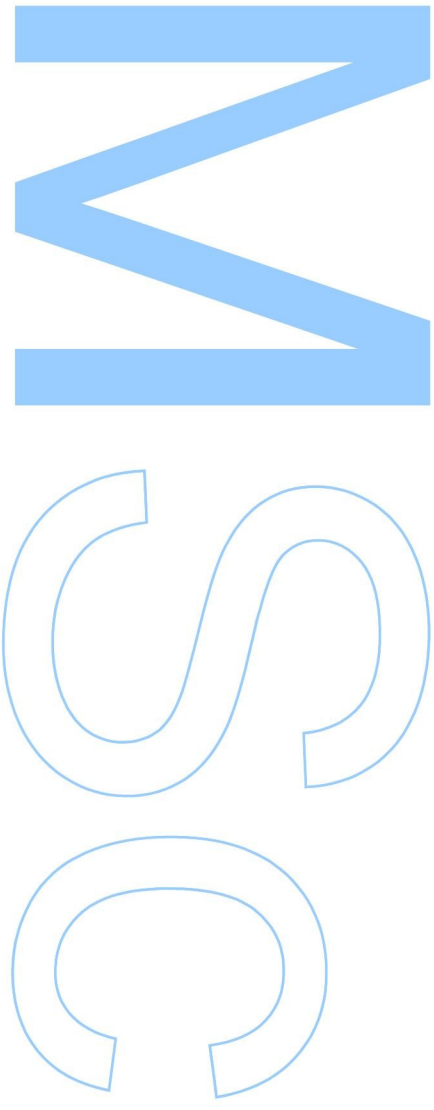
# Parser para Extração de Termos BI-RADS de Relatórios Médicos de Mamografias

João Miguel Jesus Santos Vaz de Carvalho
Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência dos Computadores
2021

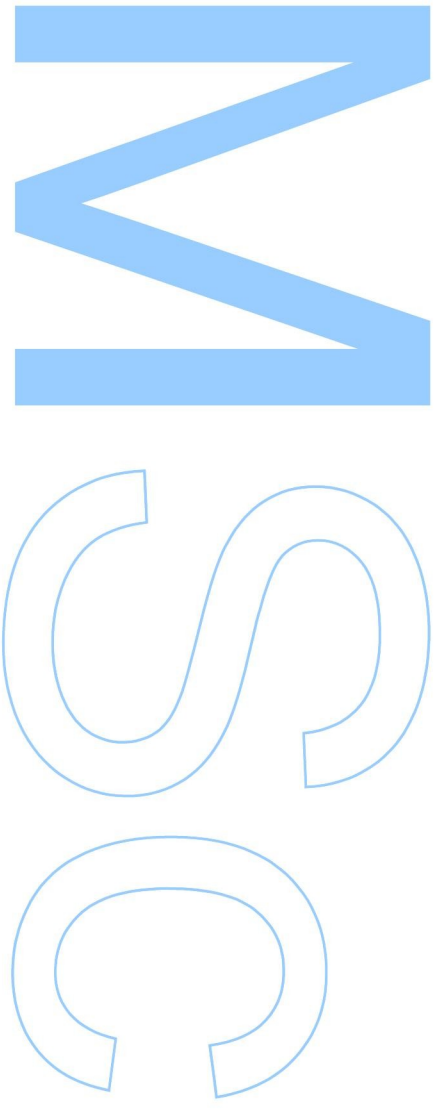
**Orientador**
Inês de Castro Dutra
Professora Auxiliar
Faculdade De Ciências da Universidade do Porto




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



# Abstract

Breast cancer is among the highest mortality rate deceases. One of the ways that has been proven most effective in reducing this rate is its detection at an early stage. The most used test to detect and prevent breast cancer is mammography.

When evaluating the image resulting from a mammographic exam, the specialist dictates or writes down the observations and describes them using the Breast Imaging Reporting and Data System (BI-RADS). This system aims to standardize the medical reports resulting from a mammogram. Depending on the medical report data, the findings severity is determined. Information in text form is a type of unstructured information. Consequently, machine learning algorithms are unable to work with medical reports written in this way. This becomes an obstacle to the development of tools that automatically classify the information contained in a medical report. It is important to develop ways to structure the information so that automated tools can work with it. In medicine, it is crucial not to have mistakes, and this type of tool, unlike human beings, is not influenced by external factors, such as tiredness.

There are already some BI-RADS descriptors extraction tools. However, most are designed to work with the English language. Extraction tools developed to receive information in Portuguese are very scarce.

Given this situation, this work develops a tool that automatically extracts descriptors from the BI-RADS lexicon that are present in medical reports written in Portuguese. A grammar is developed, for each BI-RADS descriptor, in the form of a Regular Expression. The extractor was applied to 150 medical reports divided into two sets, training and testing. The results obtained indicate that in its final version, the extractor achieved an Accuracy of 0.996 in the training set and 0.995 in the test set.

# Resumo

O cancro da mama é uma das doenças com maior taxa de mortalidade a nível mundial. Uma das formas que se tem revelado mais eficaz na redução desta taxa é a sua deteção num estado precoce. O exame mais utilizado para detetar e prevenir o cancro da mama é a mamografia.

Ao avaliar a imagem resultante de um exame mamográfico, o médico especialista dita ou redige as suas observações e descreve-as usando o sistema Breast Imaging Reporting and Data System (BI-RADS). Este sistema tem como objetivo uniformizar os relatórios médicos resultantes de uma mamografia. Consoante a informação contida no relatório médico, a gravidade dos achados é determinada. A informação em forma de texto é um tipo de informação não estruturada. Consequentemente, algoritmos de *machine learning* ficam impossibilitados de trabalhar com relatórios médicos redigidos nessa forma. Isto torna-se um obstáculo ao desenvolvimento de ferramentas que permitam classificar automaticamente a informação contida num relatório médico. É importante que se desenvolvam formas de estruturar a informação de modo a que ferramentas automáticas consigam trabalhar com ela. Na área da medicina é crucial que não se cometam erros, e este tipo de ferramentas, ao contrário do ser humano, não são influenciadas por fatores externos, como por exemplo, o cansaço.

Existem já algumas ferramentas de extração de descritores BI-RADS. No entanto, a sua maioria foi concebida para trabalhar com a língua inglesa. Ferramentas de extração desenvolvidas para receber informação em português são muito escassas.

Face a esta situação, este trabalho desenvolve uma ferramenta que extrai automaticamente os descritores do léxico BI-RADS que estejam presentes em relatórios médicos escritos na língua portuguesa. É desenvolvida uma gramática, para cada descritor BI-RADS, na forma de uma Expressão Regular (ER). O extrator foi aplicado em 150 relatórios médicos divididos em dois conjuntos, treino e teste. Os resultados obtidos indicam que na sua versão final, o extrator conseguiu uma *Accuracy* de 0,996 no conjunto de treino e 0,995 no conjunto de teste.

# Agradecimentos

Para começar, gostaria de agradecer à professora Inês, por me deixar ficar com a oportunidade de realizar este trabalho e por toda a ajuda e disponibilidade.

À minha mãe, que há muito anos que faz o papel de pai e de mãe.

À minha irmã, que já sofreu tanto como eu.

Aos meus amigos, por estarem sempre ao meu lado.

À Cláudia, que viveu isto comigo e me teve de aturar durante estes meses.

Por último, mas nunca menos importante...

Ao meu pai. Tenho a certeza que está orgulhoso.

**Dedico à minha amiga Joana, à minha mãe, e a todas as pessoas que já tiveram de lidar com o cancro...**

# Conteúdo

# Lista de Tabelas

# Lista de Figuras

# Acrónimos

**ACR** American College of Radiology

**ACS** American Cancer Society

**BI-RADS** Breast Imaging Reporting and Data System

**CHUSJ** Centro Hospitalar Universitário de São João

**ER** Expressão Regular

**EUA** Estados Unidos da América

**FN** Falsos Negativos

**FP** Falsos Positivos

**GCO** Global Cancer Observator

**IARC** International Agency for Research on Cancer

**NLP** Natural Language Processing

**PVP** Primeira Versão do Parser

**SVP** Segunda Versão do Parser

**TVP** Terceira Versão do Parser

**VN** Verdadeiros Negativos

**VP** Verdadeiros Positivos

# Capítulo 1

# Introdução

Este capítulo tem como objetivo explicar de um modo geral o problema que originou este trabalho e os principais objetivos que se pretende obter com ele.

## 1.1 Contexto

O cancro é uma das doenças mais incidentes e com maior taxa de mortalidade a nível global. Em 2020 é estimado que tenham surgido 19,3 milhões de novos casos e aproximadamente 10 milhões de mortes devido ao cancro.[1]

Um dos tipos de cancro mais frequentes é o cancro da mama. De acordo com as estatísticas da GLOBOCAN, em 2020, foi o tipo de cancro mais incidente no sexo feminino.[1]

A Radiologia é uma área bastante usada na deteção e prevenção do cancro da mama. O tipo de exame mais comum para prevenir e detetar este tipo de cancro é a mamografia. Para extrair a informação resultante deste exame o especialista usa o léxico Breast Imaging Reporting and Data System (BI-RADS) [2]. Este sistema tem como objetivo padronizar os relatórios médicos resultantes de um exame mamográfico.

Em 2010, Ferreira [3] criou um sistema de apoio à decisão, o *MammoClass*. Este sistema consiste numa aplicação que usa *machine learning* para prever o resultado de uma mamografia através de um conjunto reduzido de descritores pertencentes ao léxico BI-RADS.

## 1.2 Motivação

Grande parte dos modelos preditivos usa algoritmos de *machine learning* para obter os seus resultados. Uma das principais características deste tipo de algoritmos é que a maior parte deles apenas conseguem funcionar com informação estruturada. Quando um especialista redige o relatório médico resultante de uma mamografia, na maioria das vezes, o relatório vem na

forma de texto. Dado que texto é um tipo de informação não estruturada, isto representa um obstáculo a algoritmos de *machine learning* como por exemplo o algoritmo que é usado pelo sistema *MammoClass*. Atualmente, o *MammoClass* trabalha com uma ferramenta de extração de descritores baseada na 4ª edição do BI-RADS. É necessário desenvolver uma ferramenta atualizada e que permita extrair os descritores com base na edição mais recente do BI-RADS(5ª edição).

Subsequentemente, fatores importantes e que tornam relevante a realização deste trabalho são o facto de este se focar em estruturar informação redigida na língua portuguesa e o facto de se basear no léxico BI-RADS.

## 1.3 Objetivo

O principal objetivo deste trabalho é criar um *parser* que faça uma análise sintática de relatórios médicos resultantes de uma mamografia e redigidos em português. Com o apoio do léxico BI-RADS este *parser* vai extrair a informação útil e estruturá-la. Isto permitirá que aplicações como o *MammoClass* usem essa informação para prever os seus resultados.

## 1.4 Estrutura

Este trabalho está dividido em mais 5 capítulos:

- **Contextualização** explica o contexto geral do trabalho e os conceitos necessários à sua compreensão. Este capítulo faz também um resumo de alguns trabalhos relacionados, já publicados.
- **Pré-processamento dos Dados** aborda os conjuntos de dados utilizados para realizar este trabalho. É feita a descrição dos métodos usados no seu pré-processamento e o resultado final depois de terem sido aplicados.
- **Desenvolvimento** descreve a metodologia usada no desenvolvimento do *parser* que analisa os relatórios médicos. É explicado o desenvolvimento de cada uma das versões do *parser*.
- **Resultados e Discussão** começa por fazer uma descrição do *output* produzido pelo *parser*. São apresentados os resultados obtidos em cada uma das suas versões e no fim uma discussão sobre os mesmos.
- **Conclusão** apresenta um resumo geral do desenvolvimento, motivações e contribuições deste trabalho.

# Capítulo 2

# Contextualização

Neste segundo capítulo são explicados o contexto geral do trabalho e os conceitos necessários à sua compreensão. É também feita uma revisão aos trabalhos relacionados já publicados.

## 2.1 Cancro da Mama e Mamografia

O cancro da mama é o tipo de cancro que se desenvolve na glândula mamária. Em 2020 foi o cancro mais incidente no mundo segundo os dados da GLOBOCAN [1]. A GLOBOCAN é uma base de dados produzida pela International Agency for Research on Cancer (IARC) e que contém informação sobre as incidências de 36 tipos de cancro em 185 países. Os dados da GLOBOCAN podem ser acedidos *online* no *website* do Global Cancer Observator (GCO) (gco.iarc.fr) [4].

A Figura 2.1, retirada do *website* do GCO mostra as incidências dos tipos de cancro em ambos os sexos. Podemos observar que o cancro da mama é o mais incidente representando 11,7% dos casos. Devemos considerar o facto do cancro da mama ser praticamente inexistente nos homens representando, no ano de 2015, um valor menor que 1% do número total de casos deste tipo de cancro [5].

Se considerarmos apenas a incidência do cancro apenas no sexo feminino constatámos que o cancro da mama é o tipo mais incidente com 24,5% dos novos casos. O cancro da mama representa também a maior taxa de mortalidade no sexo feminino com 15,5% dos casos [1].

Em 2018 foi feito um estudo, nos Estados Unidos da América (EUA), comparando a taxa de incidência em diferentes regiões portuguesas [6]. Foram usados dados entre 1998 e 2011 e concluiu-se que o número de casos veio a aumentar nesse período sendo a zona sul a mais afetada. Neste estudo prevê-se ainda que, considerando o aumento de casos entre as zonas norte e sul, a zona norte ultrapasse a zona sul no número de casos.

