FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Analysis of usage patterns of medical image exams in medical environments

Nuno Martins Marques Pinto

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Luís Filipe Pinto de Almeida Teixeira

Co-Supervisor: Vera Lucia Miguéis Oliveira e Silva

Company Supervisor: António Cardoso Martins

July 23, 2019

Analysis of usage patterns of medical image exams in medical environments

Nuno Martins Marques Pinto

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Carlos Manuel Milheiro de Oliveira Pinto Soares
External Examiner: Inês de Castro Dutra
Supervisor: Luís Filipe Pinto de Almeida Teixeira
July 23, 2019

Abstract

The increasing technological developments, the significant increase in healthcare data and the lack of knowledge regarding such data, have raised serious concerns regarding medical examinations' management.

There are three distinct information systems, Hospital Information System, Radiology Information System and Picture Archive and Communications System (PACS) in a healthcare organization. In the standard workflow, after the image acquisition by the imaging modality, the image is sent and stored in PACS where it can be remotely accessed by any authorized medical staff. In Portuguese public hospitals there is not much information regarding to which cases the exams are consulted, by which doctors nor their specialties, neither for any specific purposes. The lack of information and analytic tools related to these issues are among the major concerns of the Portuguese health system stakeholders.

These problems can be tackled with the implementation of systems such as Business Intelligence allied with analytic techniques - cluster analysis and mining association rules. Therefore, in this work we intend to obtain groups of medical services of a hospital, according to its accesses to examinations' pattern, during a 112-day period. These groups were obtained by performing a cluster analysis on the data set of similar medical services, with the clusters corresponding to the targeted segments. In addition to that, association rules mining were also performed in order to uncover associations between the multiple examinations' characteristics, the specific medical services that access them and their visualization patterns.

In the end, this study enables to obtain homogeneous segments, which refer to the usage patterns of the medical services, and association rules that translate the relationships among the different components.

Resumo

Com a crescente evolução tecnológica, o aumento significativo da quantidade de dados relativos à saúde e a falta de conhecimento sobre os mesmos, criam-se sérias preocupações relativamente à gestão de exames médicos. Assim, devido a este crescimento a análise desses mesmos dados tornou-se ainda mais complexa e de difícil execução.

Num sistema hospitalar existem três sistemas de informação distintos, designadamente, o *Hospital Information System*, o *Radiology Information System* e o *Picture Archive and Communications System* (PACS). No *workflow* tradicional, após a aquisição de imagens por parte da modalidade, estas são enviadas e armazenadas no PACS onde podem ser acedidas remotamente por pessoal médico devidamente autorizado. Nos hospitais públicos portugueses existe pouca informação referente em que casos os exames são ou não consultados, por que médicos, de que especialidades e com que intuito. A falta de informação e de ferramentas de suporte relativas a esta problemática estão entre as principais preocupações dos intervenientes do sistema de saúde português.

Estes problemas podem ser colmatados com a implementação de sistemas, tal como *Business Intelligence*, aliados a métodos analíticos - análise *clustering* e regras de associação.

Posto isto, neste trabalho pretende-se obter uma segmentação de um conjunto de serviços médicos de um hospital, de acordo com o seu padrão de acesso às imagens médicas dos exames, durante um periodo de 112 dias. Esta segmentação foi obtida efetuando uma análise cluster no conjunto de serviços médicos em estudo, correspondendo os clusters aos segmentados pretendidos. Para além disso, efetuou-se também uma exploração das regras de associação, com o objetivo de encontrar associações entre as variadas características dos exames médicos, os serviços médicos a que os acedem e os seus padrões de visualização. Assim, estas estratégias permitiram-nos obter segmentos homogéneos que traduzem os perfis de utilização dos múltiplos serviços médicos, e regras de associação que traduzem as relações entre as mais variadas componentes.

Acknowledgements

Always in first place, would like to dedicate this dissertation to all my family for supporting my moody days and nights, my rights and wrongs, and for teaching me every single day that at the end of the road through all the adversity, if you can get where you wanted to be, you remember whatever doesn't kill you make you stronger, and all of the adversity was worth it.

To all my loyal and longtime friends, a huge thanks for bringing joy, happiness and childish moments to my life.

A very special thanks to my supervisors Luís Teixeira and Vera Miguéis, for the incredible guidance, partnership and availability. You were flawless.

Last but not least, I would like to express my sincere gratitude to António Martins and Carlos Cardoso for allowing me to be part of such an amazing project and team in an extraordinary environment. Also, a huge thanks to my Sectra colleagues for every single laugh and lesson.

Nuno Pinto

"Never stop learning, never stop grinding, never stop loving every single minute of your life"

Mark Cuban

Contents

List of Figures

List of Tables

Abbreviations

AD	Active Directory
BI	Business Intelligence
BFS	Breadth-First Search
CHUSJ	Centro Hospitalar Universitário São João
CT	Computed Tomography
DFS	Depth-First Search
DICOM	Digital Imaging and Communication in Medicine
DM	Data Mining
DW	Data Warehouse
EHR	Electronic Health Record
ETL	Extract Transform and Load
ETL	Extract, Transform and Load
FK	Foreign Key
GDPR	General Data Protection Regulation
HL7	Health Level 7
HIPAA	Health Insurance Portability and Accountability Act
HIS	Hospital Information System
HIT	Health Information Technology
KPI	Key Performance Indicator
MRI	Magnetic Resonance Imaging
PACS	Picture Archive and Communication System
PHI	Protected Health Information
PK	Primary Key
RIS	Radiology Information System

Chapter 1

Introduction

1.1 Context and Problem

With the lightning fast technologies and its constant evolution, health information systems and healthcare data have been massively growing over the past years. As per Lloyd et al. [Min17], from 2013 to 2020 healthcare data is expected to grow 48 percent annually reaching more than 2 000 exabytes (2 000 000 000 terabytes) in 2020. However, due to this growth, the extraction, analysis and understanding of healthcare data has become an even more complex task.

Each year more than 6 million exams based on medical imaging are made for diagnostic purposes in Portuguese public hospitals [dS18]. Still, there is a huge lack of knowledge of their utilization patterns.

It is expected each exam to be consulted, after the acquisition of the medical imaging, by a radiologist that performs the diagnosis and elaborates the respective report, followed by another doctor's consulting - typically a clinician (e.g. an orthopedist or a pediatrician). In Portugal, there is not much information regarding in which cases are the exams consulted, by which doctors, from what specialties and for what purpose. On the other hand, we have exams that are potentially consulted by non-medical staff, are never consulted and/or do not have a report.

In the Radiology department, we can distinguish two different information systems: Radiology Information System (RIS) is used for scheduling, patient tracking, results reporting and image tracking; Picture Archiving and Communication System (PACS) is used for storing and transmit medical images and reports. As previously stated, the data available in these two information systems has been significantly increasing, but there are no implemented processes capable of interpreting such data and turn it into meaningful data to help the decision making, which nowadays is considered a very valuable tool amongst hospital's management personnel.

Several studies ran by Yichuan et al. [WKB18], Osama et al. [ACJ16] and Ishola et al. [Mur12] show that data analytic solutions, such as Business Intelligence (BI), effectively provide operational and strategic improvements in healthcare. Along with it, a BI implementation allow the

final users to access meaningful and useful information through simple and interactive data presentation (dashboards, scorecards, charts, graphics and so on) [Zhe17].

As stated, the existence of large amounts of healthcare data brings new challenges. On one hand, in order to keep track of when exams are accessed and who accesses them, there is the need to create a data warehouse (DW). On the other hand, there is the need to identify utilization patterns of the exams, by grouping and segmenting them according to meaningful variables. Although there are multiple analytical techniques used for data segmentation, cluster analysis remains the most common and the most used procedure, being vastly used in utilization patterns' identification and analysis [MP87, WQK+17, AFHC13]. Mining association rules also play an important role in data mining, being used for undercover relationships between data items within large data sets [AIS93, AS94].

This work is made in cooperation with Sectra[1] and Centro Hospitalar Universitário São João (CHUSJ)[2]. Sectra is involved in the markets of healthcare and cybersecurity, being mostly known for development, deployment and support of imaging IT solutions. The main theme of the dissertation was proposed by the company which, in turn, arose from the need to provide a solution to the problem of the insufficient healthcare data analysis in Portuguese public hospitals.

1.2 Motivation and Objectives

Our main purpose in this dissertation is to try to tackle the challenges described in the previous section 1.1 applying a custom BI framework, based on the standard one - Data Sources, Data Acquisition, Data Storage, Data Analysis, Data Visualization -, and applying data analysis techniques, clustering and mining association rules, to obtain exams' visualization patterns. More specifically, develop an accurate system capable of:

1. Extracting data (structured, semi-structured and unstructured data) correctly from specific sources;
2. Processing heterogeneous data;
3. Anonymizing data classified as Protected Health Information (PHI);
4. Storing processed data in a proper data warehouse;
5. Segmenting data by applying cluster analysis;
6. Identifying relationships between data objects using association rules mining;
7. Interpreting and displaying analyzed data in several formats, promoting an interactive visualization.

[1] https://www.sectra.com/investor/about.html
[2] http://portal-chsj.min-saude.pt/

To the best of our knowledge, there are no records of any study or work focusing on the utilization patterns of exams based on medical imaging. This dissertation has also the goals to successfully complete development, deployment and validation of this system.

1.3 Dissertation Structure

In addition to this chapter, the dissertation is composed by five other chapters:

- **Chapter 2** introduces the theoretical aspects of all the subjects relative to this dissertation. This chapter starts with a brief history about the evolution of the healthcare information systems until the present, then a review of the state-of-the-art of Business Intelligence, cluster analysis and the current data protection regulations (General Data Protection Regulation).
- **Chapter 3** presents the methodology involved in the dissertation, including the proposed solution and its subsequent implementation details;
- **Chapter 4** describes and discusses the results obtained from our implemented cluster analysis;
- **Chapter 5** describes and discusses the results obtained from association rules mining;
- **Chapter 6** summarizes the results of the whole dissertation and concludes with improvements and possible future development.

Chapter 2

Background Knowledge

In this chapter different subjects related to the dissertation and to the implementation are addressed. The first step is to understand which are the different healthcare information systems, how they communicate and what information is stored in each system. These systems are going to be the main data sources. Subsequently also to understand which of these systems are implemented in Portuguese public hospitals.

The second step is to understand the Business Intelligence architecture, its framework and how the integration with the information systems can be made. Since dealing with healthcare data which can be considered sensitive - as it may directly or indirectly identify an individual -, data anonymization is required under the General Data Protection Regulation (GDPR).

Lastly, a review of the data analysis techniques used is also presented.

2.1 Health Information Systems

Understand and distinguish the different data sources to be used, what information can be extracted and how to extract from each source, is an essential first step of the project. Throughout this analysis it is possible to identify three main systems: Hospital Information System (HIS); Radiology Information System (RIS); Picture Archive and Communication System (PACS). Each system has different purposes although the same data can be found in more than one system, in result of the systems' interaction.

In order to integrate and create interoperability between different health related systems, two protocols and standards were created. Digital Imaging and Communications in Medicine (DICOM)[1] and Health Level 7 (HL7)[2] are the communication standards used in health information's exchange. DICOM is used to transmit, store, process and display medical imaging information over a network (e.g. transfer Computed Tomography (CT) images to a remote storage), and HL7

[1] https://www.dicomstandard.org/

[2] http://www.hl7.org/implement/standards/index.cfm?ref=nav

is used to exchange, integrate, share and retrieve electronic health information (e.g. patient data) between medical software. While DICOM increases interoperability, HL7 decreases the incompatibilities amongst the different medical applications. Both standards are the basis of the informational integration of software based medical processes [COV+10, RA15].

