Utilização de *Text Mining* para Apoio à Classificação de Registos de Alergias e Reações Adversas

---

Mélanie Gonçalves de Castro

2017

Mestrado em Informática Médica

Faculdade de Ciências | Faculdade de Medicina

Universidade do Porto

Orientador: Prof. Doutor José Alberto da Silva Freitas, Faculdade de Medicina da Universidade do Porto

Co-orientador: Prof. Doutor João Almeida Lopes da Fonseca, Faculdade de Medicina da Universidade do Porto

Agradecimentos

Aos meus pais, por tudo o que têm feito por mim, pelo ambiente de crescimento pessoal e intelectual a partir do qual pude crescer e me tornar na pessoa que sou hoje.
Aos meus avós e namorado que sempre me deram confiança e motivação e me apoiaram em todos os momentos.
Ao meu orientador Prof. Doutor Alberto Freitas, pelas suas recomendações e cordialidade com que sempre me recebeu, pelos seus conselhos e orientação que me permitiu evoluir profissionalmente.
Ao meu co-orientador Prof. Doutor João Fonseca, pela disponibilidade demonstrada e contributo na recolha e classificação manual dos registos clínicos de alergologia.
Ao Mestre Júlio Souza pela amizade, orientação e ajuda incansável, prestando-me apoio na parte informatizada do meu projeto, pela cordialidade e partilha de conhecimento que se tornou imprescindível para os resultados hoje apresentados.
À CUF Porto, pela contribuição e acesso dos registos clínicos de alergologia me permitiu avançar com o meu trabalho.
À Faculdade de Medicina (FMUP) e Faculdade de Ciências (FCUP) da Universidade do Porto, pela qualidade do ensino prestado.
A todas as pessoas minhas amigas, que, ainda que não tenham contribuído diretamente, o fizeram de forma indireta e me apoiaram incondicionalmente a todo o momento no decorrer desta tese.

A todos o meu mais sincero obrigada,
Mélanie Castro

Resumo

O Catálogo Português de Alergias e outras Reações Adversas (CPARA) utiliza a terminologia clínica *SNOMED CT* (*Systematized Nomenclature of Medicine - Clinical Terms*) para a classificação normalizada das alergias e reações adversas. No entanto, é sentida a necessidade de um sistema capaz de identificar e classificar automaticamente de acordo com o CPARA registos clínicos em texto livre. Assim, pretendemos utilizar técnicas de *text mining* e *machine learning* para a criação de um protótipo de uma ferramenta *web* para a classificação automática do registo clínico em texto livre para a terminologia *SNOMED CT*. Foram obtidos registos clínicos de doentes observados em consulta de Imunoalergologia da CUF Porto onde estes foram classificados com base na terminologia *SNOMED CT* por dois métodos: 1) manualmente, por um especialista em Imunoalergologia; 2) automaticamente, por técnicas de *text mining* e *machine learning*. Para a criação do protótipo da ferramenta *web* foram classificados 48 registos clínicos de doentes para o conjunto de treino. De forma a estruturar a informação presente nos textos dos registos clínicos, numa etapa de pré-processamento, foram aplicados algoritmos de *text mining* disponíveis através do *software* Rapidminer, nomeadamente os processos de *tokenização* e "*filter stopwords*". Nesta etapa, foi também utilizado um filtro para a identificação de palavras chave associadas às definições de cada código *SNOMED CT*. Foram construídos vários modelos de classificação, com base na informação estruturada e na classificação manual, através de técnicas de *machine learning*, nomeadamente algoritmos *Naive Bayes*, *Support Vector Machines* (*SVM*), J48 (árvore de decisão), *Random Forest*, *Nnge* e *k-Nearest Neighbours* (*k-NN*). Medidas de sensibilidade, especificidade e AUROC foram utilizadas para a avaliação dos classificadores obtidos. Os classificadores obtidos apresentaram alta ou média especificidade e boa sensibilidade para a maioria dos códigos *SNOMED CT* analisados. Os algoritmos *Random Forest* e *SVM* apresentaram os melhores resultados relativamente à especificidade, sendo selecionados como os melhores classificadores para 11 dos códigos da amostra, enquanto que o *k-NN* apresentou o melhor desempenho quanto à sensibilidade, sendo escolhido como o melhor algoritmo para 15 dos códigos da amostra. A árvore de decisão (J48), foi o algoritmo que apresentou os piores resultados, quer na sensibilidade como na especificidade. Os resultados deste estudo mostraram o potencial das técnicas de *text mining* e *machine learning* para apoio à identificação de casos de alergias e reações adversas, permitindo extrair de registos clínicos em texto livre a informação necessária para a classificação automática em códigos da terminologia *SNOMED CT,* conforme as normas de classificação do CPARA. Para disponibilizar efetivamente o serviço *web*, os classificadores deverão ser revistos e melhorados tendo por base a obtenção de novos registos clínicos e a melhoria contínua do dicionário de palavras chave.

Palavras-Chave: registo clínico, alergologia, imunoalergologia, classificação supervisionada, text mining, machine learning, apoio à decisão.

Abstract

The Portuguese Catalog of Allergies and other Adverse Reactions (CPARA - Catálogo Português de Alergias e outras Reações Adversas) has adopted SNOMED CT terminology to classify clinical cases of allergies and adverse reactions. Nevertheless, flagging and classifying textual clinical records according to the CPARA specifications in an automated way is highly necessary. Therefore, we aim to apply text mining and supervised machine learning techniques in order to implement a web-based application to automatically classify textual clinical records according to the SNOMED CT terminology, which is used in CPARA specifications. Forty-eight textual clinical records were obtained from outpatient consultations in allergy/immunology in a private hospital located in Porto. These records were classified into SNOMED CT codes in two different phases: 1) manually performed by a specialist in immunoallergology; 2) automatically performed by a web-based application which was developed after applying text mining and supervised machine learning techniques, based on the clinical records previously classified by specialists in phase 1. The no structured textual clinical records were submitted to a pre-processing phase, in which we applied text-mining algorithms in order to identify and filter the most relevant words and sentences, namely Tokenization and Stopwords removal. In this pre-processing phase, a filter to identify specific keywords associated with some SNOMED CT codes was also applied. Several classification models were built using supervised machine learning algorithms, such as Naive Bayes, Multinomial Naive Bayes, Support Vector Machines (SVM), J48, NNge and k-Nearest Neighbors (k-NN). Sensitivity and specificity measures, as well as the Area Under the ROC curve were analysed to assess and validate the classifiers. The obtained classifiers performed well as they presented medium to high specificity and good sensitivity for most of the SNOMED CT codes analysed in this study. Random Forest and SVM had the best results regarding specificity as they were the best performers for 11 SNOMED CT codes, whereas K-NN had the best results regarding sensitivity, being the best performer for 15 SNOMED CT codes. Decision tree-based algorithm J48 had the worst results for sensitivity and specificity. With this study, we showed that text mining and supervised machine learning techniques can potentially be useful for the proper identification of allergies and adverse reactions, as well as extract all the necessary information for the automatic classification of textual clinical records into specific SNOMED CT codes, as recommended by Portugal's CPARA. In order to provide a web platform to address this matter, the classifiers must be first improved with a larger number of new textual clinical records and count with the help of specialists in immunology/allergy to increase the number and quality of relevant keywords that can potentially be associated with specific SNOMED CT codes.

Key-Words : clinical record, allergy, immunology, supervised classification, text mining, machine learning, decision support.

Preâmbulo

Durante o meu percurso escolar no Mestrado em Informática Médica da Faculdade de Medicina e Faculdade de Ciências da Universidade do Porto, deparei-me com diversas terminologias utilizadas na Informática Médica, tal como a *ICD-9* e a *ICD-10*. Terminologias estas que permitem criar padrões e aperfeiçoar termos e siglas geralmente utilizadas pelo Ministério da Saúde, promovendo uma maior facilidade de recuperação, acesso, divulgação e partilha das informações de saúde. Posto isto, verifiquei assim que, por mais eficientes que fossem as terminologias e por mais vasto que fosse o vocabulário e experiência, quando necessária a aplicação ou a transição para novas terminologias, poucas eram as ferramentas/aplicações existentes que apoiavam essa etapa. Pelo que então lhes resolvi dedicar uma especial atenção.

Regularmente surgia o tema *SNOMED CT* nas unidades curriculares do Mestrado em Informática Médica, onde esta era supracitada com grande relevância e importância, como uma nova linguagem, uma nova terminologia, ainda pouco conhecida e abrangida em Portugal. Esta novidade, despertou um interesse em conhecer melhor quais as vantagens que esta terminologia realmente trazia e quais os desafios presentes na utilização desta terminologia pelos profissionais de saúde. Posto desperto o interesse em conhecer melhor esta terminologia, surgiram algumas reuniões com o Prof. Doutor Alberto Freitas e o Prof. Doutor João Fonseca da possibilidade em vir a contribuir neste âmbito. Foi então que tive conhecimento que a SPMS teria criado um Catálogo Português de Alergias e outras Reações Adversas (CPARA), onde a qual já estaria na 3ª versão e onde inclusivamente já estaria a ser utlizada a classificação em *SNOMED CT* na mesma. Tomei conhecimento que o método *standard* de classificação no CPARA era ainda feito de forma manual, geralmente pelos especialistas de imunoalergologia, e, era de grande interesse para os profissionais de saúde a existência de um método de classificação automatizada que sugerisse os códigos *SNOMED CT* para o CPARA. Assumindo conscientemente a importância deste projeto, aceitei o desafio, apesar de conhecer as dificuldades que poderiam surgir quer a nível informático, quer a nível temporal que estaria implicado nas burocracias para o acesso a informações de foro clínico, quer no número de registos que iria conseguir obter devidamente classificados manualmente.

Conclusivamente, espero assim contribuir para uma evolução dos serviços de saúde que utilizam o CPARA na classificação de registos de alergias e reações adversas, otimizando o serviço prestado na classificação com a terminologia *SNOMED CT*, contribuindo assim para o desenvolvimento informatizado na saúde.

Índice

Índice de Tabelas

Índice de Figuras

Índice de Acrónimos

ARFF	Attribute-Relation File Format
ATC	Anatomical Therapeutic Chemical
AUROC	Area under the ROC curve
CART	Classification and Regression Trees
CSS	Cascading Style Sheets
CPARA	Catálogo Português de Alergias e Outras Reações Adversas
EUA	Estados Unidos da América
FCUP	Faculdade de Ciências da Universidade do Porto
FMUP	Faculdade de Medicina da Universidade do Porto
FN	Falso Negativo
FP	Falso Positivo
GNU	General Public License
HIV	Human Immunodeficiency Virus
HTML	Hyper Text Markup Language
ICD	International Classification of Diseases
ICD-9-CM	International Classification of Disease, 9th edition, Clinical Modification
ICD-10	International Classification of Diseases, 10 th Edition
ICD-O3	International Classification of Diseases for Oncology (ICD-O) - third edition
ICPC-2	International Classification of Primary Care – ICPC (Classificação Internacional de Cuidados de Saúde Primários)
ID3	Iterative Dichotomiser 3
IHTSO	International Health Terminology Standards Development Organisation
Java EE	Java Enterprise Edition
K-NN	K-Nearest Neighbours
NNge	Non-Nested Generalized Exemplars
PNV	Plano Nacional de Vacinação
ROC	Receiver Operating Curve
RSE	Registo de Saúde Eletrónico
SNOMED CT	Systematized Nomenclature of Medicine-Clinical Terms
SPAIC	Sociedade Portuguesa de Alergologia e Imunologia Clínica
SPMS	Serviços Partilhados do Ministério da Saúde
SVM	Support Vector Machine
TI	Tecnologias de Informação
TVN	Taxa de Verdadeiros Negativos
TVP	Taxa de Verdadeiros Positivos
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

VPP	Valor Preditivo Positivo
VPN	Valor Preditivo Negativo
WEKA	Waikato Environment for Knowledge Analysis

"Things should be made as simple as

possible, but no simpler."

Albert Einstein

1. Introdução

1.1 Enquadramento

A identificação de reações alérgicas é de extrema relevância para os cuidados na saúde visto que podem causar sintomas respiratórios e gastrointestinais graves, complicações cardiovasculares, infarto do miocárdio ou paragem cardíaca (Ewan, 1998); (Hegvik, Johan-Arnt ; Rygnestad, 2002). Por exemplo, as reações adversas a medicamentos, ocorrem entre 10 a 15% das hospitalizações globalmente (WAO, 2016), sendo que já foram identificados riscos significativos, custos e o aumento do tempo de internamento associados a reações adversas desconhecidas (Pirmohamed et al., 2004). Adicionalmente a anafilaxia, cujas causas mais comuns decorrem de reações a certos medicamentos, alimentos e picadas de insetos, é responsável por complicações e óbitos por exemplo associadas a procedimentos cirúrgicos (Dippenaar & Naidoo, 2015); (Lindsted et al., 2014);(Y. Xu et al., 2014).

Informações importantes sobre alergias podem ser identificadas nos registos clínicos, entretanto tais informações continuam a ser pouco utilizadas pelos médicos (Iskio et al., 2006); (Slight et al., 2013). A procura manual de informações clínicas em textos de registos clínicos requere muita atenção dos médicos, médicos estes que não dispõem de tempo para esse efeito, o que pode impor dificuldades no fluxo de trabalho e na comunicação médico-paciente. Desta forma é necessário desenvolver métodos mais robustos para a identificação e apresentação de informações clínicas relevantes a partir dos registos clínicos.

1.2 Motivação

O que me motivou a trabalhar neste projeto foi a ausência de ferramentas de apoio à decisão automatizadas que sugiram classificações utilizando a terminologia *SNOMED CT* para o Catálogo Português de Alergias e outras Reações Adversas.

Ao longo do meu percurso no Mestrado em Informática Médica, tive conhecimento da existência do Catálogo Português de Alergias e outras Reações Adversas, que procura caracterizar os casos clínicos em sete domínios (na V3.0): Origem da Informação, Data da Reação, Classificação da Reação Adversa, Alergénios e outras Substâncias, Reação Adversa, Gravidade e Estado, do qual também adquiri conhecimento que atualmente o método de classificação neste registo utiliza uma terminologia muito abrangente, a *SNOMED CT*. Verifiquei que a classificação neste catálogo é feita manualmente pelos profissionais da área de imunoalergologia, identificando desde logo a necessidade de apoiar esta classificação para

estes utilizadores, criando um sistema automatizado de sugestão de classificações *SNOMED CT* para o CPARA, no domínio de alergologia. Com isto, pretende-se estimular a classificação, com uma metodologia de maior simplicidade, recorrendo a este modelo de classificação que venho apresentar.

1.3 Objetivos

O objetivo principal deste projeto é a criação de um método de sugestão de classificação automatizado de registos clínicos não estruturados, que sugira códigos *SNOMED CT* para cada domínio do CPARA, recorrendo à aplicação de técnicas de *text mining* e *machine learning*. Com este projeto, pretende-se criar uma ferramenta de apoio à decisão, mais concretamente um protótipo de uma plataforma *web*, que sugira os códigos necessários para a classificação de registos clínicos de alergologia, tarefa esta que é geralmente desempenhada (manualmente) pelos especialistas em imunoalergologia. A ferramenta informatizada que se propõe, permite apoiar a classificação dos registos clínicos de alergologia, conforme as regras de classificação do Catálogo Português de Alergias e Reações Adversas na versão 3.0.

Tendo em conta a falta de um método que auxilie a classificação do CPARA na terminologia *SNOMED CT*, o objetivo maioritário passa por colmatar essa falta e inovar, cobrindo as necessidades verificadas pelos profissionais de saúde ou utilizadores deste tipo de classificação.

1.4 Organização da Tese

Este projeto encontra-se organizado de forma a facilitar a sua leitura e encontra-se dividido nas seguintes partes:

No primeiro capítulo encontraremos a introdução. Este capítulo permite obter um enquadramento do projeto e permite dar a compreender quais os objetivos e motivações que estimularam o desenvolvimento deste trabalho.

No segundo capítulo encontraremos o estado da arte, fase de extrema relevância que permite esclarecer termos essências do tema da dissertação, tais como *text mining* e *machine learning*, os classificadores utilizados e trabalhos com diferentes abordagens desenvolvidos neste âmbito.

No capítulo 3 encontra-se a metodologia, onde são dadas a conhecer as 3 diferentes fases envolventes deste projeto: 1) Recolha da Informação e Pré Processamento; 2) Construção do Modelo de Classificação; 3) Construção do Protótipo. Aqui é possível compreender quais

as peças fulcrais para a construção do modelo final, o protótipo e quais as tecnologias utilizada para que este resultasse.

