uma determinada vizinhança K, ou seja, o número de sintomas que seriam analisados de forma a classificar um diagnóstico. Esta vizinhança foi escolhida tendo em conta o número total de sintomas que podem ser identificados, o número de registos recolhidos e os resultados obtidos. Para tal foi utilizado o classificador ibk que é fornecida pela ferramenta *Weka*.

Foi também utilizada a técnica de *cross-validation* para prever e avaliar o desempenho de um modelo. Esta técnica foi utilizada devido ao baixo volume de registos clínicos, permitindo uma divisão da informação em dados de treino e teste avaliando os modelos aleatoriamente.

Desta forma, os dados de treino continham os valores para o algoritmo identificar ou deduzir padrões, e os dados de teste para verificar a percentagem de acerto dos modelos construídos.

Desta forma, a Tabela 6 apresenta os resultados obtidos numa fase inicial do projeto, onde vários testes foram realizados para diferentes valores de K e de *cross-validation*. Estes resultados foram obtidos a partir da análise perante 19 sintomas em 18 registos, que embora

seja um número reduzido é possível retirar algumas conclusões. Assim, foi possível concluir que a melhor classificação foi K=1, mas como se pode observar existe uma possibilidade de *overfitting*, que acontece quando se tentar construir modelos complexos, onde existem mais atributos ou sintomas do que exemplos.

Tabela 6 - Desempenho do algoritmo K-NN, fase inicial

		Cross-validation fields			
		2	3	4	**5**
KNN k-value	1	88.89%	88.89%	100%	100%
	2	77.78%	88.89%	100%	100%
	3	77.78%	66.67%	77.78%	77.78%

Estes exemplos incorretamente classificados permitem uma análise de falsos negativos, onde estes exemplos classificados inicialmente como não tendo epilepsia, podem na verdade sofrer de epilepsia. Na área médica é importante evitar este tipo de erros, uma vez que é importante receber um tratamento, de forma a controlar estas crises para que estas pessoas possam viver normalmente.

Por fim, foram realizados vários testes com diferentes algoritmos para analisar resultados e retirar conclusões sobre a classificação de diagnósticos deste processo proposto, como se pode observar pela Tabela 7.

Estes testes foram realizados com um conjunto de dados de 91 registos classificados com um diagnóstico de epilepsia e 22 registos com falsos ou diagnósticos ausentes de epilepsia. Além disso, foram adicionados novos sintomas de "hipertonia"[22] ou "hipotonia"[23], e "parestesias"[24]. Foi utilizada uma *cross validation* de 20 para o algoritmo FURIA e K-NN (IBK em WEKA) e *cross validation* de 16 no algoritmo CART de forma obter uma classificação mais realista. Para o algoritmo *Nearest Neighbours* foi utilizado um K de três para o primeiro teste, e cinco para os restantes já que como foram adicionados novos sintomas e houve a necessidade de utilizar uma maior vizinhança para a classificação.

[22] Consiste no aumento do anormal do tónus muscular
[23] Consiste na diminuição anormal do tónus muscular
[24] Consiste em sensações espontâneas de frio, calor, formigueiro, pressão, entre outros

Como foi possível observar pelos resultados da Tabela 7, o conjunto de dados encontra-se ainda desequilibrado, ou seja pode-se observar que existem mais casos onde existe epilepsia, do que verdadeiros negativos, que consiste na classificação inicial de não epilepsia, quando na realidade o paciente sofre de epilepsia.

Verificou-se outro fator onde os verdadeiros negativos contêm semelhanças a certos casos verdadeiros positivos. Imaginando que um paciente que tenha sintomas similares a epilepsia, quando faz EEG o resultado pode ser normal, isto pode-se dever ao fato de a epilepsia evoluir ao longo do tempo (por exemplo um caso poderá não fazer diagnóstico dependendo da instituição). Outro fator é que um paciente poder ter sintomas similares a outro tipo doença e não serem identificados por este processo.

Tabela 7 - Resultado dos testes finais para classificação de um provável diagnóstico

	VN	FP	FN	VP	Taxa de acerto	F-Measure
SimpleCART	1	21	3	88	78.76%	72.4%
FURIA	2	20	3	88	79.6%	74.1%
IBK (K=5)	4	18	4	87	80.53%	76.7%
LibSVM	0	22	0	91	80.53%	71.8%

Embora todas estas características apresentadas é possível dizer que estes resultados são animadores, conseguindo-se uma percentagem de acerto de pelo menos 78%, como se pode verificar pela Tabela 7.