HIS is the main management system in a hospital. It is responsible for administrative and financial management, stock management (e.g. medicines) and store Electronic Health Record (EHR) of patients, which contains patient personal data and patient medical history. Generally, the HIS communicates with RIS to share patient related data. In this project, data extraction from HIS will not occur.

RIS, like the name suggests, is used by the Radiology department to better manage processes and improve the workflow. RIS allows staff to keep track of patient's workflow (e.g. patient doing a Magnetic Resonance Imaging (MRI)), scheduling appointments, do clinical reports, although nowadays it is usually done in PACS, and image management. Its advantage prevails on keeping data easily accessible turning into a more efficient workflow management [NMN13]. RIS obtains data related to patient from HIS via HL7 messages, and sends that information to the imaging modalities (e.g. CT scan) via DICOM when required.

Moreover, PACS is responsible for medical imaging storage and also has the ability to provide remote access to those images. PACS receives images from imaging modalities and the information related to patient, previously sent by RIS to the modality, and stores them in a virtual storage, via DICOM [O'C17, RA15]. In short, as pointed out by Carter et al. [CV10], a PACS serves as the file room, reading room, duplicator and courier, providing image access to multiple users at the same time. Therefore, in PACS data is stored relative to patients, exams, images, medical staff and so on. PACS will be the main data source of this project.

In summary, to better understand the schedule workflow in a radiology department, figure 2.1 describes the typical workflow since patient examination's schedule until the exam's consulting by medical staff. The patient is registered, in HIS, and an examination requisition is made. This requisition is passed to RIS, where the exam's confirmation is made and it is scheduled to a specific date. Then, when performing the examination, the images are acquired by the medical imaging modality and sent to PACS. As known, PACS will provide medical staff remote access to the images, allowing radiologist to analyze the images, make the diagnosis and write the respective report. The report will then be gathered with the patient's EHR, being able to be consulted by any authorized medical staff.

2.2 Business Intelligence

Business Intelligence has been implemented for several purposes and it brought up significant improvements in healthcare. From providing medical staff tools for decision making, to providing clinical Key Performance Indicators (KPI), BI has a wide range of applications combining data gathering, data storage, data integration and knowledge management with analytical tools to

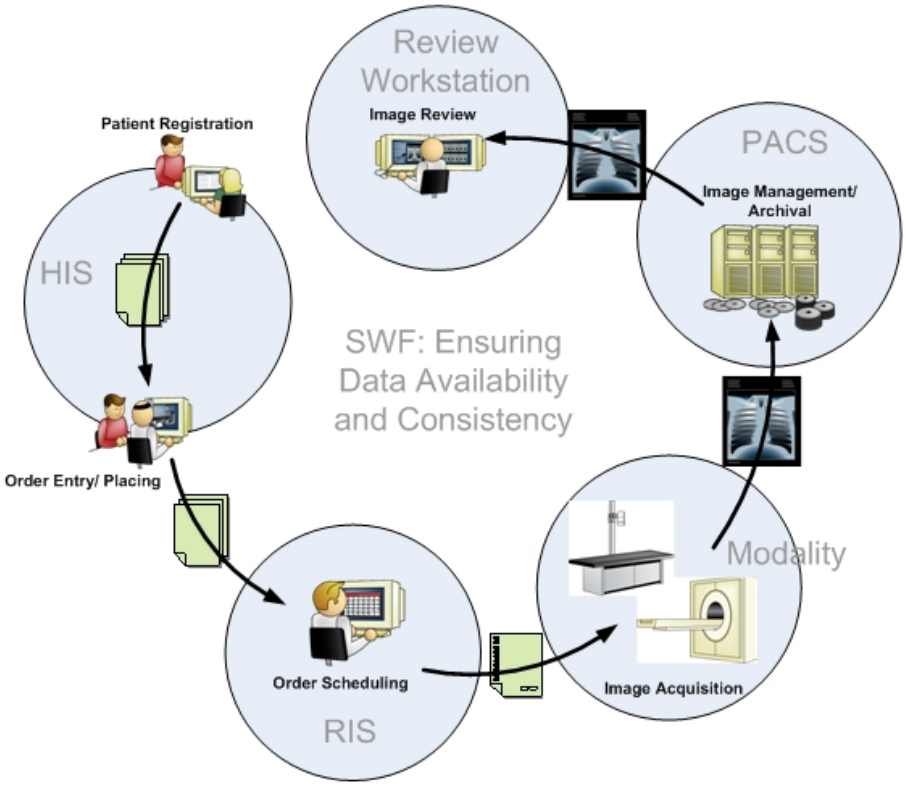


Figure 2.1: Main workflow in a radiology department [tHE]

display key information for planners and decision makers [You18]. Furthermore, BI follows a specific framework, as illustrated in figure 2.2, that includes data management, analysis, presentation and delivery.

As pointed out by Sang [You18], Ali et al. [AÖNC13] and Mark Peco [Pec11], BI applications in healthcare can be categorized in two major sets of solutions, i.e. Business Solutions and Technology Solutions.

Technology Solutions, consists on Data & Information Tools, are as follows:

- Decision Support Systems (DSS): Support managerial decision making, usually day-to-day tactical.
- Executive Information Systems: Support decision making at the senior management level which provide and consolidate metrics-based performance information.
- Online Analytical Processing (OLAP): Support analysts with the capability of performing multidimensional analysis of data.
- Query and Reporting Services: Provide quick and easy access to the data with predefined report design capabilities.
- Data Mining (Predictive Model as explained by Ali et al. [NCH11, NHC13, NCH12]): Examines data to discover hidden facts in databases using different techniques.

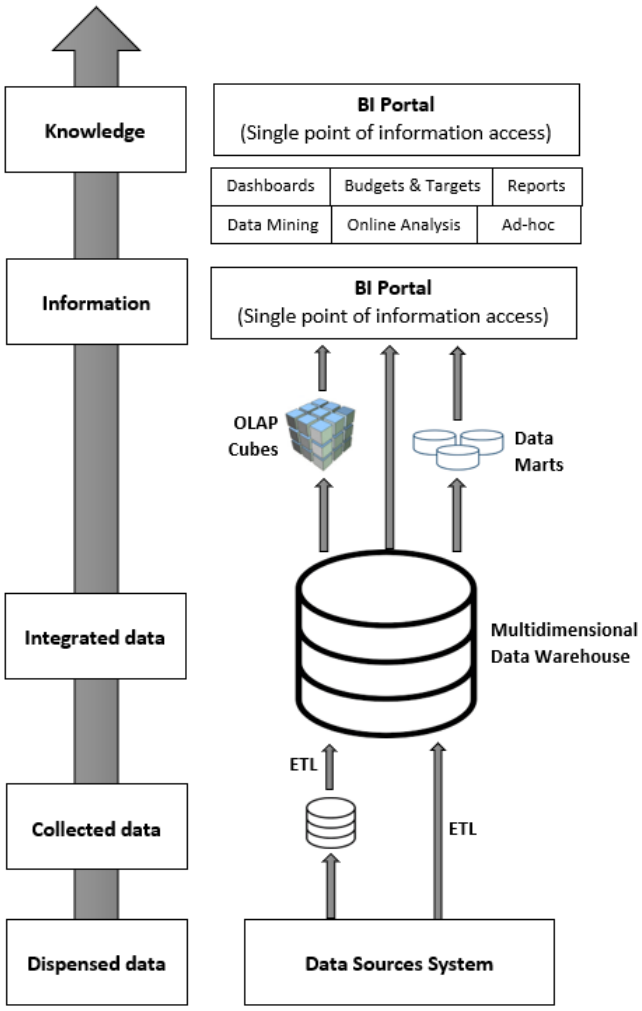


Figure 2.2: Basic architecture BI framework

- Operational Data Services: Collect data from end users, organize data, establish solid data structures and store them in different databases, retrieve data from multiple databases.

Business Solutions consists on business focused analytical applications, as follows:

- Patient Analysis: Focuses on analysis of patients' demographic and satisfaction processes.
- Electronic Health Record Analysis: Focuses on the analysis of the quality of clinical data.
- Performance Analysis: Streamline and optimize the way that a business uses its resources.
- Fund Channel Analysis: Devise, implement, and evaluate fund strategies, then use the corporate metrics to continuously monitor and enhance the fund process.
- Productivity Analysis: Focuses on building business metrics for activities such as quality improvement, risk mitigation, asset management, capacity planning, etc.

- Behavioural Analysis: Understanding and predicting trends and patterns that provides business advantage.
- Supply Chain Analysis: Monitor, benchmark, and improve supply chain activities from materials ordering through service delivery.
- Wait Time Analysis: Focuses on the factors that are associated with longer waiting times and the effects of delays in scheduling and operation.

Regarding the purpose of the project and our scope, our methodology will be processed through data mining.

2.2.1 Data Management

The first step to implement a BI system is to define which external or internal data sources are going to be explored. Data management consists on identifying structured and unstructured data sources, and extract data to temporary data store systems or directly to the final Data Warehouse (DW). If the data is first stored in a temporary data system, it must be afterwards migrated to a proper DW. This migration only occurs after a proper definition and creation of the schema. Besides migrating information in this phase, data must be processed in order to obtain homogeneous data if required, since they were extracted from different sources and may be in different formats. This whole process is called Extract, Transform and Load (ETL). The ETL process may occur directly from the data sources or from temporaries data systems.

2.2.2 Data Analysis

As stated by Schniederjans et al. [SSS14] data analytics is defined as a process that involves the use of statistical techniques, information system software, and operations research methodologies to explore, visualize, discover and communicate patterns or trends in data. We can distinguish four types of data analysis:

- Descriptive: identifies patterns but does not point to any cause or explanation.
- Diagnostic: identifies patterns and give insights into specific problems.
- Predictive: uncovers relationships and patterns within large volumes of data to detect tendencies [Eck09].
- Prescriptive: detect tendencies but suggests also actions to prevent future problems and the advantages of a promising trend.

Data analysis can be done throughout the use of multiple techniques, such as ad hoc queries or Data Mining (DM) techniques (e.g. classification or clustering) [KKS17]. In this dissertation, we will focus on the clustering techniques and on mining association rules.

2.2.2.1 Clustering

Clustering (or cluster analysis) is an exploratory technique that organizes a collection of patterns, usually represented as a point in a multidimensional space, into groups (clusters) based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster [JMF99]. Being a data analysis technique, clustering can be used as a stand-alone tool that allows to obtain the clusters' natural structure and its characteristics. It can also be used alongside other techniques, like classification, and may be a starting point for other purposes, such as data summarization [TSK06].

Cluster analysis techniques have been developed in an attempt to analyze even larger data sets. This was a result of the need to transform huge amounts of data into meaningful and useful information. Although data mining can be considered as the main origin of clustering, it is vastly used in other fields of study such as bioinformatics [MC05, ENBJ05] and machine learning [ZK04].

Until this day, there is not a consensual definition of cluster. In fact, there are many different definitions of cluster and its interpretations are so diversified that, although the broad nature of this concept may reveal positive aspects, once it is applied in different contexts, it can be considered rather vague and somewhat confusing [MS03].

There are many techniques for cluster formation what may lead to clusters with significant differences. As per Hand et al. [HSM01] these methods can identify structures in distinct groups, however it is not always transparent the description of the respective clustering algorithm. For instance, a cluster can be defined as a collection of objects where the maximum distance between every pair of objects is the smallest possible. In this case, a method that minimizes the distances between each pair of objects is applied. On the other hand, a cluster can be defined as a group of objects where each object is the closest to any other object of its group, not being necessarily close to the other members of the cluster. This way, the methods mentioned in first place would not be efficient to identify the second type of clusters.