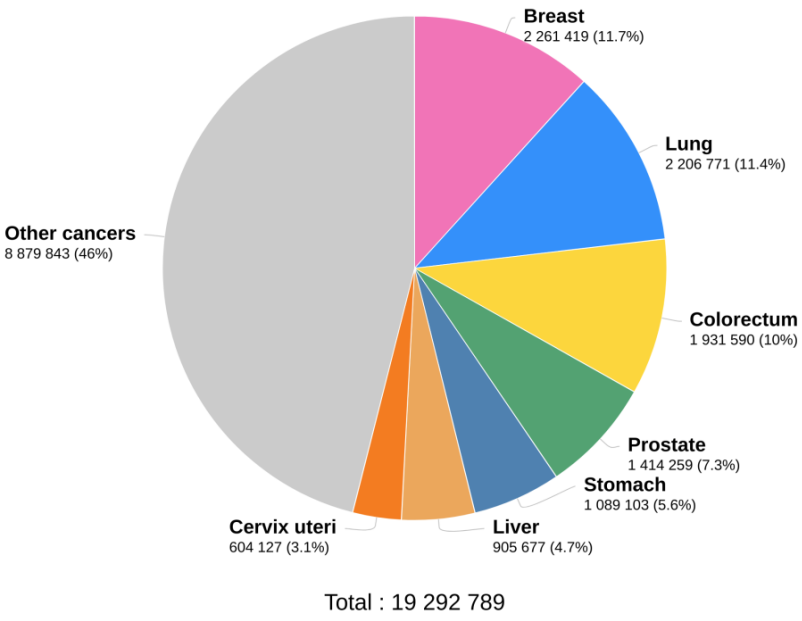



Figura 2.1: Taxa de incidência por tipo de cancro em 2020. Retirado de [4]

A mamografia é o método mais recomendado para prevenção e deteção do cancro da mama. A American Cancer Society (ACS) recomenda que este exame seja realizado anualmente em mulheres a partir dos 45 anos de idade e duas vezes por ano em mulheres com mais de 54 anos de idade [7].

Outras instituições como a American College of Radiology (ACR) consideram que a mamografia está diretamente ligada à redução da mortalidade no cancro da mama. Antes da expansão da mamografia a nível nacional nos EUA (entre 1980 e 1990) a taxa de mortalidade manteve-se relativamente constante. Desde 1990 até 2014 observou-se uma redução na taxa de mortalidade em 38% [8].

No geral, a mamografia permite verificar a existência de tumores que muitas vezes não são palpáveis, podendo estar já numa fase avançada. A deteção precoce de um tumor aumenta o número de tratamentos possíveis a serem realizados. Isto resulta numa maior probabilidade de sobrevivência do paciente.

## 2.2 Léxico BI-RADS

O léxico Breast Imaging Reporting and Data System (BI-RADS) é um sistema léxico pelo qual os radiologistas se baseiam quando descrevem os achados encontrados numa mamografia, exame de ultrassom ou numa ressonância magnética [2].

Este sistema foi criado pelo ACR e o seu principal objetivo foi homogeneizar os relatórios médicos resultantes de uma mamografia. Atualmente o BI-RADS está na sua quinta edição.

### 2.2.1 Descritores BI-RADS

Segundo o atlas da ACR para as mamografias [2, 9], existe um guia de classificação para as informações encontradas no exame mamográfico. Essa classificação é baseada na presença ou não de descritores no exame mamográfico. A Tabela 2.1 representa a lista de descritores BI-RADS, para a sua 5ª edição.

| **Breast Composition** | | |
|---|---|---|
| The breasts are almost entirely flat<br>There are scattered areas of fibroglandular density<br>The breasts are heterogeneously dense, which may obscure small masses<br>The breasts are extremely dense, which lowers the sensitivity of mammography | | |
| **Masses** | | |
| **Shape** | **Margin** | **Density** |
| Oval<br>Round<br>Irregular | Circumscribed<br>Obscured<br>Microlobulated<br>Indistinct<br>Spiculated | High Density<br>Equal Density<br>Low Density<br>Fat-containing |
| **Calcifications** | | |
| **Typically beningn** | **Suspicious morphology** | **Distribution** |
| Skin<br>Vascular<br>Coarse or “popcorn-like”<br>Large rod-like<br>Round<br>Rim<br>Dystrophic<br>Milk of calcium<br>Suture | Amorphous<br>Coarse heterogeneous<br>Fine pleomorphic<br>Fine linear or fine-linear branching | Diffuse<br>Regional<br>Grouped<br>Linear<br>Segmental |
| **Architectural Distortion** | | |
| **Skin Lesion** | | |
| **Asymmetries** | | |
| Asymmetry<br>Global Asymmetry<br>Focal asymmetry<br>Developing asymmetry | | |
| **Intramammary lymph node** | | |
| **Solitary Dilated Duct** | | |
| **Associated Features** | | |

| |
|---|
| Skin Retraction |
| Nipple Retraction |
| Skin Thickening |
| Trabecular Thickening |
| Axillary Adenopathy |
| Architectural Distorcion |
| Calcifications |
| **Location of Lesion** |
| Laterality |
| Quadrant and Clock Face |
| Depth |
| Distance from the nipple |

Tabela 2.1: BI-RADS 5ª Edição: Descritores (Mamografia)

O BI-RADS tem ainda um sistema que classifica os resultados de um exame mamográfico consoante os descritores achados. Este sistema divide-se em sete categorias e está representado na Tabela 2.2.

| **BI-RADS Assessment Categories** | | |
|---|---|---|
| **Category 0** | Incomplete | |
| **Category 1** | Negative | |
| **Category 2** | Benign | |
| **Category 3** | Probably Benign | |
| **Category 4** | Suspicious | 4A: Low suspicion for malignancy |
| | | 4B: Moderate suspicion for malignancy |
| | | 4C: High suspicion for malignancy |
| **Category 5** | Highly Suggestive of Malignancy | |
| **Category 6** | Known Biopsy-Proven Malignancy | |

Tabela 2.2: BI-RADS 5ª Edição: Categorias de Classificação

### 2.2.2 Alterações BI-RADS 5ª edição

Em 2013 a ACR lançou o atlas para a 5ª edição do BI-RADS. Esta nova edição veio substituir a anteriormente utilizada (4ª edição) lançada em 2003. No site da ACR está um documento onde se podem encontrar resumidas as alterações que a 5ª edição veio introduzir no léxico BI-RADS [10].

Existiram alterações gerais que são aplicadas a todo o léxico BI-RADS, independentemente do tipo de exame, e depois alterações mais específicas e diferenciadas para cada tipo de exame (mamografia, ultrassom e ressonância magnética).

Em 2017, Spak et al. [11] resumiram as principais mudanças que a 5ª edição introduziu.

A tabela 2.3 descreve resumidamente as alterações provocadas pela quinta edição do BI-RADS na secção do exame mamográfico e que têm relevância para realização deste trabalho.

| **2003 BI-RADS Atlas (4th Edition)** | **2013 BI-RADS Atlas (5th Edition)** |
|---|---|
| Masses/Shape/Lobular | Masses/Shape/Lobular eliminated |
| Calcifications/Typically Benign/Lucent-Centered Calcifications | Calcifications/Typically Benign/Lucent-Centered eliminated (incorporating it into Calcifications/Typically Benign/Rim) |
| Calcifications/Typically Benign/Eggshell or Rim Calcifications | Calcifications/Typically Benign/Rim ("eggshell"eliminated) |
| Calcifications/Intermediate Concern, Suspicious Calcifications<br>Calcifications/Higher Probability Malignancy (2 categories) | Calcifications/Suspicious Morphology (merged the 2 categories into a single one) |
| Special Cases/Global Asymmetry<br>Special Cases/Focal Asymmetry | Asymmetries/Asymmetry<br>Asymmetries/Global Asymmetry<br>Asymmetries/Focal Asymmetry<br>Asymmetries/Developing Asymmetry<br>(expanded and grouped under separate category) |
| Special Cases/Intramammary Lymph Node | Intramammary Lymph Node (separate category) |
| Special Cases/Asymmetric Tubular Structure-Solitary Dilated Duct | Solitary Dilated Duct (separate category) |
| Associated Findings/Skin Lesion | Skin Lesion (separate category) |

Tabela 2.3: BI-RADS: Comparação entre 4ª Edição e 5ª Edição

Conclui-se que existiram alterações por parte da 5ª edição do BI-RADS, nomeadamente a remoção, renomeação e adição de descritores e categorias.

## 2.3 *MammoClass*

O *MammoClass* é um sistema que usa modelos preditivos para dar apoio à decisão clínica com base nos achados retirados de um exame mamográfico [12–14].

Segundo [12], o *MammoClass* permite que os especialistas selecionem descritores BI-RADS específicos e os seus valores, que serão usados para fazer a previsão da categoria a atribuir. Na sua mais recente versão (versão 2) permite que o utilizador use uma das seguintes opções:

- Dite o seu relatório médico que irá ser convertido em texto
- Insira o seu relatório médico na forma de texto
- Preencha o formulário manualmente

A Figura 2.2 apresenta a interface atual do MammoClass apresentada no seu website.

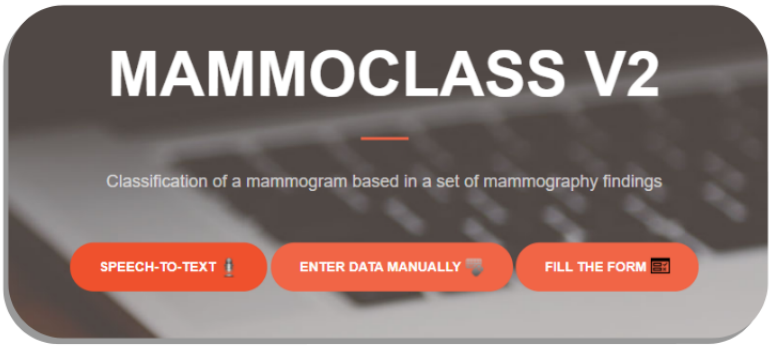



Figura 2.2: Interface MammoClass V2. Retirado de [14].

Tanto a opção em que o utilizador dita o relatório ("SPEECH-TO-TEXT"), como a opção em que o utilizador escreve o relatório ("ENTER DATA MANUALLY"), resultam num ficheiro de texto que vai passar por um processo de extração para retirar as informações necessárias e estruturá-las. Caso o *parser* não consiga extrair alguma variável necessária do ficheiro de texto, o sistema emite um alerta e o utilizador pode escolher o seu valor no formulário manualmente.

Sendo que uma extração manual tem um custo de tempo maior que uma extração automática, sistemas como o *MammoClass* têm como objetivo aplicar uma extração e classificação automática. Este processo, além de reduzir o custo de tempo, garante que todo o processo não é susceptível a erros causados por fatores externos, como por exemplo o cansaço.

No entanto, para o *MammoClass* apresentar resultados fiáveis, necessita de ter uma ferramenta de extração para a versão atual do BI-RADS. Este trabalho pretende desenvolver essa ferramenta.

## 2.4 Pré-processamento e Mineração de Dados

O pré-processamento de dados é um passo importante na mineração de dados. Analisar dados que não tenham sido preparados através do pré-processamento pode causar resultados enganadores [15]. O seu principal objetivo é reduzir o tamanho dos dados, encontrar relações entre eles e normalizá-los [16].

Em García et al. [15] e Alasadi and Bhaya [16] são referidos e descritos alguns passos importantes para a preparação dos dados:

- **Limpeza dos dados**:
  - No seu estado ”bruto”, dados podem estar incorretos e inconsistentes. A limpeza dos dados é o primeiro passo para encontrar valores em falta e corrigir os dados inconsistentes. Permite ainda reduzir todos dos dados que não sejam necessários. Alguns exemplos de métodos usados para lidar com os valores em falta são:
    - Ignorá-los/Removê-los

    * Preenchê-los usando uma constante global
    * Preenchê-los usando a média do valor dessa variável no conjunto de dados

- **Transformação dos dados**:
  - Este processo consiste na conversão ou consolidação dos dados para que o processo de mineração possa ser aplicado.
- **Integração dos Dados**:
  - Este passo é bastante útil quando os dados são provenientes de fontes diferentes. O seu objetivo é juntá-los num único conjunto de dados. Tem que ser cuidadosamente aplicado caso contrário o conjunto de dados final pode ficar redundante ou inconsistente.
- **Normalização dos Dados**:
  - Depois deste processo ser aplicado todos os atributos do conjunto de dados ficarão expressos nas mesmas unidades e usarão uma escala comum. O objetivo deste passo é deixar todos os atributos com o mesmo "peso" .
- **Redução dos Dados**:
  - Esta técnica é usada para fazer uma representação do conjunto de dados num volume mais pequeno mantendo a integridade do conjunto de dados original.

Após estes passos serem aplicados, o conjunto de dados é suposto ser uniforme, compacto e sem valores em falta.

## 2.5 Validação de Resultados

Após a aplicação de um algoritmo de aprendizagem automática é necessário perceber se os resultados que esse algoritmo produziu são fiáveis. Para se conseguir perceber isso são usadas métricas de validação estatística.

O método *Cross Validation* é o método mais comum para validar resultados. Consiste em fazer a divisão do conjunto de dados em duas partes: treino e teste. O conjunto de treino, tal como o nome indica, é onde se vai treinando o modelo que se está a desenvolver para depois o seu desempenho ser avaliado no conjunto de teste [17].

Existem vários métodos possíveis para fazer a divisão do conjunto de dados original:

- ***Hold Out Validation***
  - Neste método, o conjunto de dados é simplesmente dividido em duas partes, uma para treino e uma para teste.