6.3 Classificação ICD

Foram também realizados testes de classificação baseados em códigos ICD para obter uma classificação *standard* do tipo de epilepsia apresentado no registo médico.

Desta forma, foram efetuados testes iniciais com 19 registos clínicos para analisar os resultados mediante as diferentes classificações encontradas, como por exemplo crises focais

simples, crises focais complexas e crises epilética generalizadas, como se pode observar pela Tabela 8.

Tabela 8 - Resultados iniciais relativamente à classificação do tipo de crise

Seizure Type	FP	FN	VP	VN	F-Measure
Complex focal seizure	1	5	9	4	73%
Generalized convulsive epilepsy	4	10	2	3	62.2%
Simple focal seizure	3	15	0	1	N/A

Estes resultados foram então obtidos utilizando o algoritmo *K Nearest Neighbor*, com K=3 e utilizando a técnica de *cross-validation* com o valore 3.

Analisando estes valores, é possível concluir que estes registos com classificação de crise focal simples são muito escassos, para que os algoritmos consigam construir um modelo, sendo impossível a verificação de uma correta classificação para este tipo focal simples, como para a correta dedução e classificação para outros tipos de crises.

Assim sendo, foi necessário utilizar outros testes onde este tipo de classificação fosse removida, como se pode observar pela Tabela 9. Com esta modificação já foi possível verificar uma ligeira melhoria perante os resultados iniciais, demonstrando que se o conjunto de dados tiver um número considerável e distribuído perante os diferentes tipos de classificação é possível alcançar uma precisão bastante aceitável.

Tabela 9 - Resultados preliminares relativamente à classificação de cada crise

Seizure Type	FP	FN	VP	VN	F-Measure
Complex focal seizure	1	3	9	3	74%
Generalized convulsive epilepsy	3	8	3	2	68.1%
Weighted Average					71.05%

Como se pode observar é possivel verificar que o número de falsos negativos foi significativamente reduzido, especialmente para o tipo de crise focal complexo (de 5 para 3), que é extremamente relevante para a área médica.

Embora estes resultados sejam apenas preliminares, com poucos registos clínicos, poucos tipos de crises epiléticas e com o risco de ocorrência de *overfitting*, é ainda possível efetuar

uma classificação aceitável obtendo uma F-Measure média de 71,05%, no entanto é necessário mais registos com diferentes tipos de crises para obter resultados mais confiantes.

Em seguida, foram realizados testes com 51 registos, verificando se é possível obter uma boa taxa de acerto perante um maior número de registos. Assim sendo, estes registos foram identificados com ajuda de profissionais médicos em três diferentes classificações, entre as quais, parciais complexas, parciais simples e generalizadas convulsivas. Estes tipos de classificações têm diferentes características, por exemplo as convulsões parciais complexas epiléticas incidem na generalização, com provável perda de consciência, geralmente focadas nos lobos temporais do cérebro e associadas a problemas psicomotores. As pessoas que sofrem de epilepsia parcial simples não têm movimentos ou convulsões generalizadas, não sofrem de perda de consciência, poderão conter alucinações, perda de controlo de esfíncteres e são focais. A classificação generalizada convulsiva possui movimentos generalizados e pode conter perda de consistência.

Desta forma, foram efetuados testes utilizando diferentes algoritmos, tais como *IBK* com k=5, *SimpleCART*, *Furia* e *Library for Support Vector Machines* (LibSVM) com um *cross validation* de 20, como é apresentado na Tabela 10. Embora LibSVM não seja um algoritmo *white box*, é importante testar e fornecer diversas perspetivas de decisão médica. Este algoritmo permite utilizar técnicas de classificação para obter resultados aceitáveis mais rapidamente [86].

Tabela 10 - Resultados obtidos para classificação dos registos segundo os códigos ICD-9

	Classificada como Parcial Simples	Classificada como Parcial Complexa	Classificada como Generalizada Convulsiva	Taxa de acerto	F-Measure
IBK	94%	33,3%	57,1%	60.78%	59%
SimpleCART	94%	27%	80%	68,6%	65,3%
Furia	86,7%	33,3%	76,1%	66,7%	64,5%
LibSVM	73,3%	33,3%	85,7%	66,7%	64,5%

É possível verificar pelos resultados obtidos que a classificação parcial complexa tem uma reduzida percentagem de acerto. A classificação parcial complexa contem muitos sintomas semelhantes à classificação generalizada convulsiva, tornando-se difícil de identificar sem exames complementares. Assim sendo, cada vez que uma destas classificações é deduzida,

devem ser sugeridos exames complementares para uma melhor classificação, aumentando assim a eficácia na classificação baseada em códigos ICD-9.