In order to choose the technique for cluster formation, the real problem and the analysis goals must be taken into consideration. The analysis should not be limited by assumptions and we should look for new techniques and new clusters in order to possibly identify patterns that were not taken into account previously [HSM01].

In summary, the goal of clustering is to split the objects in groups, so that the objects in a group are closer to them than to the members of any other group. This way, we can guarantee that the method of the cluster formation depends on a distance measure between objects, and also between groups.

Although each cluster analysis has its specifics problems, features and goals, it is possible identify a sequence of steps. As per Jain et al. [JMF99, JD88] these five steps are:

1. Object representation (optionally including feature extraction and/or selection).

2. Definition of an object proximity measure appropriate to the data domain.

3. Cluster analysis.

4. Data abstraction (if needed).

5. Assessment of output (if needed).

Object representation

Refers to the number of classes, the number of available objects, and the number, type, and scale of the features available to the clustering algorithm. On one hand, feature (or variable) extraction is the use of one or more transformations of the input features to produce new salient features. On the other hand, feature selection is the process of identifying the most effective subset of the original features to use in clustering. These techniques are used to obtain an appropriate set of features to use in cluster analysis.

This stage is the equivalent to data preprocessing in clustering, which also includes several procedures [HK06]:

1. Handling missing values: real data sets tend to be incomplete, contain noisy or inconsistent data. Data cleansing is the first step to complete missing values, reduce noisy data by detecting outliers and fix inconsistencies.

2. Integration: gather data from different sources, processing heterogeneous and eliminating redundancies.

3. Transformation: transformation of the variables (standardization, discretization and variables' construction) in order to improve algorithms efficiency and/or identify patterns. Standardization is used when variables are in different units or scales and they are changed to the same unit, scale or format. Discretization is, sometimes, used in variables with low levels of abstraction (e.g. age), so that those variables are substituted with ones with high levels of abstraction (e.g. adult). Construction of new variables is used to better interpret and understand data and the relation between variables.

4. Data reduction: reduce data set size by cutting down the number of objects or the variables in study.

The analyst's role in the object representation process is to gather facts and conjectures about the data, optionally perform variable extraction, and design the subsequent elements of the clustering system. The user also should decide which aspects are relevant in its particular problem and context. A good object representation can often yield a simple understood clustering, as a poor one may yield a complex clustering.

Object proximity

Consider that the data set has n objects, which may represent either a physical object (e.g. a car) or an abstract notion (e.g. a style of writing). Cluster analysis algorithms are mainly applied on two data sets [HK06]:

- Dissimilarity matrix: A nxn matrix where each entrance matches the dissimilarities between each pair of objects. Dissimilarity between two objects is a measure that represents the difference between those objects.

- Data matrix: A $n\,x\,p$ matrix with n objects and p variables $(X_1, X_2, ..., X_p)$ that define them.

Hereby, the dissimilarity is lower for a pair of objects more similar. The dissimilarity measure D between object x_1 and object x_2 satisfies the following [Gos12]:

1. Non negativity: $D(x_1, x_2) \geq 0$
2. Reflexivity: $D(x_1, x_2) = 0\,,\, x_1 = x_2$
3. Symmetry: $D(x_1, x_2) = D(x_2, x_1)$
4. Triangle Inequality: $D(x_1, x_2) + D(x_2, x_3) \geq D(x_1, x_3)$

Choosing the right measure of dissimilarity depends on the type of variables - numerical and categorical. We will only focus on the numerical ones.

The Euclidean distance is commonly used to calculate the proximity of objects in two or three-dimensional space, being the most used metric. This distance measure can be defined as follows:

$$D_2(x_i, x_j) \;=\; (\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2\,)^{1/2} \;=\; ||x_i - x_j||_2\,,$$

which, in turns, is a special case ($q = 2$) of the Minkowski metric

$$D_q(x_i, x_j) \;=\; (\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^q\,)^{1/q} \;=\; ||x_i - x_j||_q\,,$$

where p is a positive integer and $x_i \;=\; |x_{i,1}, ..., x_{i,p}|\,,| : i = 1, ..., n$ are the data set's objects with p variables.

Another usual distance is the Manhattan distance, also known as City Block, usually compared to the Euclidean distance. Manhattan distance can be defined as follows:

$$D(x_i, x_j) \;=\; \sum_{k=1}^{p} |x_i, k - x_j, k|\,,$$

If the variables have different significance, each variable assigned to a weight w_i, we can use the weighted Euclidean distance

$$D(x_i, x_j) \;=\; \sqrt{\sum_{k=1}^{p} w_i (x_{i,k} - x_{j,k})^2}\,,$$

or also use the same weighted distance variation to the Manhattan or Minkowski [Gos12, HK06, JMF99, SAW15, TSK06].

Cluster Algorithms

Cluster analysis includes a variety of techniques, each one used for a different purpose, as discussed before. There are innumerous distinct approaches, however in this work we only focus prototype-based and density-based clusters, i.e. K-means and DBSCAN respectively.

First and foremost, it is important to understand what is partitional clustering and what's its difference from hierarchical clustering. Partitional clustering is a division of the set of data into non-overlapping clusters such that each data object is in exactly one subset, and hierarchical clustering is a set of nested clusters that are organized as a tree - each node (cluster) in the tree, except for the leaf nodes, is the union of its children, and the root of the tree is the cluster containing all the objects. Thus, the subsequent output clustering can be hard (a partition of the data into groups) or fuzzy (each object has a variable degree of membership in each of the output clusters) [JMF99, Kol01].

Partitional methods require prior knowledge of the number K of clusters and group all the object into those K clusters. These algorithms have advantages when used with large data sets. The objects in the same cluster should be similar while objects in different clusters should be different (be far away in a dimensional representation). The most known partitional methods are K-means, K-medoids (or Partitioning Around Medoids) and its variations, which are based on prototypes - every object relative to a certain prototype is closer to it than to any other prototype. In K-means, the prototype is called centroid (the center of the cluster), while in K-medoids the prototype is called medoids. Unlike centroids, medoids must be an object in the data set [KJR90, HK06].

Besides being easy to implement, **K-means** is one of the most used algorithms, with time complexity of $O(n*k*i)$ where n is the number of objects in K clusters and i is the number of iterations needed to find the optimum solution.

As explained by Pavel Berkhin [Ber02], Pang-Ning et al. [TSK06] and Jiawei Han et al. [HK06], the first step of K-means is to choose K initial centroids, each object is then assigned to the closest centroid, according to the Euclidean distance between the object and the centroid, and each group of objects assigned to its respective centroid is a cluster. The centroid of each cluster is then updated based on the mean value of its objects, and this process is repeated until there are no changes in the centroids (vide algorithm 1).

Algorithm 1: K-means partitioning algorithm

```
Choose K objects as initial centroids;
while centroids change do
    Assign each object to its closest centroid forming K clusters;
    Update the clusters' centroids;
end
```

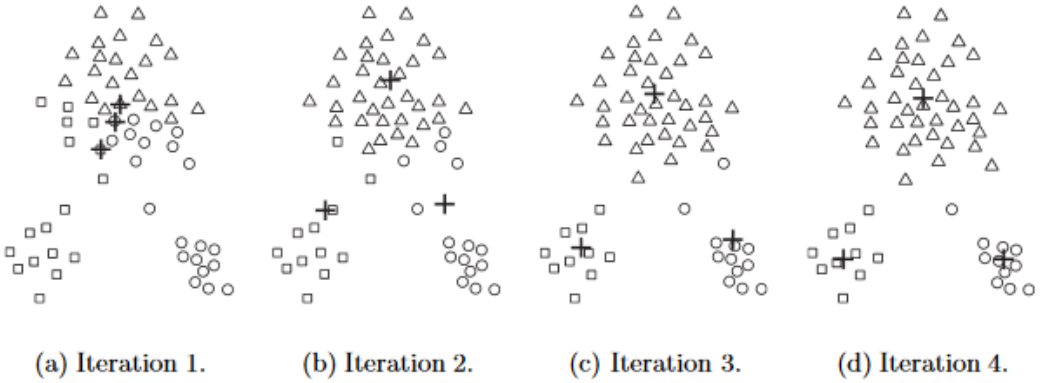
There are some variants of the K-means algorithm, which can differ on the selection of the number of clusters, on the proximity measure or on the objective function.

As a centroid-based partitioning technique, the center of a cluster is represented by a centroid. The dissimilarity between an object x_j and a cluster's centroid c_i is measured by $dist(x_j, c_i)$, where $dist(y,z)$ is the Euclidean distance between two points y and z. The objective function that measures the quality of a cluster, is the sum of squared error between all objects in a cluster and its centroid c_i. This measure can be defined as follows:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in K_i} dist(x_j, c_i)^2 ,$$

where SSE is the sum of the squared error for all objects in the data set, x_j a given object and c_i a centroid of cluster K_i. In short, for each object in each cluster, the distance from the object to its centroid is squared and the distances are summed. The objective function is to make K clusters as compact and separate as possible [HK06].

A simple example of the **K-means** is shown in figure 2.3, using the mean as the centroid. The crosses represent the centroids, the triangles, squares and circles the objects in a data set (similar shape in different objects means they are assigned to the same cluster). When the algorithm terminates, in iteration 4, the centroids have grouped its respective points.



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 2.3: Using K-means to find three clusters in sample data (page 498 [TSK06]).

On the other hand, **density based algorithms** are generally related to finding non-linear shapes structure clusters [RSJM10]. One of the most widely and used density based algorithms is **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise). DBSCAN produces a partitional clustering, in which the number of clusters is automatically determined by the algorithm. However, DBSCAN can not produce a complete clustering because points in low-density regions are classified as noisy data and are omitted.

Initially we must take into consideration the key notion of density. There are several distinct methods for defining density, like center-based approach on which DBSCAN is based. In this approach, density is estimated for a particular object in the data set by counting the number of objects within a specified radius of that object, including the object itself. An example of this approach is illustrated in figure 2.4 where the number of objects, within an Eps radius, of object A is seven [EFKS00, Kol01].

Thus, based on center-based approach, according to Pang-Ning et al. [TSK06] the objects can be classified as:

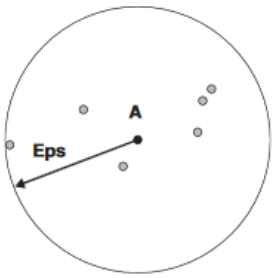


Figure 2.4: Example of center-based density (page 528 [TSK06]).

- Core points: if the number of objects, within a given neighborhood (with a certain radius), around the object exceeds a certain threshold ($MinPts$ is the number of minimum points required as a point to be considered as core point). In the following example 2.5 object A is a core point, for a radius Eps if $MinPts \leq 7$.
- Border points: is in a core's point neighborhood but it is not classified as core point. In the following example 2.5 object B is a border point.
- Noise points: any object that is neither a core point nor a border point. In the following example 2.5 object C is a noise point.

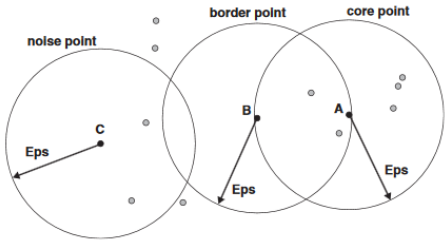


Figure 2.5: Illustration of classification of points (page 528 [TSK06]).