- ***K-fold Cross-validation***:
  - o conjunto de dados é divido em ”k” partes iguais. O modelo é treinado em ”k-1” partes e testado na parte que sobra. O conjunto de teste vai trocando a cada iteração até todas as partes terem sido alvo de teste.
- ***Leave One Out Cross-validation***:
  - Neste método, que é um caso particular do *k-fold cross validation*, ”k” é igual ao número de instâncias do conjunto de dados.
- ***Repeated k-fold Cross-validation***:
  - Aplica-se o método *k-fold cross-validation* várias vezes, sendo que a cada repetição os dados são baralhados e divididos de novo.

As métricas de avaliação ou desempenho são consideradas o método para avaliar a performance de um classificador [18]. Para os problemas de classificação binários pode ser usada uma matriz de confusão, como por exemplo, a representada na Tabela 2.4.

| | Positivos Reais | Negativos Reais |
|---|---|---|
| Positivos Previstos | **Verdadeiro Positivo (VP)** | **Falso Positivo (FP)** |
| Negativos Previstos | **Falso Negativo (FN)** | **Verdadeiro Negativo (VN)** |

Tabela 2.4: Matriz de confusão binária

Nesta matriz:

- Verdadeiros Positivos (VP): Número de previsões corretas para um caso positivo;
- Verdadeiros Negativos (VN): Número de previsões corretas para um caso negativo;
- Falsos Positivos (FP): Número de previsões erradas para um caso negativo;
- Falsos Negativos (FN): Número de previsões erradas para um caso positivo.

A Tabela 2.5 apresenta algumas métricas que se podem calcular através da matriz de confusão.

| Métrica | Fórmula | Descrição |
|---|---|---|
| ***Accuracy*** | $\frac{VP+VN}{VP+VN+FP+FN}$ | Percetagem de previsões corretas no total de instâncias avaliadas |
| ***Error Rate*** | $\frac{FP+FN}{VP+VN+FP+FN}$ | Pergentagem de previsões erradas no total de intâncias avaliadas |
| ***Precision*** | $\frac{VP}{VP+FP}$ | Percentagem de previsões positivas corretas de todas as instâncias positivas previstas |
| ***Recall*** | $\frac{VP}{VP+VN}$ | Fração de positivos previstos corretamente |

Tabela 2.5: Métricas de avaliação possíveis usando valores da matriz de confusão

## 2.6 Expressões Regulares

As Expressões Regulares (ERs) são padrões de texto que definem a forma que um conjunto de caracteres deve ter [19]. O seu nome foi atribuído por Stephen Kleene no seu *paper "A Logical Calculus of the ideas immanent in nervous activity"* [20].

Também chamadas de "regex" ou "regexp", as ERs são muitas vezes usadas para fazer a verificação se um determinado padrão se encontra num texto, extrair esse padrão ou substituir o seu conteúdo. Exemplos de entidades que são passíveis de receber essa função de extração, podem ser os endereços de email, números de cartões de crédito ou nomes de genes e proteínas [21].

### 2.6.1 Sintaxe das Expressões Regulares

Considere-se o seguinte exemplo de uma ER que representa qualquer string começada por "a":

**/a\w*/**

Nesta expressão encontram-se dois tipos de caracteres:

- **Literais:** "a"
- **Metacaracteres:** "\w" e "*"

Os caracteres literais são a forma mais simples de uma ER e vão sempre obter uma correspondência quando esse caractere for encontrado [19].

Por exemplo, se se aplicar a ER "/fox/" à frase *"The quick brown fox jumps over the lazy dog"* encontra-se uma correspondência na palavra "fox". No entanto, também é possível obter várias correspondências de uma ER numa frase: "/be/" obtém duas correspondências se aplicada na frase *"To be or not to be"* [19].

Os metacaracteres existem para facilitar a criação de uma ER que consiga corresponder a um maior número de correspondências. Por exemplo, se o objetivo for encontrar uma correspondência das palavras ”campo” ou ”campos” numa *string*, usando apenas caracteres literais, teriam que ser feitas duas pesquisas nessa *string*, uma para cada palavra. No entanto, com a possibilidade de se usarem metacaracteres as seguintes ERs dariam a mesma correspondência com apenas uma pesquisa: ”campos?/” e ”/camp[o|os]/”

A Tabela 2.6 descreve alguns dos elementos presentes nas ERs e a sua função.

| Elemento | Descrição |
|---|---|
| [ | Obtém correspondência no conjunto de caracteres seguinte |
| ] | Termina conjunto de caracteres |
| **0-9** | Obtém correspondência entre 0 e 9 (0,1,2,3, ... , 9) |
| **a-z** | Obtém correspondência entre ”a” e ”z” (a,b,c, ... ,z) |
| **.** | Obtém correspondência em qualquer caractere excepto ”\n” |
| **\w** | Obtém correspondência em qualquer caractere alfanumérico |
| **\W** | Obtém correspondência em qualquer caractere que não seja alfanumérico |
| \| | Obtém correspondência com o caractere anterior ou com o caractere seguinte |
| **?** | Caractere anterior é opcional (0 ou 1 repetição) |
| ***** | Caractere anterior aparece 0 ou mais vezes |
| + | Caractere anterior aparece no mínimo uma vez |
| **{n,m}** | Caractere anterior aparece entre n e m |

Tabela 2.6: Alguns elementos usados em expressões regulares

## 2.7 Trabalhos Relacionados

Para encontrar trabalhos relacionados com este trabalho, foi executada uma pesquisa sistémica utilizando a *query* descrita na Figura 2.3

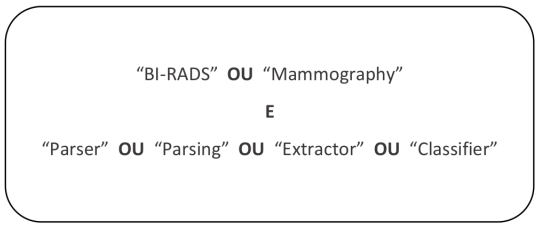



Figura 2.3: Esquema da pesquisa sistémica realizada neste trabalho

Os resultados relevantes encontrados estão descritos nos próximos parágrafos.

Em 2009, Nassif et al. [22] desenvolveram o primeiro *parser* dedicado a extrair descritores BI-RADS de relatórios médicos. Recorrendo às ERs e utilizando relatórios de um hospital em São Francisco, EUA obtiveram uma taxa de acerto de 97,7%. O *parser* desenvolvido neste trabalho extrai descritores de relatórios escritos na língua inglesa.

Em 2016, Gao et al. [23] desenvolveram um *parser* para extrair quatro categorias BI-RADS de relatórios médicos. Este *parser* foi desenvolvido em *perl*, usando ERs e para a língua inglesa,

conseguindo uma taxa de acerto entre 96% e 99% quando testado nos 100 relatórios médicos usados no trabalho.

Em 2016, Bozkurt et al. [24] desenvolveram um *parser* dedicado a extrair elementos BI-RADS de relatórios médicos de mamografias, utilizando técnicas de Natural Language Processing (NLP). No total foram utilizados 300 relatórios e foi conseguida uma taxa de acerto de 98%.

Mais recentemente, em 2018, Miao et al. [25] desenvolveram um *parser* para extrair descritores BI-RADS de relatórios médicos, neste caso, de exames de ultrassom. O *parser* desenvolvido neste trabalho foi desenvolvido para a língua Chinesa e testado em 540 relatórios. Conseguiram uma *F-measure* de 0.904.

Dada a especificidade do *parser* que este trabalho pretende desenvolver, não foram encontrados muitos trabalhos semelhantes. A maioria dos trabalhos utilizam um *parser* que não é desenvolvido para a língua portuguesa ou é proprietário.

Em 2011, Cunha [26], baseando-se no trabalho desenvolvido por [22] desenvolveu um *parser* com o mesmo objetivo, mas dedicado à língua portuguesa. No seu trabalho, Cunha extraiu informações de 153 relatórios médicos.

O *parser* desenvolvido por Cunha conseguiu uma taxa de acerto entre 87% e 91%. Ou seja, conseguiu extrair com essa taxa a mesma informação que a radiologista que o acompanhou no seu trabalho extraiu ”manualmente” dos mesmos relatórios médicos [26].

O trabalho que Cunha desenvolveu foi desenvolvido com recurso a ERs e tendo por base a 4ª edição do BI-RADS. Tendo já sido apresentadas anteriormente as alterações que a 5ª edição introduziu, nomeadamente a remoção, renomeação e adição de descritores, o trabalho desenvolvido por Cunha pode estar desatualizado face aos relatórios médicos atuais que se baseiam na 5ª edição do BI-RADS.

# Capítulo 3

# Pré-Processamento dos Dados

Este capítulo aborda os conjuntos de dados utilizado para a realização deste trabalho. É feita a descrição do pré-processamento desses dados, e dos conjuntos finais resultantes desse mesmo processo.

## 3.1 Conjuntos de Dados

Para a realização deste trabalho foram cruciais dois conjuntos de dados diferentes:

- Os relatórios médicos para o *parser* analisar
- Uma avaliação por uma especialista da área desses mesmos relatórios como meio de comparação com os resultados obtidos pelo *parser*

Esses dois conjuntos de dados são provenientes de dois ficheiros: `MAMOGRAFIA.txt` e `AvalClinica&Relatorio-2Aval.xls`

O ficheiro `MAMOGRAFIA.txt` é um ficheiro de texto com um conjunto de registos clínicos (realizados entre 2008 e 2009) e facultados pelo Centro Hospitalar Universitário de São João (CHUSJ), no Porto. O ficheiro `AvalClinica&Relatorio-2Aval.xls` contém os resultados(para cada Descritor BI-RADS) relativos à avaliação do *parser* utilizado por Cunha [26], comparando-os com a avaliação por parte duma especialista (Inês Moreira, radiologista no Centro de Mama do CHUSJ). Neste ficheiro estão contidas as avaliações de 153 relatórios. As Figuras 3.1 e 3.2 apresentam o conteúdo de cada um dos ficheiros.

```
1  ID;ID_DOENTE;C_ADIPOSA;C_FIBRO_GLANDULARES;C_HETEROGENEAS;C_DENSA;ID_RESPONSAVEL;DATA_INSEI
2  271;817616;'2';'2';'2';'2';8027;7/21/2008 10:38:28 AM;'No actual estudo, observamos padrão
3  A pele e o tecido celular subcutâneo apresentam aspectos mamográficos normais.Não se indiv:
4  A pele e o tecido celular subcutâneo apresentam aspectos mamográficos normais.Não se indiv:
5  265;8458;'2';'2';'2';'2';7038;7/21/2008 10:21:23 AM;'Paciente submetida cirurgia conservad
6  Ao actual estudo mamográfico, assinalamos relativa escassez de parênquima mamário e padrão
7  A pele e o tecido celular subcutâneo apresentam aspectos mamográficos normais.Não se indiv:
8
9
10 Alterações provavelmente benignas - Bi-Rads - 3.';;;;0;;;;;;;0;0;'0';;;;;7/21/2008;'0';;;;
11
12 235;473671;'2';'2';'2';'2';7038;7/15/2008 11:26:28 AM;'ECOGRAFIA MAMÁRIA
13
14 Na transição dos quadrantes externos da mama esquerda continuamos a observar nódulo sólido
15 Não são aparentes outras formações nodulares sólidas ou líquidas, áreas focais de atenuaçã
16 230;369081;'2';'2';'2';'2';100;7/15/2008 9:29:00 AM;'Exame mamografico  do exterior: sem a
17 270;817616;'2';'2';'2';'2';8027;7/21/2008 10:38:27 AM;'No actual estudo, observamos padrão
18 A pele e o tecido celular subcutâneo apresentam aspectos mamográficos normais.Não se indiv:
```

Figura 3.1: Ficheiro `MAMOGRAFIA.txt`

| ID_DOENTE | INES 10 | DenHigh | INES 11 | DenIso | INES 12 | DenLow | INES 13 | DenFat |
|---|---|---|---|---|---|---|---|---|
| 28152 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 35421 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 35421 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 60768 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100668 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 126347 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 154663 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 170010 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 175893 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 176027 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 197482 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 201087 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 212119 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 223047 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 240155 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 250738 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 252099 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 |
| 252099 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 260020 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 261150 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 261735 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 265306 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 270881 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 275145 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 317075 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figura 3.2: Ficheiro `AvalClinica&Relatorio-2Aval.xls`

## 3.2 Pré-processamento

Para se poder obter um bom conjunto de dados que o *parser* vai analisar é necessário primeiro extrair as informações necessárias de cada um dos dois ficheiros já mencionados no sub-capítulo anterior.

### 3.2.1 MAMOGRAFIAS.txt

Como se pode verificar na Figura 3.1 o ficheiro `MAMOGRAFIAS.txt` é um ficheiro onde a informação está desorganizada. Para ser possível extrair a informação relevante para este trabalho aplicaram-se algumas das técnicas mencionadas no capítulo 2.4. Todas as alterações efetuadas ao conjunto de dados foram realizadas através de programas, em *Python*, desenvolvidos para o efeito.