6.4 Discussão dos Resultados Obtidos

Os resultados obtidos tanto na sugestão de um provável diagnóstico, como na sua classificação de acordo com a normalização ICD-9, sugerem um bom desempenho. Desta forma, conclui-se que apesar do reduzido número de registos é possível sugerir e classificar diagnósticos de uma forma significativa.

É possível identificar diferenças significativas ao longo dos testes realizados, onde a análise de diferentes tipos de epilepsia, a introdução de novos sintomas e de um maior número de registos permitiram para resultados mais seguros nesta área.

Foi necessário utilizar uma abordagem *white box* para que os médicos conhecessem as razões para uma determinada classificação. Além disso, como podem ser utilizados algoritmos para propor uma classificação, é possível obter várias opiniões para a análise de um registo clinico, poupando tempo na discussão com outros médicos. A utilização de diferentes algoritmos também permitiu uma melhor análise dos resultados, sendo possível verificar as razões para um determinada classificação, ajudando a desenvolver abordagens para o tratamento das classificações erradas.

Devido à dúvida perante a classificação de um diagnóstico parcial complexo e generalizado convulsivo, é necessário guardar a evolução de um paciente para verificar características relevantes para uma classificação mais correta. Além disso, a possível utilização de exames complementares (EEG e RM) para a classificação de diagnósticos, proporciona uma melhor precisão.

Estes resultados foram analisados de registos médicos reais e anónimos fornecendo uma maior confiança a esta abordagem na sugestão e classificação de diagnósticos.

7. Conclusões

Neste capítulo são referidos os principais contributos que este processo permite fornecer à comunidade médica. Além disto, é apresentado um resumo do trabalho realizado, são apresentadas conclusões deste projeto bem como o trabalho futuro a desenvolver.

7.1 Síntese do trabalho realizado

O principal objetivo desta dissertação foi desenvolver uma abordagem que permitisse a análise e extração de contexto relevante de textos médicos para sugerir um diagnóstico, bem como uma classificação baseada em códigos *standard*, de modo a aconselhar os melhores procedimentos e tratamentos dependendo da classificação encontrada.

Desta forma, foi necessário estudar as diferentes áreas de conhecimento envolvidas bem como a análise das soluções existentes. Verificou-se que a classificação na área de epilepsia analisando registos médicos portugueses seria uma mais-valia. Além disso, poucas aplicações permitiam um suporte à decisão de classificações ICD-9.

Foi então apresentada uma solução utilizando técnicas de *text mining* para atingir os objetivos de apoio ao diagnóstico e sua classificação. Este foi um processo iterativo onde foram identificadas e testadas várias ferramentas que ao longo deste projeto foram sendo modificadas e adaptadas, de forma a atingir os objetivos propostos. Foi também necessário proceder a uma recolha de informação, efetuando um protocolo com o Hospital Santo André de Leiria, uma vez que estes registos são considerados confidenciais.

Desta forma, construiu-se um conjunto de dados baseado em casos reais, onde considerações e abordagens relevantes foram consideradas para que esta informação pudesse ser utilizadas por diferentes algoritmos de aprendizagem.

Em seguida foram selecionadas os métodos de avaliação que melhor permitissem analisar resultados tanto para a classificação do apoio ao diagnóstico como para a classificação ICD.

7.2 Principais Contributos

Nesta secção são apresentados os contributos mais significativos do processo proposto nesta dissertação, como:

- Proposta de abordagem que visa a redução do erro na sugestão e classificação de diagnósticos na área de epilepsia infantil;
- Revisão da literatura e estruturação dos conceitos e projetos de investigação relacionados com esta área de *text mining* aplicada à área de epilepsia, que poderá ajudar outros futuros trabalhos;
- A identificação de questões de investigação também são bastante importantes para a realização de futuros trabalhos, permitindo uma visão das áreas de relevante investigação.
- O desenvolvimento de ontologias desenvolvidas que possibilitam a identificação e classificação das palavras em textos. A abordagem utilizada permite ainda adaptação a área de investigação em causa;
- Recolha do conjunto de dados. Este conjunto de dados poderá ser utilizado em trabalhos de investigação futuros que necessitem de material na área da epilepsia infantil;
- Identificação da lista de sintomas que permitem a identificação de epilepsia infantil e a classificação de um diagnóstico segundo a normalização ICD-9;
- O plataforma de integração das ferramentas que permitissem utilizar técnicas de *text mining* para identificar e extrair conhecimento relevante de registos médicos portugueses;
- A proposta de abordagem que se efetuou para classificar um diagnóstico segundo os diferentes códigos ICD-9.