Based on the previous settings, DBSCAN groups two core points that are close enough to each other, with a distance lower than its Eps. The same thing is applied with border points, every one that is close enough to a core point is assigned to the same cluster as the core point. Noise points do not belong to any cluster (vide algorithm 2).

Algorithm 2: DBSCAN algorithm

Classify all objects;
Discard noise points;
Match all core points with a distance from each other lower than a given Eps ;
Assign different clusters to every matched core point;
Assign the respective core point's cluster to every border point;

In order to determine Eps and $MinPts$, the main approach used is to look at the distance from an object to its k^{th} nearest neighbor, named $k-dist$. Considering all objects in a cluster, the value of $k-dist$ will be small if k is not larger than the cluster size. This may differ depending on the cluster's density and the distribution of the objects. For objects not assigned to a cluster (noise points) the $k-dist$ will be large. So computing $k-dist$ for all the data set for a given k, sort the objects in increasing order and then plot them, it is expected to observe an abrupt change of the $k-dist$ value. Assuming this measure as Eps and the value of k as $MinPts$, then objects for which $k-dist$ is less than Eps will be classified as core points, while the remaining are border or noise points [GT15].

Figures 2.6 and 2.7 show an example of a data representation with the respective $k-dist$ graph. Although the value of Eps depends on k, its change is not significant. It is crucial to give a right value to k. Giving a low value to k will incorrectly assign noise points as clusters and giving high value, then small clusters can be classified as noise. Generally, the default value of k is 4, which appears to be the most suitable value according to Ester et al. [EFKS00].

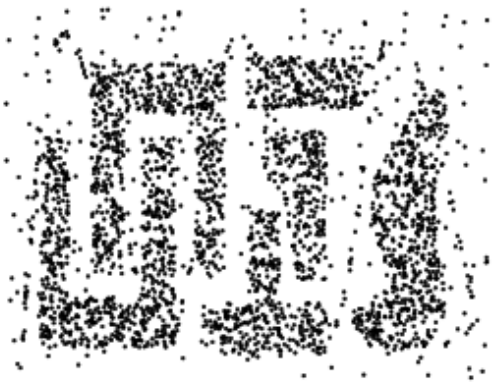


Figure 2.6: Sample data set.

To demonstrate the DBSCAN application, we will continue the previous example 2.6 composed by 3000 two-dimensional objects. The value of Eps was found by plotting the $k-dist$ graph 2.7 and identifying the value when the abrupt change occurs ($Eps = 10$). Using this k value and $MinPts = 4$, the clusters resulted from the DBSCAN algorithm are illustrated in the following figure 2.8. The crosses represent the noise, and all the different shapes represent each cluster.

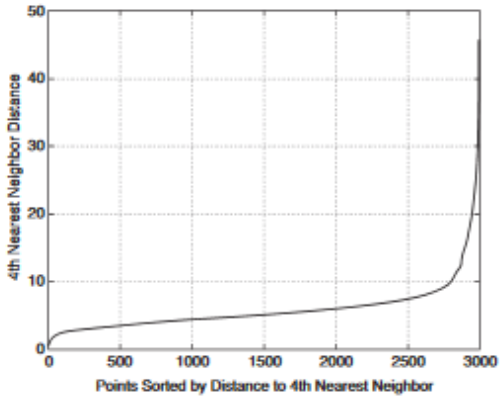


Figure 2.7: $K - dist$ plot for the data set.

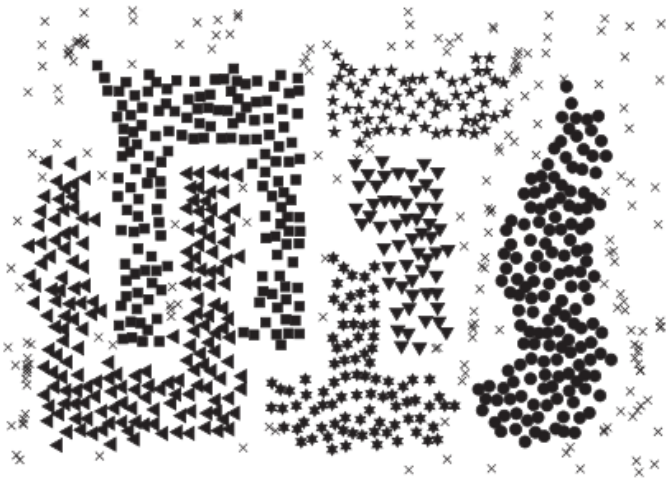


Figure 2.8: Cluster resulted from DBSCAN [TSK06].

Although being capable of identifying many clusters that K-means could not do, DBSCAN algorithm has complications when clusters have varying densities or with high levels of multi-dimensional data, where densities are more difficult to define [Ber02, Kol01, EFKS00].

Data abstraction

This stage is sometimes included in the previous step. It includes a simplistic representation of the data set (e.g. plotting), promoting an intuitive comprehension and understanding of the patterns [DS76].

Assessment of output

Represents the evaluation of the output. Questions like *"Are the clustering results good or poor? What characterizes a good clustering and a poor one?"*. Every cluster algorithm will produce clusters, however some clustering algorithms may produce better results depending on the context, on the data set and on the goal. The study of cluster tendency (vide Yizong Cheng [Che95]

and Richard Dubes [Dub87]) is used, prior to the cluster analysis, to examine the data set to check if there are meaningful clusters through statistical and visual analysis methods.

2.2.2.2 Association Rules

Overview

Association rules are an important class of regularities within data which have been extensively studied by the data mining community. The general objective is to find frequent co-occurrences, called associations, of items within a set of transactions. According to Bing Liu [Liu06] the idea of discovering such rules is derived from market basket analysis, where the goal is to mine patterns describing the customer's purchase behaviour. Table 2.1 illustrates an example of the market basket transactions, where each row corresponds to a transaction and a set of items bought by a given customer.

Transaction ID	Items
1	{Rice, Bread}
2	{Rice, Apple, Beer, Eggs}
3	{Bread, Apple, Beer, Cookies}
4	{Rice, Bread, Apple, Beer}
5	{Rice, Bread, Apple, Cookies}

Table 2.1: An example of market basket transactions.

Table 2.2 illustrates the binary representation of table 2.1, where an item is treated as a binary variable whose value is 1 if the item is present in a transaction and 0 if it is not present. There are also methods for handling categorical and quantitative data, however we will only focus the binary representation [AIS93].

Transaction ID	Rice	Bread	Apple	Beer	Cookies	Eggs
1	1	1	0	0	0	0
2	1	0	1	1	0	1
3	0	1	1	1	1	0
4	1	1	1	1	0	0
5	1	1	1	0	1	0

Table 2.2: An example of binary representation of the market basket transactions.

The uncovered relationships can be represented as follows: *{Apple}→{Beer}*. This rule suggests that a strong relationship exists between the sale of apples and beers, because many customers who buy apples also buy beer.

Association analysis has widely range of applications domains such as retailing, Web mining, bioinformatics, medical diagnosis and scientific data analysis. When applying association analysis it is important to take into consideration that some discovered patterns are potentially spurious because they may happen just by chance [TSK06].

Problem Definition

The problem of mining association rules is stated as follows: $T = \{t_1, t_2, ..., t_n\}$ is a set of transactions, where each transaction contains items of the item set $I = i_1, i_2, ..., i_k$. Thus, an association rule is an implication of the form: $X \rightarrow Y$, where X is a set of some items in I ($X \subset I$) and Y is a single item in I ($Y \subset I$) that is not present in X ($X \cap Y = 0$). The strength of an association rule can be measured in terms of its support, confidence and lift. Support indicates how often a certain rule is applicable to a given data set

$$Support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

where $\sigma(X \cup Y)$ is the support count - number of transactions which contain X and Y - and N is the total number of transactions. Confidence determines how frequently items in Y appear in transactions that contain X

$$Confidence, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

where $\sigma(X)$ is the number of transactions that contain X. Lift is simply the ratio of these values: target response divided by average response.

$$Lift(X \rightarrow Y) = \frac{\frac{Support(X \cup Y)}{N}}{\frac{Support(X)}{N} * \frac{Support(Y)}{N}}$$

Generally, if the lift value is greater than one then X and Y are positively correlated. If the lift is less than one then X and Y are negatively correlated [AIS93, HGN00].

For instance, considering the rule {Bread, Apple}→{Beer} the support count for {Bread,Apple,Beer} is 2 and the total number of transactions is 5, so the rule's support is $2/5 = 0.4$. The rule's confidence is obtained by dividing the support count for {Bread,Apple,Beer} by the number of transactions that contain {Bread,Apple}. So, the confidence for this rule is $2/3 = 0.67$.

As stated by Pang-Ning et al. [TSK06], these measures interpretation is crucial to understand if a rule is interesting or not, if a rule happens by chance or not. When a rule has very low support it may occur simply by chance, and it is also an indicator of low significance. For these reasons, support is often used to eliminate uninteresting rules. On the other hand, confidence measures the reliability of the inference by a rule, what means the higher the confidence, more likely it is for Y to be present in transactions that contain X.

As per Agrawal et al. [AIS93] mining association rules can be divided in two major subtasks:

1. Discover frequent itemset: whose objective is to find all itemsets that are greater or equal than a given support value - called minimum support (*minSup*).

2. Rule Generation: whose objective is to extract all the rules, with high confidence values, from itemsets found in the previous step.

Is known that the computational requirements are much more expensive for the search of frequent itemset than those for rule generation.

Apriori Algorithm

Since the introduction of the Apriori Algorithm, several algorithms focusing either frequent items' search or rule generation, have been proposed [AIS93]. However, Apriori provides solution for both problems.

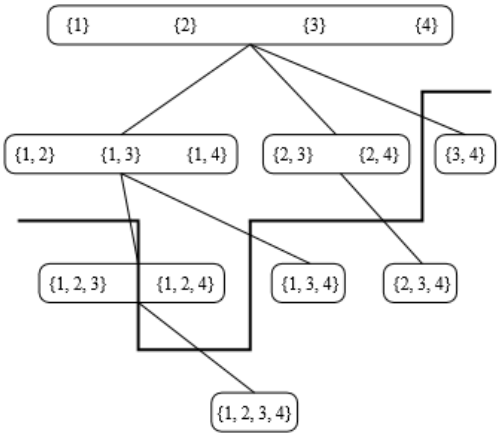


Figure 2.9: An example of itemsets representation [HGN00].

In figure 2.9, the bold line represents the border between frequent and infrequent itemsets. All items above the border fulfill the minimum support requirements - called frequent itemsets - while the ones bellow the line do not fulfill such requirements - called infrequent itemsets. The algorithm has the goal to determine that line in order to find the frequent itemsets.

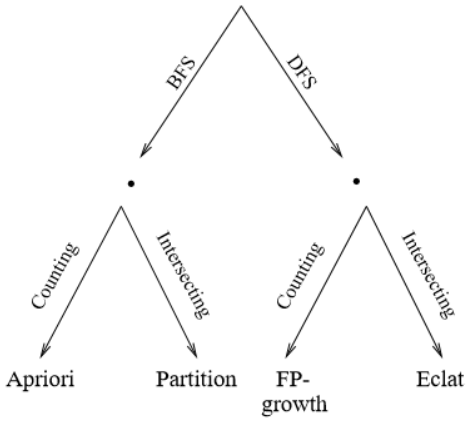


Figure 2.10: Association rules types of algorithms [HGN00].

There are two types of association rules algorithms: i.e. Breadth-First Search (BFS) and Depth-First Search (DFS) [HGN00]. In BFS algorithms, such as Apriori, the value of support for

each itemset is determined for each specific level of depth, whereas DFS recursively descends the graph. Mining association rules algorithms can be summarized as shown in figure 2.10.