No capítulo 4 encontra-se os resultados obtidos com o apoio do software *WEKA*, que nos permitiu efetivamente construir e obter o *feedback* da classificação, a viabilidade e a análise dos resultados obtidos.

No capítulo 5 encontram-se as limitações deste trabalho. Quais os obstáculos com que nos deparamos e dificuldades enfrentadas na implementação deste tipo de técnicas de *machine learning*.

No capítulo 6 e 7, temos a conclusão deste trabalho e os trabalhos futuros, onde é feita uma análise conclusiva do trabalho desenvolvido no âmbito desta dissertação e são dados a conhecer quais os projetos que ambicionávamos para o futuro, que permitissem otimizar estes modelos de classificação de registos clínicos de alergologia.

2. Estado da Arte

2.1 Registo de Saúde

Quando um paciente recorre a um serviço de saúde, é essencial a elaboração de um registo que armazene toda a informação do processo clínico, denominado este como: registo clínico. Este registo é efetuado por profissionais de saúde (médicos, enfermeiros, entre outros) ou até mesmo pelo utente ou acompanhante do utente, e, contém informações relevantes quer de caráter médico quer de caráter administrativo (Estado, 2009).

Ao longo dos anos, a informatização dos registos clínicos tem sido uma realidade presente. Inicialmente estes registos eram realizados de forma manuscrita, surgindo com este método alguns problemas, tal como a necessidade de integração de toda a informação de um processo clínico resultante de diversas fontes. Outros pontos que afetavam os registos em papel eram: a ilegibilidade do autor do registo, a falta de estruturação, a duplicidade de informação e até mesmo a sua perda, comprometendo a informação no registo e comunicação. De forma a colmatar estes pontos negativos do registo em papel, surgiu o RSE, registo de saúde eletrónico (Medicina, 2003). Este novo método digital de registo permitiu assim resolver problemas de estruturação que lhe estavam, associados, permitiu um fácil acesso e consulta, permitiu a consulta e interoperabilidade entre diferentes sistemas, possibilitou inclusivamente a prestação de auxílio na decisão clínica, na avaliação dos cuidados prestados e permitiu diminuir o espaço ocupado no depósito dos registos clínicos. No entanto, surgiram também alguns problemas, tais como a segurança dos dados no sistema informático.

Em suma, este tipo de registo veio inovar, pois contribuiu para a evolução clínica, permitindo a consulta dos dados com diversos formatos, promoveu a partilha de informação entre diversos utilizadores e possibilitou a restrição do acesso aos dados.

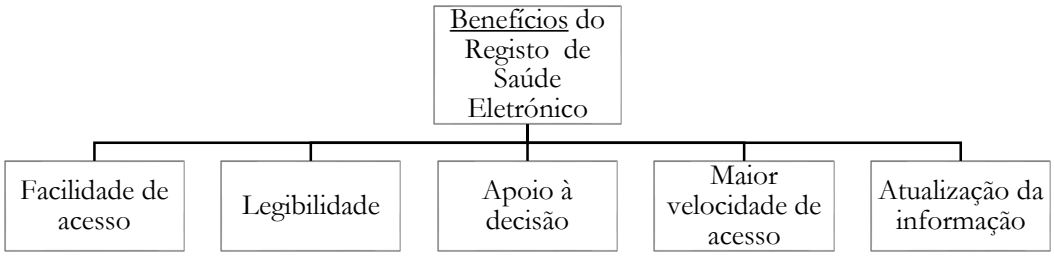


Figura 1: Descrição de alguns benefícios do RSE (Registo de Saúde Eletrónico)

Figura 2: Alguns Problemas do RSE (Registo de Saúde Eletrónico)

2.2 CPARA (Catálogo Português de Alergias e outras Reações Adversas)

Estima-se que em 2012, mais de 2 milhões de Portugueses, detinham alguma forma de doença alergénica, tais como, a título exemplificativo, a rinite ou asma (Comissão para a Informatização Clínica et al., 2012). Define-se o conceito de alergia como uma resposta exagerada do sistema imunológico, quando exposto a uma ou mais substâncias estranhas ao organismo. Quando o organismo é exposto a estas substâncias, surge uma reação de hipersensibilidade no organismo, manifestando-se de diversas formas: urticária, eczema, asma, rinite, etc. É definida como reação adversa, quando ocorre qualquer reação inesperada após um contato a um estímulo definido.

De encontro com a ausência verificada de uma norma nacional que permitisse definir um registo estruturado eficaz e correto, promovendo a partilha da informação clínica e da sua normalização, foi criado o CPARA, que atualmente se encontra com a versão 3.1.1.(Comissão para a Informatização Clínica et al., 2016). O CPARA - Catálogo Português de Alergias e Outras Reações Adversas, foi criado em 2012, com a versão inicial V1.0, com o intuito de registar e partilhar informações de reações alérgicas em todo o sistema de saúde português, permitindo inclusivamente a obtenção de uma perspetiva de saúde pública e a diminuição do número de incidentes clínicos associados a um registo incompleto de alergias e/ou reações adversas.

O CPARA permite aumentar o conhecimento associado às alergias e reações adversas contribuindo em diferentes vertentes, sendo que uma delas é a capacidade de promover quais os melhores alimentos ou medicamentos que determinado paciente deve tomar, de forma a evitar episódios de alergia associados a exposição a um alergénio. Para além das vantagens anteriormente referidas, o CPARA trás vantagens visíveis como a facilidade de acesso e consulta do histórico do paciente, a otimização do tempo de ação e rapidez, a prevenção de episódios agravados perante determinado alergénico e uma maior eficácia de partilha e interoperabilidade de sistemas.

2.2.1 Versões do CPARA

Atualmente encontra-se na versão 3.1.1. A primeira e segunda versão surgiram ambas em 2012. A terceira versão do CPARA (V3.0) foi implementada em 2015 e permitiu a codificação de casos clínicos de imunoalergologia em *SNOMED CT*, ajustando o modelo de informação com maior especificidade, bem com a atualização do conteúdo (Comissão para a Informatização Clínica et al., 2012). A versão 3 permitiu ultrapassar problemas de partilha a nível nacional da informação e contribuiu para a promoção e qualidade dos cuidados de saúde prestados a cidadãos portadores de alergias e/ou reações adversas. Em 2016, criaram uma atualização da versão 3.0, a 3.1 que visa abordar com maior detalhe o ponto 6 da estrutura do registo (gravidade) (Comissão para a Informatização Clínica et al., 2016).

2.2.2 Estrutura do CPARA

O CPARA é de implementação obrigatória, encontrando-se disponível nos sistemas de informação das instituições de saúde. Desde esta nova atualização da versão 3.0 para a 3.1., o CPARA encontra-se estruturado em oito domínios: 1- Origem da informação 2- Data da reação; 3- Classificação; 4- Reação adversa; 5- Gravidade; 6- Alergénio; 7- Estado; 8- Confirmação (Comissão para a Informatização Clínica et al., 2016). Cada domínio permite uma identificação pormenorizada, dos dados mais relevantes de um registo clínico de saúde, tal como podemos observar seguidamente:

Origem da Informação: Este domínio identifica qual a pessoa ou entidade responsável pela informação prestada quanto à reação adversa ou intolerância, se foi pelo médico ou pelo alergologista ou até mesmo pelo próprio utente. Foi criado com o objetivo de validar a fonte da informação, permitindo assim identificar a pessoa, bem como inferir se há qualidade e validade na informação prestada no registo de alergologia. Pode ser registada pelo utente/paciente, médico, imunoalergologista, outro profissional de saúde, cuidador, acompanhante, tutor ou familiar.

Data da Reação: O domínio "data da reação", permite que seja feito o registo da data mais recente em que o paciente apresentou um episódio de reação alérgica. De forma a ser completa, é importante que contenha os três campos: dia, mês e ano.

Classificação da Reação Adversa: permite dar a conhecer qual o tipo de agente responsável pela reação alérgica. Com esta classificação, pode identificar-se se a reação se encontra relacionada com fármacos, alimentos ou outra substância ou agentes. Na versão 3 do CPARA, apenas é possível classificar algumas das vacinas do Plano Nacional de Vacinação (PVN) que se encontram na tabela ATC, sendo que as restantes não podem ser classificadas com esta versão.

Reação Adversa: Este campo codifica o conjunto de valores que identifica o tipo de reação ao alergénio, por exemplo: anafilaxia, eczema, asma, etc.

Gravidade: este domínio codifica o grau de gravidade da reação alérgica.

Alergénio: este domínio está relacionado com a categoria da reação alérgica. Apresenta três categorias: alergénios alimentares, outros alergénios, medicamentos.

Estado: permite compreender qual o estado da atividade da reação adversa. Existem dois estados: ativo e inativo. O estado "ativo" indica se há suspeita ou confirmação de alergia ou reação adversa que leva a uma evicção de um alergénio; o estado "inativo" carateriza quando a suspeita não foi confirmada ou se ocorreu alguma tolerância.

Confirmação: o estado "confirmado", confirma que o diagnóstico foi validado pelo imunoalergologista através de testes clínicos ou laboratoriais e o estado "não confirmado" indica que o diagnóstico ainda não foi validado pelo imunoalergologista.

Note-se que este trabalho foi desenvolvido com a versão 3.0 do CPARA.

2.3 *SNOMED CT*

SNOMED CT - Systematized Nomenclature of Medicine-Clinical Terms, é uma terminologia clínica multilinguística, utilizada para a classificação na saúde em mais de 50 países. Tem uma origem internacional e é considerada a terminologia em saúde mais ampla do mundo, abrangendo o seu conteúdo mais de 300 mil termos, desde diagnósticos a procedimentos administrativos (ITHSO, 2014). É sustentada e atualizada pela *IHTSO- International Health Terminology Standards Development Organisation*, organização esta sem fins lucrativos.

O objetivo da *SNOMED CT*, passa por ser a mais extensa e abrangente terminologia clínica multilíngue do mundo, representando de forma precisa informações na área de saúde.

Esta terminologia facilita a estruturação e interoperabilidade entre sistemas de informação, como também permite o registo de dados e consulta com maior precisão. Permite a codificação, bem como o armazenamento, a troca e a agregação de dados clínicos e tem como principal foco contribuir para o melhoramento do atendimento do paciente, facilitar o acesso e recuperação da informação registada e respetiva comunicação na saúde (Ciolko, Lu, & Joshi, 2010).

Esta terminologia acarreta inúmeras vantagens tais como o aumento das oportunidades de apoio à decisão em tempo real, relatórios retrospetivos mais precisos para a pesquisa e gestão e redução de custos associados à duplicação de informação (Skeppstedt, Kvist, & Dalianis, 2007). A *SNOMED CT* permite inclusivamente especificar episódios alérgicos, contribuindo desta forma para um maior conhecimento dos dados dos utentes, trazendo uma maior perceção do estado de saúde.

A *ITHSO* demonstrou num documento recente de 2014, quem, e como, se pode beneficiar do uso eficiente da terminologia *SNOMED CT*, citando as três partes interessadas e beneficiadas com este uso: benefício populacional, benefício dos pacientes e clínicos, e, benefício no suporte à saúde baseada em evidências (ITHSO - International Health Terminology Standards Development Organisation, 2014).

Benefícios associados aos Pacientes e Clínicos:

- Acesso à informação clínica
- Acesso a sistemas de apoio à decisão permitindo verificar o registo e capacidade de prestar aconselhamento em tempo real, a título exemplificativo: os alertas clínicos
- Apoio na partilha de informações adequadas com outros profissionais de saúde, promovendo a interoperabilidade e o aumento da capacidade de compreensão e interpretação
- Permitindo pesquisas com precisão e abrangência, e mudanças de tratamento da paciente baseada na revisão do histórico
- Eliminando as barreiras linguísticas, dado esta terminologia ser multilingue

Benefício Populacional:

- Facilita a identificação de problemas de saúde, acompanha a saúde da população e as respostas na mudança de práticas clinicas
- Reduz as duplicações e erros
- Permite o fornecimento de dados relevantes para apoiar a investigação clínica e contribuir para futuras melhorias no tratamento
- Reforça as auditorias da prestação de cuidados com opções para análise detalhada dos registros clínicos para investigar

Beneficio no suporte à saúde baseada em evidências:

- Permite ligações entre registos clínicos e diretrizes clínicas e protocolos avançados
- Permite melhorar a qualidade dos cuidados experienciados pelos indivíduos
- Permite a redução dos custos de testes e tratamento inadequado e duplicação
- Limita a frequência e o impacto de eventos adversos de saúde
- Eleva o custo-benefício e a qualidade dos cuidados prestados às populações

Para além destes benefícios, a *SNOMED CT* permite dar uma maior abrangência de especialidades clínicas. Quando são utilizadas diferentes terminologias ou classificações, existe uma certa barreira linguística, que cria uma fronteira entre diferentes terminologias, barreira esta que a *SNOMED CT* consegue eliminar, promovendo a partilha de informação de uma forma mais simplificada e reutilização da informação clínica de forma estruturada (Ihtsdo, 2014). A classificação médica em diferentes países tem suscitado alguns obstáculos quando cruzados, em diferentes países. Com esta terminologia, os registos clínicos podem ser processados de diversas formas, no âmbito do paciente, como exemplificativamente: auditoria clínica, pesquisa, epidemiologia e gestão.

O processo da organização da informação com a *SNOMED CT* decorre com uma identificação de conceitos pré-definidos como parte da terminologia. Apresentado uma forma hierárquica de gravação, com esta terminologia é possível gravar de forma mais precisa dependendo do conteúdo e do detalhe, por exemplo: pneumonia (classificação mais básica); pneumonia bacteriana (maior grau de complexidade e detalhe); ou até mesmo pneumonia pneumocócica (descrevendo assim um elevado grau de detalhe). Isto é, dentro de vários tipos de pneumonias a *SNOMED CT* permite ir mais além, especificando um subtipo, por exemplo "pneumocócica" que possibilita assim especificar qual o agente causador e contribuir para a sua análise (Ihtsdo, 2014).

2.4 *Text Mining*

Com a grande evolução tecnológica a nível mundial, deparamo-nos com a recolha de informações em massa. Informações estas que muitas vezes não apresentam qualquer tipo de estruturação. Posto isto, tornou-se essencial a organização, mas acima de tudo a classificação destes blocos informativos. Para colmatar esta deficiência, era primordial a existência da sumarização automática de dados, extraindo a essência dos dados armazenados e a descoberta de padrões de grandes dimensões de conjuntos de dados (Zaïane, 1999). Quando somos confrontados com a simplicidade na procura da informação pretendida, surgiram ferramentas de auxílio neste sentido, que permitiram aumentá-la de forma significativa, agilizando o processo com inteligência, tal como o *text mining* (Report & Vincent, 2005). Inspirado no *Data Mining*, o *text mining* é definido como a ciência responsável pelo tratamento do processamento de informação/texto.

Une diferentes áreas tais como a estatística, linguística, informática e ciência cognitiva. Por outras palavras, o *text mining* consiste, através da identificação de diferentes tendências ou padrões, na extração de dados regulares de textos estruturados ou semi-estruturados (Aranha & Passos, 2006).

As competências que nos trás o *text mining* são diversas, dentro das quais a extração de dados, a recuperação de informações, a classificação, e inclusivamente a extração de padrões, de resumos e de textos (Bezerra, 2010). Como objetivo principal, o *text mining* tem o processamento de informações textuais que não apresentam qualquer tipo de estruturação, extraindo índices numéricos significativos a partir do texto para que fiquem disponíveis para *text mining*. Através deste método, as informações contidas num texto podem ser extraídas para que possam ser obtidas através dos resumos, palavras ou inclusivamente para o cálculo de resumos. Assim, é possível a análise de palavras, de conjuntos de palavras e/ou até mesmo a análise de documentos e consecutivamente a determinação de semelhanças entre eles ou quais as relações entre eles, quais as variáveis (Report & Vincent, 2005).

Para este trabalho, recorreu-se ao *text mining* com o intuito de através das suas técnicas nos fosse possível processar e trabalhar registos clínicos de alergologia, de forma a ser-nos possível trabalhar estes textos sem estruturação, filtrando a informação essencial que nos possibilitasse através das técnicas de *machine learning* a sugestão de classificações *SNOMED CT* para o CPARA.