O primeiro passo foi organizar o ficheiro `MAMOGRAFIAS.txt`. O conjunto de dados resultante foi colocado no ficheiro `mamografia_completo.csv`. A Figura 3.3 representa esse ficheiro.

```
1  ID;ID_DOENTE;C_ADIPOSA;C_FIBRO_GLANDULARES;C_HETEROGENEAS;C_DENSA;ID_RESPONSAVEL;DATA_INSERCAO;F
2  271;817616;'2';'2';'2';'2';8027;7/21/2008 10:38:28 AM;'No actual estudo, observamos padrão mamoç
3  265;8458;'2';'2';'2';'2';7038;7/21/2008 10:21:23 AM;'Paciente submetida cirurgia conservadora à
4  235;473671;'2';'2';'2';'2';7038;7/15/2008 11:26:28 AM;'ECOGRAFIA MAMÁRIANa transição dos quadrar
5  230;369081;'2';'2';'2';'2';100;7/15/2008 9:29:00 AM;'Exame mamografico  do exterior: sem alteraç
6  270;817616;'2';'2';'2';'2';8027;7/21/2008 10:38:27 AM;'No actual estudo, observamos padrão mamoç
7  240;901659;'2';'2';'2';'2';8027;7/15/2008 11:36:18 AM;'No actual estudo, observamos grande quant
8  243;854350;'2';'2';'2';'2';100;7/15/2008 11:50:10 AM;'Ao actual estudo mamográfico, assinalamos
9  262;8458;'2';'2';'2';'2';7038;7/21/2008 10:21:10 AM;'Paciente submetida cirurgia conservadora à
10 286;329228;'2';'2';'2';'2';8027;7/21/2008 12:52:18 PM;'Exame efectuado a pedido da Radiologia.Nc
11 277;125640;'2';'2';'2';'2';8027;7/21/2008 11:12:38 AM;'Observamos padrão mamário extremamente de
12 96;288865;'0';'2';'2';'2';7038;7/3/2008 3:37:57 PM;'Glândulas mamárias predominantemente adiposa
13 251;175662;'2';'2';'2';'2';8027;7/17/2008 11:10:17 AM;'Trata-se de doente submetida a cirurgia c
14 188;424567;'2';'2';'2';'2';7038;7/14/2008 10:29:25 AM;'Releitura de estudos do exterior, datados
```

Figura 3.3: Ficheiro `mamografias_completo.csv`

Este ficheiro contém um total de 1144 registos onde cada registo apresenta 36 atributos. Os atributos estão descritos na Tabela 3.1

| **Atributos mamografia_completo.csv** | | | |
|---|---|---|---|
| 1 | ID | 19 | MEDICO_RESP2 |
| 2 | ID_DOENTE | 20 | MEDICO_RESP1 |
| 3 | C_ADIPOSA | 21 | TECNICO |
| 4 | C_FIBRO_GLANDULARES | 22 | BENIGNA |
| 5 | C_HETEROGENEAS | 23 | PROV_BENIGNA |
| 6 | C_DENSAOLA | 24 | BAIXAS_MALIGNA |
| 7 | ID_RESPONSAVEL | 25 | SUSPEITA_MALIGNA |
| 8 | DATA_INSERCAO | 26 | MALIGNA |
| 9 | RELATORIO | 27 | DATA_EXAME |
| 10 | BIRADS_0 | 28 | TIPO1A |
| 11 | BIRADS_1 | 29 | TIPO1B |
| 12 | BIRADS_2 | 30 | TIPO1C |
| 13 | BIRADS_3 | 31 | TIPO2 |
| 14 | BIRADS_4A | 32 | TIPO3 |
| 15 | BIRADS_4B | 33 | TIPO4 |
| 16 | BIRADS_4C | 34 | INF_EXAM_POS |
| 17 | BIRADS_5 | 35 | INF_QTD |
| 18 | BIRADS_6 | 36 | INF_QTE |

Tabela 3.1: Atributos ficheiro `mamografia_csv`

Os atributos relevantes para o desenvolvimento deste trabalho são o atributo número 1, "ID", e o atributo número 9, "RELATORIO". Após a observação do ficheiro `mamografia_completo.csv`

foi possível constatar que:

- Existem registos onde o atributo ”RELATORIO” se repete;
- Existem registos onde o atributo ”RELATORIO” é nulo, ou seja, uma *string* vazia;

Procedeu-se então à técnica ”limpeza dos dados”, referida no capítulo 2. Foram eliminados todos os atributos que não são relevantes para o trabalho e removeram-se todos os registos repetidos ou com o atributo ”RELATORIO” em falta. O resultado final foi o ficheiro `mamografia_versao_final.csv` que está representado na Figura 3.4.

```
1   28152;'Opacidade nodular no QSE da mama direita, com cerca de 3cm, de contornos espicul
2   35421;'antecedentes de mamoplastia de reduçãoachados imagiológicos benignos, compativei
3   35421;'antecedentes de mamoplastia de redução bilateralNo actual estudo, observamos pad
4   60768;'No actual estudo, observamos grande quantidade de parênquima mamário, com padrão
5   100668;'No actual estudo, observamos grande quantidade de parênquima mamário, com padrã
6   126347;'Efectuada releitura imagiológica de exames de 17/09/2008 e 23/09/2008.Mantém-se
7   154663;'No actual estudo, observamos padrão mamográfico de densidades fibroglandulares
8   170010;'Efectuamos estudo mamográfico,no actual estudo mamográfico, assinalamos relativ
9   175893;'Massa envolvendo praticamente todo o globo mamário, à esquerda - ca avançado.MB
10  176027;'Opacidade nodular, com 25mm, região central, retro-areolar à direita.Achados im
11  197482;'Estudo mamográfico sem achados de suspeição, nomeadamente a aréa de densificaçã
12  201087;'No actual estudo, observamos padrão mamográfico de densidades fibroglandulares
13  212119;'Por estudo ecográfico observamos na transição dos quadrantes externos da mama e
14  223047;'No actual estudo, observamos padrão mamográfico de densidades fibroglandulares
```

Figura 3.4: Ficheiro `mamografia_versao_final.csv`

O ficheiro `mamografia_versao_final` é composto pelos dois únicos atributos relevantes para este trabalho: ”ID” e ”RELATORIO”. No total tem 150 registos. Sendo este número diferente dos 153 registos no ficheiro que contém a avaliação da especialista Inês, foi feita uma comparação dos registos verificando-se que dois registos eram repetidos e um outro tinha o atributo ”RELATORIO” vazio, ou seja, sem relatório.

O conjunto de dados a ser utilizado neste trabalho ficou com 150 registos e guardado no ficheiro `mamografia_versao_final.csv`. A Tabela 3.2 compara os registos do conjunto de dados inicial com o conjunto de dados final.

| | **MAMOGRAFIA.txt** | **mamografia_versao_final.csv** |
|---|---|---|
| **Registos** | 1144 | 150 |
| **Atributos** | 36 | 2 |

Tabela 3.2: Comparação entre `MAMOGRAFIA.txt` e `mamografia_versao_final.csv`

### 3.2.2 AvalClinica&Relatorio-2Aval.xls

Este ficheiro, apresentado na Figura 3.2 foi criado por Cunha [26] enquanto realizava a sua dissertação.

Foram criadas duas colunas para cada descritor BI-RADS. Uma coluna correspondente à avaliação feita pelo *parser* de Cunha (coluna branca) e uma coluna para a avaliação da especialista Inês (coluna a azul). Se fosse encontrado um descritor num determinado registo, a célula correspondente a esse descritor e a esse registo seria preenchida com um ”1”. Caso contrário, seria preenchida com um ”0”.

A Tabela 3.3 descreve o resultado em cada uma das quatro possíveis combinações de valores entre as duas colunas de um descritor. As combinações possíveis estão todas também representadas na Figura 3.2.

| **Descritor BI-RADS** | | **Resultado** |
|---|---|---|
| **Avaliação Inês** | ***Parser* Filipe** | |
| 0 | 0 | Nada acontece (são concordantes) |
| 1 | 0 | Células das duas colunas vermelhas (são discordantes) |
| 0 | 1 | Células das duas colunas vermelhas (são discordantes) |
| 1 | 1 | Células das duas colunas verdes (são concordantes) |

Tabela 3.3: Resultados possíveis no ficheiro `AvalClinica&Relatorio-2Aval.xls`

Neste ficheiro, Cunha atribuiu um nome a cada descritor BI-RADS e introduziu no início da coluna que pertencia à avaliação do seu *parser*. Por exemplo para o descritor *High Density* foi escolhido o nome *DenHigh*, como se pode observar na Figura 3.2.

Para a realização deste trabalho, a informação relevante é a informação presente nas colunas azuis. Ou seja, as correspondências encontradas pela especialista Inês. Extraiu-se então, para cada registo, o ”ID” correspondente e os achados encontrados nesse registo pela especialista. Ou seja, o nome do descritor de todas as colunas azuis que tivessem o número ”1”.

Essa informação foi extraída para o ficheiro `resultados_ines.csv` e está representada na Figura 3.5.

```
 8   170010;
 9   175893;AxAdnp;
10   176027;
11   197482;SCFoc;
12   201087;
13   212119;MgCrc;
14   223047;
15   240155;MgSpi;DenHigh;
16   250738;DenHigh;MrphSut;SCAsym;SkThck;ArcDst;
17   252099;DenLow;
18   252099;DenIso;
19   260020;SkThck;
20   261150;DenIso;
21   261735;ArcDst;
22   265306;MgSpi;DenHigh;
23   270881;DenHigh;NpRtr;ArcDst;
24   275145;
25   317075;AxAdnp;
26   321725;
27   324616;DiClstd;
28   328707;SCAsym;SkThck;ArcDst;
```

Figura 3.5: Ficheiro `resultados_ines.csv`

Como demonstra essa Figura, para o registo com "ID" igual a "197482" foi encontrado o descritor "SCFoc" que representa o descritor "Assimetria Focal".

Estando o pré-processamento dos dados completo, resultam dois ficheiros que são cruciais para o desenvolvimento do *parser* deste trabalho:

- `mamografia_versao_final.csv`
  - ficheiro que contém os "ID" e "RELATORIO" de cada registo. Está ordenado por "ID".
- `resultados_ines.csv`
  - ficheiro que contém os "ID" e descritores achados (pela especialista) em cada registo. Está ordenado por "ID".

# Capítulo 4

# Desenvolvimento

Neste capítulo é descrita a metodologia no desenvolvimento do *parser* que vai fazer a análise dos relatórios médicos.

## 4.1 Metodologia

Para o desenvolvimento do *parser* foi decidido que este seria desenvolvido em *Python* e que teria três versões:

- Versão 1:
  - Esta versão usa única e exclusivamente uma gramática base desenvolvida através da tradução dos descritores BI-RADS para português e usando sinónimos desses descritores.
- Versão 2:
  - Esta versão refina a gramática usada na versão anterior, com base nos relatórios médicos e na informação contida no ficheiro `resultado_ines.csv`.
- Versão 3:
  - Versão com a gramática da segunda versão mas que inclui uma função de negação. Ou seja, verifica se uma correspondência detetada num relatório está a ser negada.

Em relação ao conjunto de dados, o método de validação utilizado será o método de validação *Hold-out* mencionado no capítulo 2. 70% dos registos serão usados como conjunto de treino, restando 30% que serão usados como conjunto de teste. O facto dos dados serem provenientes da mesma fonte e o facto da proximidade com que dois registos foram feitos, não influenciar os descritores presentes no relatório médico, ajudaram na escolha deste método de validação.

Perante esta situação, procedeu-se à divisão dos ficheiros `mamografia_versao_final.csv` e `resultados_ines.csv` mencionados no capítulo 2. Os quatro ficheiros resultantes estão representados na Tabela 4.1.

| Ficheiro Original | Ficheiro Após Divisão | Conteúdo |
|---|---|---|
| mamografia_versao_final.csv | treino.csv | Conjunto de treino |
| | teste.csv | Conjunto de teste |
| resultados_ines.csv | resultados_ines_treino.csv | Resultados para o conjunto de treino |
| | resultados_ines_teste.csv | Resultados para o conjunto de teste |

Tabela 4.1: Tabela dos ficheiros usados pelo *parser*

## 4.2 Gramática para Extração de Atributos

Antes de se iniciar o desenvolvimento do *parser*, foi necessário definir-se a gramática a utilizar na verificação de que descritores estão presentes no relatório médico. Começou-se por fazer uma tradução direta dos descritores para português. No entanto a língua portuguesa é muito complexa e há muitos fatores a ter em conta.

Por exemplo, considere-se o descritor *"Masses:Shape:Round"*, onde *"Round"* é a palavra-chave. A sua tradução para português pode ter todas as seguintes opções: redondo, redonda, arredondado, arredondada. E ainda todos os plurais dessas opções. Ou seja, no processo de tradução de uma palavra é necessário verificar-se o seu plural, masculino e feminino, se existirem.

Outro caso a considerar no processo de criação de uma gramática são os sinónimos. Eles podem ser palavras parecidas ou não e ser uma única palavra ou não também. Considere-se o descritor *"Masses:Density:Equal Density"*, onde *"Equal density"* são as palavras-chave. Na sua tradução para português, este descritor pode apresentar-se em pelo menos três formas distintas: igual densidade, isodenso e densidade homogénea.

No desenvolvimento da gramática foram feitas as respetivas alterações gramaticais de modo a adaptar-se a gramática à 5ªa Edição do BI-RADS. Por exemplo, no caso do descritor *"Calcifications:Tipically Benign: Punctate"*, que na 5ª edição juntou-se ao descritor *"Calcifications:Tipically Benign: Round"* [27] foi introduzido, na sua tradução para português, no respetivo descritor.