7.3 Conclusões

Foram apresentados vários trabalhos no capítulo da Revisão da Literatura, relacionados com a área de epilepsia, que apenas classificavam diagnósticos depois do diagnóstico realizado pelo médico. Não foi encontrado nenhum projeto que classificasse registos clínicos portugueses. Além disso, nenhum destes projetos abordavam a área de epilepsia infantil, no qual é uma área de grande importância, para que as crianças possam viver e compreender o mundo.

Desta forma, foi elaborada uma abordagem que permitisse a sugestão de diagnósticos e a sua classificação de acordo com a normalização ICD-9. Assim, é possível reduzir o erro médico na prescrição, nos procedimentos e aumentar a eficácia do processo de diagnóstico médico. Este trabalho permite ainda uma fonte de conhecimento para futuros projetos nesta área.

Foi também possível realizar uma investigação nas áreas em aberto de maior relevância, baseado na extração de *text mining* de registos clínicos eletrónicos, proporcionando um estudo para a realização de possíveis trabalhos futuros.

Os resultados obtidos tanto na sugestão de um provável diagnóstico e na sua classificação sugerem uma precisão segura.

O diagnóstico e sua classificação é um processo complexo e lento. É necessário ter em conta um grande número de características e fatores para diferentes doenças. Desta forma, são necessários mais registos de diferentes tipos de epilepsia, para analisar mais sintomas e características que podem ser uteis para uma melhor classificação.

Verificou-se que os médicos ainda estão bastante apreensivos quanto a este tipo de tecnologia. Desta forma, é importante utilizar uma abordagem *white box*, onde é possível identificar as razões para uma determinada classificação para que o médico possa compreender. Além disso, foi possível determinar que a utilização de vários algoritmos ajudaria no processo de decisão, expondo sintomas, ou até mesmo outros pontos de vista para um determinado diagnóstico.

7.4 Trabalho Futuro

O processo proposto nesta dissertação sugere um bom desempenho, mas poderá ser melhorado.

Uma característica relevante para este sistema é a aprendizagem pelas decisões e opiniões do médico, por outras palavras este sistema aprendia novos sintomas ou adaptava a classificação realizada de acordo com o médico. Para tal é necessário uma funcionalidade que permitisse ao médico especificar com facilidade a palavra ou o conjunto de palavras que formam esse sintoma.

Verifica-se que na identificação de uma provável epilepsia é necessário guardar informação sobre a evolução de um paciente. Esta evolução pode permitir uma melhor precisão nesta área, mas também uma melhor eficácia na análise de tratamentos e procedimentos a tomar.

Embora este processo tenha sido restringido à área de epilepsia infantil, poderá ser aplicado a toda a área de epilepsia com algumas modificações. Além disso, poderá também ser adaptado a outras, adicionando os sintomas e características importantes às ontologias perante a categoria respetiva, e construindo algumas regras necessárias para uma correta classificação.

Existem ainda diversos contextos e sinónimos que podem ainda ser identificados e adicionados às ontologias. Esta ontologia apenas foi desenvolvida consoante os registos e a área apresentada, contudo seria útil a integração e tradução de algumas ontologias fornecidas pela UMLS para melhor classificar e conseguir analisar outras áreas.

Bibliografia

1. Cohen, A.M. and W.R. Hersh, *A survey of current work in biomedical text mining.* Briefings in Informatics, 2004. **6**(1): p. 57-71.

2. Ludwick, D.A. and J. Doucette, *Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries* International journal of medical informatics, 2008. **78**: p. 22-31.

3. Din, Z.M.U., S.H. Woo, W. Qun, J.H. Kim, and J.H. Cho, *HEN Simulation of a Controlled Fluid Flow-Based Neural Cooling Probe Used for the Treatment of Focal and Spontaneous Epilepsy.* Sensor Science and Technology, 2011. **20**(1): p. 19-24.