2.5 Diferentes Abordagens utilizadas na área de *Text Mining* e *Machine Learning* na Saúde

A utilização de *text mining* e *machine learning* para a análise de registos clínicos integra uma área complexa que exige tempo e esforço consideráveis. Desenvolver métodos para o processamento de registos clínicos pode ser desafiante dado que geralmente estes registos encontram-se num formato de texto livre, não estruturado. Desta forma, o processamento destes registos torna-se difícil pois os episódios clínicos, sintomas e procedimentos são descritos com linguagens e formas distintas, o que requer conhecimento e ferramentas adicionais para a interpretação de termos específicos de modo a extrair informação semântica dos registos clínicos. Apesar dos desafios encontrados, esta é uma área de investigação com um grande potencial, tendo em conta a crescente digitalização de dados provenientes de registos clínicos. (Luis et al., 2015).

O processamento de dados de registos clínicos pode ser útil em diversas aplicações médicas. Existem inúmeros tipos de dados médicos em registos que incluem dados desde: informações demográficas de pacientes, história clínica e testes laboratoriais. Estes dados são uteis para a gestão de serviços e recursos, tal como o apoio ao diagnóstico e o tratamento de pacientes. Para além disso é importante ter em consideração que os médicos utilizam

terminologias clínicas normalizadas para descrever diagnósticos, sintomas, procedimentos ou tratamentos. As terminologias *ICD* e *SNOMED CT* são dois exemplos de terminologias clínicas normalizadas para descrever diagnósticos, sintomas e procedimentos em registo clínicos. Esta codificação é crucial para a partilha de informação clínica e como este processo não é direto, representa outro desafio importante para o processamento de registos clínicos (Schulz et al., 2012).

Na subsecção seguinte serão abordados os estudos relacionados com este trabalho de investigação, nomeadamente artigos publicados sobre a utilização de técnicas de *machine learning* e *text mining* na classificação de textos livres de registos clínicos. Esta revisão também inclui trabalhos que aplicaram técnicas de *text mining* e *machine learning* para a identificação de casos associados a imunoalergologia em registos clínicos.

2.5.1 Estudos de *Machine Learning* e *Text Mining* na Saúde

Nos últimos anos, tem se verificado a utilização de técnicas de *machine learning* e *text mining* para a identificação de doenças, sintomas ou condições clínicas a partir de textos livres, não estruturados de registos clínicos. Inúmeros algoritmos de *machine learning* têm sido testados com esta finalidade, encontrando-se o *SVM*, o *Naive Bayes*, as Árvores de Decisão, o *Random Forest* e o *K-NN* nos algoritmos mais frequentes para a classificação de textos livres provenientes de registos clínicos. Visto que este trabalho propõe a atribuição de códigos clínicos (códigos *SNOMED CT*) a registos de saúde, limitamos a pesquisa a trabalhos que procuraram desenvolver sistemas de suporte à decisão baseados na classificação de textos livres de registos clínicos em códigos de terminologias clínicas, tal como a *ICD* e a *SNOMED CT*.

A aplicação de técnicas de *machine learning* no estudo de (Rijo, Silva, & Gonçalves, 2014) apresenta algumas semelhanças com o nosso trabalho. Os autores classificam textos livres de registos clínicos para o apoio no diagnóstico de epilepsia em crianças através da atribuição de códigos *ICD-9*. Para esta finalidade, os autores utilizaram o algoritmo *k-NN* para mapear cada registo, ao código *ICD-9* correspondente. Embora tenha sido utilizada uma pequena amostra para a construção dos classificadores, os resultados obtidos a partir da aplicação do sistema em registos médicos reais demonstraram ter um bom desempenho quanto aos métodos utilizados. Outro estudo semelhante ao nosso trabalho foi recentemente conduzido por (Horng et al., 2017), em que classificadores baseados no algoritmo *SVM* foram aplicados para a identificação de pacientes com suspeita de infeção através de textos livres de registos clínicos, pelo que o tipo de infeção foi devidamente classificado com a terminologia de *ICD-9-CM*. Este trabalho demonstrou que a abordagem de classificação de textos livres apresentou melhor capacidade discriminatória em identificar infeções quando comparado com trabalhos anteriores que apenas utilizaram dados estruturados. (Ruch, Gobeill, Tbahriti, & Geissbühler, 2008), propôs um sistema de apoio à codificação clínica através da classificação automática de textos de registos clínicos. O sistema de classificação considerado neste trabalho foi a *ICD* e apenas considerou casos de hospitalizações. Neste trabalho foi aplicado a combinação de três classificadores baseados no algoritmo *k-NN* com um sistema

de classificação denominado como *data-poor classifier* que propõem um conjunto de códigos a um dado caso clínico com base em semelhanças lexicais.

O processamento de textos livres através de técnicas de *machine learning* também foi verificada em relatórios clínicos de diferentes especialidades médicas. (Jouhet et al., 2012) e (Nguyen et al., 2010), utilizaram relatórios patológicos para identificar os diferentes estágios do cancro. Enquanto o primeiro recorreu a classificadores baseados nos algoritmos *SVM* e *Naive Bayes* para classificar automaticamente a topografia e morfologia dos casos de cancro através da terminologia *ICD-O3*, o segundo optou por utilizar apenas o algoritmo *SVM* para aprender as relações entre conceitos da terminologia *SNOMED CT* e cada estágio do cancro. (Connolly et al., 2014), também utilizou classificadores baseados no algoritmo *SVM* para identificar o tipo de epilepsia a partir do processamento de textos livres de relatórios de progresso em pacientes diagnosticados epilepsia.

Existem também vários estudos que utilizaram técnicas de *machine learning* para classificação de textos livres de registos de autopsias. (Mujtaba et al., 2017), propôs um sistema para classificação automática de causas de morte provocadas por acidente através de atribuição de códigos *ICD-10* a registos de óbitos. Neste trabalho, o autor combinou várias técnicas de seleção de atributos para considerar apenas os termos mais relevantes para detetar a causa de morte com cinco algoritmos de *machine learning* distintos, nomeadamente *Naive Bayes*, *SVM*, *k-NN*, Árvore de Decisão (*J48*) e o *Random Forest*. Os resultados deste trabalharam indicaram que o sistema proposto é pratico e exequível para a classificação automática para causas de morte associadas a acidentes principalmente quando são utilizados os algoritmos *J48* e *Random Forest* em conjunto com técnicas de seleção de atributos. (Koopman, Zuccon, Nguyen, & Bergheim, 2015) utilizou textos livres de registos de óbitos para a identificação automática e classificação de cancros e outras quatro doenças de interesse, nomeadamente a diabetes, *influenza*, pneumonia e infeção por *HIV*. Em ambos os estudos, o autor considerou apenas as caraterísticas de maior relevância para a classificação, como termos simples e compostos e conceitos *SNOMED CT*, juntamente com classificadores *SVM*. Os resultados demonstraram que os métodos de classificação baseados em técnicas de *machine learning* podem ser utilizados para a identificação automática para as doenças investigadas. (Butt et al., 2013), utilizou vários algoritmos de *machine learning* como o *SVM*, *Naive Bayes*, Árvore de Decisão e *Boosting* para a classificação automática para casos de cancro como causa de morte a partir de textos de registos de óbitos. O autor verificou que os métodos baseados em *machine learning* puderam efetivamente identificar os casos de cancro, sendo que o *SVM* apresentou o melhor desempenho.

A aplicação de *text mining* para identificação de reações alérgicas a partir de textos livres de registos clínicos tem sido utilizada principalmente para reconhecimento de reações a medicamentos (Casillas, Gojenola, Perez, & Oronoz, 2016; Oronoz, Gojenola, Pérez, Ilarraza, & Casillas, 2015; Sarker & Gonzalez, 2015; Sohn, Kocher, Chute, & Savova, 2011; R. Xu & Wang, 2015a, 2015b), apesar de alguns trabalhos terem utilizado *text mining* para classificar automaticamente doenças crónicas como asma em doentes pediátricos ou até para a construção de sistemas que permitam identificar sintoma doenças ou alergias em diversas outras áreas para além da reação a medicamentos (Baranov et al., 2016)(Oronoz et al., 2015).

Segundo a revisão sistemática conduzida por (Ford, Carroll, Smith, Scott, & Cassell, 2016), que incluía estudos sobre extração de informação e deteção de condições clínicas específicas a partir de textos livres em registos de saúde eletrónicos. A maioria dos estudos selecionados na revisão sistemática foram conduzidos nos Estados Unidos, sendo que registos eletrónicos completos ou partes destes registos, como resumos de alta hospitalar e relatórios sobre patologias, foram utilizados como fonte de informação. Técnicas de *machine learning*, pesquisa por palavras-chave e algoritmos baseados em regras foram as técnicas utilizadas para extração de informação, onde a exatidão (ou precisão conforme a perspetiva) destes métodos foi no geral boa, no entanto com algum grau de variabilidade, com alguns algoritmos a apresentar valores de sensibilidade e especificidade acima de 90%. Não houve nenhum algoritmo específico cujo desempenho fosse, em geral, superior aos restantes.

Relativamente à avaliação dos algoritmos utilizados, foram identificadas duas formas distintas de apresentar resultados relativamente à avaliação dos algoritmos utilizados, nomeadamente o cálculo de medidas como a precisão, *recall* e *F-Measure* na área de informática, e, sensibilidade e especificidade na área da Medicina (Ford et al., 2016).

Em cerca de 43% dos estudos analisados por Ford et al., 2016, a informação de textos livres dos registos foi combinada com códigos de terminologias clínicas, resultados laboratoriais ou medicamentos a partir da utilização de algoritmos baseados em regras, modelos de regressão logística e técnicas de *machine learning* para a deteção de condições clínicas. Estes métodos apresentaram um bom desempenho, refletindo-se nos valores médios da sensibilidade e especificidade e *AUROC* (respetivamente 80%, 95% e 24%).

As técnicas de *machine learning* mais utilizadas foram *Support Vector Machine*, Árvores de Decisão, *Naive Bayes*, *Random Forest* e algoritmos baseados em Regras.

Em síntese, a revisão sistemática concluiu que textos livres de registos eletrónicos podem ser utilizados para a deteção de inúmeros e diversos tipos de condições clínicas, variando desde doenças infeciosas e eventos agudos, a condições psicológicas, com graus variados de sucesso. Conclui-se inclusivamente que será uma grande valia para os investigadores na área, que haja um consenso na apresentação de resultados relativos ao desempenho dos algoritmos, tais como apresentar sempre medidas de sensibilidade e precisão e tornar métodos de extração de informação mais compatíveis e comparáveis entre estudos.

2.6 Algoritmos de *Machine Learning*

Os algoritmos de *machine learning* podem ser utilizados para encontrar um modelo ou função que descreva diferentes classes de dados, o que nos permite classificar automaticamente novas instâncias de bases de dados com uma determinada classe aplicando assim o modelo ou a função que resultou da aprendizagem (Han e Kamber, 2011). Ao longo desta secção encontramos a explicação dos diferentes classificadores utilizados no decorrer deste projeto. Os algoritmos selecionados referem-se aos algoritmos mais comuns na literatura, conforme descrito na secção anterior, nomeadamente nos trabalhos relacionados com esta dissertação e que tenham aplicado técnicas de *machine learning* para processar textos

livres de registos de saúde com a finalidade de classificar os documentos em terminologias que descrevam doenças ou eventos clínicos.

2.6.1 Classificador *Naive Bayes*

O *Naive Bayes* é considerado um dos algoritmos mais simples e pode ser facilmente aplicável a grandes conjuntos de dados. Neste algoritmo, o modelo de classificação baseia-se num conjunto de probabilidades, probabilidades estas que são estimadas pela contagem da frequência de cada valor de caraterística (atributo) para as instâncias (exemplos) dos dados de treino (António Cardoso Martins, João Miguel Marques, 2009). O classificador estima qual a probabilidade de determinada instância pertencer a uma classe específica em função do produto das probabilidades condicionadas individuais para o valor de cada característica daquela instância. Na figura 3, podemos observar a classe e os atributos, em que a classe é descrita como a informação que se deseja aprender e as variáveis são os atributos. Estes atributos, tal como referido anteriormente, são variáveis auxiliares em que os valores podem ser determinados e podem auxiliar na previsão da classe (Costa, Carlos, & Filho, 2014).

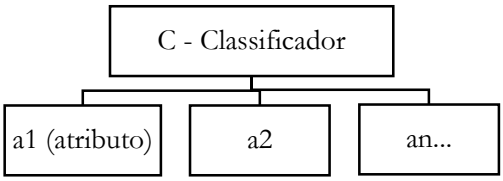


Figura 3: Classificador Naive Bayes

O teorema de Bayes é aplicado para o cálculo das probabilidades tendo em conta que todos os atributos são independentes (por isso a denominação "naive"), isto é, as características não interferem no valor umas das outras, o que não é válido para a maioria dos problemas reais. Apesar de sempre considerar a independência entre os atributos, este algoritmo apresenta normalmente um bom desempenho para a maioria das aplicações (Wu et al., 2008).

2.6.2 *Support Vector Machines* (*SVM*)

Este algoritmo adquire como entrada um determinado conjunto de instâncias, cujo os atributos e classificações são conhecidas a priori, e prediz para cada uma dessas instâncias, qual de duas classes possíveis essa instância faz parte. Tudo isto faz do SVM um classificador linear binário não probabilístico. O SVM encontra uma linha de separação entre as instâncias das duas categorias, denominado como o hiperplano. O hiperplano de separação ótimo é escolhido de modo a maximizar a distância de separação entre as classes (Ford et al., 2016)(Reis, Aires, Reis, Silva, & Lima, 2015), tal como podemos observar na figura 4:

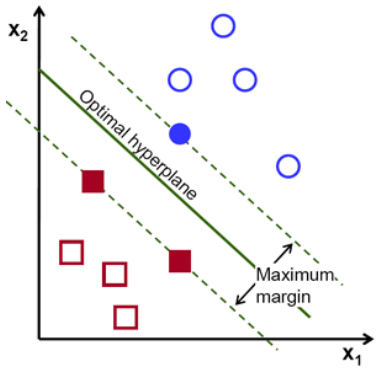


Figura 4: SVM - Adaptada de "Introduction to Support Vector Machines", (García, 2017)

2.6.3 J48 (árvore de decisão)

O algoritmo J48 é um classificador do tipo árvore de decisão. Uma árvore de decisão é um modelo de *machine learning* preditivo em formato de árvore, em que os nós internos da árvore correspondem aos diferentes atributos e os ramos entre os nós indicam quais são os valores possíveis que os atributos podem adquirir na amostra em observação. Os nós terminais indicam a classe. Existem vários algoritmos do tipo árvore de decisão, tal como o CART, ID3 e C4.5 (Dangare, 2015). O algoritmo J48, que consiste na versão C4.5 adaptada e implementada pelo *software Weka*, constrói uma arvore de decisão com base no valor dos atributos dos dados de treino (Afzal et al., 2013). Deste modo, a partir do conjunto de dados de treino, o algoritmo identifica o atributo e faz uma descriminação das várias instâncias. Este método permite indicar e dar a conhecer mais informações sobre as instâncias mais claramente, originando assim uma melhor classificação. No caso de entre os valores possíveis dessas caraterísticas, surgirem quaisquer instâncias que abranjam a sua categoria então esse ramo é finalizado e é atribuído para ele o valor alvo que foi obtido (Reis et al., 2015).

2.6.4 *Random Forest*

Este algoritmo resulta na predição de diversas árvores de decisão, pelo que o conjunto de árvores de decisão é denominado "floresta". Uma *random forest*, apresenta um número aleatório de árvores de decisão simples para determinar a classe final de uma instância. Em problemas de classificação, o algoritmo disponibilizará a classe que apresentar maior número de votos obtidos com as árvores de decisão da floresta. (Reis et al., 2015).

2.6.5 *NNge* (baseado em regras)

Introduzido em 1995 por Brent, o NNge (*Nearest Neighbor with generalisation*) é um algoritmo baseado em regras que constrói hiperectângulos no espaço de atributos que representam regras de decisão. Este método foi proposto como solução para melhorar o desempenho de classificação do algoritmo *k-NN*. Embora este algoritmo tenha ganhado popularidade nos últimos anos, nomeadamente por apresentar um custo reduzido em termos de tempo de execução e implementação, o *k-NN* apresenta algumas limitações importantes, como dificuldade em aplicar a função de distância ideal e o viés que pode ser introduzido por construir regras menores e simples e genéricas com base apenas na semelhança de exemplos (semelhança esta medida pela função de distância), o que pode reduzir o desempenho de classificação. A solução para estas limitações é a generalização implementada pelo algoritmo NNge, em que apenas exemplos com a mesma classe são agrupados a fim de produzir regras mais completas. Esta característica reduz o papel da função de distância enquanto aumenta a utilização de regras mais completas, melhorando o desempenho de classificação do sistema k-NN ao mesmo tempo que reduz o tempo de classificação. (Martim, 1995)

Este algoritmo classifica uma instância quando encontra um ou mais hiperectângulos associados a ela. Uma vez classificada a nova instância é generalizada através da junção com o exemplo mais próximo e que pertence à mesma classe. Exemplo este que pode ser uma instância ou hiperectângulo. (Devasena L., Sumathi T., Gomathi V., 2011) Em estudos anteriores, o NNge foi integrado num sistema para a classificação de texto livre de documentos clínicos para identificar casos de depressão (Zhou et al., 2015).