Com base neste tipo de situações, foi criada uma gramática que serviu como base para a Primeira Versão do Parser (PVP), e como base para as duas versões seguintes. Está representada na Tabela 4.2. Na gramática está representada apenas uma versão de cada palavra para cada descritor, e os seus sinónimos. Não estão representados todos os casos possíveis envolvendo masculino e feminino ou plurais.

| ***Masses*** | | |
|---|---|---|
| Descritor Original | | Gramática |
| ***Shape*** | ***Round*** | Redonda; Arredondada |
| | ***Oval*** | Oval |
| | ***Irregular*** | Irregular; Mal Definida |
| ***Margins*** | ***Circumscribed*** | Circunscrita; Bem Definida; Regular |
| | ***Obscured*** | Obscura |
| | ***Microlobulated*** | Microlobular |
| | ***Indistinct*** | Indistinta; Imprecisa; Irregular; Indefinida |
| | ***Spiculated*** | Espiculada |
| ***Density*** | ***High*** | Alta; Elevada; Hiperdensa |
| | ***Equal*** | Igual; Homogénea; Isodensa |
| | ***Low*** | Baixa; Pouca |
| | ***Radiolucent/Fat containing*** | Radiolucente; Contém Gordura; Gordurosa |
| ***Calcifications*** | | |
| Descritor Original | | Gramática |
| ***Tipically Benign*** | ***Skin*** | Dérmicas; Pele |
| | ***Vascular*** | Vasculares |
| | ***Coarse or "popcorn-like"*** | Grosseiras; Pipoca |
| | ***Large rod-like*** | Bastonete |
| | ***Round*** | Redondas; Punctiformes |
| | ***Rim*** | Periferia; Periféricas |
| | ***Dystrophic*** | Distróficas |
| | ***Milk of Calcium*** | "Leite Cálcio" |
| | ***Suture*** | Cicatriciaais; Cicatriz |
| ***Suspicious Morphology*** | ***Amorphous*** | Amórficas |
| | ***Coarse Heterogeneous*** | Grosseiras heterogéneas |
| | ***Fine Pleomorphic*** | Pleomórficas |
| | ***Fine Linear or fine-linear branching*** | Finas; Lineares |
| ***Distribution*** | ***Diffuse*** | Difusas; Dispersas |
| | ***Regional*** | Regionais |
| | ***Grouped*** | Agrupadas; Grupos |
| | ***Linear*** | Lineares |
| | ***Segmental*** | Segmentais |
| ***Asymmetries*** | | |
| Descritor Original | | Gramática |
| ***Asymmetry*** | | Assimetria |

| | |
|---|---|
| ***Global Asymmetry*** | Assimetria Global |
| ***Focal Asymmetry*** | Assimetria Focal |
| ***Developing Asymmetry*** | Desenvolver assimetria |
| Associated Features | |
| Descritor Original | Gramática |
| ***Skin retratction*** | Retração Cutânea; Pele; Dérmica |
| ***Nipple Retraction*** | Retração do Mamilo |
| ***Skin thickening*** | Espessamento Cutâneo |
| ***Trabecular thiekening*** | Espessamento Trabecular |
| ***Axillary adenopathy*** | Adenopatia Axilar; Gânglio Axilar |
| Descritor Original | Gramática |
| ***Architectural Distorcion*** | Distorção Arquitetural; Desorganização arquitetural |
| ***Intramammary Lymph Node*** | Gânglio linfático intramamário; Mamário |
| ***Skin Lesion*** | Lesão Cutânea; Pele; Dérmica |
| ***Solitary Dilated Duct*** | Ducto Dilatado |

Tabela 4.2: Gramática usada na PVP

## 4.3 Primeira Versão do Parser

Como referido no inicio deste capítulo, o objetivo principal desta versão do *parser* é basear-se na gramática apresentada na Tabela 4.2.

Verificar cada combinação possível para cada descritor presente na gramática, além de ser um processo longo, não seria produtivo. Para solucionar este problema procedeu-se ao uso das Expressões Regulares (ERs) mencionadas no Capítulo 2. Verificando-se que todos os descritores são compostos por no mínimo duas palavras desenvolveu-se uma ER que fosse aplicada a todos os descritores, e que permitisse encontrar uma combinação de palavras num texto.

```
/\b(palavra_1)\w+(?:\W+\w+){0,5}?\W+(palavra_2)\w+\b/
```

Figura 4.1: ER que encontra uma combinação de duas palavras

Considerando a Figura 4.1 verifica-se o seguinte significado da ER:

- **"\b"**:
  - Marca o início da palavra.
- **"(palavra_1)\w+"**
  - Primeira combinação de palavras a ser procurada. Terminando em "\w+" permite a que as palavras que estão dentro de parênteses não estejam completas. Por exemplo, "(redond)\w+" encontrará correspondência em "redondo"e "redonda"e nos seus respetivos plurais.
- **"(?:\W+\w+){0,5}?\W+"**
  - Isto indica que entre a primeira e a segunda palavra podem existir entre 0 e 5 palavras. Para modificar o intervalo de palavras a usar basta alterar os valores "0" e "5" sendo que o primeiro representa o número mínimo de palavras e o segundo o número máximo.

Esta ER permite que se consiga obter correspondência com palavras derivadas de um certo radical. Com base nesta ideia, a gramática dos descritores Breast Imaging Reporting and Data System (BI-RADS) representada na Tabela 4.2, começou a ser desenvolvida na forma de ERs. Foi construída uma ER para cada um dos descritores. Por não existir ainda bem uma percepção da estrutura de um relatório médico, foi considerada a existência de 0 a 5 palavras, entre cada par de palavras pertencente a um descritor. Este valores foram aplicados a todos os descritores.

Por exemplo, para o descritor "Massas:Forma:Redonda" foram tidas em conta as palavras "massa" e "redonda". Começando por ver sinónimos e palavras possíveis na Tabela 4.2 foi-se construindo a sua ER até chegar à expressão representada na Figura 4.2.

```
/\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+((ar)?(redon))\w+\b/
```

Figura 4.2: ER do descritor "Massas:Forma:Redonda" na PVP

A ER da figura 4.2 permite encontrar diversas correspondências para o mesmo descritor como por exemplo "massa arredondada", "nódulo redondo" ou "densidade redonda". Permite ainda encontrar combinações em que as palavras não são vizinhas. São aceites também portanto combinações como "nódulo com forma redonda" ou "massa densa e arredondada".

Na Figura 4.3 está representada a gramática, de ERs, utilizada na PVP para a categoria "Massas". A gramática completa desta versão pode ser consultada no Anexo A.

```
# Masses Shape:
\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+((ar)?(redon))\w+\b # Round
\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+(ova|ov(ó|o))\w+\b # Oval
\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+(irregu)\w+\b # Irregular

# Masses Margins:
# Circumscribed
\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+((bem )?(circunsc|definid|(de)?limit||regular))\w+\b
\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+((obscur|escu|sem brilh))\w+\b # Obscured
\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+(microlobul)\w+\b) # Microlobulated
\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+(irregu|indistin|mal (definid)\w*)\w*\b # Indistinc
\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+(espicul)\w+\b # Spiculated

# Masses Density:
\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+(alta densi|hiperden)\w*\b # High
\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+((normal|m(é|e)dia) densi|isoden)\w*\b # Medium
\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+((pouca|baixa) densi)\w*\b # Low
\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+(gordur)\w+\b # Fat-containing
```

Figura 4.3: Gramática usada na PVP para a categoria "Massas"

Olhando para a gramática descrita na figura 4.3 podemos concluir que foi usada uma enorme variedade de combinações possíveis em cada ER. Isto aumentará as hipóteses de correspondência para cada descritor BI-RADS.

Ficou assim concluída a Primeira Versão do Parser que este trabalho desenvolveu, sendo que foi baseada numa tradução dos descritores originais para português.

## 4.4 Segunda Versão do Parser

Após a concretização da PVP desenvolveu-se, com base na primeira gramática, uma gramática mais sofisticada e adaptada à realidade dos relatórios médicos de mamografias.

Depois de se consultar os registos no conjunto de treino, foi possível obter uma maior percepção da estrutura dos relatórios médicos. Com base nisso, foi decidido reduzir o número de palavras possíveis entre dois termos de um descritor. Na gramática da primeira versão, o número de palavras variava entre 0 e 5, sendo que nesta versão varia entre 0 e 3. Esta alteração foi aplicada a todos os descritores, com excepção dos descritores pertencentes à categoria "Calcificações", que se mantiveram iguais à gramática da primeira versão.

Foi necessário perceber se para caso falhado o motivo era um dos seguintes:

- Erro ortográfico.
- Gramática da PVP não prevê o caso específico.
- Descritor está a ser negado.

Os três casos foram verificados em diversos registos. São exemplificados aqui alguns exemplos dos dois primeiros casos para explicar como foram criadas as alterações que permitissem à gramática da Segunda Versão do Parser (SVP) obter melhores resultados que a gramática da

PVP. Um descritor estar a ser negado apenas foi resolvido com a Terceira Versão do Parser (TVP), que tem como objetivo não se obter uma correspondência positiva onde é suposto ser negativa.

### 4.4.1 Erros ortográficos

Este tipo de erro acontece quando, no relatório médico, existem palavras que não estão escritas como devem estar. São diversas as razões para isto acontecer. Pode faltar um acento na palavra ou o especialista que redigiu se ter enganado a escrever, ou então ainda usar o antigo acordo ortográfico. De qualquer modo, detetar-se erros ortográficos nos relatórios foi um acontecimento comum.

Estes fatores resultaram que alguns descritores, na PVP, não fossem detetados quando deveriam ter sido. Considere-se a Figura 4.4:

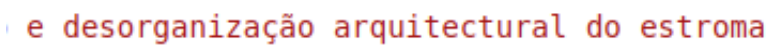



Figura 4.4: Excerto retirado de um registo do conjunto de treino

A expressão presente na Figura 4.4 foi retirada de um dos relatórios médicos presentes no conjunto de treino, e revela a presença do descritor ”Desorganização Arquitetural” nesse mesmo relatório. No entanto, a PVP não foi capaz de o detetar e então procedeu-se à respetiva verificação do motivo.

```
/\b(arquitetur|distor|desorganiza)\w*(?:\W+\w+){0,5}?\W+(arquitetur|desorganiz|distor)\w+\b/
```

Figura 4.5: ER do atributo ”Desorganização Arquitetural” na PVP

A expressão representada na Figura 4.5 foi a ER usada na gramática da PVP, para o descritor "Desorganização Arquitetural".

Após se consultar a expressão, foi possível perceber que o motivo pelo qual o *parser* não detetou o descritor foi porque não estava prevista a palavra ”arquitectura”. Ou seja, não foi previsto que existisse a letra ”c” na palavra, derivado do antigo acordo ortográfico. Foi então feita uma atualização na ER de modo a que este erro estivesse previsto e ainda se evitasse alguns outros possíveis. A ER atualizada esta representada na Figura 4.6.

```
/\b(arquite|distor|desorgani|estr?o|altera)\w*(?:\W+\w+){0,3}?\W+(arquite|desorgani|distor|(eco)?estr|seq|estr?o)\w+\b
```

Figura 4.6: ER do atributo ”Desorganização Arquitetural” na SVP

Através da expressão representada na Figura 4.6 é possível ver que na segunda versão, o *parser* deixou de procurar correspondência em palavras com o radical ”arquitetu” e passou a procurar correspondências em palavras com o radical ”arquite”. Isto permite obter correspondência positiva nas palavras ”arquitetura” e ”arquitectura”, entre outras.

Conclui-se então que uma das soluções possíveis para lidar com erros ortográficos é aplicar o processo de *stemming*, ou seja, reduzir o radical das palavras. Quanto menor for o radical, maior o número de palavras derivadas.

No entanto, reduzir o tamanho do radical nem sempre é uma solução eficaz.

```
Distorção do estoma.
```

Figura 4.7: Excerto presente num registo do conjunto de treino.

A Figura 4.7 mostra uma frase presente no relatório de um dos registos no conjunto de treino. Neste caso, a palavra ”estroma” está escrita de forma errada. Também foi um erro específico detetado em alguns registos. No entanto, reduzir o seu radical de modo a evitar este erro, não foi uma solução viável. Ao reduzir os seu radical para ”est” provocou correspondências positivas em casos em que a palavra é ”estudo”.

Perante este tipo de situações em que o erro é comum, optou-se por alterar a gramática de um modo específico que aceite o caso correto e o erro comum. Para este caso específico do erro ”estoma” alterou-se o seu radical para ”estr?o” como está representado na Figura 4.6.

Estas duas técnicas permitiram corrigir os erros mais comuns nos relatórios médicos presentes no conjunto de treino.

### 4.4.2 Casos não previstos

Os casos não previstos pela gramática da PVP são o tipo de casos em que os relatórios usam termos que não estão nessa gramática.

Esta é uma situação que já estava prevista, dado que a gramática da primeira versão é baseada apenas na tradução direta dos descritores em inglês para português e seus sinónimos.

```
com prisão da placa areolar
```

Figura 4.8: Expressão presente num registo do conjunto de treino

A Figura 4.8 representa um destes casos. O descritor representado é ”Repuxamento do Mamilo”. Como pode ser verificado na Tabela 4.2, ”placa areolar” não consta nas hipóteses do descritor ”Repuxamento do Mamilo”, e como tal não está presente na gramática da PVP.

Para solucionar esta situação, bastou adicionar ”placa areolar” à ER do descritor ”Repuxamento do Mamilo”. No entanto, ao modificar a expressão regular foi tido em conta que ”placa areolar” pode também ser representada simplesmente por ”aréola”, existindo assim por exemplo, uma maneira de descrever este descritor como ”repuxamento da aréola”. Com base nesta informação alterou-se a ER de modo a aceitar as duas versões. A expressão está representada na Figura 4.9.

```
/\b(retra(c)?|repuxam|fixa|invers|pris)\w*(?:\W+\w+){0,3}?\W+(mami|plac|ar[e|é]ol)\w+\b/
```

Figura 4.9: ER para o descritor ”Repuxamento do Mamilo” na SVP

Como pode ser constatado, o radical presente na Figura 4.9, ”ar[e|é]ola’, permite obter uma correspondência positiva nas palavras ”areolar” ou ”aréola”.