4. Meacham, J., *A Storm In The Brain.* Newsweek, 2009.

5. Brown, R.J. and M.R. Trimble, *Dissociative psychopathology, non-epileptic seizures, and neurology.* J Neurol Neurosurg Psychiatry, 2000. **69**(3): p. 285-9.

6. Fogoros, R.N. *The Misdiagnosis of Epilepsy*. 2009 May 20th 2013]; Available from: http://heartdisease.about.com/b/2009/08/07/the-misdiagnosis-of-epilepsy.htm.

7. Engel, J., *Seizures and Epilepsy*. 2 ed2012: Oxford University.

8. Berg, A.T., et al., *Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology.* Epilepsia, 2010. **51**(4): p. 676-685.

9. Warman, M.L., et al., *Nosology and classification of genetic skeletal disorders: 2010 revision.* Am J Med Genet A, 2011. **155A**(5): p. 943-68.

10. Armstrong, D., *Diagnosis and nosology in primary care.* Sociology of Diagnosis, 2011. **73**(6): p. 801–807.

11. Coonan, K.M., *Medical informatics standards applicable to emergency department information systems: making sense of the jumble.* Academic Emergency Medicine, 2004. **11**(11): p. 1198-1205.

12. Software, A. *The International Classification of Diseases, 9th Revision, Clinical Modification*. May 30th 2013]; Available from: http://www.icd9data.com/2013/Volume1/320-389/340-349/345/default.htm.

13. Hoerbst, A. and E. Ammenwerth, *Electronic Health Records.* Methods of Information in Medicine, 2010. **49**(4): p. 320-36.

14. Tsumoto, S. and S. Hirano, *Clustering-based Analysis in Hospital Information Systems.* International Conference on Granular Computing, 2011: p. 669-674.

15. Luo, J.S., *Electronic Medical Records.* Primary Psychiatry, 2006. **2**(13): p. 20-23.

16. Elmasri, R. and S. Navathe, *Fundamentals of Database Systems* 2010: Pearson Education.

17. Pawlak, Z., *Rough sets and intelligent data analysis.* Informatics and Computer Science, 2002. **147**(4): p. 1-12.

18. Piatetsky-Shapiro, G., *Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop.* AI Magazine, 1990. **11**(4).

19. Tan, P.N., M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Vol. 1. 2006: Pearson Education.

20. Ngaia, E.W.T., L. Xiub, and D.C.K. Chaua, *Application of data mining techniques in customer relationship management: A literature review and classification.* Expert Systems with Applications, 2009. **36**(2): p. 2592–2602.

21. Kantardzic, M., *Data mining: concepts, models, methods and algorithms* 2011, United States of America: John Wiley & Sons, Inc.

22. Chaovalit, P. and L. Zhou, *Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches.*, in *Proceedings of the 38th Hawaii International Conference on System Sciences* 2005.

23. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases.* American Association for Artificial Intelligence, 1996: p. 37-54.

24. Fayyad, U. and G. Piatetsky-Shapiro, *From Data Mining to Knowledge Discovery in Databases.* AIntelligence, 1996: p. 37-54.

25. Azevedo, A. and M.F. Santos, *KDD, SEMMA And CRISP-DM: A Parallel Overview* Computer Science and Information Systems, 2008.

26. Chapman, P. and J. Clinton, *CRISP-DM 1.0.* 2000.

27. Wirth, R. and J. Hipp, *CRISP-DM: Towards a Standard Process Model for Data Mining.* 2000.

28. Kadav, A., J. Kawale, and P. Mitra *Data Mining Standards*.

29. Han, J. and M. Kamber, *Data Mining Concept and Techniques Second Edition*. 2 ed, ed. J. Gray and M. Research2006, San Francisco: Morgan Kaufmann.

30. Rohanizadeh, S.S. and M.B. Moghadam, *A Proposed Data Mining Methodology and its Application to Industrial Procedures* Journal of Industrial Engineering, 2009. **4**(1): p. 37-50.

31. Zhao, Y. *R*. 2011; Available from: http://www.rdatamining.com/r.

32. Williams, G.J., *Rattle: A Data Mining GUI for R.* The R Journal, 2009. **1**(2): p. 45-55.

33. Hearst, M.A., *Untangling text data mining* in *Association for Computational Linguistics on Computational Linguistics* 1999 Stroudsburg. p. 3-10.