2.6.6 *K-Nearest Neighbours* (*k-NN*)

O k-vizinhos mais próximos (k-NN) é um dos algoritmos de *machine learning* mais bem-sucedidos em problemas de classificação, além de ser muito utilizado devido à sua simplicidade, o mesmo é capaz de obter uma exatidão considerável para a classificação de diversas instâncias. (Shouman, Turner, & Stocker, 2012)

Durante a fase treino, este algoritmo identifica e aprende os casos semelhantes com base numa função de distância a ser determinada pelo investigador. Desta forma, são construídas regras com base nos grupos de casos semelhantes. Este classificador contem um parâmetro K, também a ser determinado pelo investigador, que corresponde ao numero de vizinhos semelhantes a ser considerado para a classificação.

Os exemplos são mantidos na memória para posterior classificação de casos novos.(Shouman et al., 2012) Para k=1, o k-NN atribui a uma instância (caso) a classe do vizinho mais próximo. Para k>1, inicialmente são calculadas as distâncias do novo caso a ser classificado de modo a identificar os k exemplos mais próximos no conjunto de treino. Assim, o novo caso é classificado como pertencente à classe mais frequente entre os k exemplos mais próximos do conjunto de treino. (Medjahed, 2013)

3. Metodologia

Este projeto encontra-se dividido em três fases. A primeira consiste na recolha da informação e pré processamento do registo eletrónico clínico; a segunda compreende a fase da construção do modelo de classificação e a terceira na implementação do protótipo.

3.1 Primeira Fase – Recolha da Informação e Pré-Processamento

A recolha de registos clínicos não estruturados (estritamente anonimizados), foi uma tarefa de primeira instância, para que possibilitasse a aquisição dos conteúdos necessários, de forma a que posteriormente fossem aplicados os filtros de pré-processamento. Foram recolhidos 48 casos clínicos de alergologia, com o apoio da CUF Porto. Recolhidos os registos clínicos de alergologia, classificaram-se os registos por dois métodos, o manual e o automatizado, onde se recorreu às técnicas de *text mining*.

Classificação Manual – Etapa de grande interesse, permitiu que fosse criado um "dicionário", que recolhesse o maior número de informação possível para o processo de construção do modelo. Este método foi concretizado por um especialista em imunoalergologia. Cada registo foi classificado de acordo com o método utilizado no CPARA, com a terminologia utilizada na última versão, a *SNOMED CT*. A relação dos códigos *SNOMED CT* utilizados nos diferentes eixos de classificação especificados pelo catálogo do CPARA pode ser consultada no anexo B. Desta forma organizou-se em excel uma tabela com todos os campos necessários para a classificação, ocupando assim o seguinte formato que se pode ver na Figura 5:

Registo Clínico de Alergologia	Origem da Informação	Classificação	Alergénio Alimentar	Reação Adversa	Gravidade	Estado
Vem para PPO com camarão cozido. Refere edema do lábio e prurido oral e cutâneo da face <1h após ingestão de camarão (+ outros mariscos); previamente noção de algum edema labbial e prurido oral, sem outros sintomas. Sem outros sintomas acompanhantes. Tem testes cutâneos por picada com extrato comercial de camarão negativos. Sem infeção aguda recente. Sem anti-H1 há cerca de 1 semana. Sem medicação habitual.EO: AP N / RA HCI, mucosa pálida, algumas secreções / pele okExplicado procedimento e assinado consentimento informado. Administradas doses crescentes de camarão cozido, atingindo a dose cumulativa total de 106g, sem reações imediatas. Fica com indicações e contactos em caso de reação tardia.	408439002	414285001	278840001	41291007	24484000	73425007

Figura 5: Estrutura da folha excel com os diferentes eixos de classificação do CPARA - Modelo de Classificação Manual

Classificação Automática – Este método apenas pôde ser feito após recolha da informação anterior obtida pelo método de classificação manual. Foi realizado através de técnicas de *machine learning*, técnicas estas que permitem a análise e o tratamento da informação e consecutivamente a aprendizagem do programa com base nessa informação. As técnicas de *machine learning* recorrem a métodos probabilísticos para a escolha do melhor código de classificação *SNOMED CT* que pode estar associado ao registo em análise.

Visto que os registos manualmente classificados pelo especialista de imunoalergologia não se encontravam estruturados para a aplicação das técnicas de *machine learning*, verificou-se crucial aplicar filtros de pré-processamento de texto, recorrendo a técnicas de *text mining*. Estas técnicas permitiram extrair do registo clínico de alergologia não estruturado as informações de maior relevo e impacto, para a construção do modelo de classificação automatizado. Para esta fase de pré-processamento, foram utilizados três filtros: o *Tokenize*, o Filtro de *Stopwords* e o Filtro de Palavra-Chave.

A aplicação do algoritmo T*okenize* foi escolhida por ser essencial para explorar as palavras individualmente e por ser um processo indispensável para a produção de símbolos léxicos que podem ser manipulados pelo computador. Considerando que muitas das palavras presentes em documentos apresentam pouco ou nenhuma relevância para o processo de classificação, palavras estas denominadas "*stopwords*", foi aplicado o filtro *Stopwords* com a finalidade de reduzir e eliminar o conteúdo de menor relevância para o processo de classificação, aumentando a eficiência do classificador.

Por fim, o terceiro filtro, denominado filtro de Palavras-chave, foi criado separadamente através da elaboração de um dicionário cujas palavras (e os vários sinónimos possíveis) foram derivadas das definições dos códigos *SNOMED CT*. Também foram considerados, para a elaboração do dicionário de palavras-chave, alguns termos comuns para descrever os aspetos considerados para a classificação, nomeadamente termos que descrevem sintomas, reações adversas, episódios anafiláticos. No seu conjunto, estes atuaram filtrando a vasta informação contida num registo clínico de alergologia, deixando no final apenas as palavras chave de maior relevo para a classificação, eliminando toda a informação desnecessária.

3.1.1 Filtros de Pré-Processamento

Neste sub-capítulo são explicadas as definições de cada filtro empregue em cada etapa de pré-processamento.

Tokenize – este filtro, de fácil implementação, permitiu separar as palavras, identificando os termos um a um, separando por palavas ou carateres, tal como podemos observar na figura 6. Assim, permitiu que fosse feita uma análise termo a termo e não de um todo, obtendo cada palavra a sua própria identidade.

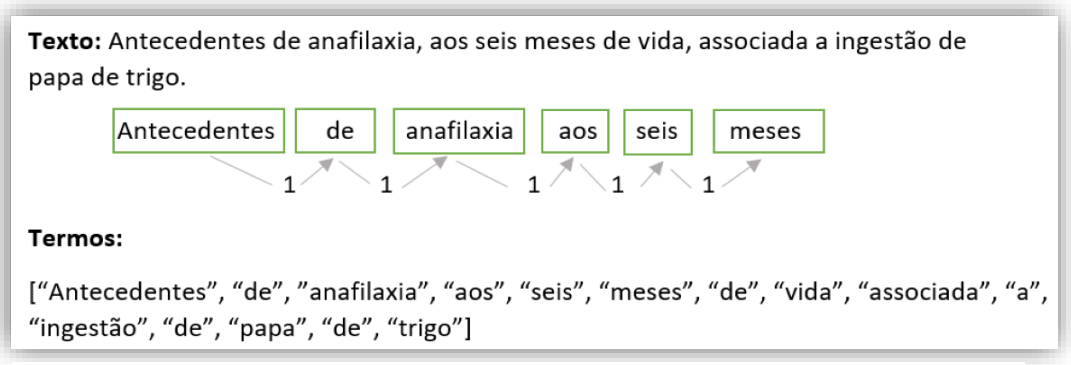


Figura 6: Exemplo da aplicação de Filtro Tokenize

Filtro *Stopwords* – Geralmente um texto contém sempre palavras de menor importância para a classificação, palavras estas que não determinam nem interferem na classificação, mas sim na construção das frases. Este filtro deteta e elimina estas palavras "*stop*", tais como artigos definidos e não definidos, as preposições e até mesmo conjugações. A figura 7 ilustra a aplicação do filtro para a remoção de *stopwords*.

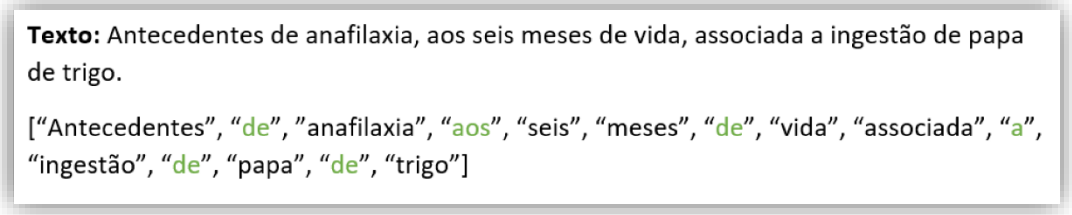


Figura 7: Exemplo da aplicação do Filtro de *Stopwords*

Neste projeto, foi utlizado o software *Rapidminer*, tal como referido anteriormente. Na implementação deste filtro, verificou-se um problema, a ausência desta lista de *stopwords*, no idioma Português. Posto isto, este obstáculo foi contornado recorrendo à pesquisa da lista das "*stopwords* Portuguesas". Identificadas estas palavras, criou-se um documento ***txt.*** com elas. Como este programa permite a adição de novas palavras/novas *stopwords* de diferentes

idiomas, foi inserido na biblioteca deste software, o ficheiro com as nossas palavras *stop*. Com isto tornou-se possível a sua identificação através do *Rapidminer*.

Filtro por Palavras-chave – Este filtro permite ensinar ao programa sinónimos e quando presentes num registo clínico diversas palavras-chave dentro do mesmo grupo, o programa sugere a classificação geral do termo. Na figura 8, podemos observar que dentro de cada classificador encontramos subgrupos que podiam ser identificados num grupo, como por exemplo:

grupo – classificação: marisco
subgrupo – classificação: camarão, lagosta, marisco, lagostim

Marisco 44027008	Lagosta 230031005
	Ostra/Ostras 230032003
	Camarão 278840001
	Moluscos/Delícias do mar/ Crustáceos 227146005

Figura 8: Exemplo de filtros por palavra-chave

O objetivo deste mapeamento é abranger os casos em que foram identificados mais do que um alergénio alimentar, pertencente a um mesmo grupo. No caso de existirem diferentes significados para uma sigla ou palavra, este filtro é muito relevante pois permite a identificação manual destas palavras, ultrapassando possíveis erros de sugestões erradas de classificação.

Na tabela 1, encontramos o resultado de um registo clínico após as técnicas de pré-processamento utilizadas no decorrer deste trabalho, aqui conseguimos obter maior perceção de como atuam os filtros. Com estas técnicas, a redução de um texto livre à sua essência é fundamental para a construção do modelo de classificação.

Tabela 1 - Exemplo de um Registo Clínico de Alergologia Antes e Após aplicar os Filtros de Pré-Processamento

Registo Clínico sem Filtros de Pré-Processamento de texto	**Registo Clínico após os Filtros de Pré-Processamento de texto**
Dermatite atópica. Asma alérgica. Asma de esforço: Rinite alérgica. Anafilaxia noz. Fez IT subcutânea Allergovac polimerizado 100% DT PT 5 anos (terminou em 2013). Este ano tem tido crises pontuais de asma e "constipações" sic. R/ Relvar e kestine lio SOS. Anapen. 13-09-2013 - peço Iges especificas frutos secos - sensação de dificuldade e dor em engolir (estava a jantar no restaurante) Sushi, antes tinha tido episódio semelhante com caju.	Dor Asma Rinite Anafilaxia Frutos Secos Caju Sushi

3.2 Segunda Fase – Construção do Modelo de Classificação

Nesta fase, recorreu-se a técnicas de *machine learning* com o método de aprendizagem supervisionada. Neste tipo de aprendizagem, o modelo é treinado com uma amostra de registos clínicos já classificados de modo a que possa aprender as características específicas mais importantes para a atribuição de um dado código *SNOMED CT*. Por outras palavras, considerando-se que a classe já era conhecida (neste caso a classe é referente a um código *SNOMED CT*), os algoritmos de *machine learning* podem reconhecer os termos mais relevantes e frequentes que estarão associados a um determinado código *SNOMED CT*.

Para a aplicação destas técnicas foi utilizado o *software WEKA* 3.8.0, que traz como vantagens ser um software *open source* gratuito, disponibilizar um grande número algoritmos de *data mining*, como também é um dos softwares de *data mining* mais utilizados nos últimos anos (Jovi, Brki, & Bogunovi, 2014). Apesar do *Weka* não ser tão utilizado quanto outros *softwares* de *data mining*, tal como o *Rapidminer* e o *R, o Weka* disponibiliza inúmeras *API's* em *Java*, o que permite incluir as suas funcionalidades em qualquer aplicação a ser implementada em *Java*. Desta forma é possível integrar os algoritmos de *data mining* implementados pelo *WEKA,* nomeadamente os algoritmos de classificação supervisionada, na ferramenta a que nos propusemos a desenvolver neste trabalho.

Entre os diferentes algoritmos de *machine learning* utilizados neste trabalho, temos o *Support Vector Machines* (*SVM*), as Árvores de Decisão (J48 e *Random Forest*), o *Naive Bayes*, classificadores baseados em regras (*Nnge*) e o *K-Nearest Neighbours* (*K-NN*). Os algoritmos foram aplicados separadamente de modo a identificar cada código *SNOMED CT* através de um processo de classificação binária ("sim", caso o registo clínico esteja associado a determinado código, "não", caso contrário). A técnica de validação cruzada foi utilizada para dividir toda a amostra dos registos clínicos disponíveis em subgrupos, em que, parte dos

subgrupos foram utilizados para o processo de treino e a outra parte para testar e avaliar o desempenho do modelo de classificação.

Durante o processo de construção do modelo de classificação, nomeadamente a etapa de treino, foi necessário a adaptação dos registos clínicos ao formato ARFF (*attribute – relation file format*), para a utilização do software *WEKA*. Esta adaptação exigiu que cada registo fosse convertido num vetor de números binários que especificava a presença ou ausência das palavras chave (definidas para o filtro de palavras-chave proposto na primeira etapa) de maior relevância para a classificação (0 quando a palavra estava ausente ou não era identificada no registo, e 1 quando a palavra aparecia no registo). Desta forma, cada linha contém um vetor binário que representa um episódio clínico distinto e o seu código *SNOMED CT* correspondente identificado pelos termos *yes* ou *no*. Assim sendo, para cada um dos 48 códigos *SNOMED CT* foi criado um ficheiro ARFF distinto em que os registos pertencentes a aquele código eram classificados como *yes,* enquanto que os restantes eram classificados como *no*.

Com a finalidade de analisar e obter uma maior perceção da fiabilidade induzida neste projeto de sugestão de classificação automatizada de códigos *SNOMED CT*, avaliamos o modelo de classificação através de medidas de desempenho fornecidas pelo *software WEKA*, tais como a área sob a curva ROC, a precisão, o *recall*, a taxa de verdadeiros positivos, a taxa de falsos negativos e *F-measure*. Entretanto, as principais medidas para avaliação que utilizamos para a escolha dos melhores modelos de classificação foram a sensibilidade (taxa de verdadeiros positivos) e especificidade (taxa de verdadeiros negativos), juntamente com a área sob a curva ROC.

A **sensibilidade** (*recall*, Taxa de Verdadeiros Positivos - TVP) mede a proporção/percentagem dos positivos reais que são identificados corretamente. Por exemplo: qual a percentagem de doentes que são diagnosticados com uma doença, tendo mesmo a doença.

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

A **especificidade** (Taxa de Verdadeiros Negativos - TVN) mede qual a proporção/percentagem de negativos, que foram identificados como tal. Por exemplo: mede qual a percentagem de pessoas saudáveis que não têm mesmo a doença, que não verificam a condição.