Um dos descritores que sofreu mais alterações foi o descritor ”Distorção Arquitetural”. É um descritor que nos relatórios médicos, é redigido de várias formas diferentes. As Figuras 4.5 e 4.6, apresentadas anteriormente, representam respetivamente as ERs deste descritor na primeira e na segunda versão do *parser*. É possível perceber que ao primeiro conjunto de radicais, foram adicionados 2 radicais e ao segundo foram adicionados 3. Foi necessário fazer estas adições porque este descritor foi referido de várias formas distintas. ”Distorção sequelar”, ”Alteração ecoestrutural” e ”Distorção do estroma” representam algumas dessas formas. A gramática da PVP não previu nenhum destes casos e portanto, foi necessário efetuar as respetivas alterações.

Com base neste tipo de situações, foi-se adaptando a gramática da primeira versão de modo a que se encaixasse na estrutura e no tipo de termos usados pelos especialistas ao redigirem os relatórios. Também foi a gramática usada para os descritores na Terceira Versão do Parser e pode ser consultada no Anexo B.

## 4.5 Terceira Versão do Parser

Após a conclusão da Segunda Versão do Parser, obteve-se uma gramática que cumpria os requisitos de correspondência para cada descritor BI-RADS. No entanto, não previa uma situação em que o descritor esteja a ser negado. O objetivo desta versão é corrigir esta situação, de modo a que o *parser* não detete um descritor quando, num relatório médico, a existência dele esteja a ser negada.

A Figura 4.10 representa uma negação do descritor ”Distorção Arquitetural”. Com base na ER deste descritor usada na SVP, e representada na Figura 4.6, verifica-se que é encontrada

```
Não se confirma área de desorganização do estroma, nomeadamente na glândula mamária esquerda.
```

Figura 4.10: Expressão presente num registo do conjunto de treino

uma correspondência positiva entre a ER e ”desorganização do estroma”. Isto constituí um erro dado que no início da frase está uma negação.

Para resolver as situações em que um descritor está a ser negado, foram consideradas três situações de negação:

- A palavra ”não” aparecer até seis palavras antes do descritor
- A palavra ”ausência” aparecer até cinco palavras antes do descritor
- As palavras ”sem” ou ”nem” aparecerem até duas palavras antes do descritor

Estas três situações deram origem à expressão representada na Figura 4.11

```
/\b(n(a|ã)o(?:\W+\w+){0,6}?\W+|aus(ê|e)ncias?(?:\W+\w+){0,5}?\W+|(sem|nem)(?:\W+\w+){0,2}?\W+)\b/
```

Figura 4.11: Expressão regular para a negação de um descritor

Através da ER representada na figura anterior, é possível constatar que esta obtém uma correspondência positiva nas palavras ”não” e ”ausência”, mesmo que estas estejam escritas sem o acento.

Sempre que um descritor obtiver uma correspondência positiva num relatório, é feita uma verificação se a ER da Figura 4.11, junto com a ER do descritor, obtém uma correspondência negativa no relatório em questão. Caso seja obtida uma correspondência negativa, é então dada como positiva a presença do descritor no relatório. Caso contrário, o *parser* assume que o descritor está a ser negado e a sua presença no relatório é dada como negativa.

Após a conclusão da expressão representada na Figura 4.11 estar concluída, ficou também concluída a gramática e o *parser* ficou concluído.

## 4.6 Metodologia de Avaliação

Para se proceder à avaliação dos resultados e do desempenho do parser desenvolvido foi utilizada uma matriz de confusão, já mencionada no Capítulo 2 e representada na Tabela 2.4.

Foi portanto criada, para cada versão do *parser*, uma tabela como a representada na Tabela 4.3 onde:

- "*Parser* - P" representa o número de casos que o *parser* classificou o descritor como sendo positivo
- "*Parser* - N" representa o número de casos que o *parser* classificou o descritor como sendo negativo
- "Inês - P" representa o número de casos em que a especialista Inês classificou o descritor como sendo positivo
- "Inês - N" representa o número de casos em que a especialista Inês classificou o descritor como sendo negativo
- VP representa o número de casos em que tanto o *parser* como a médica especialista classificaram o descritor como positivo
- FP representa o número de casos em que o *parser* classificou o descritor como sendo positivo e a médica especialista classificou como sendo negativo
- FN representa o número de casos em que o *parser* classificou o descritor como sendo negativo e a médica classificou como sendo positivo
- VN representa o número de casos em que tanto o *parser* como a médica especialista classificaram o descritor como sendo negativo

| | Inês - P | Inês - N |
|---|---|---|
| *Parser* - P | **VP** | **FP** |
| *Parser* - N | **FN** | **VN** |

Tabela 4.3: Matriz de confusão usada na classificação do *parser*

Na extração de um descritor, um erro pode trazer grandes consequências:

- Se um descritor positivo não for detetado, uma pessoa pode estar a ser classificada como não doente, mesmo estando doente.
- Se um descritor negativo for detetado, podemos submeter uma pessoa a tratamentos que não necessita.

Com base nestes casos, foi considerado que é tão importante classificar uma instância negativa corretamente, como uma positiva. Com base neste fator a métrica escolhida para classificar o desempenho do *parser* foi a *Accuracy*. Esta métrica já foi mencionada no Capítulo 2 e é calculada através da fórmula:

$$\frac{VP + VN}{VP + FP + FN + VN}$$

Para cada uma das versões serão apresentadas a sua matriz de confusão e uma tabela com a descrição do tipo de erro. Este processo será efetuado apenas no conjunto de treino, dado que não se verificou os registos do conjunto de teste. Por fim será apresentada a *Accuracy* para cada versão do *parser*.

No próximo capítulo são apresentados os resultados obtidos em cada versão e a sua respetiva discussão.

# Capítulo 5

# Resultados e Discussão

Neste capítulo é feita uma descrição do *output* produzido pelo *parser* e são apresentados os resultados relativos a cada uma das suas versões. É feita também uma discussão sobre o significado desses resultados.

## 5.1 *Output* do *parser*

Em cada versão do *parser* são produzidos três ficheiros de *output*:

- `parser_resultados.csv`
- `parser_matriz.csv`
- `parser_resumo.txt`

O ficheiro `parser_resultados.csv` é um ficheiro que contém as informações dos descritores detetados em cada registo. Para cada registo é apresentado o "ID" e os descritores achados. É um ficheiro com uma estrutura igual ao ficheiro `resultados_ines.csv`, mencionado no Capítulo 3 e representado na Figura 3.5.

O ficheiro `parser_matriz.csv` junta as informações dos descritores encontrados pelo *parser* e pela especialista Inês. Constitui uma matriz onde cada descritor está dividido em duas colunas, representando cada uma dessas colunas, a avaliação do *parser* e a avaliação da especialista Inês para esse descritor. Está representado na Figura 5.1.

| Descritores | MaShIr | MaShIr | MaMaCi | MaMaCi | MaMaOb | MaMaOb |
|---|---|---|---|---|---|---|
| Relatório | João | Ines | João | Ines | João | Ines |
| 154663 | 0 | 0 | 0 | 0 | 0 | 0 |
| 170010 | 0 | 0 | 0 | 0 | 0 | 0 |
| 175893 | 0 | 0 | 0 | 0 | 0 | 0 |
| 176027 | 0 | 0 | 0 | 0 | 0 | 0 |
| 197482 | 0 | 0 | 0 | 0 | 0 | 0 |
| 201087 | 0 | 0 | 1 | 0 | 0 | 0 |
| 212119 | 0 | 0 | 1 | 1 | 0 | 0 |
| 223047 | 0 | 0 | 0 | 0 | 0 | 0 |
| 240155 | 0 | 0 | 0 | 0 | 0 | 0 |
| 250738 | 0 | 0 | 0 | 0 | 0 | 0 |
| 252099 | 0 | 0 | 0 | 0 | 0 | 0 |
| 252099 | 0 | 0 | 0 | 0 | 0 | 0 |
| 260020 | 0 | 0 | 0 | 0 | 0 | 0 |
| 261150 | 0 | 0 | 1 | 0 | 0 | 0 |
| 261735 | 0 | 0 | 0 | 0 | 0 | 0 |
| 265306 | 0 | 0 | 0 | 0 | 0 | 0 |

Figura 5.1: Ficheiro `parser_matriz.csv`

Através da matriz representada na Figura 5.1, conclui-se que para o relatório com "ID" igual a "201087", o *parser* detetou o descritor "MaMaCi"(Massas:Forma:Circunscrita) como sendo positivo, ao contrário da especialista Inês, que não detetou a presença do descritor no relatório. Já no relatório com "ID" igual a "212119", tanto o *parser* como a especialista Inês deram o descritor como presente.

O último ficheiro gerado pelo *parser* tem como objetivo apresentar os erros encontrados que precisam de ser corrigidos. Ou seja, apresenta os relatórios avaliados pelo *parser* que não tiveram os mesmos resultados que a avaliação feita pela especialista Inês. Além disso, no fim do ficheiro, apresenta os resultados totais da avaliação. O ficheiro está representado na Figura 5.2.

```
########################################
1046499
Falsos Positivos: ['CaTbPo']
########################################
1077725
Falsos Positivos: ['CaDiGr']
############ Totais ####################

Total de corretas: 4499
Total de erros: 16
Total de falsos negativos: 6
Total de falsos positivos: 10
Total de verdadeiros negativos: 4421
Total de verdadeiros positivos: 78
```

Figura 5.2: Ficheiro `parser_resumo.txt`

Através da Figura 5.2 retiram-se as seguintes informações:

- Para o relatório com "ID" igual a "1046499" foi detetado incorretamente o descritor "CaTbPo" que representa o descritor "Calcificações Benignas: Grosseiras"
- Para o relatório com "ID" igual a "1077725" foi detetado incorretamente o descritor "CaDiGr" que representa o descritor "Calcificações Benignas: Agrupadas"
- Em 6 situações o *parser* não detetou um descritor que está presente no relatório
- Em 10 situações o *parser* detetou um descritor que não está presente no relatório
- Em 4421 situações o *parser* não detetou um descritor que não está presente no relatório
- Em 78 situações o *parser* detetou um descritor que está presente no relatório

Os resultados produzidos por cada versão do *parser* foram retirados do ficheiro `parser_resumo.txt`

## 5.2 Resultados Primeira Versão do Parser (PVP)

Os resultados obtidos na classificação da PVP no conjunto de treino e no conjunto de teste estão representados na Figura 5.3 e Figura 5.4 respetivamente.

```
#############  Totais #############

Total de corretas: 4464
Total de erros: 51
Total de falsos negativos: 37
Total de falsos positivos: 14
Total de true negativos: 4417
Total de true positivos: 47
```

Figura 5.3: Resultados da PVP no conjunto de treino

```
############ Totais ###########

Total de corretas: 1904
Total de erros: 31
Total de falsos negativos: 23
Total de falsos positivos: 8
Total de true negativos: 1893
Total de true positivos: 11
```

Figura 5.4: Resultados da PVP no conjunto de teste

Com base nos dados apresentados nas Figuras 5.3 e 5.4 foi construída a respetiva matriz de confusão que está representada na Tabela 5.1.

| | **Treino** | | **Teste** | |
|---|---|---|---|---|
| | **I - P** | **I - N** | **I - P** | **I - N** |
| **E - P** | 47 | 14 | 11 | 8 |
| **E - N** | 37 | 4417 | 23 | 1893 |

Tabela 5.1: Matriz de confusão da PVP

A classificação do tipo de erro obtido no conjunto de treino da PVP está representada na Tabela 5.2.

| **Tipo de Erro** | **Total de Casos** |
|---|---|
| Erro ortográfico | 6 |
| Gramática não previa caso específico | 44 |
| Descritores negados | 1 |

Tabela 5.2: Erros encontrados para a PVP no conjunto de treino

Os dados apresentados nas Tabelas 5.1 e 5.2 revelam que a primeira versão classificou erradamente 51 descritores para o conjunto de treino, e 31 descritores para o conjunto de teste. Foram ainda classificados 6 descritores errados como sendo provenientes de um erro ortográfico, 44 descritores errados como sendo provenientes de uma combinação gramatical não presente nas expressões regulares, e 1 descritor errado como sendo proveniente de um caso em que o descritor está a ser negado.

Por fim concluiu-se que a Primeira Versão do Parser obteve uma *Accuracy* de 0,989 para o conjunto de treino e 0,984 para o conjunto de teste.

## 5.3 Resultados Segunda Versão do Parser (SVP)

Os resultados obtidos na classificação da SVP, no conjunto de treino e no conjunto de teste, estão representados na Figura 5.5 e Figura 5.6 respetivamente.

```
############# Totais #############

Total de corretas: 4496
Total de erros: 19
Total de falsos negativos: 6
Total de falsos positivos: 13
Total de true negativos: 4418
Total de true positivos: 78
```

Figura 5.5: Resultados da SVP no conjunto de treino

```
############ Totais ############

Total de corretas: 1924
Total de erros: 11
Total de falsos negativos: 3
Total de falsos positivos: 8
Total de true negativos: 1893
Total de true positivos: 31
```

Figura 5.6: Resultados da SVP no conjunto de teste

Com base nos dados apresentados nas Figuras 5.5 e 5.6 foi construída a respetiva matriz de confusão que está representada na Tabela 5.3.

| | **Treino** | | **Teste** | |
|---|---|---|---|---|
| | **I - P** | **I - N** | **I - P** | **I - N** |
| **E - P** | 78 | 13 | 31 | 8 |
| **E - N** | 6 | 4418 | 3 | 1893 |

Tabela 5.3: Matriz de confusão da SVP

A classificação do tipo de erro obtido no conjunto de treino da SVP está representada na Tabela 5.4.

| **Tipo de Erro** | **Total de Casos** |
|---|---|
| Erro ortográfico | 0 |
| Gramática não previa caso específico | 16 |
| Descritores negados | 3 |

Tabela 5.4: Erros encontrados para a SVP no conjunto de treino

Os dados apresentados nas Tabelas 5.3 e 5.4 revelam que a segunda versão classificou erradamente 19 descritores para o conjunto de treino e 11 descritores para o conjunto de teste. Foram ainda classificados 0 descritores errados como sendo provenientes de um erro ortográfico, 16 descritores errados como sendo provenientes de uma combinação gramatical não presente nas expressões regulares, e 3 descritores errados como sendo provenientes de um caso em que o descritor está a ser negado.