34. Grimes, S. *A Brief History of Text Analytics*. 2007 08-04-2013]; Available from: http://www.b-eye-network.com/view/6311.

35. Hammouda, K.M. and M.S. Kamel, *Efficient Phrase-Based Document Indexing for Web Document Clustering* in *IEEE Transanctions on knowledge and data engineering* 2004. p. 1279-1296.

36. Witten, I.H., K.J. Don, M. Dewsnip, and V. Tablan, *Text mining in a digital library.* International Journal on Digital Libraries, 2004. **4**(1).

37. Feldman, R. and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press, 2006.

38. Krallinger, M., R.A. Erhardt, and A. Valencia, *Text-mining approaches in molecular biology and biomedicine.* Drug discovery today: biosilico, 2005. **10**(6).

39. Tseng, Y.H. and C.J. Lin, *Text mining techniques for patent analysis.* Information Processing and Management, 2006. **43**(5).

40. Han, J. and M. Kamber, *Data Mining Concept and Techniques* 2006, San Francisco: Morgan Kaufmann.

41. Wongthongtham, P. and E. Chang, *Development of a Software Engineering Ontology for Multi-site Software Development*, in *IEEE Transactions on knowledge and Data Engineering*2008.

42. Selden, C.R. and B.L. Humphreys, *Unified Medical Language System: Current Bibliographies in Medicine, January 1986 - December 1996,* 1997: Diane Publishing.

43. Mandl, K.D., P. Szolovits, and I.S. Kohane, *Public standards and patients' control: how to keep electronic medical records accessible but private.* British Medical Journal, 2001. **322**(7281): p. 283-286.

44. Tremblay, M.C., D.J. Berndt, S.L. Luther, P.R. Foulis, and D.D. French, *Identifying fall-related injuries: Text mining the electronic medical record.* Information Technology and Management, 2009. **10**(4): p. 253-265.

45. Molina, A. and F. Pla, *Shallow parsing using specialized HMMs.* Journal of Machine Learning Research, 2002. **2**(4): p. 595-613.

46. Punyakanok, V., D. Roth, and W.-t. Yih, *The importance of syntactic parsing and inference in semantic role labeling.* Computational Linguistics, 2008. **34**(2): p. 257-287.

47. Zhou, X., H. Han, I. Chankai, A. Prestrud, and A. Brooks. *Approaches to Text Mining for Clinical Medical Records*. in *Association for Computing Machinery*. 2006. New York.

48. Friedlin, J., S. Grannis, and J.M. Overhage, *Using natural language processing to improve accuracy of automated notifiable disease reporting.* American Medical Informatics Association Annu Symp Proc., 2008: p. 207-211.

49. Friedman, C., G. Hripcsak, W. DuMouchel, S.B. Johnson, and P.D. Clayton, *Natural language processing in an operational clinical information system.* Natural Language Engineering, 1995. **1**(1): p. 83 - 108.

50. Liu, H. and C. Friedman, *CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML.* Studies in health technology and informatics, 2004. **107**(1): p. 639.

51. Fiszman, M., P.J. Haug, and P.R. Frederick, *Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports.* Proc American Medical Informatics Association Symp, 1998: p. 860-864.

52. Fiszman, M., W.W. Chapman, D. Aronsky, R.S. Evans, and P.J. Haug, *Automatic detection of acute bacterial pneumonia from chest X-ray reports.* Journal of the American Medical Informatics Association, 2000. **7**(6): p. 593-604.

53. Day, S., L.M. Christensen, J. Dalto, and P. Haug, *Identification of trauma patients at a level 1 trauma center utilizing natural language processing.* Journal of Trauma Nursing, 2007. **14**(2): p. 79-83.

54. Trick, W., W. Chapman, M. Wisniewski, B. Peterson, S. Solomon, and R. Weinstein, *Electronic interpretation of chest radiograph reports to detect central venous catheters.* Infection Control and Hospital Epidemiology, 2003. **24**(12): p. 950-954.

55. Pestian, J.P., C. Brew, P. Matykiewicz, D.J. Hovermale, N. Johnson, K.B. Cohen, and W. Duch, *A shared task involving multi-label classification of clinical free text.* BioNLP '07, 2007: p. 97-104.