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

A relação entre a sensibilidade e especificidade pode ser representada pela curva ROC, por outras palavras a curva ROC representa a taxa de verdadeiros positivos em função da taxa de falsos positivos. Esta curva é uma representação gráfica que permite ilustrar a performance de um classificador binário, como o seu limiar de discriminação é variado. Note-se que a área sob a curva ROC, varia entre 0 e 1. O 1, representa o classificador ideal, e o 0, o classificador errado. Quando a área é de 0,5, indica que o classificador é praticamente aleatório.

Para cada código presente nos registos clínicos do conjunto de treino, foram selecionados dois modelos de classificação: o modelo de mais alta sensibilidade e o modelo de mais alta especificidade. O primeiro é aplicado para sugerir códigos *SNOMED CT,* dentro de cada um dos sete eixos de classificação do CPARA (que consideramos)", que possam estar associados ao caso clínico em avaliação. A importância de um classificador altamente sensível foi importante no contexto do problema em questão, pois é de extremo interesse que os casos reais positivos a determinado alergénio sejam prontamente e devidamente sugeridos, visto que a sugestão será confirmada *à posteriori* pelo especialista de imunoalergologia. O segundo modelo é utilizado para confirmar ou validar os códigos sugeridos pelo primeiro modelo, tarefa esta que também é útil no processo de confirmação a ser feito pelo especialista de imunoalergologia. Um modelo de alta especificidade é o ideal para a tarefa de confirmação tendo em vista a sua grande capacidade em descartar falsos positivos ("falsos alarmes").

3.2.1 Ferramentas Utilizadas

- ***Weka***

Waikato Environment for Knowledge Analysis (*WEKA*) surgiu em 1993, na Universidade de Waikato, na Nova Zelândia. Desenvolvido em linguagem java, este *software* foi posteriormente adquirido por uma empresa, no final de 2006. Possui uma licença de código aberto, *GNU - General Public License*, que permite a aquisição do código fonte e inclusivamente o estudo dele (Witten et al., 2009). O *Weka* permite a organização de dados através da identificação de alterações e ou padrões que apresentem determinada anomalia relevante. Envolve a área de inteligência artificial, que se dedica ao estudo de técnicas de *machine learning*. Desta forma, permite assim que um computador possa "aprender" intuitivamente ou dedutivamente, através de técnicas de *text mining* e *machine learning*.

De uma forma matemática, este *software* procede a uma análise computacional e estatística da informação fornecida através das técnicas de *text mining*, tendo como principal objetivo decifrar a partir dos padrões encontrados, gerar hipóteses e teorias mais assertivas relativamente aos dados em questão (Schiefelbein, Moiano, & Livinalli, 2015).

- ***Rapidminer***

O *Rapidminer,* é considerada uma das melhores plataformas de análise de dados, *open source*, que contém um vasto conjunto de ferramentas para *text mining*. É um *software* que permite realizar análises avançadas de textos, com pouca ou até nenhuma codificação necessária (Ramamohan, Vasantharao, Chakravarti, & Ratnam, 2012).

Este *software* de implementação em Java permite a criação de modelos de análise preditiva. Trás com ele inúmeras vantagens: não só é um *software* de código aberto, como também possibilita obter a licença de negócio para a o desenvolvimento de aplicações. Existe a versão *free* e a versão paga, sendo que esta última é uma versão para profissionais. Atualmente os utilizadores e programadores desta ferramenta tem vindo a aumentar, apresenta, pois, uma interface intuitiva, possibilita a leitura de diversificados formatos de arquivos, codificação simples e de fácil utilização. O Rapidminer apresenta cerca de 250 algoritmos diversificados para *text mining* e inclusive inúmeros de pré-processamento (Burget, Karasek, Smekal, Uher, & Dostal, 2010).

Esta tecnologia foi escolhida para utilizar no projeto de dissertação, baseada em alguns requisitos mínimos (Rapid-I GmbH, n.d.) :

- *Software* para classificação
- Versão gratuita
- Bom desempenho
- Permita fácil associação
- Intuitivo
- Existência de algoritmos de pré-processamento
- Acessibilidade e grande flexibilidade no suporte ao utilizador

3.3 Terceira Fase – Protótipo

Para implementar na prática os modelos de classificação e disponibilizar a ferramenta para a utilização, foi criado um protótipo de serviço *web* para a classificação de registos de alergologia. Com este serviço pretendeu-se criar uma ferramenta para auxiliar o processo de classificação de alergologia em Portugal que até à data ocorrem exclusivamente de forma manual. Este serviço *web* foi desenvolvido em parceria com o CINTESIS da Faculdade de Medicina do Porto.

3.3.1 Tecnologias Utilizadas no Protótipo

Para a implementação deste protótipo, utilizou-se a linguagem Java, em conjunto com o *Java EE* (*Enterprise Edition*), que consiste num conjunto de especificações destinadas à construção de aplicações *web*. A linguagem *Java* foi escolhida por ser das linguagens mais utilizadas para a construção de aplicações *web*, pelo caráter *open source* e pela portabilidade, isto é, a aplicação pode ser executada em qualquer *Browser*. A tecnologia *Java Servlet* também foi utilizada para a implementação das funcionalidades de um servidor, nomeadamente o processamento e as respostas das requisições *online* feitas pelo utilizador. A interface gráfica foi construída recorrendo a linguagem *HTML* (*Hyper Text Markup Language), CSS* (*Cascading Style Sheets*) e *Java Script*. O ambiente de desenvolvimento utilizado para a codificação do protótipo foi *Eclipse IDE* com suporte ao desenvolvimento *Java EE*. A aplicação é executada através do servidor *apache tomcat 6.0,* que faz a simulação do servidor *web* real.

3.3.2 Funcionalidades e Interface

Nesta sub-capítulo encontra-se a especificação da interface gráfica do protótipo e as funcionalidades inerentes à aplicação. Na figura 9, podemos encontrar a interface inicial do serviço *web*. Esta interface foi criada de modo a ser de *user-friendly*, intuitiva e simples. A estrutura da página e a disposição das informações foram pensadas de modo a permitir uma maior usabilidade, sobretudo porque o objetivo desta ferramenta, consiste em facilitar e apoiar uma tarefa que deverá ser executada sistematicamente. As opções de navegação da plataforma podem ser visualizadas no menu superior da página inicial. Esta barra de navegação contém o acesso a informações do sistema de classificação do CPARA incluindo, a ligação direta com a página do CPARA e da SPAIC, além de instruções do uso da ferramenta. Para iniciar a classificação dos registos clínicos através da ferramenta, o utilizador pode selecionar a opção "Iniciar Classificação" ou aceder diretamente na barra de navegação através da opção "Classificador".

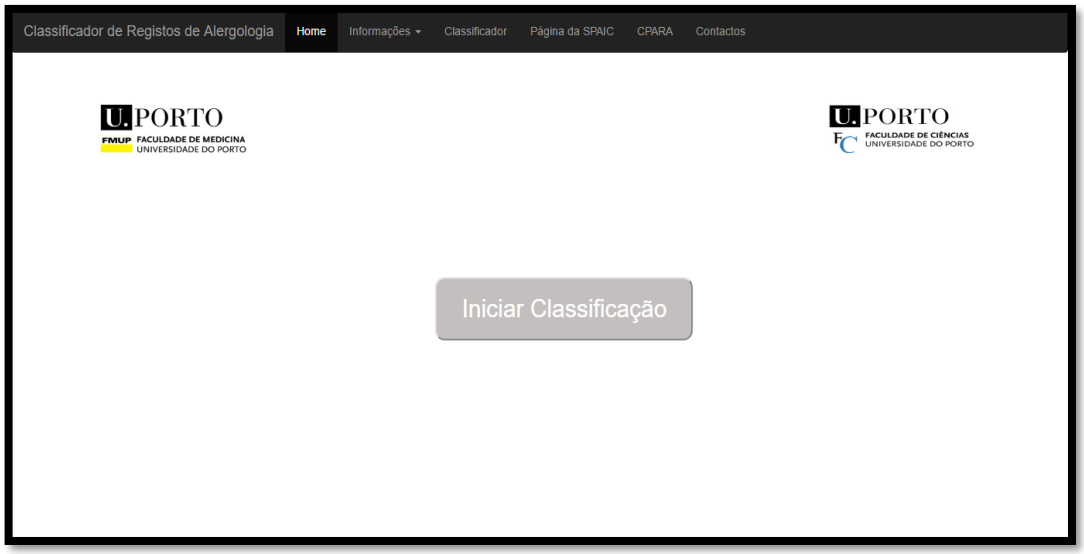


Figura 9: Protótipo - Classificador Automático de Registos Clínicos de Alergologia, *SNOMED CT*

As figuras 10 e 11 mostram, respetivamente, a interface da classificação para a entrada de múltiplos registos clínicos no formato excel (.xlsx) e a entrada de registos individuais através da introdução de um texto livre. Conforme podemos observar na figura 11, o utilizador pode submeter um conjunto de registos clínicos ao fazer o *upload* do ficheiro em formato excel cujo modelo pode ser consultado ao premir a opção "Modelo de ficheiro para submissão". Ao clicar na opção "Enviar ficheiro Excel" será aberta uma janela com o campo para o *upload* e submissão do ficheiro, conforme mostra a figura 11.

Figura 10: Interface do Classificador com opção de submissão de múltiplos registos (ficheiro excel)

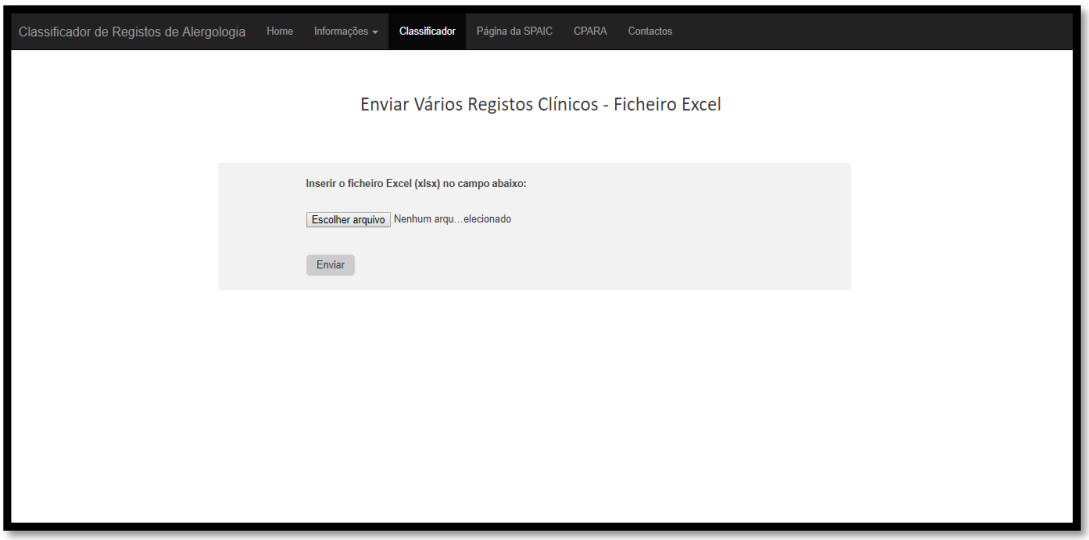


Figura 11: Interface do Classificador para entrada de múltiplos registos (ficheiro xlsx)

Alternativamente, o utilizador pode submeter um registo clínico individual, de modo a obter uma sugestão de classificação imediata, conforme mostra a figura 12. Ao clicar na opção "Enviar Registo", é aberta uma nova janela com um campo caixa de texto para inserir diretamente o registo, como podemos observar na figura 13. A resposta com a sugestão de códigos *SNOMED CT* é posteriormente retornada ao utilizador em tempo real.

Figura 12: Opção para a Submissão dos Registos Clínicos individuais (formulário)

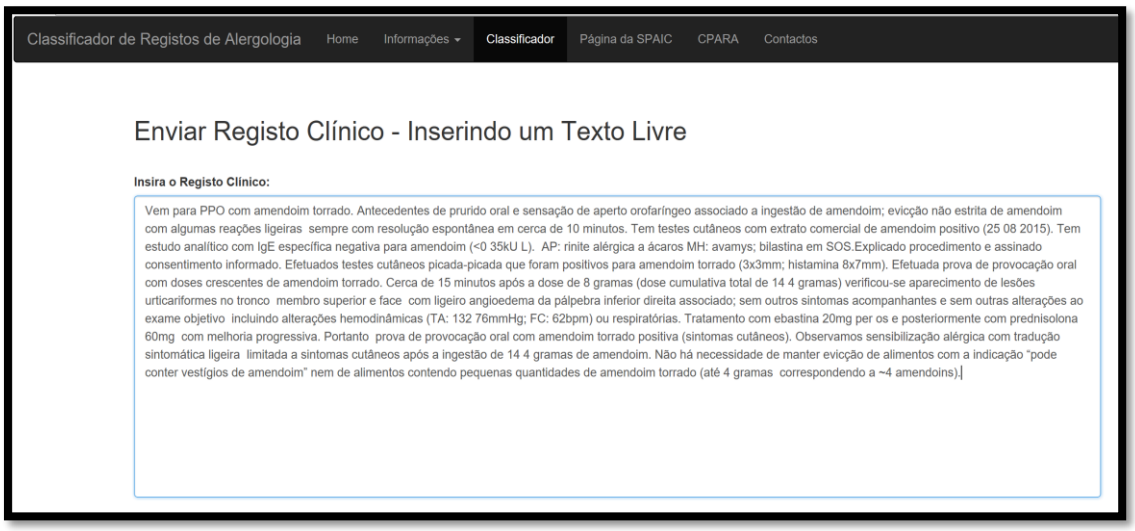
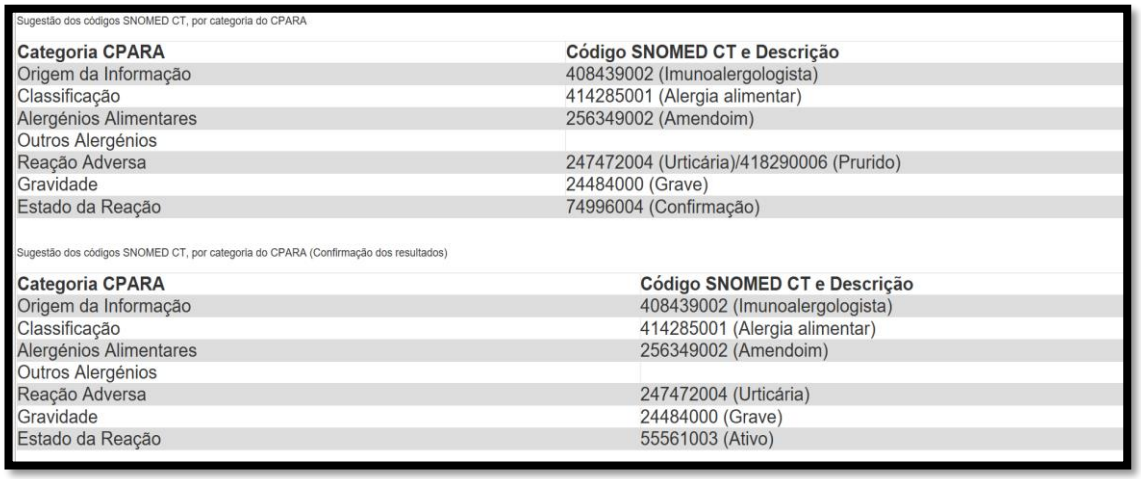


Figura 13: Caixa de Texto para inserir texto livre para a Submissão do Registo Clínico

Após a submissão do registo clínico individual, a ferramenta *web* deve retornar uma página que apresenta, para cada domínio proposto pelo CPARA, uma sugestão de código *SNOMED CT*, conforme ilustrado pela figura 14.



Sugestão dos códigos SNOMED CT, por categoria do CPARA

Categoria CPARA	Código SNOMED CT e Descrição
Origem da Informação	408439002 (Imunoalergologista)
Classificação	414285001 (Alergia alimentar)
Alergénios Alimentares	256349002 (Amendoim)
Outros Alergénios	
Reação Adversa	247472004 (Urticária)/418290006 (Prurido)
Gravidade	24484000 (Grave)
Estado da Reação	74996004 (Confirmação)

Sugestão dos códigos SNOMED CT, por categoria do CPARA (Confirmação dos resultados)

Categoria CPARA	Código SNOMED CT e Descrição
Origem da Informação	408439002 (Imunoalergologista)
Classificação	414285001 (Alergia alimentar)
Alergénios Alimentares	256349002 (Amendoim)
Outros Alergénios	
Reação Adversa	247472004 (Urticária)
Gravidade	24484000 (Grave)
Estado da Reação	55561003 (Ativo)

Figura 14 – Resultado da Classificação Automática - Sugestão de Códigos SNOMED CT

Para o caso de o utilizador submeter vários registos para a classificação, o resultado com a sugestão dos códigos *SNOMED CT* será enviado através de um ficheiro excel (.xlsx), após o término do processamento do pedido que deverá retornar uma página *web* com o *link* para aceder ao ficheiro com as respostas, conforme mostra a figura 15. O envio do resultado com a resposta através do ficheiro excel (.xlsx) é conveniente pois permite ao utilizador editar e modificar as sugestões com facilidade.