Por fim concluiu-se que a SVP obteve uma *Accuracy* de 0,995 para o conjunto de treino e 0,994 para o conjunto de teste.

## 5.4 Resultados Terceira Versão do Parser (TVP)

Os resultados obtidos na classificação da Terceira Versão do Parser no conjunto de treino e no conjunto de teste, estão representados na Figura 5.7 e Figura 5.8, respetivamente.

```
############ Totais ################

Total de corretas: 4499
Total de erros: 16
Total de falsos negativos: 6
Total de falsos positivos: 10
Total de verdadeiros negativos: 4421
Total de verdadeiros positivos: 78
```

Figura 5.7: Resultados da TVP no conjunto de treino

```
############ Totais ############

Total de corretas: 1926
Total de erros: 9
Total de falsos negativos: 4
Total de falsos positivos: 5
Total de true negativos: 1896
Total de true positivos: 30
```

Figura 5.8: Resultados da TVP no conjunto de teste

Com base nos dados apresentados nas Figuras 5.7 e 5.8 foi construída a respetiva matriz de confusão que está representada na Tabela 5.5.

| | **Treino** | | **Teste** | |
|---|---|---|---|---|
| | **I - P** | **I - N** | **I - P** | **I - N** |
| **E - P** | 78 | 10 | 30 | 5 |
| **E - N** | 6 | 4421 | 4 | 1896 |

Tabela 5.5: Matriz de confusão da TVP

A classificação do tipo de erro obtido no conjunto de treino da TVP está representada na Tabela 5.6.

Os dados apresentados nas Tabelas 5.5 e 5.6 revelam que a terceira versão classificou erradamente 16 descritores para o conjunto de treino e 9 descritores para o conjunto de teste. Foram ainda classificados 0 descritores errados como sendo provenientes de um erro ortográfico, 16 descritores errados como sendo provenientes de uma combinação gramatical não presente nas expressões regulares e 0 descritores errados como sendo provenientes de um caso em que o descritor está a ser negado.

Por fim concluiu-se que a Terceira Versão do Parser obteve uma *Accuracy* de 0,996 para o conjunto de treino e 0,995 para o conjunto de teste.

| Tipo de Erro | Total de Casos |
|---|---|
| Erro ortográfico | 0 |
| Gramática não previa caso específico | 16 |
| Descritores negados | 0 |

Tabela 5.6: Erros encontrados para a TVP no conjunto de treino

A Tabela 5.7 faz um sumário dos resultados obtidos em cada versão do *parser*.

| | PVP | | SVP | | TVP | |
|---|---|---|---|---|---|---|
| | Treino | Teste | Treino | Teste | Treino | Teste |
| **VP** | 47 | 11 | 78 | 31 | 78 | 30 |
| **VN** | 4117 | 1893 | 4418 | 1893 | 4421 | 1896 |
| **FP** | 14 | 8 | 13 | 8 | 10 | 5 |
| **FN** | 37 | 23 | 6 | 3 | 6 | 4 |
| ***Accuracy*** | 0,989 | 0,984 | 0,995 | 0,994 | 0,996 | 0,995 |

Tabela 5.7: Sumário dos resultados para a PVP, SVP e TVP

## 5.5 Resultados após segunda avaliação

Após a conclusão do *parser* foram revistos os relatórios que continham descritores que o *parser* classificou erradamente, no conjunto de treino. Estes relatórios foram enviados de novo para a especialista Inês fazer uma segunda avaliação e verificar se os relatórios continham ou não os descritores que o *parser* classificou erradamente.

Com base nesta segunda avaliação da especialista Inês, os descritores dos respetivos relatórios foram atualizados e procedeu-se a uma nova classificação do *parser*. A Figura 5.9 apresenta os resultados obtidos nesta classificação.

```
############### Totais #################

Total de corretas: 4511
Total de erros: 4
Total de falsos negativos: 2
Total de falsos positivos: 2
Total de verdadeiros negativos: 4417
Total de verdadeiros positivos: 94
```

Figura 5.9: Resultados após segunda avaliação da especialista

A Tabela 5.8 faz um sumário dos resultados obtidos pela TVP antes e após a segunda avaliação da especialista, no conjunto de treino.

| | TVP Primeira Avaliação Especialista | TVP Segunda Avaliação Especialista |
|---|---|---|
| **VP** | 78 | 94 |
| **VN** | 4421 | 4417 |
| **FP** | 10 | 2 |
| **FN** | 6 | 2 |
| ***Accuracy*** | 0.996 | 0.999 |

Tabela 5.8: Resultados obtidos pela TVP antes e após segunda avaliação

## 5.6 Discussão

Através dos resultados obtidos na Primeira Versão do Parser, representados na Tabela 5.1 verifica-se que, no conjunto de treino, 73% dos descritores mal classificados são falsos negativos. Ou seja, são casos em que o *parser* classificou o descritor como sendo negativo, e a especialista Inês classificou como sendo positivo. Esta percentagem é justificada pelo facto da primeira gramática ser construída apenas traduzindo os descritores do inglês para português. Esta situação contribuiu para que a quantidade de falsos positivos fosse maior porque a gramática não previa certas expressões usadas pelos radiologistas como sendo um descritor.

A segunda Segunda Versão do Parser corrigiu grande parte dos problemas encontrados na primeira versão. Enquanto na primeira versão, no conjunto de treino, 51 descritores foram mal classificados, na segunda versão apenas se verificaram 19 erros. Como seria de esperar, a segunda versão reduziu a percentagem de falsos negativos de 73% para 32%. Esta redução é explicada com o facto de a gramática da segunda versão já incluir expressões usadas pelos radiologistas e presentes nos relatórios médicos. As Tabelas 5.2 e 5.4 mostram que o número de erros provenientes de casos que a gramática não prevê foi reduzido de 44 para 16. É possível ainda verificar-se que, no conjunto de treino, houve uma redução em 100% dos erros provenientes de casos em que o descritor contém um erro ortográfico. Já no conjunto de teste, o número de erros foi reduzido de 31 para 11.

Por fim, na Terceira Versão do Parser esperava-se uma redução do número de descritores mal classificados provenientes de uma negação. As Tabelas 5.4 e 5.6 mostram que na SVP existem 3 descritores mal classificados provenientes de uma negação. Já na terceira versão verifica-se que este número foi reduzido a 0. Verifica-se portanto, que no conjunto de treino, a Expressão Regular (ER) que representa a negação reduziu em 100% os erros provenientes de uma negação. Já no conjunto de teste, através das Tabelas 5.3 e 5.5 é possível verificar-se que que o número de erros total na segunda versão são 11 erros e na terceira versão são 9. Houve uma redução de 2 erros. No geral, pode concluir-se que a implementação da expressão regular negativa obteve

ótimos resultados.

Através da Tabela 5.8 é possível verificar-se que o *parser* apresentou uma redução de 75% dos seus erros, no conjunto de treino, após uma segunda avaliação dos relatórios por parte da especialista. Antes da segunda avaliação a TVP classificou erradamente 16 descritores e após a segunda avaliação classificou erradamente apenas 4. Estes resultados aumentaram a *Accuracy* da TVP de 0,996 para 0,999.

Por fim, é possível verificar-se que nenhuma versão apresentou, em nenhum dos conjuntos de dados, uma *Accuracy* inferior a 0,98. Pode concluir-se que o *parser*, nos conjuntos de dados utilizados, obteve um ótimo desempenho.

# Capítulo 6

# Conclusão

O cancro da mama é o cancro com maior taxa de incidência e de mortalidade no sexo feminino. Uma das formas mais eficazes de reduzir a sua taxa de mortalidade é a prevenção por rastreio. O tipo de exame mais comum e utilizado na deteção e prevenção do cancro da mama é a mamografia. Através da imagem gerada neste exame, o médico especialista elabora um relatório onde descreve o tipo de achados que verificou.

Para limitar a diversidade na descrição dos achados resultantes de uma mamografia, o American College of Radiology (ACR) desenvolveu um conjunto de termos, ou descritores, o léxico Breast Imaging Reporting and Data System (BI-RADS), para serem utilizados pelos médicos especialistas. O objetivo desta gramática, é homogeneizar o tipo de relatórios médicos resultantes de uma mamografia.

No entanto, existem especialistas que preferem utilizar sinónimos para alguns descritores do léxico BI-RADS. Muitos dos problemas seriam evitados se os especialistas pudessem preencher um formulário eletrónico onde selecionassem valores pré-preenchidos com os descritores. Contudo, o preenchimento de um formulário deste tipo pode atrapalhar o fluxo regular de trabalho, limitando o seu desempenho. A maior parte dos especialistas prefere ditar textos ou escrevê-los. A ferramenta desenvolvida no contexto desta dissertação e a sua baixa taxa de erros resolve este problema ao permitir que os especialistas continuem a poder escrever ou ditar os seus relatórios.

Atualmente o léxico BI-RADS está na sua 5ª edição e tem três variantes: mamografia, exame ultrassom e ressonância magnética. A variante da mamografia é composta por 53 descritores. Neste trabalho foi desenvolvida uma gramática com o objetivo de extrair os descritores presentes nos textos resultantes dos relatórios médicos.

O conjunto de dados utilizados representa um total de 150 relatórios médicos resultantes de uma mamografia e uma extração dos atributos BI-RADS de cada um desses relatórios, por parte de uma especialista. Os relatórios foram divididos num conjunto de treino e teste.

Na primeira avaliação foi usada uma gramática que consiste única e exclusivamente na tradução dos descritores BI-RADS da sua língua original para português, e introduzindo alguns

sinónimos. Os resultados foram comparados com os da especialista e foram verificadas algumas classificações erradas por parte da gramática desenvolvida. Verificou-se que a a limitação da gramática era a principal causa desses resultados.

Na segunda avaliação, a gramática foi alterada de modo a incluir termos em português usados frequentemente pelos especialistas para descrever certos descritores. Como seria de esperar, os resultados foram melhores que na avaliação anterior, tendo sido eliminados grande parte dos erros.

A terceira avaliação teve o objetivo de resolver os casos em que um descritor está presente nos textos mas está a ser negada a sua existência. Nesta avaliação usou-se a mesma gramática que na avaliação anterior e adicionada uma Expressão Regular (ER) que correspondesse a uma negação. Os resultados revelaram uma redução no número de más classificações. Foi ainda possível perceber que, no conjunto de treino, esta gramática reduziu os erros causados por uma negação do descritor em 100%. Na sua versão final, o *parser* conseguiu uma *Accuracy* entre os 0,995 e os 0,996.

Após uma segunda avaliação, por parte da especialista, dos relatórios que contém descritores que a Terceira Versão do Parser (TVP) classificou erradamente foi possível reduzir em 75% o número de erros e aumentar a *Accuracy* de 0,996 para 0,999, no conjunto de treino. Estes resultados demonstram que o *parser* detetou descritores que numa primeira avaliação não foram detetados pelo especialista, mostrando assim que ferramentas de extração automáticas não estão sujeitas a fatores externos como fadiga ou distração e que contribuem para o erro humano.

É ainda importante referir que os relatórios utilizados têm origem numa única fonte: o Centro Hospitalar Universitário de São João (CHUSJ). Desta forma, a gramática construída pode estar enviesada para os tipos de frases utilizadas por especialistas deste hospital. Seria interessante aplicar o *parser* aqui desenvolvido em relatórios de outras fontes.

Em suma, conclui-se que é possível obter uma ferramenta de extração automática de descritores BI-RADS, em português, com um elevado grau de confiança e que pode contribuir significativamente para a área do cancro da mama.