The Apriori principle defines that if an itemset is frequent, then all of its subsets must also be frequent. Conversely, if an itemset is infrequent, then all of its supersets must be infrequent too. In the following figure 2.11, we can observe the these two theorems. Assuming {c,d,e} is a frequent itemset, then its subsets ({c,d},{c,e},{d,e},{c},{d},{e}) are also frequent. On the other hand, if a itemset {a,b} is infrequent, then all the itemsets containing {a,b} are infrequent and can be discarded immediately [TSK06].

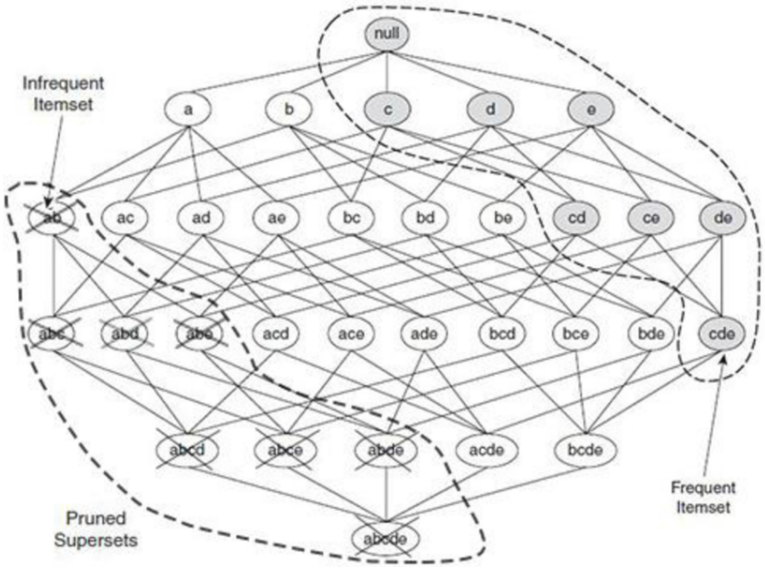


Figure 2.11: Illustration of the Apriori Principle [TSK06].

Concerning the discover of frequent itemset, in the first place every item is considered a 1-itemset candidate. Using the market basket example and using support threshold of 60% (*minSup=0.6*), after determining the support value for each candidate, {Cookies} and {Eggs} are discarded because they appear in less than 3 transactions -i.e. $0.6 * 5 = 3$. In next iteration, 2-itemsets candidates are generated using only the frequent 1-itemsets ({Rice},{Bread},{Apple},{Beer}), due to the infrequent theorem. Because there are four frequent 1-itemsets, the number of 2-itemsets candidates generated by the algorithm is $\binom{4}{2} = 6$ - {Beer,Rice}, {Beer,Apple}, {Beer,Bread}, {Rice,Apple}, {Rice,Bread}, {Apple,Bread}. Two these six candidates ({Beer,Rice},{Beer,Bread}) are also labeled as infrequent because their support count is two. Without the pruning method there are $\binom{6}{3} = 20$ 3-itemset candidates, with the Apriori principle, as we only consider 3-itemsets whose subsets are frequent, there is only one 3-itemset candidate - {Rice,Apple,Bread}.

A bruce-force strategy of enumerating all itemsets will produce

$$\binom{6}{1} + \binom{6}{2} + \binom{6}{3} = 41$$

candidates. Applying the Apriori principles, this number decreases massively to

$$\binom{6}{1} + \binom{4}{2} + 1 = 13$$

candidates, which means a 68% reduction in the number of candidates [AS94, CR06]. This Apriori algorithm is described, in detail, in the following pseudocode.

Algorithm 3: Apriori Algorithm pseudocode

k=1 ,{Find all frequent 1-itemset};
$F_k = (i | i \in I \wedge \sigma(i) \geq N * minSup$;
while $F_k \neq 0$ **do**
  k = k+1;
  $C_t = apriori - gen(F_{k-1})$,{Generate itemsets candidates} ;
  for *each transaction* $t \in T$ **do**
    $C_t = subset(C_k, t)$,{Identify all candidates that belong to it} **for** *each candidate itemset* $c \in C_t$ **do**
      $\sigma(c) = \sigma(c) + 1$,{Increment support count};
    end
  end
  $F_k = \{c | c \in C_k \wedge \sigma(c) N * minSup\}$,{Extract the frequent k-itemsets}
end
Result = $\cup F_k$

The *apriori-gen* function is the one responsible for grouping the itemsets into new itemsets, with one more item than the previous itemset, and removing the infrequent itemsets. The candidates are being stored in a hash-tree, where each node contains an itemset candidate. The *subset* functions starts from the root node going towards the leaf nodes in order to find all the candidates contained in a given transaction t.

Regarding generating association rules in Apriori algorithm, each frequent k-itemset, Y, can produce up to $2^k - 2$ association rules, discarding rules with empty antecedents or consequents (0→Y or Y→0). An association rule can be extracted by partitioning the itemset Y into two non-empty subsets $X \rightarrow Y$ and $Y - X$, such that $X \rightarrow Y - X$ satisfies the confidence threshold (e.g. assume Y={1,2,3} is a frequent itemset, so six candidates can be generated {1,2}→{3}, {1,3}→{2}, {2,3}→{1}, {1}→{2,3}, {2}→{1,3} and {3}→{1,2}, discarding afterwards the ones that do not fulfill the minimum confidence requirement). The Aprori also applies another theorem that states *If a rule* $X \rightarrow Y - X$ *does not satisfy the confidence threshold, then any rule* $X' \rightarrow Y - X'$, *where* X' *is a subset of* X, *must not satisfy the confidence threshold as well* [TSK06]. The figure 2.12 illustrates this theorem.

2.2.3 Presentation and Delivery

The results of the many data mining techniques can be hard to interpret and understand. For instance, it is impractical to analyze 90 pages of rules generated by a data mining algorithm. Thus,

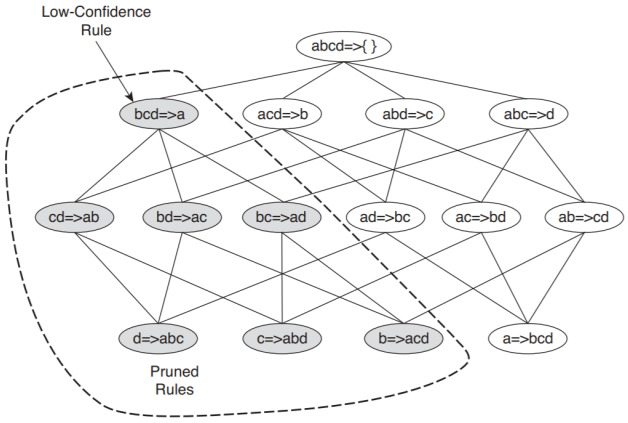


Figure 2.12: Illustration of the Apriori pruning using the confidence measure [TSK06].

visual representation of such rules heightens its comprehensibility [VP06].

Business data visualization has two explicit objectives, visualize key metrics for an easy and fast comprehension of the data which directly facilitates decision-making, and provide a visual and interactive way to explore data [VW07]. As concluded by Iris Vessey [Ves91] and Ben Shneiderman [Shn96] data visualization has a significant value and contribution to the information-seeking process and to the decision process.

As per Jack Zheng [Zhe17], visualization helps data comprehension and enhances problem-solving capabilities in the following ways:

- Eases the cognitive load of information processing [BVB+13].
- Data visualization techniques provide a visual overview of complex data sets to identify patterns, structures, relationships and trends at a high level.
- Provides visual cues that draw people's attention to quickly focus on areas of interest or difference [Teg99].
- Exploits the human visual system to extract additional (implicit) information and meaning, sometimes referred to as intuition.

Typically, the Business Intelligence results are presented in the form of reports, dashboards and analytical tools. Dashboards are mostly data visualization driven, whereas reports focus more detailed data with embedded visuals or charts/diagrams, which enhances reports' readability. It is the goal of the Business Intelligence manager to understand the features, the strengths and weaknesses of each form of data visualization.

2.3 Data Anonymization

The quantity of clinical data which is electronically available is setting higher standards on what concerns to its complexity and diversity. Even though this growth carries many benefits, it also

raises some barriers and challenges, such as concerns over the sensitive information security and privacy [AbhKS17].

Health Insurance Portability and Accountability Act (HIPAA) has formulated a regulation for data de-identification in the United States of America (USA), named HIPAA Compliance Guide[3]. There are listed 18 identifiers considered protected health information (PHI) which must be removed in order to data do be de-identified, according to the Safe Harbor method[4]. A PHI is any kind information that can be used to identify an individual, so when de-identification is achieved it is not possible to identify an individual according to that data. These 18 identifiers are:

1. Names
2. All geographic data smaller than a state
3. Dates, besides year, directly related to an individual (e.g. date of birth)
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social Security numbers
8. Medical record numbers
9. Health insurance plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers including license plates
13. Device identifiers and serial numbers
14. Web URLs
15. Internet Protocol (IP) addresses
16. Biometric identifiers (e.g. fingerprints)
17. Full face photos and comparable images
18. Any unique identifying number, characteristic or code

In the European Union (E.U.), recital 26 of the GDPR states that data that has been truly anonymized lies outside the scope of the regulation:

[3] https://www.hipaajournal.com/wp-content/uploads/2015/05/HIPAAJournal-com-HIPAA-Compliance-Guide.pdf
[4] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

> "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes." [REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT[5], Recital 26]

It is possible to assume that if the Safe Harbor method is strictly complied, there are not any possible violations of both European and U.S. regulations.

2.4 Related Works

Presently, there are many BI implementations in healthcare. For example, Gaardboe, Nyvang and Sandalgaard [GNS17] implemented a BI system in Danish hospitals to forecast the load on hospitals' resources and to allow employees to follow up on Key Performance Indicators (KPI). This implementation successfully validated DeLone and McLean's IS Success Model [DM92] applied to HIS in the public sector. Another study proposed by Ali, Crvenkovski and Johnson [ACJ16] points to the implementation of BI data analytical solution in healthcare. The paper focus the improvements in the rehabilitation workflow for patients with hip fracture, with the BI system. Many other papers present a literature review on Business Intelligence in healthcare sector [FK14, Mur12, AÖNC13].

There are also papers that focus on data anonymization in healthcare, how to preserve privacy and the challenges faced when dealing with big amounts of data under the General Data Protection Regulation [GMVJ18, Pag17, Zha18, HRP+16, SAS+17]. A study ran by Gruschka, Mavroeidis, Vishi and Jensen [GMVJ18] discusses the current legislation (GDPR), the challenges of dealing with large volumes of sensitive data and the anonymization methods commonly used. A similar study [RP17] also focus the GDPR, anonymization and consent regarding data sharing.

Last but not least, a master's thesis [Fer12] regarding usage patterns of credit cards of a financial institution, presents the literature review on the different clustering techniques, and applies K-means and K-medoids algorithms to a data set. Several papers present the literature review of the clustering techniques [SAWH14, Ber02, Kol01, MS03] and of the dissimilarity measures [SAW15, PWSTE03, GL86].

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Chapter 3

Business Intelligence Solution Outline

In this chapter we explain our proposed Business Intelligence system, which uses cluster analysis and association rules to identify medical exams' utilization patterns. The following sections explain in detail each of the BI's steps, the technologies and techniques used to perform this project. A preliminary analysis of the extracted data is also presented.