Figura 15: Interface após a finalização do processamento do pedido da Classificação

O ficheiro com o resultado contém um identificador por registo clínico que coincide com o número da linha em que o registo aparece no ficheiro de submissão. Para além disso o ficheiro com as respostas inclui em cada linha o próprio registo clínico com as sugestões dos códigos *SNOMED CT* juntamente com as definições correspondentes, por domínio de classificação do CPARA conforme mostra a figura 16.

ID Registo Clínico (Nº Linha)	Registo Clínico	Código SNOMED-CT (Origem)	Código SNOMED-CT (Classificação)	Código SNOMED-CT (Alimentação)	Código SNOMED-CT (Outros Alergénios)	Código SNOMED-CT (Reação Adversa)	Código SNOMED-CT (Gravidade)	Código SNOMED-CT (Estado)
2	Edema da face (palpebras) eritema local com i	408439002 (Imunoalergologista)	419199007 (Reação alérgica)	278840001 (Camarão)		41291007 (Angioedema)	24484000 (Grave)	55561003 (Ativo)
3	Vem para PPO com amoxicilina+ácido clavulâni	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)			41291007 (Angioedema)	24484000 (Grave)	
4	2016 07 14 Vem para PPO com farinha de trigo.	408439002 (Imunoalergologista)	414285001 (Alergia alimentar)	412071004 (Trigo)		39579001 (Anafilaxia)	24484000 (Grave)	
5	Vem para PPO com camarão cozido Refere ede	408439002 (Imunoalergologista)	414285001 (Alergia alimentar)	278840001 (Camarão)		41291007 (Angioedema)	24484000 (Grave)	
6	2016 07 04 Vem para PPO com alabote (peixe)	408439002 (Imunoalergologista)	414285001 (Alergia alimentar)	227037002 (Peixe),278840001 (Camarão)			24484000 (Grave)	
7	Vem para PPO com amoxicilina+ácido clavulâni	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)			73442001 (Síndrome de S	24484000 (Grave)	
8	2016 06 29 Vem hoje fazer início de Metibasol	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)			247472004 (Urticária)	255604002 (Ligeiro)	
9	Contacto telefónico a 27 06 2016 - teve dejeçã	408439002 (Imunoalergologista)	235719002 (Intolerância alimentar)	278840001 (Camarão)		62315008 (Darreia)	255604002 (Ligeiro)	
10	Pai médico Aos 14 anos - com augmentim eden	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)				24484000 (Grave)	38434005 (Não confirmação)
11	Doente de 5 anos e 1 mÊs.Enviada pelo Dra. En	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)				255604002 (Ligeiro)	38434005 (Não confirmação)
12	Vem para provocação oral com Yoco framboes	408439002 (Imunoalergologista)	414285001 (Alergia alimentar)	226760005 (Lacticínios)		247472004 (Urticária)	255604002 (Ligeiro)	
13	2016 06 09 Vem para provocação oral com pre	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)			247472004 (Urticária)	255604002 (Ligeiro)	
14	2016 06 08 Vem para PPO com amoxicilina+ácic	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)			247472004 (Urticária)	24484000 (Grave)	
15	2016 08 26 Vem para PPO com aspirina por sus	408439002 (Imunoalergologista)	416098002 (Alergia medicamentosa)			39579001 (Anafilaxia)	24484000 (Grave)	55561003 (Ativo)
16	Apicultor amador desde sábado. Picado no don	408439002 (Imunoalergologista)	419199007 (Reação alérgica)			39579001 (Anafilaxia)	24484000 (Grave)	55561003 (Ativo)

Figura 16 – Ficheiro Excel Descarregado com o Resultado da Sugestão de Classificação pelo serviço *web*

4. Resultados e Discussão

Neste capítulo, são dados a conhecer os resultados deste trabalho, incluindo os dados obtidos com a aplicação e avaliação do classificador criado no decorrer deste projeto, demonstrando numericamente quais os valores concretos obtidos que permitiram avaliar a viabilidade desta ferramenta computacional. Foram desenvolvidos classificadores para cada uma das categorias de classificação do CPARA (origem da informação, classificação, alergénios alimentares, reação adversa, estado e gravidade), à exceção da categoria "outros alergénios", em que não foram obtidos registos clínicos classificados neste domínio. Para o domínio de classificação "origem da informação", todos os 48 registos clínicos classificados apresentavam o código 408439002 – Imunoalergologista/Médico especialista em alergias, tal deve-se ao facto de os dados recolhidos terem todos origem na CUF Porto. Assim sendo, como não existia variabilidade de códigos para o domínio "origem de informação" o classificador obteve naturalmente um desempenho ótimo.

A tabela 2 apresenta a distribuição do número de registos clínicos entre os diferentes domínios de classificação. Apenas as categorias "origem da informação" e "classificação" estavam presentes em todos os registos clínicos, enquanto que a categoria "alergénios alimentares" estava presente em apenas 15 registos de alergologia. É importante salientar que para alguns domínios de classificação, houve a sugestão de mais do que um código *SNOMED CT* possível, sendo que a possibilidade de classificações múltiplas foi considerada durante a etapa de construção dos modelos de classificação. Na coluna "Número de Casos com mais do que uma sugestão" da tabela 2, encontram-se quantos registos de alergologia (da base de dados fornecida pela CUF Porto) apresentavam por domínio mais do que um código. A título exemplificativo, o mesmo caso clínico pode ser classificado no domínio "Estado" como 55561003 - "ativo" quando ocorre a confirmação de alergia ou reação adversa, e como 74996004 - "Confirmado" quando o diagnóstico foi confirmado e validado pelo imunoalergologista. Neste caso, a atribuição de ambos os códigos é considerada correta.

Tabela 2 - Número de Casos por eixo de Classificação do CPARA e número de casos com mais do que um código por tipo de classificação

Tipos de Classificação	Número de Registos Clínicos existentes por Classe	Número de Casos com mais do que uma sugestão
Origem da Informação	48	0
Classificação	48	1
Alergénios Alimentares	15	0
Outros Alergénios	0	0
Reação Adversa	28	10
Gravidade	34	0
Estado da Reação	34	16

A tabela 3 especifica o tamanho do conjunto de dados de treino utilizado durante o processo de construção do modelo de classificação. Em *machine learning,* cada elemento do conjunto é denominado "instância" e cada instância deve ser classificada com apenas um código *SNOMED CT* por domínio. Desta forma, pode existir mais do que uma instância para representar um mesmo registo clínico de alergologia, por exemplo registos com mais do que um código *SNOMED CT* no mesmo domínio poderá ser representado em mais do que uma instância. Por esta razão o número de instâncias presente no conjunto de treino tende a ser maior do que o número de registos clínicos originais.

Tabela 3 - Número de Instâncias utilizadas na Fase de Treino

Tipos de Classificação	Número de Instâncias
Origem da Informação	73
Classificação	73
Alergénios Alimentares	25
Outros Alergénios	0
Reação Adversa	55
Gravidade	59
Estado da Reação	59

Os resultados que serão apresentados a seguir foram obtidos através do *software WEKA*. Este permitiu obter medidas estatísticas referentes ao desempenho dos classificadores após a fase de treino, tais como a sensibilidade, especificidade e AUROC. Para calcular estas estatísticas, utilizou-se a técnica de validação cruzada para n=2 tendo em conta o número reduzido de registos clínicos disponíveis, além da vantagem de que esta técnica permite obter um conjunto de teste maior.

Inicialmente, foi utilizada a técnica de classificação multinomial, que consiste na solução de problemas que envolvem a classificação de uma instância dentro de várias classes

possíveis. Por outras palavras, para cada um dos sete domínios de classificação do CPARA, foi construído um classificador para sugerir um código *SNOMED CT* dentro de todos os outros códigos disponíveis para aquele domínio. Na tabela 4 podemos observar os resultados obtidos com a classificação multinomial, nomeadamente aqueles obtidos pelo melhor classificador dentro de cada domínio. Neste caso, excluíram-se os domínios "origem da informação" e "outros alergénios", sendo que o primeiro não apresentava variedade de classes (apenas um código *SNOMED CT* para a "origem de informação" estava presente na amostra dos registos clínicos) enquanto que o segundo domínio não tinha códigos atribuídos na amostra de registos clínicos.

O modelo para áreas maior do que 0,7 (Classificação e Alergénios Alimentares) apresentou ter uma boa capacidade discriminatória, com taxas de acerto relativamente altas (variando de 79% a 89%). Para áreas menores do que 0,7 (Reação adversa, Gravidade e Estado) o modelo apresentou uma baixa capacidade discriminatória e taxas de acerto, que variavam entre os 35% e 52%. Como o número do conjunto de registos disponíveis era relativamente baixo, o número de classes possíveis (variedade de códigos) também era baixo. No entanto, futuramente, com o aumento do tamanho da amostra de registos clínicos para a classificação, maior será o número de classes disponíveis, o que implicará uma dificuldade acrescida para a obtenção de bons modelos de classificação com a abordagem multinominal, dado que existirão mais classes disponíveis para a aprendizagem.

Tabela 4 - Resultados de Avaliação Multinomial

Domínio de Classificação	Classificadores	Área sob a Curva ROC	Percentagem de Acertos
Classificação	NNge (rule based)	0,780	79,5%
Alergénios Alimentares	Naive Bayes	0,969	88,5%
Reação Adversa	Random Forest (Decision Tree)	0,656	36,4%
Gravidade	NNge (rule based)	0,504	52,5%
Estado	Random Forest (Decision Tree)	0,499	35,6%

Entretanto, o contexto do problema em questão assemelha-se a problemas comuns das práticas médicas, visto que há interesse em determinar se um paciente apresenta ou não uma dada condição clínica, pelo que seria de maior interesse obter os melhores classificadores para identificar uma única condição específica. Visto que problemas associados a testes médicos são tipicamente solucionados com a construção de classificadores binários, optou-se por adotar o método de classificação binária de modo a gerar classificadores para determinar se um dado registo clínico de alergologia pertence ou não a determinado código *SNOMED CT*. Desta forma, foi construído um modelo de classificação distinto para cada

código *SNOMED CT* a partir da utilização de seis algoritmos de *machine learning,* que de acordo com a literatura foram mais utilizados em problemas de *text mining* para a classificação de registos clínicos.

Para cada código foram então selecionados os dois algoritmos com melhor sensibilidade e especificidade: o primeiro, selecionado com o objetivo de sugerir um código e o segundo com o objetivo de confirmar ou não a sugestão dada pelo primeiro algoritmo. A área sob a curva ROC (AUROC) também foi utilizada para avaliar o algoritmo, de modo a também considerar a capacidade discriminatória do modelo. Apenas foram considerados os códigos *SNOMED CT* que apareceram nos registos clínicos disponibilizados para este trabalho. Desta forma, conforme forem obtidos novos registos clínicos, os modelos de classificação deverão ser adaptados de modo a reconhecer um maior número de códigos *SNOMED CT*.

As tabelas de 5 a 9 apresentam os resultados obtidos pelos melhores algoritmos e que por sua vez foram selecionados para construir a versão inicial da ferramenta informática. Em cada tabela iremos encontrar as medidas de sensibilidade e especificidade obtidas pelos melhores algoritmos. A título exemplificativo, para o código 235719002 (intolerância alimentar), os algoritmos de maior sensibilidade foram *K-NN*, *Naive Bayes*, *NNge*, *Random Forest*, *SVM*. Todos estes algoritmos apresentaram uma sensibilidade de 1, o que significa que todos os casos positivos de intolerância alimentar foram devidamente identificados. No entanto, é importante ressaltar que existiam apenas dois casos classificados com este código, o que de certa forma explica a sensibilidade máxima observada nestes algoritmos. Os mesmos algoritmos também apresentaram os melhores resultados relativamente à especificidade, também com um valor igual a 1, o que indica que todos os casos negativos de intolerância alimentar foram devidamente identificados.

Os resultados obtidos relativamente ao domínio "Classificação" podem ser consultados na tabela 5. Neste domínio, o classificador apresentou baixa capacidade discriminatória e foi incapaz de classificar devidamente os casos positivos de reação alérgica (AUROC= 0,536; sensibilidade= 0), enquanto apresentou capacidade discriminatória relativamente boa e um desempenho médio (AUROC= 0,786; sensibilidade =0,652) para a identificação de casos positivos de alergias alimentares.

Para além da obtenção de novos registos clínicos, a melhoria destes resultados requer um conjunto maior e com maior variedade de palavras chave que permitam descrever e identificar melhor os casos de reação alérgica e alergias alimentares e consequentemente aumentar a sensibilidade do modelo de classificação associado a estes códigos. Assim, é fulcral aperfeiçoar a etapa de pré-processamento em conjunto com um especialista de alergologia, nomeadamente para atualizar as definições para o filtro de palavras-chave.

Tabela 5 - Resultados do *WEKA* referentes à coluna "Classificação" do CPARA, por código *SNOMED CT*

Classificação					
Código *SNOMED CT*	**Definição**	**Número de Instâncias por Código**	**AUROC**	**Taxa de Verdadeiros Positivos VP/(VP+FN)**	**Algoritmos de Maior Sensibilidade**
235719002	Intolerância Alimentar	2	1	1 [2/(2+0)]	K-NN, Naive Bayes, NNge, Random Forest, SVM
414285001	Alergia Alimentar	15	0,786	0,652 [15/(15+8)]	SVM
416098002	Alergia Medicamentosa	42	0,806	0,933 [42/(42+3)]	K-NN
419199007	Reação Alérgica	0	0,536	0 [0/(0+3)]	Random Forest
Código *SNOMED CT*	**Definição**		**AUROC**	**Taxa de Verdadeiros Negativos VN/(VN+FP)**	**Algoritmos de Maior Especificidade**
235719002	Intolerância Alimentar		1	1 [71/(71+2)]	K-NN, Naive Bayes, NNge, Random Forest, SVM
414285001	Alergia Alimentar		0,753	0,940 [47/(47+3)]	K-NN
416098002	Alergia Medicamentosa		0,826	0,786 [22/(22+6)]	NNge
419199007	Reação Alérgica		0,500	1 [70/(70+0)]	SVM

Os resultados para o domínio "Alergénios Alimentares" são mostrados na tabela 6. Podemos observar que a maior parte dos classificadores apresentaram alta sensibilidade e especificidade, à exceção dos códigos que classificam a alergia à proteína do leite de vaca e a alergia a avelã. Para estes alergénios, o modelo de classificação apresentou ter uma baixa capacidade discriminatória relativamente ao problema da identificação da alergia a proteína do leite de vaca, contudo revelou ter uma alta capacidade discriminatória quando considerada a alergia a avelã (proteína do leite de vaca: AUROC= 0,598 e sensibilidade=0; avelã: AUROC= 0,870 e sensibilidade=0). Apesar disso, não foi capaz de identificar devidamente nenhum caso positivo de alergia a estes alimentos. É importante salientar que este cenário pode ser explicado pelo tamanho reduzido da amostra de registos clínicos, visto que nos registos disponíveis para este trabalho, existia apenas um caso classificado como alergia a avelã e dois casos classificados como alergia a proteína do leite de vaca, o que tende a induzir o classificador a obter uma melhor performance na identificação de verdadeiros negativos, isto é, todos os casos que não são classificados como alergia à avelã ou alergia à proteína do leite de vaca.