# Anexo A

## Gramática Primeira Versão do Parser (PVP)

```
# Masses Shape:
'\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+((ar)?(redon))\w+\b' # Round
'\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+(ova|ov(ó|o))\w+\b' # Oval
'\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,5}?\W+(irregu)\w+\b' # Irregular

# Masses Margins:
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+((bem )?
(circunsc|definid|limit|delimitad|regular))\w+\b' # Circumscribed
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+((obscur|escu|sem brilh))\w+\b' # Obscured
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+(microlobul)\w+\b' # Microlobulated
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+(irregu|indistin|mal (definid)\w*)\w*\b' #
Indistinc
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,5}?\W+(espicul)\w+\b' # Spiculated

# Masses Density:
'\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+(alta densi|hiperden)\w*\b' # High
'\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+((normal|m(é|e)dia) densi|isoden)\w*\b' #
Medium
'\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+((pouca|baixa) densi)\w*\b' # Low
'\b(((á|a)re|mass|assim|n(ó|o)du|regi)\w*)?(?:\W+\w+){0,5}?\W+(gordur)\w+\b' # Fat-containing

# Calcifications Tippically Bening:
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(pel|d(é|e)rmi)\w+\b' # Skin
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(vascular)\w+\b' # Vascular
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(grossei)\w+\b' # Coarse / Popcorn
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(bastone|bastão)\w*\b' # Large rod-like
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(redond|punctifor)\w+\b' # Round
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(perifer)\w+\b' # Rim
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(distr(ó|o)fi)\w+\b' # Dystrophic
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(ducta)\w+\b') # Milk
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(cicatrici|citoestea)\w+\b' # Suture

# Calcifications Suspicious Morphology:
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(am(ó|o)rfic)\w+\b' # Amorphous
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(((grosseir)\w+)? (hetero)\w+)\b'# Coarse heterogeneous
'\b(calcificaç(ão|ões))\w+(?:\W+\w+){0,5}?\W+(pleomórficas finas)\w+\b' # Fine pleomorphic
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(fina|lineares)\w+\b' # Fine linear

# Calcifications Distribution:
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(difus|dispers)\w+\b' # Diffuse
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(regi(o|a))\w+\b' # Regional
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(n(ú|u)cl|grupos|agrupa)\w+\b' # Grouped
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(fina|line)\w+\b' # Linear
'\b(calcificaç(ão|ões))(?:\W+\w+){0,5}?\W+(segment)\w+\b' # Segmental

# Architectural Distortion:
'\b(arquitetur|distor|desorganiza)\w*(?:\W+\w+){0,5}?\W+(arquitetur|desorganiz|distor)\w+\b'

# Asymetries:
'\b(assimetrias)\b') # Asymmetries
'\b(assimetr|(á|a)re|foc)\w+(?:\W+\w+){0,5}?\W+(foc|loc|densif|n(ó|o)du)\w+\b' # Focal
'\b(assimetr)\w+(?:\W+\w+){0,5}?\W+(glob|gera|mam(á|a))\w+\b' # Global
'\b(desenvol|assimetr)\w+(?:\W+\w+){0,5}?\W+(assimetr|desenvol)\w+\b' # Developing
```

```
# Intramammary lymph node
'\b(gângl)\w+(?:\W+\w+){0,5}?\W+(intramam(á|a))\w+\b'

# Skin lesion
'\b(les(ão|ões)|ulceraç(ão|ões))(?:\W+\w+){0,5}?\W+(pel|d(é|e)rmicas)\w+\b'

# Solitary dilacted duct
'\b(carcino|duct)\w+(?:\W+\w+){0,5}?\W+(duct|dilata)\w+\b'

# Associated Features
'\b(retra(c)?|repuxam|fixa|pris)\w*(?:\W+\w+){0,5}?\W+(pel|cutân|d(é|e)rmic)\w+\b' # Skin retraction
'\b(retra(c)?|repuxam|fixa|invers)\w*(?:\W+\w+){0,5}?\W+(mami)\w+\b' # Nipple Retraction
'\b(espessam|gross|pel)\w+(?:\W+\w+){0,5}?\W+
(gross|cutân|pel|espess)\w+\b|\b(edem|erite|masti)\w+\b' # Skin thickening
'\b(trabec|espessam)\w+(?:\W+\w+){0,5}?\W+(espess|trabec)\w+\b' # Trabecular thickening
'\b(adenopat|axil|adenomega|cavad|gângli)\w+(?:\W+\w+){0,5}?\W+(axil|adenop|cavad|positiv)\w+\b' #
Axillary adenopathy
```

# Anexo B

**Gramática Terceira Versão do Parser (TVP)**

```
# Masses Shape:
'\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,3}?\W+((ar)?(redon))\w+\b' # Round
'\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,3}?\W+(ova|ov(ó|o))\w+\b' # Oval
'\b(dens|mass|estru|assi|n(ó|o)du|quis|(á|a)re)\w+(?:\W+\w+){0,3}?\W+(irregu)\w+\b' # Irregular

# Masses Margins:
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,3}?\W+((bem )?(circunsc|bem[- ]?
definid|limit|delimitad|regular))\w+\b'# Circumscribed
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,3}?\W+((obscur|escu|sem brilh))\w+\b' # Obscured
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,3}?\W+(microlobul)\w+\b' # Microlobulated
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,3}?\W+(indistin|mal[- ]?(definid)\w*|irregular)\w*\b'
# Indistinc
'\b(marge|contor|bord|limit|n(ó|o)dul)\w*(?:\W+\w+){0,3}?\W+(espicul)\w+\b' # Spiculated

# Masses Density:
'\b((((á|a)re|mass|assim|n(ó|o)du|regi)\w*)(?:\W+\w+){0,3}?\W+)?(alta densi|hiperden)\w+\b' # High
'\b((((á|a)re|mass|assim|n(ó|o)du|regi)\w*)(?:\W+\w+){0,3}?\W+)?((normal|m(é|e)dia)
densi|isoden|homog(é|e))\w*\b' # Medium
'\b((((á|a)re|mass|assim|n(ó|o)du|regi|tén)\w*)(?:\W+\w+){0,3}?\W+)?((pouca|baixa) (densi)\w+|ténu\w*
(micro)?calci)\w*\b' # Low
'\b((((á|a)re|mass|assim|n(ó|o)du|regi)\w*)(?:\W+\w+){0,3}?\W+)?(gordur)\w+\b' # Fat-containing

# Calcifications Tippically Bening:
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(pele|d(é|e)rmi)\w+\b' # Skin
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(vascular)\w+\b' # Vascular
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(grossei)\w+\b' # Coarse / Popcorn
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(bastone|bastão)\w*\b' # Large rod-like
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+redond\w+|punct?\w+)\b' # Round
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(perifer)\w+\b' # Rim
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(distr(ó|o)fi)\w+\b' # Dystrophic
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(ducta)\w+\b' # Milk
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(perifer)\w+\b' # Rim
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(distr(ó|o)fi)\w+\b' # Dystrophic
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(ducta)\w+\b' # Milk
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(cicatrici|citoestea)\w+\b' # Suture

# Calcifications Suspicious Morphology:
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(am(ó|o)rf)\w+\b' # Amorphous
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(((grosseir)\w+)? (hetero)\w+)\b' # Coarse
heterogeneous
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(pleomórficas finas)\w+\b' # Fine pleomorphic
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(fina|lineares)\w+\b' # Fine linear

# Calcifications Distribution:
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(difus|dispers)\w+\b' # Diffuse
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(regi(o|a))\w+\b' # Regional
'\b(((micro)?(calcific\w+|calci\w*\.)|agrupa)\w*(?:\W+\w+){0,5}?\W+(n(ú|u)cl|grupos|agrupa|(micro)?
calcifi)\w+\b' # Grouped
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(fina|line)\w+\b' # Linear
'\b(((micro)?(calcific\w+|calci\w*\.))(?:\W+\w+){0,5}?\W+(segment)\w+\b' # Segmental

# Architectural Distortion:
'\b(arquitec?tur|distor|desorganiza|estr|altera)\w*(?:\W+\w+){0,3}?\W+(arquite|desorganiz|distor|(eco)?
estr|seq|estr?om)\w+\b'
```

```
# Asymetries:
'\b(assimetrias)\b' # Asymmetries
'\b(assim(e|é)tr|(á|a)re|foc)\w+(?:\W+\w+){0,3}?\W+(focal|local)\b' # Focal
'\b(assimetr|densifica)\w+(?:\W+\w+){0,3}?\W+(glob|gera|mam(á|a)|assim(é|e))\w+\b' # Global
'\b(desenvol|assimetr)\w+(?:\W+\w+){0,3}?\W+(assimetr|desenvol)\w+\b' # Developing

# Intramammary lymph node
'\b(gângl)\w+(?:\W+\w+){0,3}?\W+(intramam(á|a))\w+\b'

# Skin lesion
'\b(les(ão|ões)|ulceraç(ão|ões))(?:\W+\w+){0,3}?\W+(pel|d(é|e)rmicas)\w+\b'

# Solitary dilacted duct
'\b(carcino|duct)\w+(?:\W+\w+){0,3}?\W+(duct|dilata)\w+\b'

# Associated Features
'\b(retra(c)?|repuxam|fixa|pris)\w*(?:\W+\w+){0,3}?\W+(pel|cut(â|a)n|d(é|e)rmic)\w+\b' # Skin retraction
'\b(retra(c)?|repuxam|fixa|invers|pris)\w*(?:\W+\w+){0,3}?\W+(mami|plac|areol)\w+\b' # Nipple Retraction
'(\b(espessam|gross|pel)\w+(?:\W+\w+){0,3}?\W+
(gross|cut(â|a)n|pel|espess)\w+\b|\b(edem|erite|masti)\w+\b)' # Skin thickening
'\b(trabec|espessam)\w+(?:\W+\w+){0,3}?\W+(espess|trabec)\w+\b' # Trabecular thickening
'\b(adenopat|axil|adenomega|cavad|gângli)\w+(?:\W+\w+){0,3}?\W+
(axil|adenop|cavad|positiv|gângli)\w+\b' # Axillary adenopathy

# Negation
'\b(n(a|ã)o(?:\W+\w+){0,6}?\W+|aus(ê|e)ncias?(?:\W+\w+){0,5}?\W+|(sem|nem )(?:\W+\w+){0,2}?\W+)\b'
```

# Bibliografia


[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Sickles EA, D'Orsi CJ, and Bassett LW et al. *ACR BI-RADS ® Mammography. In: ACR BI-RADS ® Atlas, Breast Imaging Reporting and Data System.* American College of Radiology, Reston, VA, 2013.

[3] Pedro Ferreira. Aplicação de Algoritmos de Aprendizagem Automática para a Previsão de Cancro de Mama. Master's thesis, Faculdade de Ciências da Universidade do Porto, 2010.

[4] Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. https://gco.iarc.fr/today. Acedido em Setembro, 2021.

[5] Raina M Ferzoco and Kathryn J Ruddy. The epidemiology of male breast cancer. *Current oncology reports*, 18(1):1–6, 2016.

[6] Gonçalo Forjaz de Lacerda, Scott P Kelly, Joana Bastos, Clara Castro, Alexandra Mayer, Angela B Mariotto, and William F Anderson. Breast cancer in Portugal: Temporal trends and age-specific incidence by geographic regions. *Cancer epidemiology*, 54:12–18, 2018.

[7] Robert A Smith, Kimberly S Andrews, Durado Brooks, Stacey A Fedewa, Deana Manassaram-Baptiste, Debbie Saslow, and Richard C Wender. Cancer screening in the United States, 2019: a review of current American Cancer Society guidelines and current issues in cancer screening. *CA: a cancer journal for clinicians*, 69(3):184–210, 2019.

[8] Debra L Monticciolo, Mary S Newell, R Edward Hendrick, Mark A Helvie, Linda Moy, Barbara Monsees, Daniel B Kopans, Peter R Eby, and Edward A Sickles. Breast cancer screening for average-risk women: recommendations from the ACR commission on breast imaging. *Journal of the American College of Radiology*, 14(9):1137–1143, 2017.

[9] American College of Radiology. ACR BI-RADS® Atlas Fifth Edition: QUICK REFERENCE. https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-Poster.pdf, . Acedido em Janeiro,2021.

[10] American College of Radiology. BI_RADSV5Changes. https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-V5-Changes.pdf, . Acedido em Janeiro,2021.

[11] David Allen Spak, JS Plaxco, L Santiago, MJ Dryden, and BE Dogan. Bi-rads® fifth edition: A summary of changes. *Diagnostic and interventional imaging*, 98(3):179–190, 2017.

[12] Pedro Ferreira, Nuno A Fonseca, Inês Dutra, Ryan Woods, and Elizabeth Burnside. Predicting malignancy from mammography findings and image–guided core biopsies. *International journal of data mining and bioinformatics*, 11(3):257–276, 2015.

[13] Ricardo Sousa Rocha, Pedro Ferreira, Inês Dutra, Ricardo Correia, Rogerio Salvini, and Elizabeth Burnside. A speech-to-text interface for mammoclass. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE, 2016.

[14] MamoClass V2. https://mammoclass.dcc.fc.up.pt/.

[15] Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*. Springer, 2015.

[16] Suad A Alasadi and Wesam S Bhaya. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16):4102–4107, 2017.

[17] Sanjay Yadav and Sanyam Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016.

[18] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[19] Félix López and Víctor Romero. *Mastering python regular expressions*. Packt Publishing Ltd, 2014.

[20] Stephen Cole Kleene. *Representation of events in nerve nets and finite automata*. Princeton University Press, 2016.

[21] Yunyao Li, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and HV Jagadish. Regular expression learning for information extraction. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 21–30, 2008.

[22] Houssam Nassif, Ryan Woods, Elizabeth Burnside, Mehmet Ayvaci, Jude Shavlik, and David Page. Information extraction for clinical data mining: a mammography case study. In *2009 IEEE International Conference on Data Mining Workshops*, pages 37–42. IEEE, 2009.

[23] Hongyuan Gao, Erin J Aiello Bowles, David Carrell, and Diana SM Buist. Using natural language processing to extract mammographic findings. *Journal of biomedical informatics*, 54:77–84, 2015.

[24] Selen Bozkurt, Francisco Gimenez, Elizabeth S Burnside, Kemal H Gulkesen, and Daniel L Rubin. Using automatically extracted information from mammography reports for decision-support. *Journal of biomedical informatics*, 62:224–231, 2016.

[25] Shumei Miao, Tingyu Xu, Yonghui Wu, Hui Xie, Jingqi Wang, Shenqi Jing, Yaoyun Zhang, Xiaoliang Zhang, Yinshuang Yang, Xin Zhang, et al. Extraction of bi-rads findings from breast ultrasound reports in chinese using deep learning approaches. *International journal of medical informatics*, 119:17–21, 2018.

[26] Filipe Cunha. Extração de Atributos de Textos Clínicos Sobre Mamografias. Master's thesis, Faculdade de Ciências da Universidade do Porto, 2011.

[27] Ajay Aroor Rao, Jennifer Feneis, Chloe Lalonde, and Haydee Ojeda-Fournier. A pictorial review of changes in the BI-RADS fifth edition. *Radiographics*, 36(3):623–639, 2016.

[28] Jeffrey EF Friedl. *Mastering regular expressions*. "O'Reilly Media, Inc.", 2006.

[29] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.