56. Friedman, C., C. Knirsch, L. Shagina, and G. Hripcsak, *Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries.* Proc American Medical Informatics Association Symp, 1999: p. 256-260.

57. Karanikolas, N.N. and C. Skourlas, *Shifting from legacy systems to a data mart and computer assited information resources navigation framework.* Centre for Economic

and International Studies 2003 - Databases And Information Systems Integration, 2003: p. 300-305.

58. Ruch, P. and J. Gobeill, *From clinical narratives to ICD codes: automatic text categorization for medico-economic encoding.* Standard Schedules Information Manual 2007.

59. Chuang, C.-C. and Taipei, *Robust support vector regression networks for function approximation with outliers.* Neural Networks, IEEE Transactions onx\, 2002. **13**(6): p. 1322 - 1330

60. Roque, F.S., et al., *Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts.* Public Library of Science Computational Biology, 2011.

61. Holden, E.W., et al., *Developing a Computer Algorithm to Identify Epilepsy Cases in Managed Care Organizations.* Disease Management, 2005. **8**(1): p. 1-14.

62. Piazza, P., *Health Alerts to Fight Bioterror: New Web-Based Applications Collect Health-Related Data and Search for Patterns That Might Indicate That a Bioterror Attack Is Underway.* Security Management, 2002. **46**(5).

63. Carrington, M.J., S. Kok, K. Jansen, and S. Stewart, *The Green, Amber, Red Delineation of Risk and Need (GARDIAN) management system: a pragmatic approach to optimizing heart health from primary prevention to chronic disease management.* European Journal of Cardiovascular Nursing, 2013. **12**(4): p. 337-45.

64. Cheng, S., M.H. Azarian, and M.G. Pecht, *Sensor systems for prognostics and health management.* Sensors (Basel), 2010. **10**(6): p. 5774-97.

65. Pearson, S.A., A. Moxey, J. Robertson, I. Hains, M. Williamson, J. Reeve, and D. Newby, *Do computerised clinical decision support systems for prescribing change practice? A systematic review of the literature (1990-2007).* Biomedical Central Health Services Research, 2009. **9**(1).

66. Looi, K.L. and P.N. Black, *How often do physicians review medication charts on ward rounds?* BMC Clin Pharmacol, 2008. **8**(9): p. 8-9.

67. Gatenby, R.A., *A change of strategy in the war on cancer.* Nature, 2009. **459**(7246): p. 508-9.

68. Ongenaert, M. and L. Dehaspe, *Integrating automated literature searches and text mining in biomarker discovery.* BMC Bioinformatics, 2010. **11**(5).

69. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2095-128.

70. McPhee, S.J., M.A. Papadakis, and M.W. Rabow, *Current medical diagnosis & treatment 2010* 2010: McGraw-Hill Medical.

71. Weiden, M., D. Khosla, and M. Keegan. *Electroencephalographic detection of visual saliency of motion towards a practical brain-computer interface for video analysis*. in *ICMI '12 Proceedings of the 14th ACM international conference on Multimodal interaction*. 2012.

72. Topić, E., I. Watson, E. Homšak, and J.L. Krleža, *New Trends In Classification, Monitoring And Management Of Gastrointestinal Diseases* 2012, Dubrovnik.

73. Escorpizo, R., N. Kostanjsek, C. Kennedy, M.M. Nicol, G. Stucki, and T.B. Ustun, *Harmonizing WHO's International Classification of Diseases (ICD) and International Classification of Functioning, Disability and Health (ICF): importance and methods to link disease and functioning.* BMC Public Health, 2013. **13**(1): p. 742.

74. Berwick, D.M. and A.D. Hackbarth, *Eliminating Waste in US Health Care.* JAMA, 2012. **307**(14): p. 1513-6.

75. Izadi, M.T. and D.L. Buckeridge, *Decision theoretic analysis of improving epidemic detection.* American Medical Informatics Association Annu Symp Proc, 2007: p. 354-8.

76. Pereira, L., R. Rijo, C. Silva, and M. Agostinho, *Using Text Mining to Diagnose and Classify Epilepsy in Children*, in *IEEE HealthCom*2013: Lisbon.

77. DuPaul, G.J. and G. Stoner, *ADHD in the Schools: Assessment and Intervention Strategies*. Vol. 2. 2004. :.

78. Kao, A. and S.R. Poteet, *Natural Language Processing and Text Mining*2007: Springer.

79. Bukhari, A.C. and Y.-G. Kim, *Ontology-assisted automatic precise information extractor for visually impaired inhabitants.* Artificial Intelligence Review, 2012. **38**(1): p. 9-24.