Tabela 6 - Resultados do *WEKA* referentes à coluna "Alergénios Alimentares" do CPARA, por código *SNOMED CT*

Alergénios Alimentares					
Código *SNOMED CT*	**Definição**	**Número de Instâncias por Código**	**AUROC**	**Taxa de Verdadeiros Positivos VP/(VP+FN)**	**Algoritmos de Maior Sensibilidade**
303300008	Proteína do Ovo	3	1	1 [3/(3+0)]	K-NN, SVM, Random Forest, Naive Bayes
227037002	Peixe	2	1	1 [2/(2+0)]	K-NN, Naive Bayes, NNge, SVM, Random Forest
256349002	Amendoim	4	1	1 [4/(4+0)]	J48, Naive Bayes, SVM
412071004	Trigo	2	1	1 [2/(2+0)]	K-NN, Naive Bayes, NNge, Random Forest
256353000	Avelã	0	0,870	0 [0/(0+1)]	Naive Bayes
226760005	Lacticínios	2	1	1 [2/(2+0)]	Naive Bayes, NNge, Random Forest
264295007	Proteína do Leite de Vaca	0	0,598	0 [0/(0+2)]	Random Forest
Código *SNOMED CT*	**Definição**		**AUROC**	**Taxa de Verdadeiros Negativos VN/(VN+FP)**	**Algoritmos de Maior Especificidade**
303300008	Proteína do Ovo		1	1 [22/(22+0)]	K-NN, Naive Bayes, Random Forest, SVM
227037002	Peixe		1	1 [23/(23+0)]	K-NN, Naive Bayes, SVM, Random Forest, NNge
256349002	Amendoim		1	1 [21/(21+0)]	J48, Naive Bayes, SVM
412071004	Trigo		1	1 [23/(23+0)]	K-NN, Naive Bayes, NNge, Random Forest
256353000	Avelã		0,870	1 [23/(23+0)]	SVM
226760005	Lacticínios		1	1 [23/(23+0)]	NNge, Random Forest
264295007	Proteína do Leite de Vaca		0,598	1 [23/(23+0)]	Random Forest

A tabela 7 apresenta os resultados obtidos para o domínio "Reação Adversa". A avaliação do desempenho dos modelos de classificação para este domínio foi caraterizada por resultados heterogéneos em que foi observada uma sensibilidade ótima para os códigos das classes "asma", "*Síndrome de Stevens- Johnson*" e "outras", enquanto que não foram identificados os casos positivos para as classes "dispneia", "broncospasmo" e "diarreia".

Ainda o classificador apresentou um fraco desempenho para as classes "angiodema" (AUROC=0,556 e sensibilidade=0,200), "prurido" (AUROC=0,568 e sensibilidade=0,429) e "urticária" (AUROC= 0,512 e sensibilidade=0,471), no entanto o resultado foi melhor para a identificação dos casos de anafilaxia (AUROC=0,880 e sensibilidade=0,800). Assim como para os outros domínios de classificação, o tamanho reduzido da amostra dos registos clínicos teve impacto na sensibilidade dos modelos de classificação.

Tabela 7 - Resultados do *WEKA* referentes à coluna "Reação Adversa" do CPARA, por código *SNOMED CT*

Reação Adversa					
Código *SNOMED CT*	**Definição**	**Número de Instâncias por Código**	**AUROC**	**Taxa de Verdadeiros Positivos (VP/VP+FN)**	**Algoritmos de Maior Sensibilidade**
195967001	Asma	4	1	1 [4/(4+0)]	K-NN
247472004	Urticária	8	0,512	0,471 [8/(8+9)]	NNge
267036007	Dispneia	0	0,741	0 [0/(0+1)]	Naive Bayes
39579001	Anafilaxia	4	0,880	0,800 [4/(4+1)]	K-NN, NNge
41291007	Angiodema	2	0,556	0,200 [2/(2+8)]	K-NN
418290006	Prurido	3	0,568	0,429 [3/(3+4)]	KNN
4386001	Broncospasmo	0	0,750	0 [0/(0+1)]	Naive Bayes
62315008	Diarreia	0	0,981	0 [0/(0+2)]	Random Forest
73442001	Síndrome de Stevens-Johnson	2	0,849	1 [2/(2+0)]	K-NN
74964007	Outra	2	0,981	1 [2/(2+0)]	Naive Bayes
Código *SNOMED CT*	**Definição**		**AUROC**	**Taxa de Verdadeiros Negativos (VN/VN+FP)**	**Algoritmos de Maior Especificidade**
195967001	Asma		1	1 [51/(51+0)]	K-NN, Random Forest
247472004	Urticária		0,638	0,842 [32/(32+6)]	Random Forest
267036007	Dispneia		0,741	1 [54/(54+0)]	Naive Bayes
39579001	Anafilaxia		0,830	0,980 [49/(49+1)]	Naive Bayes
41291007	Angiodema		0,584	1 [45/(45+0)]	Naive Bayes
418290006	Prurido		0,521	0,958 [46/(46+2)]	J48
4386001	Broncospasmo		0,750	1 [54/(54+0)]	Naive Bayes
62315008	Diarreia		0,981	1 [53/(53+0)]	Random Forest
73442001	Síndrome de Stevens- Johnson		0,840	1 [53/(53+0)]	Random Forest
74964007	Outra		0,981	1 [53/(53+0)]	Random Forest

Conforme mostra a tabela 8, referente ao domínio "Estado", podemos observar que o classificador apresentou uma sensibilidade máxima para a identificação dos casos "não confirmado" e "ativo" enquanto que o oposto foi observado para os casos cujo o estado da

reação era "inativo", o que indica que nenhum caso positivo foi discriminado pelo classificador. Relativamente aos casos em que o estado da reação foi confirmado, o classificador apresentou um desempenho médio (AUROC=0,550 e sensibilidade=0,600).

Tabela 8 - Resultados do *WEKA* referentes à coluna "Estado" do CPARA, por código *SNOMED CT*

Estado					
Código *SNOMED CT*	**Definição**	**Número de Instâncias por Código**	**AUROC**	**Taxa de Verdadeiros Positivos (VP/VP+FN)**	**Algoritmos de Maior Sensibilidade**
38434005	Não Confirmado	43	0,500	1 (43/(43+0)]	K-NN
55561003	Ativo	25	0,500	1 [25/(25+0)]	K-NN
74996004	Confirmado	3	0,550	0,600 [3/(3+2)]	K-NN
73425007	Inativo	0	0,500	0 [0/(0+13)]	K-NN, SVM, NNge
Código *SNOMED CT*	**Definição**		**AUROC**	**Taxa de Verdadeiros Negativos (VN/VN+FP)**	**Algoritmos de Maior Especificidade**
38434005	Não Confirmado		0,500	1 [43/(43+0)]	SVM
55561003	Ativo		0,500	1 [34/(34+0)]	SVM
74996004	Confirmado		0,500	1 [54/(54+0)]	SVM
73425007	Inativo		0,500	1 [46/(46+0)]	K-NN, SVM, NNge

A tabela 9 apresenta os resultados do domínio "Gravidade", em que apenas dois códigos possíveis estavam disponíveis na amostra. Para ambos os casos, o classificador apresentou alta sensibilidade (grave: AUROC=0,500 e sensibilidade=1; ligeiro: AUROC=0,489 e sensibilidade=0,952).

Tabela 9 - Resultados do *WEKA* referentes à coluna "Gravidade" do CPARA, por código *SNOMED CT*

Gravidade					
Código *SNOMED CT*	**Definição**	**Número de Instâncias por Código**	**AUROC**	**Taxa de Verdadeiros Positivos (VP/VP+FN)**	**Algoritmos de Maior Sensibilidade**
24484000	Grave	38	0,500	1 [38/(38+0)]	SVM
255604002	Ligeiro	20	0,489	0,952 [20/(20+1)]	K-NN
Código *SNOMED CT*	**Definição**		**AUROC**	**Taxa de Verdadeiros Negativos (VN/VN+FP)**	**Algoritmos de Maior Especificidade**
24484000	Grave		0,489	0,952 [20/(20+1)]	K-NN
255604002	Ligeiro		0,500	1 [38/(38+0)]	SVM

Apesar da reduzida amostra de registos clínicos disponibilizada, o facto de alguns modelos apresentarem um bom desempenho, representa que as técnicas de *machine learning* e *text mining* podem ser utilizadas para a construção de um sistema de classificação automática na área de alergologia. No geral, não houve nenhum algoritmo que apresentasse um desempenho superior aos outros, pelo que todos podem ser utilizados para implementar o modelo final, onde apenas deverá ser considerado qual o melhor para cada situação.

Os modelos deverão ser melhorados e devidamente atualizados quando adquiridos novos registos clínicos para a amostra. Construídos através das técnicas de *text mining* e *machine learning* estes modelos poderão ser implementados numa aplicação computacional, nomeadamente uma ferramenta *web*, para uso nos serviços de saúde de Portugal, reduzindo o tempo e esforço da comunidade médica e da área de saúde, de interesse. Para além disso permitirá e facilitará a troca de informação, construção de relatórios e estatísticas a serem utilizadas na área de investigação.

5. Limitações

O projeto apresenta algumas limitações, sendo que uma delas se encontra diretamente associada ao baixo número de registos clínicos adquiridos para a fase de treino. Este baixo número de registos clínicos, devidamente classificados manualmente pelo especialista em imunoalergologia para utilização na aprendizagem recorrendo às técnicas de *machine learning*, condiciona diretamente os resultados pois quanto maior for o seu número e a sua diversificação, maior será a probabilidade de ocorrerem acertos. Este problema tentou mitigar-se apelando à comunidade médica da área de alergologia para o envio de registos clínicos de alergologia devidamente classificados com os campos do CPARA, durante uma apresentação oral na 37º Reunião Anual da SPAIC – Sociedade Portuguesa de Alergologia e Imunologia Clínica.

Um dos filtros criados (filtro de palavras chave) tinha como objetivo a inserção de diferentes termos, que traduzissem o mesmo significado, criando assim um "dicionário" e outra limitação que este modelo de classificação enfrenta, parte da dificuldade de identificar todas as siglas e os diversos significados que as mesmas palavras possam conter, limitação esta que se encontra diretamente relacionada com as próprias técnicas de *machine learning*. Uma possível resolução a longo prazo desta questão, pode passar pelo acréscimo manual destes diferentes e novos termos, ensinando consecutivamente ao programa os diferentes significados possíveis, enriquecendo-o.

Atualmente a base de dados disponível para preenchimento e carregamento para a etapa de aprendizagem, apenas pode conter em cada linha/domínio, um e apenas só um código associado e na eventualidade de surgirem diversificados alergénios num só registo clínico de alergologia (por exemplo), devem ser indicados separadamente os códigos correspondentes em linhas distintas, um campo deve conter unicamente um só código.

6. Conclusão

O trabalho realizado teve como objetivo principal criar um método para sugestão de códigos do CPARA recorrendo a técnicas de *text mining* e *machine learning*, e como objetivo secundário criar um protótipo de uma aplicação *web* para utilização dos profissionais de saúde que necessitam de classificar os diferentes domínios do CPARA com a terminologia clínica *SNOMED CT*. Com a construção deste serviço *web*, esperava-se minimizar os esforços dos profissionais de saúde de alergologia que utilizam a classificação do CPARA regularmente.

As técnicas de *text mining* e *machine learning* podem ser ferramentas muito úteis para extração de informações a partir de registos clínicos, informações estas que podem apoiar processos de apoio à decisão e facilitar tarefas que requerem a classificação sistemática de condições clínicas. Existe uma grande variedade de estudos que utilizaram técnicas de extração de informação a partir de textos de registos clínicos, nomeadamente registos de saúde eletrónicos, para a identificação de doenças e condições clínicas. Técnicas de procura por palavras chave, algoritmos baseados em regras e técnicas de *machine learning* são exemplos de ferramentas utilizadas por estes estudos, que obtiveram diferentes graus de sucesso.

A construção dos modelos de classificação para sugestão de códigos *SNOMED CT* recorreu a algoritmos de *machine learning*, ao conceito de aprendizagem supervisionada e a métodos de procura por palavra chave. Relativamente à avaliação dos classificadores obtidos com o uso destas técnicas, foram utilizadas a AUROC conjuntamente com medidas de sensibilidade e especificidade. A utilização destas métricas vai de encontro com a forma de apresentação de resultados adotada pelos estudos que utilizaram textos de registos clínicos para a extração de informação e identificação de condições clínicas, nomeadamente estudos na área médica. A escolha de tais medidas é compatível com o estudo de revisão sistemática conduzido por (Ford et al., 2016), que recomendou uma maior padronização no uso e apresentação de métricas associadas ao desempenho de algoritmos de classificação de textos de registos clínicos, em particular o uso de valores preditivos positivos e sensibilidade.

Considerando a amostra de registos clínicos utilizada neste trabalho, os resultados referentes aos classificadores obtidos revelaram uma boa performance para a identificação da maior parte dos códigos *SNOMED CT*. Os classificadores apresentaram alta ou média especificidade e boa sensibilidade para a maior parte dos códigos analisados. Relativamente

aos classificadores que apresentaram um baixo desempenho quanto à sensibilidade, concluímos que os potenciais fatores associados são o baixo número de registos clínicos disponibilizados para o treino e limitações relacionadas à plenitude das palavras-chave. Neste contexto, verifica-se a possibilidade de algumas intervenções, nomeadamente a obtenção de mais registos clínicos de forma a complementar a otimizar os resultados obtidos com os modelos de classificação, além da criação de um dicionário mais amplo de palavras-chave, que deverá ter maior participação de especialistas em imunoalergologia.

A título informativo, este é o primeiro trabalho que utiliza técnicas de *machine learning* e *text mining* em textos livres de registos clínicos em Português para a deteção automática de códigos *SNOMED CT,* códigos estes que são utilizados para caraterizar os sete domínios de classificação definidos pelo CPARA em Portugal. A identificação de informações associadas a alergias em registos clínicos é uma área desafiante, nomeadamente porque textos livres em registos clínicos apresentam uma grande variedade de estruturas gramaticais e diferentes formas de se caraterizar um quadro clínico. Métodos baseados em *machine learning* para a classificação automática de registos de imunoalergologia promovem uma solução parcial para os desafios associados à utilização de registos com diferentes vocabulários clínicos. Conforme foram empregues neste trabalho, certas características sobre a distribuição das palavras no texto, nomeadamente a contagem e ocorrência de termos e expressões relevantes, permitiram construir modelos de classificação automática. Com base nos resultados obtidos até agora, tenciona-se otimizar os modelos de classificação obtidos e implementar um serviço *web* que sugira códigos *SNOMED CT* de alergologia a partir de textos livres, conforme é recomendado pelo CPARA. Atualmente, o maior desafio para concluir esta proposta, desafio este que se encontra geralmente associado à maior parte dos métodos baseados em *machine learning* explorados na área da saúde, é a dependência de grandes amostras de registos manualmente categorizados para o treino de modelos de classificação.

7. Trabalhos Futuros

O maior interesse neste projeto, é a evolução do protótipo para um serviço *web*, que permita que esta ferramenta chegue a todos os profissionais de saúde de imunoalergologia de Portugal. Em investigações futuras, seria relevante que fossem feitos testes do utilizador, com o intuito de validar a aplicação em termos de usabilidade, tempo e satisfação, e que fosse inclusivamente feita a atualização dos classificadores com a obtenção de novos registos clínicos.

Referências

(WAO), T. W. A. O. (2016). No Title. Retrieved January 31, 2017, from http://www.worldallergy.org/professional/allergic_diseases_center/drugallergy/

Afzal, Z., Schuemie, M. J., Blijderveen, J. C. Van, Sen, E. F., Sturkenboom, M. C. J. M., & Kors, J. A. (2013). Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records, 1–11.

António Cardoso Martins, João Miguel Marques, P. D. C. (2009). Estudo Comparativo de Três Algoritmos de Machine Learning na Classificação de Dados Electrocardiográficos, 16. Retrieved from https://www.dcc.fc.up.pt/~ines/aulas/0910/MIM/trabs_ano_anterior/noname-1.pdf

Aranha, C., & Passos, E. (2006). A Tecnologia de Mineração de Textos. *RESI-Revista Eletrônica de Sistemas de Informação*, *2*, 1–8. https://doi.org/10.5329/171

Baranov AA , Namazova-Baranova LS , Smirnov IV , Devyatkin DA , Shelmanov AO , Vishneva EA , Antonova EV, S. V., & Nauk, V. R. A. M. (2016). Technologies for Complex Intelligent Clinical Data Analysis.

Bezerra, E. (2010). A Tarefa de Classificação em Text Mining. *Revista de Sistemas de Informação Da FSMA*, *5*, 42–62.

Burget, R., Karasek, J., Smekal, Z., Uher, V., & Dostal, O. (2010). RapidMiner Image Processing Extension:A Platform for Collaborative Research. *International Conference on TELECOMMUNICATIONS AND SIGNAL PROCESSING*, (November 2015).

Butt, L., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N., & Butt, L. (2013). Classification of Cancer-related Death Certificates using Machine Learning What this study adds :, 292–299.