3.1 Overview

The system's architecture, illustrated in figure 3.1, is based on the main BI architecture 2.2, where three main stages can be identified: Extract, Transform and Load (ETL); Data Analytics; Presentation. The first one is where data extraction, data cleansing, data process and data migration to the main DW occurs. Although data anonymization is represented apart from ETL, it occurs within the ETL process. Then, clustering and association rules mining take place in data analytic stage. Finally, data is presented to the final users through interactive graphs, dashboards and other relevant formats.

3.2 Data Sources

Initially there was the need to identify what data and which data sources were going to be targeted. After discussing what data was required with the stakeholders, we established that data regarding medical staff, medical examinations and accesses to medical exams was needed to accomplish the goals of this dissertation. Such data is represented in table 3.1.

Regarding medical staff data, we discard any personal information (e.g. name, age, sex, address) and only focus on the professional's unique identifier, his/hers respective groups in the hospital's domain and his/hers respective medical service (or medical specialty), that will be extracted from the domain groups. As to examinations, information regarding the type of examination (e.g.

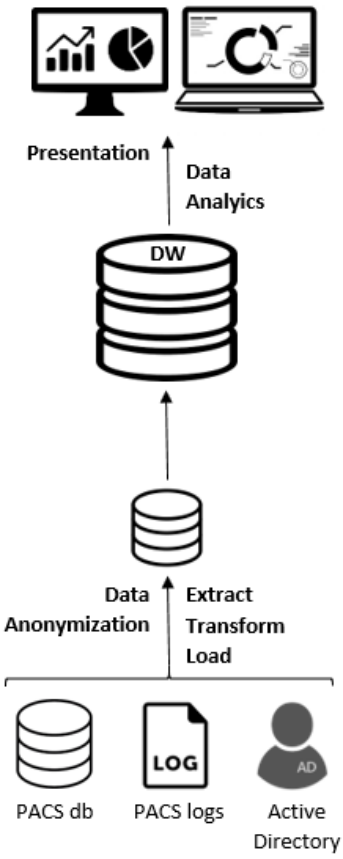


Figure 3.1: Proposed BI system architecture

Thorax CT Scan), the medical imaging modality (e.g. CT), the body part of the exam (e.g. Thorax), if it has a report and its date, the examination's date and the module of the exam (e.g. URG refers to emergency). Finally, concerning accesses to the exams we target who accesses the examination - physician ID -, the action type (access or modify), the examination accession number and when the exam was accessed - access date. The accession number is the ID of the examination, so we will call it exam ID.

Since working under the General Data Protection Regulation (vide section 2.3), we identify two fields that need to be encrypted: medical staff ID and exam's accession number.

PACS has its associated software to visualize medical imaging, so information relative to accesses to medical images are stored in logs. From there we can extract information regarding who accesses the exam, what exam is accessed and when the access occurs. Subsequently the data sources for this work were defined:

- Active Directory (AD): extract medical staff data.
- PACS database: extract information regarding the examinations;
- PACS logs: extract information regarding accesses to the examinations.

Medical staff	**Examinations**	**Accesses to exams**
ID	Accession Number (ID)	Medical staff ID
Active Directory Groups	Exam Code	Action Type
Medical Specialty	Modality	Exam Accession Number
	Body Part	Access Date
	Body Description	
	Report	
	Report Date	
	Exam Date	
	Module	

Table 3.1: Required data for extraction.

3.3 Data Extraction, Transformation and Load

After properly defining what data was needed and where such data was stored, two scripts were made to extract the desired data. Since the whole process of data extraction was made on the client-side, there were many limitations regarding the technologies to be used.

The first script was written using PowerShell[1] to extract information from the active directory. Powershell is the main scripting language used at Sectra due to the constraints previously mentioned, and due to its versatility of the administration of applications that run on Windows Server environment and of its commands - cmdlets[2]. This script consists on matching the Sectra's software users with the domain users and extract data - AD groups and the respective medical service. The information will be then stored in a sqlite3 database[3].

As shown in the flow chart 3.2, the script first creates the proper database table with the required fields - medical staff ID, domain groups and service -, then lists all the users in the software, with a specific command, that is not allowed to be revealed due to privacy agreements, and lists the domain users with the command *net user /domain*. We will go through each line in the listed software users and match each line referring to a user's ID with the regular expression *(?<=login:)(.*?)(?=$)*. The users in Sectra's software may or may not directly match a user in the domain, since in the hospitals' domain there is a certain prefix aggregated to the medical staff ID. Whenever there is no direct match between those two users, a certain prefix is added to the ID - final ID. Users may appear in both formats, resulting in replication of users. So for each user's ID matched in the users software list, this verification will occur and the final ID will be matched with the listed domain users. If the final ID exists in the domain, then information regarding that user will be retrieved with the command *net user <finalID> /domain* and a regular expression is used to obtain the domain groups $((?<=Global\,Group\,memberships)([\backslash s\backslash S]*)(?=The\,command))$. Finally, the medical service is obtained from the domain groups, the final ID is encrypted in MD5[4] and those three fields are inserted in the sqlite3 database.

[1] https://docs.microsoft.com/en-us/powershell/scripting/overview?view=powershell-3.0
[2] https://docs.microsoft.com/en-us/powershell/developer/cmdlet/cmdlet-overview
[3] https://social.technet.microsoft.com/wiki/contents/articles/30562.powershell-accessing-sqlite-databases.aspx
[4] https://en.wikipedia.org/wiki/MD5

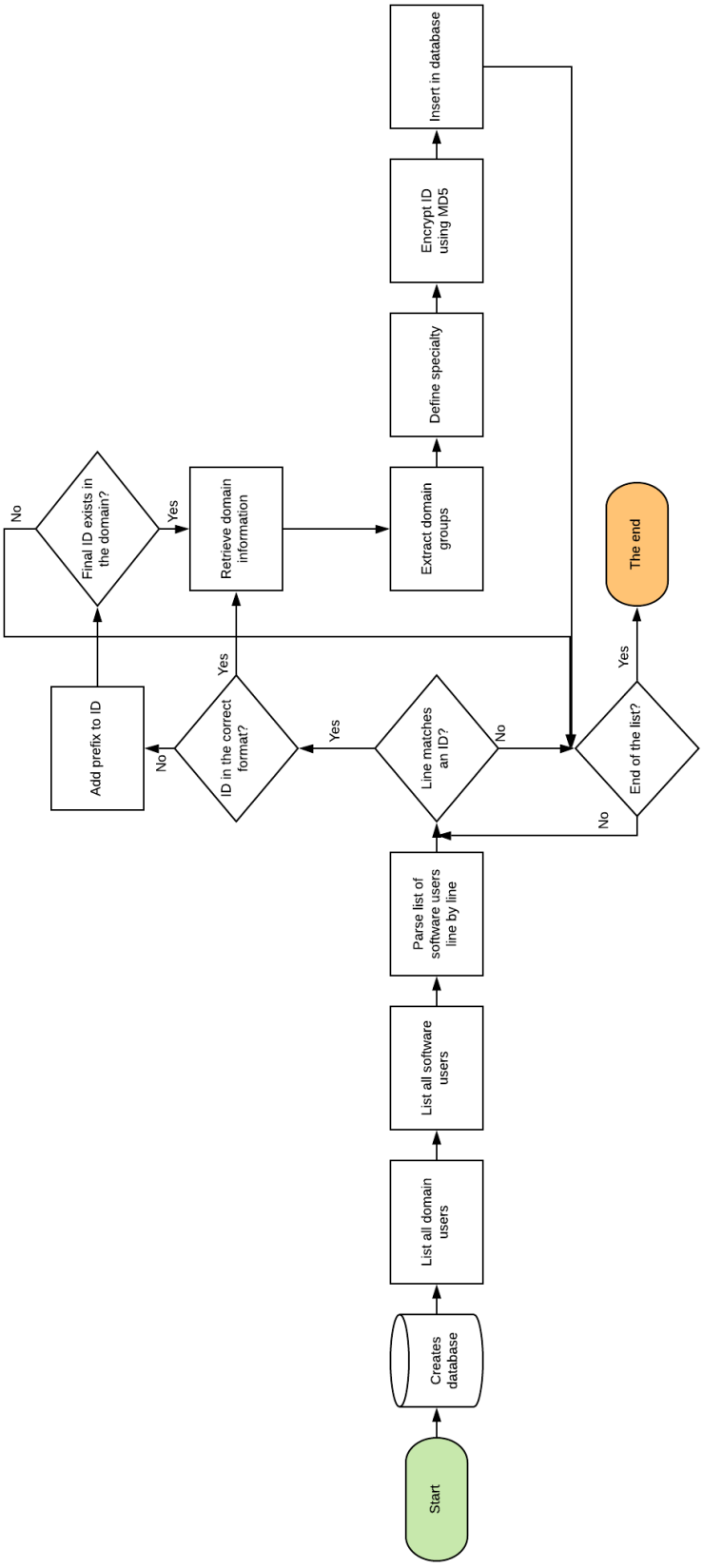


Figure 3.2: User data extraction script flow chart

In its turn, the second script extracts the information relative to accesses by parsing the logs and extracts exams' information by accessing the PACS database. To achieve such goal a script was written in Python 2.7[5] and several modules and libraries were used to parse XML files (Minimal DOM[6]), to connect to SQL server (pyobdc[7]), to encrypt (hashlib[8]) and to connect to the sqlite3. In order to overcome the impossibility of running Python scripts directly on the client-side, cx-freeze[9] was also used for freezing the Python script into an executable, allowing this way the execution of the script on the client-side.

The script was ran everyday when the hospital has less activity - between 1 a.m. and 6 a.m. - to avoid any potential conflicts and it parses the log of the previous day. The log is composed of XML messages, each message referencing a single action by a specific user or by the system.

As shown in the flow chart 3.3, first the script connects to the SQL Server of the PACS database and then reads the last created log (the log of the previous day) and starts parsing the log. Different XML messages have different tags; since as we are only focusing on accesses to the exams, we look for specifics tags. Every XML message that does not have the required tags is discarded. For every XML message that matters, we extract data regarding the access - user (physician) ID, action type, access date, exam ID - and two verifications are done: if the user ID tag is referring to the system (occurs when a exam is created); if the exam ID already exists in our database. In the first case the XML message is discarded and the log parsing continues. In the second case since we already have the data regarding that exam, we only encrypt the user ID using MD5 and insert information regarding the access to the exam in the database. When none of these cases verify, we use the exam ID to query the PACS database and collect exam data (vide table 3.1). Then we encrypt both exam ID and user ID using MD5 and insert access data and exam data in the respective sqlite3 database. The program ends when we reach the end of the log.

After gathering all the necessary data, the sqlite3 database on the client-side was migrated to a PostgreSQL[10] in our local machine. The final schema consists of three tables, similar to the table 3.1 (vide table 3.2).