80. O'Connor, M., H. Knublauch, S. Tu, B. Grosof, M. Dean, W. Grosso, and M. Musen, *Supporting Rule System Interoperability on the Semantic Web with SWRL.* Computer Science, 2005. **3729**: p. 974-986.

81. Cunningham, H., et al. *Developing Language Processing Components with GATE Version 7 (a User Guide)*. 2013; Available from: http://gate.ac.uk/sale/tao/splitch18.html#chap:ml.

82. Aha, D.W., D. Kibler, and M.K. Albert, *Instance-based learning algorithms.* Machine Learning, 1991. **6**(1): p. 37-66.

83. Lavanya, D. and K.U. Rani, *Performance Evaluation of Decision Tree Classifiers on Medical Datasets.* International Journal of Computer Applications, 2011. **26**(4).

84. Dong, F., P.D. Mitchell, V.M. Davis, and R. Recker. *Impact of Atrazine on the Sustainability of Weed Management in Wisconsin Corn Production*. in *2013 Annual Meeting, August 4-6, 2013, Washington, DC*. 2013. Agricultural and Applied Economics Association.

85. Tsoumakas, G. and I. Katakis, *Multi-Label Classification: An Overview.* International Journal of Data Warehousing & Mining, 2007. **3**(3): p. 1-13.

86. Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification*, 2003.

Anexo 1

Mãe refere que ele tem grande dificuldade em
dormir
lipotimia na escola (noção de movimentos
convulsivos) Amnésia para o aconteci
mento
à entrada neurologicamente bem glasgow 15 com
cefaleias
Peso=65kg
TT=36,1°c
En= 5
vai ficar na zona das macas

TRIAGEM PEDIÁTRICA...
Atribuição de Prioridade
Prioridade: Urgente/Amarelo
Discriminador: Urgente

Ped./Presc.:	Código	Descrição	Qtd.	Tipo	Destino	Estado
	->	->	->	Pedido Esp.	PEDIATRIA	->

PEDIATRIA (1ª observação)

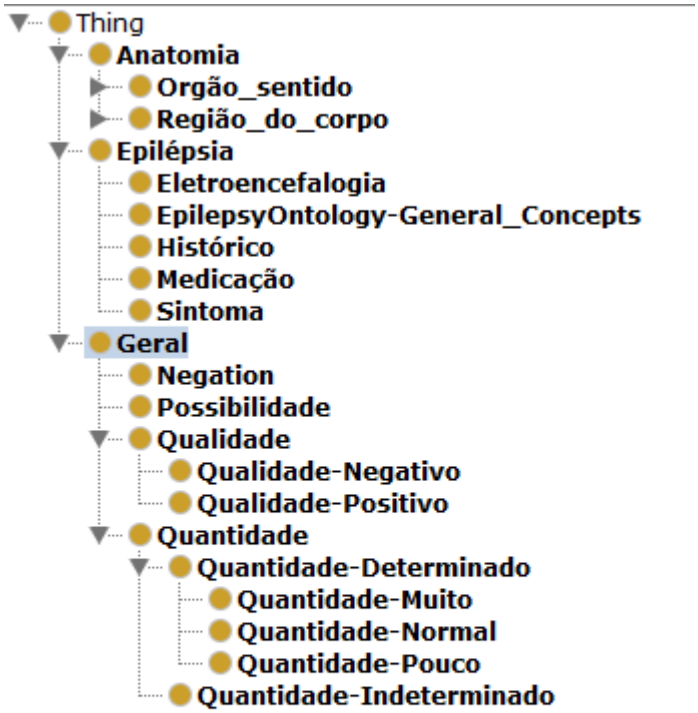
Entrada:	Saída:
2012-11-30 13:04:46	2012-11-30 13:28:23

Observações:

Adolescente de 17 anos
Vem por:
-lipotimia na escola, estava no computador, não tem
memoria para o sucedido, mas
segundo quem estava com ele terá tido movimentos
anomalos dos membros. Sem perd
a de continencia dos esfincteres
-segundo o proprio ja teve episodios, com inicio ha 1

Anexo 1 - Exemplo de um registo clínico eletrónico

Anexo 2



Anexo 2 - Fase inicial da ontologia desenvolvida

(Inicia em página impar)