Casillas, A., Gojenola, K., Perez, A., & Oronoz, M. (2016). Clinical text mining for efficient extraction of drug-allergy reactions, 946–952.

Ciolko, E., Lu, F., & Joshi, A. (2010). Intelligent clinical decision support systems based on SNOMED CT. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2010*, 6781–4. https://doi.org/10.1109/IEMBS.2010.5625982

Comissão para a Informatização Clínica, Sociedade Portuguesa de Alergologia e Imunologia Clínica, First Solutions, Faculdade Medicina Universidade Porto, Serviços Partilhados do Ministério da Saúde, & Direcção Geral de Saúde. (2012). Catálogo Português de Alergias e outras Reações Adversas, 1–21.

Comissão para a Informatização Clínica, Sociedade Portuguesa de Alergologia e Imunologia Clínica, First Solutions, Faculdade Medicina Universidade Porto, Serviços Partilhados do Ministério da Saúde, & Direcção Geral de Saúde. (2016). Catálogo Português de Alergias e outras Reações Adversas, 1–21. Retrieved from https://interop-pt.atlassian.net/wiki/display/CTCPT/CPARA+V3.1?preview=/58884219/73895998/CPARA_Especificacoes_V3.1_31.12.2016.pdf

Connolly, B., Matykiewicz, P., Cohen, K. B., Standridge, S. M., Glauser, T. A., Dlugos, D. J., … Pestian, J. (2014). Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals, 866–870. https://doi.org/10.1136/amiajnl-2013-002601

Costa, V. O., Carlos, A., & Filho, D. P. (2014). APLICAÇÃO DO CLASSIFICADOR NAIVE BAYES PARA, *6*, 888–895.

Dangare, C. S. (2015). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, (June 2012). https://doi.org/10.5120/7228-0076

Devasena L., Sumathi T., Gomathi V., H. M. (2011). Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set, *1*(December), 1–3.

Dippenaar, J. M., & Naidoo, S. (2015). Allergic Reactions and Anaphylaxis During Anaesthesia. *Current Allergy & Clinical Immunology*, *28*(1), 18–22.

Estado, D. De. (2009). RSE – Registo de Saúde Electrónico.

et al., L. P. (2015). Text Mining Applied to Electronic Medical Records :, *6*(September). https://doi.org/10.4018/IJEHMC.2015070101

Ewan, P. W. (1998). Anaphylaxis Aetiology. *Management*, *316*(May), 1442–1445.

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association*, *23*(5), 1007–1015. https://doi.org/10.1093/jamia/ocv180

García, F. I. (2017). Introduction to Support Vector Machines. Retrieved January 13, 2017, from http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

Han, Jiawei; Kamber, M. (2011). *Data Mining. San Francisco, CA, itd: Morgan Kaufmann* (Vol. 12).

https://doi.org/10.1007/978-3-642-19721-5

Hegvik, Johan-Arnt ; Rygnestad, T. (2002). Behandling av alvorlige allergiske reaksjoner. *Johan*, (10), 1018–1020.

Horng, S., Sontag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., & Nathanson, L. A. (2017). Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning, *67*, 1–16.

Ihtsdo. (2014). SNOMED CT Starter Guide. *Snomed*, (July), 1–56. Retrieved from http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_StarterGuide_Current-en-US_INT_20141202.pdf?ok

Iskio, J. U. M. F., Uperman, G. I. J. K., Lumenfeld, B. A. B., Ecklet, E. L. G. R., Ates, D. A. W. B., Ms, C., & Andhi, T. E. K. G. (2006). Improving Acceptance of Computerized Prescribing Alerts in Ambulatory Care, *13*(1), 5–11. https://doi.org/10.1197/jamia.M1868.Computerized

ITHSO - International Health Terminology Standards Development Organisation. (2014). SNOMED CT Starter Guide, 56. Retrieved from http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_StarterGuide_Current-en-US_INT_20140222.pdf

Jouhet, V., Defossez, G., Burgun, A., Beux, P., Levillain, P., Ingrand, P., & Claveau, V. (2012). Automated Classification of Free- text Pathology Reports for Registra - tion of Incident Cases of Cancer, 242–251. https://doi.org/10.3414/ME11-01-0005

Jovi, A., Brki, K., & Bogunovi, N. (2014). An overview of free software tools for general data mining.

Koopman, B., Zuccon, G., Nguyen, A., & Bergheim, A. (2015). International Journal of Medical Informatics Automatic ICD-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics*, *84*(11), 956–965. https://doi.org/10.1016/j.ijmedinf.2015.08.004

Lindsted, Gerda; Larsen; Kroigaard, M.; Garvey, L. . . . . N. (2014). Transfusion-associated anaphylaxis during anesthesia and surgery - retrospective study, 158–165.

Martim, B. (1995). Instance-Based Learning: Nearest Neighbour with Generalisation. Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Medicina, F. De. (2003). Registos Clínicos Electrónicos Novembro 2003 Motivação. *Medicina*, 1–19.

Medjahed, S. A. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules, *62*(1), 1–5.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Al-garadi, M. A. (2017). Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection, 1–27. https://doi.org/10.1371/journal.pone.0170242

Nguyen, A. N., Lawley, M. J., Hansen, D. P., Bowman, R. V, Clarke, B. E., Duhig, E. E., & Colquist, S. (2010). Symbolic rule-based classification of lung cancer stages from free-text pathology reports. https://doi.org/10.1136/jamia.2010.003707

Oronoz, M., Gojenola, K., Pérez, A., Ilarraza, A. D. De, & Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish : Mining adverse drug reactions. *JOURNAL OF BIOMEDICAL INFORMATICS*, *56*, 318–332. https://doi.org/10.1016/j.jbi.2015.06.016

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., … Breckenridge, A. M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Bmj*, *329*(7456), 15–19. https://doi.org/10.1136/bmj.329.7456.15 329/7456/15 [pii]

Ramamohan, Y., Vasantharao, K., Chakravarti, C. K., & Ratnam, A. S. K. (2012). A Study of Data Mining Tools in Knowledge Discovery Process. *International Journal of Soft Computing and Engineering*, *2*(3), 191–194.

Rapid-I GmbH. (n.d.). RapidMiner and RapidAnalytics. *Marketplace*. Retrieved from http://www.rapid-i.com

Reis, F., Aires, K., Reis, F., Silva, R. R. V. e, & Lima, K. (2015). Detecção do uso de capacete utilizando máquinas de comitê. *Revista de Sistemas E Computação*, *5*(1), 60–70.

Report, T., & Vincent, K. P. (2005). Text Mining Methods for Event Recognition in Stories, (April), 22. Retrieved from http://kmi.open.ac.uk/publications/pdf/kmi-05-2.pdf

Rijo, R., Silva, C., & Gonçalves, D. (2014). Decision Support System to Diagnosis and Classification of Epilepsy in Children, *20*(6), 907–923.

Ruch, P., Gobeill, J., Tbahriti, I., & Geissbühler, A. (2008). From Episodes of Care to Diagnosis Codes : Automatic Text Categorization for Medico-Economic Encoding, 636–640.

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, *53*, 196–207. https://doi.org/10.1016/j.jbi.2014.11.002

Schiefelbein, U. H., Moiano, I. C., & Livinalli, T. (2015). Mineração de Dados a partir do Currículo Lattes com a Ferramenta WEKA, 291–294.

Schulz, S., Rector, A., Rodrigues, J., & Spackman, K. (2012). Competing Interpretations of

Disorder Codes in SNOMED CT and ICD, 819–827.

Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients, *2*(3), 220–223.

Skeppstedt, M., Kvist, M., & Dalianis, H. (2007). Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. *Eighth International Conference on Language Resources and Evaluation, LREC 2012*, 1250–1257.

Slight, S. P., Seger, D. L., Nanji, K. C., Cho, I., Maniam, N., Dykes, P. C., & Bates, D. W. (2013). Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. *PLoS ONE*, *8*(12), 1–7. https://doi.org/10.1371/journal.pone.0085071

Sohn, S., Kocher, J. A., Chute, C. G., & Savova, G. K. (2011). Drug side effect extraction from clinical narratives of psychiatry and psychology patients. https://doi.org/10.1136/amiajnl-2011-000351

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (n.d.). Weka : Practical Machine Learning Tools and Techniques with Java Implementations.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). *Top 10 algorithms in data mining*. *Knowledge and Information Systems* (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2

Xu, R., & Wang, Q. (2015a). Comparing a knowledge-driven approach to a supervised machine learning approach in large- scale extraction of drug-side effect relationships from free-text biomedical literature. *BMC Bioinformatics*, *16*(Suppl 5), S6. https://doi.org/10.1186/1471-2105-16-S5-S6

Xu, R., & Wang, Q. (2015b). Large-scale automatic extraction of side effects associated with targeted anticancer drugs from full-text oncological articles. *Journal of Biomedical Informatics*, *55*, 64–72. https://doi.org/10.1016/j.jbi.2015.03.009

Xu, Y., Kastner, M., Harada, L., Xu, A., Salter, J., & Waserman, S. (2014). Anaphylaxis-related deaths in Ontario: a retrospective review of cases from 1986 to 2011. *Allergy, Asthma & Clinical Immunology*, *10*(1), 38. https://doi.org/10.1186/1710-1492-10-38

Zaïane, O. R. (1999). Introduction to Data Mining. *Principles of Knowledge Discovery in Databases*, 1–15.

Zhou, L., Baughman, A. W., Lei, V. J., Lai, K. H., Navathe, A. S., & Chang, F. (2015). Identifying Patients with Depression Using Free-text Clinical Documents. https://doi.org/10.3233/978-1-61499-564-7-629

Anexos

A. Lista de *Stopwords* Portuguesas

de	à	me
a	seu	esse
o	sua	eles
que	ou	estão
e	ser	você
do	quando	tinha
da	muito	foram
em	há	essa
um	nos	num
para	já	nem
é	está	suas
com	eu	meu
não	também	às
uma	só	minha
os	pelo	têm
no	pela	numa
se	até	pelos
na	isso	elas
por	ela	havia
mais	entre	seja
as	era	qual
dos	depois	será
como	sem	nós
mas	mesmo	tenho
foi	aos	lhe
ao	ter	deles
ele	seus	essas
das	quem	esses
tem	nas	pelas

este
fosse
dele
tu
te
vocês
vos
lhes
meus
minhas
teu
tua
teus
tuas
nosso
nossa
nossos
nossas
dela
delas
esta
estes
estas
aquele
aquela
aqueles
aquelas
isto
aquilo
estou
está
estamos
estão
estive
esteve
estivemos
estiveram
estava
estávamos
estavam
estivera

estivéramos
esteja
estejamos
estejam
estivesse
estivéssemos
estivessem
estiver
estivermos
estiverem
hei
há
havemos
hão
houve
houvemos
houveram
houvera
houvéramos
haja
hajamos
hajam
houvesse
houvéssemos
houvessem
houver
houvermos
houverem
houvéreis
houverá
houveremos
houverão
houveria
houveríamos
houveram
sou
somos
são
era
éramos
eram

fui
foi
fomos
foram
fora
fôramos
seja
sejamos
sejam
fosse
fôssemos
fossem
for
formos
forem
serei
será
seremos
serão
seria
seríamos
seriam
tenho
tem
temos
têm
tinha
tínhamos
tinham
tive
teve
tivemos
tiveram
tivera
tivéramos
tenha
tenhamos
tenham
tivesse
tivéssemos
tivessem

tiver	terá	teríamos
tivermos	teremos	teriam
tiverem	terão	
terei	teria	

B. Catálogo com os Termos *SNOMED CT* utilizados no CPARA - V3.0

Origem da Informação	
Código *SNOMED CT*	**Descrição**
116154003	Utente/Doente
112247003	Médico
408439002	Imunoalergologista
223366009	Outro profissional de Saúde
133932002	Cuidador
420058008	Acompanhante
58626002	Tutor
125677006	Familiar

Classificação	
Código *SNOMED CT*	**Descrição**
416098002	Alergia medicamentosa
59037007	Intolerância medicamentosa
414285001	Alergia alimentar
235719002	Intolerância alimentar
419199007	Reação alérgica
29544009	Intolerância
160244002	Sem conhecimento de alergias

Alergénios Alimentares	
Código *SNOMED CT*	**Descrição**
11526002	Aspartame
13577000	Noz
15838006	Caracol
24515005	Especiarias
25631100	Maçã
28647000	Carne
44027008	Marisco
51905005	Mostarda
72511004	Frutas
226026007	Bebida de cacau
226760005	Lacticínios
226934003	Porco
227313005	Vegetais
227449005	Frutos secos
230031005	Lagosta
256309005	Pêssego
256319004	Cenoura
256326004	Aipo
256355007	Rebentos de Soja
260176001	Kiwi
264295007	Proteína do leite de vaca
412068007	Centeio
412357001	Milho
430503006	Glutamato
230032003	Ostra
23182003	Cereais
278840001	Camarão
303300008	Proteína do Ovo
418504009	Aveia
59533004	Aditivos alimentares
102259006	Citrinos
102261002	Morango
227037002	Peixe
227146005	Moluscos
227388008	Canela
256310000	Cerejas
256327008	Tomate
256349002	Amendoim
260170007	Batata
264337003	Semente
406771001	Sulfitos
412071004	Trigo
80237000	Manteiga de cacau

226915003	Carne vermelha
28230009	Carne de frango
227144008	Atum
226359003	Óleo de Peixe
7791007	Proteína da soja
227252000	Cogumelos
419420009	Melancia
256353000	Avelã
260177005	Melão
406774009	Ácidos gordos Omega 3 derivados do peixe
418504009	Aveia
227425007	Figos
102264005	Queijo
412061001	Mirtilos
421013007	Framboesas

Outros Alergénios	
Código *SNOMED CT*	**Descrição**
418920007	Adesivos
276310004	Epitélio de animais
288328004	Veneno de abelha
260152009	Pelo de gato
11894001	Toxina botulínica
33396006	Níquel
289122001	Cosméticos
260154005	Pêlo de Cão
61789006	Tintas para têxteis
115589000	Etanolamina
256435007	Penas
256277009	Ervas (pólenes de ervas)
422304003	Gramíneas
256417003	Pelo de Cavalo
128488006	Ácaros
111088007	Látex
255667006	Parafina
418785009	Perfume
256259004	Pólen
43230003	Borracha
303314008	Veneno de escorpião
51420009	Silicone e seus derivados
303315009	Aranha
43735007	Enxofre
256260009	Árvores
256440004	Veneno de vespa
14402002	Madeira
311984009	Agentes físicos
74964007	Outros agentes ou substâncias
260153004	Pelo de Vaca
2309006	Ouro
261243003	Latão
66925006	Cobre
3829006	Ferro
420111002	Metais de contacto
41967008	Prata
12503006	Alumínio
412161004	Lã
256305004	Resina
415710007	Terpeno
256303006	Pólen de artemísia e tasneira
57126000	Cola
83619009	Polyvinyl alcohol (substance)
387398009	Podofilina
59545008	Pesticidas
410853002	Perfluorinato
119417004	Organofosfatos
116549003	Pesticida organoclorado
31006001	Fetos
59351004	Citrato
387293003	Antralina
9021002	Carbaril
14241002	Barata
128489003	Areia
412145001	Tinta para o cabelo

288841007	Substância de teste cutâneo
412160003	Material de sutura
42416001	Lanolina
91598004	Peróxido de benzoílo
2799001	Cloreto de metilbenzetónio
256504004	Material dentário em policarboneto
412156001	Seda
386936005	Ácido azelaico
412153009	Pelo de Coelho
12510000	Óleo de eucalipto
419604006	Ervas daninhas

Reação Adversa	
Código *SNOMED CT*	**Descrição**
95361005	Mucosite
39579001	Anafilaxia
41291007	Angioedema
43116000	Eczema
267036007	Dispneia
40275004	Dermatite de Contacto
418290006	Prurido
247472004	Urticária
1985008	Vómitos
9826008	Conjuntivite
74964007	Outra
24079001	Dermatite atópica
4386001	Broncospasmo
195967001	Asma
70076002	Rinite
115664001	Fotossensibilidade
410430005	Paragem cardiorrespiratória
62315008	Diarreia
73442001	Síndrome de Stevens- Johnson
271759003	Exantema bolhoso
31996006	Vasculite

Gravidade	
Código *SNOMED CT*	**Descrição**
24484000	Grave
255604002	Ligeiro

Estado	
Código SNOMED CT	**Descrição**
55561003	Ativo
73425007	Inativo
74996004	Confirmação
38434005	Não confirmação