3.4 Preliminary Data Analysis

3.4.1 Research Variables

The research data set is composed by 442 158 objects which are characterised by 12 variables. Each object consists on a specific access, made by a health professional - exclusively physicians - to a medical examination. The research variables reference not only accesses, but also the examinations and the health professionals. We were able to identify the medical service of 4804 doctors, and gather data in a 112 day period. The data set also corresponds to the medical exams

[5] https://docs.python.org/2/
[6] https://docs.python.org/2/library/xml.dom.minidom.html
[7] https://pypi.org/project/pyodbc/
[8] https://docs.python.org/2/library/hashlib.html
[9] https://cx-freeze.readthedocs.io/en/latest/index.html
[10] https://www.postgresql.org/about/

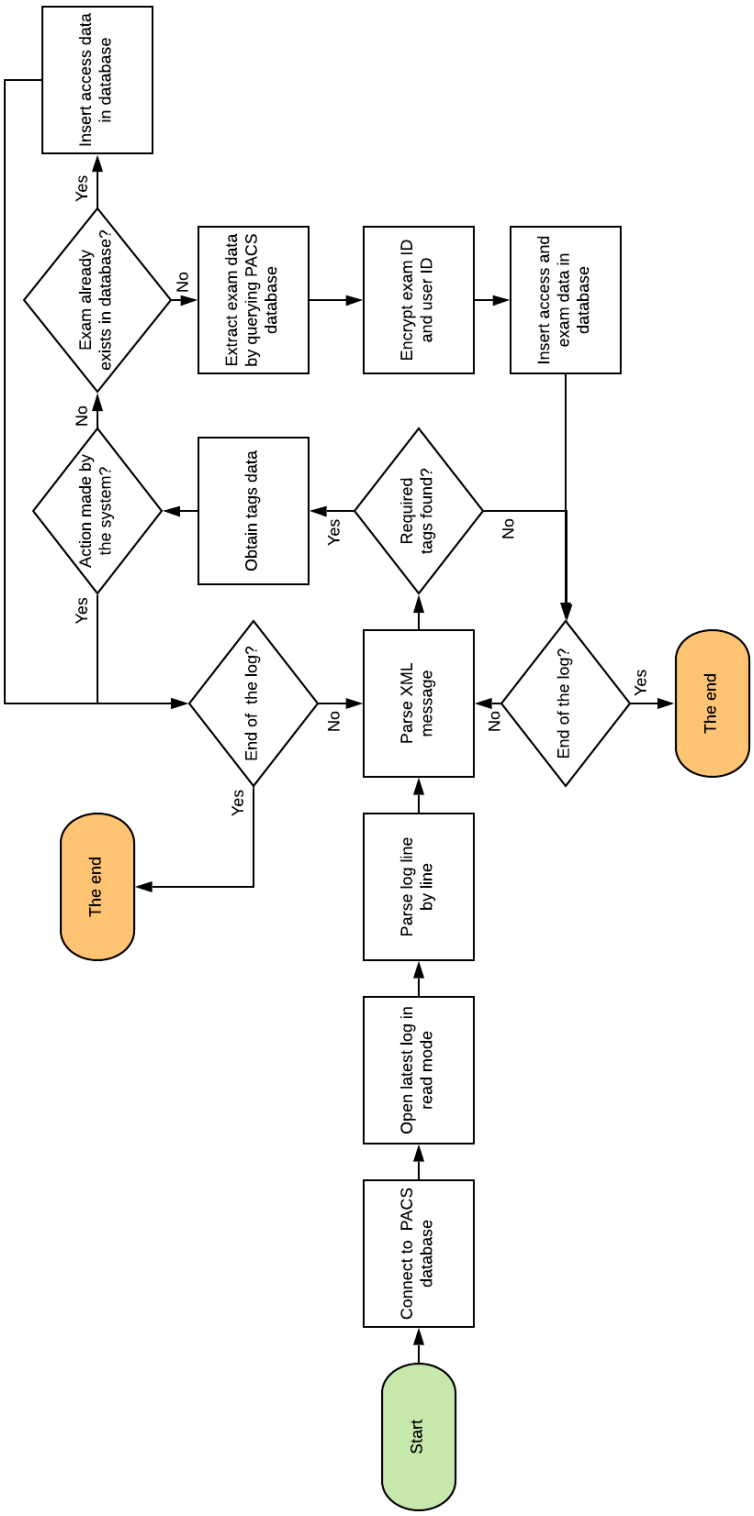


Figure 3.3: Flow chart of Log Parsing Script

ADgroups	**examInfo**	**auditInfo**
pacsUserLogin (PK)	accessionNr (PK)	pacsUserLogin (FK)
groupName	examCode	actionType
service	modality	accessionNr (FK)
	bodyPart	accessDate
	bodyDescription	
	report	
	reportDate	
	examDate	
	module	

Table 3.2: Final schema of the data extracted.

performed in the first 101 days, giving a 11-day window to only gather its respective accesses. Each object is characterized by two subsets of variables that describe: (1) the access; (2) the exam that is accessed.

1. Variables regarding the access:

   - Physician ID: consists on a unique 32 digit hash that identifies a health professional.
   - Health Professional Service: indicates the service that the health professional is assigned. It is a categorical variable with 92 categories (vide appendix A.1).
   - Date of access: indicates the date when the access occurs, in the format MM/DD/YYYY. The date, in this study, can be between January 11 of the current year (01/11/2019) and May 2 (02/05/2019).
   - Access Interval: indicates the time interval of when the access to the exam occurs, starting from the date when the exam is done. It is a categorical variable with 6 categories: within the first week, within the second week, within the first month, within the second month, within the third month, within the forth month. For instance, if an examination is done on January 10 and the access occurs on January 15, we consider that the access occurs within the first week. If the access occurs on January 25, we consider that the access occurs within the first month. These categories are mutual exclusive. That means we only consider an access within the second month when the access happens after 30 days, from the exam date, and before 60 days.

2. Variables regarding the exam:

   - Exam accession number: consists on a unique 32 digit hash that identifies an exam.
   - Exam date: indicates the date when the exam is done, in the format MM/DD/YYYY. In this study, it ranges from 01/11/2019 to 04/21/2019.
   - Exam code: consists on a number that specifies the type of exam. There are 758 different exam codes.

- Body Part: indicates a categorical variable which represents the specific body part of an exam. There are 661 different categories.
- Modality: indicates the medical imaging modality, based on the DICOM library[11]: BMD, CR, CT, DR, ECG, ES, MG, MR, NM, OT, US, XA, XC. BMD refers to Bone Densitometry, CR to Computed Radiography, CT to Computed Tomography, DR to Digital Radiography, ECG to Electrocardiography, ES to Endoscopy, MG to Mammography, MR to Magnetic Resonance, NM to Nuclear Medicine, OT to Other, US to Ultrasound, XA to X-Ray Angiography and XC to External-camera Photography.
- Module: indicates if the exam was made due to a medical appointment (*CON*), due to emergency (*URG*), regarding inpatient (*INT*) and regarding surgery (*BLO*). There is also the *HDI* and *RAD* modules for ambulatory care and for exams that do not belong to any of the previous cases, respectively.
- Report: consists on a boolean variable that indicates if a exam has a report or not - 0 has not a report; 1 has a report.
- Report Date: indicates the date of the report, if there is one. If the exam does not have a report, the date is null.

3.4.2 Exploratory Data Analysis

In order to get a perspective of the gathered data, we start by exploring each of the available variables, which were organized according to two strands of data: accesses to examinations and the examinations in study.

The number of health professionals in this study is 6414, which is much smaller than the number of total accesses, since it is expected that a physician to do more than one access throughout the data gathering period. As said previously in section 3.2, the same health professional may have more than one ID, so for 6414 health professionals there are 10285 different IDs. In the following table is displayed how the different IDs are associated with the different medical services.

Table 3.3: Number of physicians IDs and number of accesses per medical service.

Medical Service	Associated IDs	Number of accesses
Anatomia Patológica	156 (1.52%)	1161
Anestesiologia	463 (4.50%)	8547
Anestesiologia/Reanimação	68 (0.66%)	380
Bloco Operatório Central	8 (0.08%)	571
Bloco Operatório Serviço Urgência	6 (0.06%)	5
C.Estomatologia	41 (0.40%)	332
Cardiologia	112 (1.09%)	432

Continued on next page

[11] https://www.dicomlibrary.com/dicom/modality/

Table 3.3 – continued from previous page

Medical Service	Associated IDs	Number of accesses
Cardiologia Pediátrica	95 (0.92%)	1785
Cirurgia A	66 (0.64%)	178
Cirurgia B	67 (0.65%)	7
Cirurgia Cardiotorácica	139 (1.35%)	3373
Cirurgia Geral	230 (2.24%)	2950
Cirurgia Pediátrica	133 (1.29%)	2264
Cirurgia Plástica Reconstrução Estética e Maxilo Facial	171 (1.66%)	4829
Cirurgia Vascular	136 (1.32%)	3128
Cuidados Paliativos	36 (0.35%)	703
Dermatovenereologia	167 (1.62%)	2225
Doenças Infeciosas	250 (2.43%)	3060
Endocrinologia	149 (1.45%)	906
Estomatologia	55 (0.53%)	370
Estomatologia-Valongo	4 (0.04%)	19
Gastroenterologia	95 (0.92%)	708
Ginecologia	197 (1.92%)	1155
Hospital Dia/Quimioterapia	14 (0.14%)	18
Imagiologia	5 (0.05%)	1273
Imunoalergologia	77 (0.75%)	751
Imunohemoterapia	72 (0.70%)	135
Internato Medico Geral	67 (0.65%)	2424
Interno Ano Comum	231 (2.25%)	925
Medicina 1	2 (0.02%)	2
Medicina A Internamento	491 (4.77%)	2663
Medicina B Internamento	243 (2.36%)	4916
Medicina do Trabalho	24 (0.23%)	163
Medicina Fisica e Realibilitação	162 (1.58%)	5454
Medicina Fisica e Realibilitação-Valongo	3 (0.03%)	4
Medicina Intensiva	101 (0.98%)	160
Medicina Interna	266 (2.59%)	2144
Medicina Nuclear	41 (0.40%)	19163
Medicina Outros Serviços	10 (0.10%)	132
Nefrologia	135 (1.31%)	4004
Neonatalogia	54 (0.53%)	147
Neurocirurgia	176 (1.71%)	2626
Neurofisiologia/EEG	25 (0.24%)	2407
Neurologia	242 (2.35%)	5161

Continued on next page

Table 3.3 – continued from previous page

Medical Service	Associated IDs	Number of accesses
Neurologia Pediátrica	4 (0.04%)	3
Neurologia UCI	8 (0.08%)	349
Neurorradiologia	61 (0.59%)	64200
Obstetrícia Unidade Cuidados Intermédios	6 (0.06%)	37
Obstetrícia	72 (0.70%)	294
Obstetrícia Piso 4	7 (0.07%)	116
Obstetrícia Piso 5	6 (0.06%)	27
Obstetrícia/Ginecologia	127 (1.23%)	2921
Oftalmologia	155 (1.51%)	1886
Oncologia Médica	16 (0.16%)	8
Ortopedia	168 (1.63%)	10758
Ortopedia-Valongo	2 (0.02%)	11
Otorrinolaringologia	239 (2.32%)	450
Patologia Clínica	84 (0.82%)	577
Pediatria	720 (7.00%)	7899
Pediatria Médica	31 (0.30%)	14
Pneumologia	192 (1.87%)	2433
Psiquiatria	415 (4.04%)	4529
Psiquiatria-Valongo	1 (0.01%)	30
Radiologia	289 (2.81%)	219745
Radiologia-Valongo	7 (0.07%)	5355
Radioterapia	38 (0.37%)	903
Reumatologia	153 (1.49%)	586
Unidade Hemato-Oncologia	138 (1.34%)	1025
UCI Cardiologia	8 (0.08%)	94
UCI Cirurgia Programada	23 (0.22%)	21
UCI Pediatria	156 (1.52%)	2284
UCI de Neurocriticos	250 (2.43%)	2309
UCI Polivalente Geral	147 (1.43%)	1615
UCI Polivalente da Urgência	189 (1.84%)	4801
Unidade Exploração Funcional Respiratória	16 (0.16%)	22
Unidade Convalescença-Valongo	6 (0.06%)	10
Unidade Cuidado Intermédios	45 (0.44%)	522
Unidade AVC	34 (0.33%)	294
Unidade de Cuidados Paliativos	2 (0.02%)	11
Unidade de Queimados	19 (0.18%)	29
Unidade Doentes Neutropénicos	6 (0.06%)	13

Continued on next page

Table 3.3 – continued from previous page

Medical Service	Associated IDs	Number of accesses
Unidade Integrada Processos	2 (0.02%)	8
Unidade Oncologia	6 (0.06%)	6
Unidade Reumatologia	30 (0.29%)	804
UPC Intermédios Geral	98 (0.95%)	1709
UPC Intermédios Urgência	17 (0.17%)	5
Urgência-Adultos	424 (4.12%)	4454
Urgência-Valongo	37 (0.36%)	2
Urgência Geral/SO	76 (0.74%)	328
Urgência Pediátrica	333 (3.24%)	4382
Urgência/Outras	10 (0.10%)	17
Urologia	127 (1.23%)	499

By looking at the table we can do a straightforward analysis, observing an asymmetric distribution of associated IDs and number of accesses between medical services, and identifying several services with a very reduced number of associated IDs: *Bloco Operatório Central, Bloco Operatório Serviço Urgência, Estomatologia-Valongo, Imagiologia, Medicina 1, Medicina Fisica e Realibilitação-Valongo, Neurologia Pediátrica, Neurologia UCI, Obstetrícia Unidade Cuidados Intermédios, Obstetrícia Piso 4, Obstetrícia Piso 5, Ortopedia-Valongo, Psiquiatria-Valongo, Radiologia-Valongo, UCI Cardiologia, Unidade Convaslescença-Valongo, Unidade de Cuidados Paliativos, Unidade Doentes Neutropénicos, Unidade Integrada Processos, Unidade Oncologia, Urgência/Outras*. Generally, the number of accesses are proportional to the number of associated IDs, however there are cases where the number of accesses in a medical service is much higher when comparing to its number of associated IDs - *Imagiologia, Neurorradiologia, Radiologia, Radiologia-Valongo* -, and where the number of accesses is lower than the number of associated IDs - *Bloco Operatório Serviço Urgência, Cirurgia B, Neurologia Pediátrica, Oncologia Médica*.

By performing a pareto analysis, illustrated in figure 3.4, we can observe such asymmetric distribution related to the number of accesses. *Radiologia* is the medical service with most accesses (49.70% of the total number of accesses) and we can also conclude that the top three medical services with most accesses, *Radiologia, Neurorradiologia, Medicina Nuclear*, represent 68.55% of the total number of accesses. The remainder 89 medical services represent 31.45%.

Table 3.4 shows that 50% of the medical services have less or equal than 70 associated IDs, and all the medical services average approximately 112 associated IDs. On the right hand side, we can see that 50% of the medical services have a total number of accesses less or equal than 644, and the medical services average a total of around 4806 accesses. We can clearly observe a very irregular distribution.

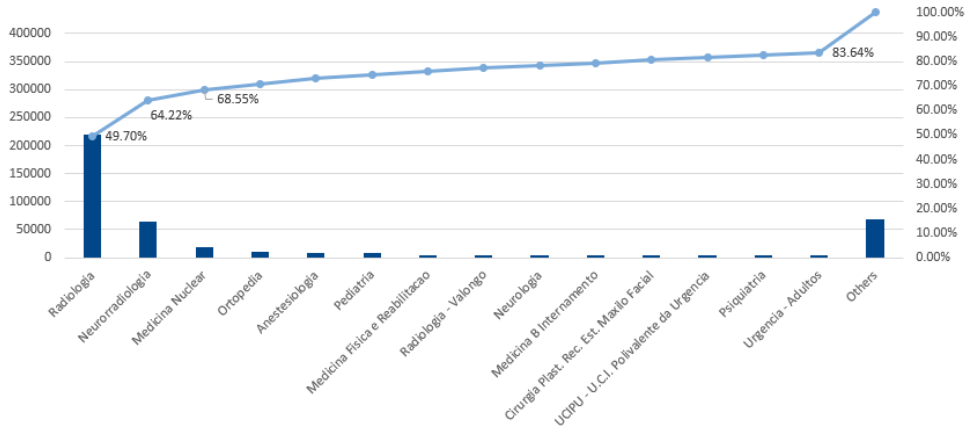


Figure 3.4: Pareto analysis of the number of accesses per medical service.

Physician ID		Number of accesses	
Min	1	Min	2
Max	720	Max	219745
Average	111.79	Average	4806.03
Median	70	Median	644.5

Table 3.4: Measures that summarize the variable Physician ID and the number of accesses per medical service.

Next, to understand how the accesses occur during the time, we analyze the access interval variable. As the chart 3.5 shows, the accesses are done most of the times (82.8%) within the first week, followed by accesses within the second week (6.8%) and within the first month (5.4%). The last two categories have a smaller importance, with less than 2%, in the accesses.

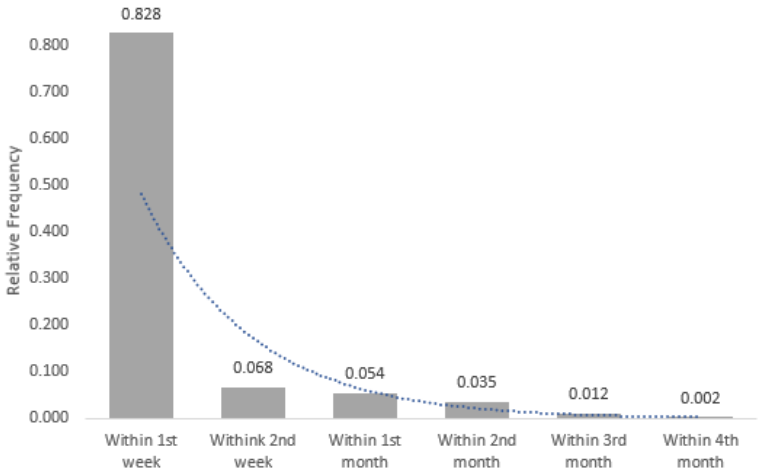


Figure 3.5: Distribution of the Access Interval.

As far as the characterization of the examinations is concerned, of the total number of examinations (52 643) approximately 18% have the exam code 10108, 9% the exam code 10033 and 3% the exam code 13099, referring to Chest (Thorax) X-Ray, Brain CT, and Electrocardiogram respectively. The remainder 755 exam codes correspond to around 70% of total examinations, as illustrated in figure 3.6.

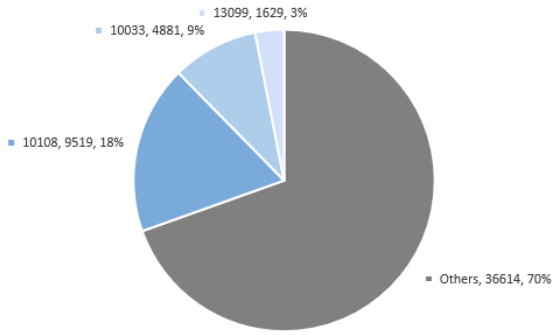


Figure 3.6: Exam Code frequency.

Figure 3.7 allows to conclude that the majority (52.58%) of the exams done are X-Rays (CR), followed by CT scans (21.43%) and Ultrasounds (11.25%). We observe a huge discrepancy between the different modalities. Digital Radiography (DR), Endoscopies (ES), X-Ray Angiography (XA) and External-camera Photography (XC) only represent 2%, approximately, of the total number of examinations.

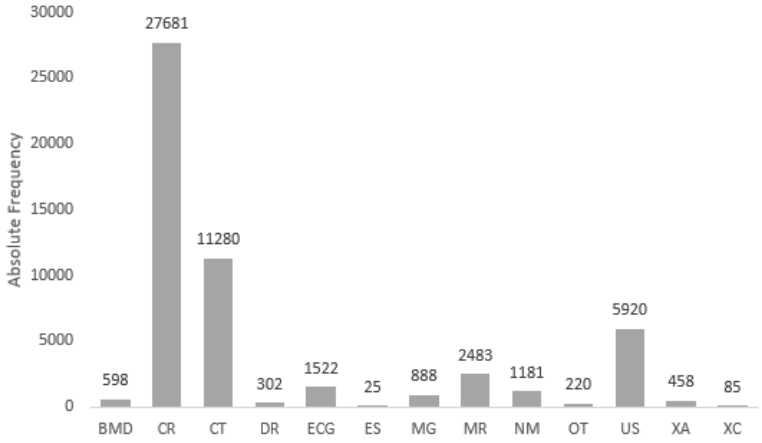


Figure 3.7: Modality distribution.

Furthermore, from figure 3.8 we conclude that the Module variable has also big variations in its distribution. Emergency (URG) is the most representative module with 38.4% of the number of examinations, medical appointments (CON), with 37.5%, is the second most representative and inpatient care (INT), the third with 23.4%. The modules RAD, BLO and HDI represent less than 1% of the examinations.

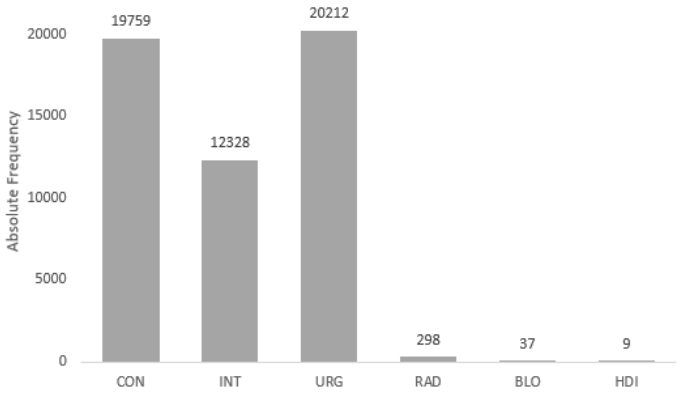


Figure 3.8: Module distribution.

We can also conclude that 97.4% of the examinations have report and 2.6% do not have report, as figure 3.9 shows. The boolean variables 1 and 0 were replaced by Yes and No, respectively, to better illustrate the Report variable.

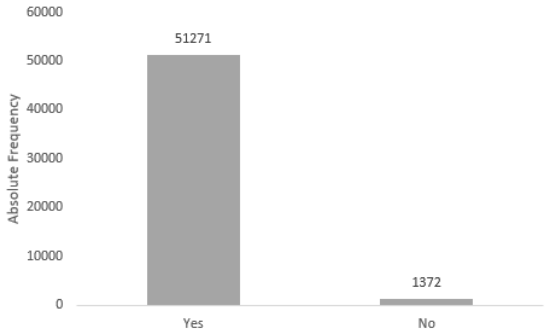


Figure 3.9: Distribution of the Report variable.

After an overview analysis of the different variables, to understand how the different medical services are related with the other variables, an analysis of the accesses per examination was done.

From table 3.5 we conclude that 50% of the examinations were accessed 5 or less times. The minimum number of times an examination was accessed is 1 and the maximum is 197. In average, an exam was accessed around 8 times.

Min	1
Max	197
Average	8.40
Median	5
Standard Deviation	10.44
Variance	109.09

Table 3.5: Measures that summarize the number of accesses per examination.

Then, all the physicians were grouped in two major supersets: Radiologists, include doctors from the medical services *Neurorradiologia, Radiologia, Radiologia-Valongo*; Other physicians, include the doctors from the remainder medical services. This was done due to the workflow in a radiology department, previously explained in chapter 2.1 and demonstrated in figure 2.1. Then for each exam, the number of accesses of each of the supersets was defined, and we conclude that Radiologists did not access 13047 examinations (24.8%) while Other physicians did not access 25481 examinations (48.4%). However, an examination was always seen, at least, by either a Radiologist or by Other physician.

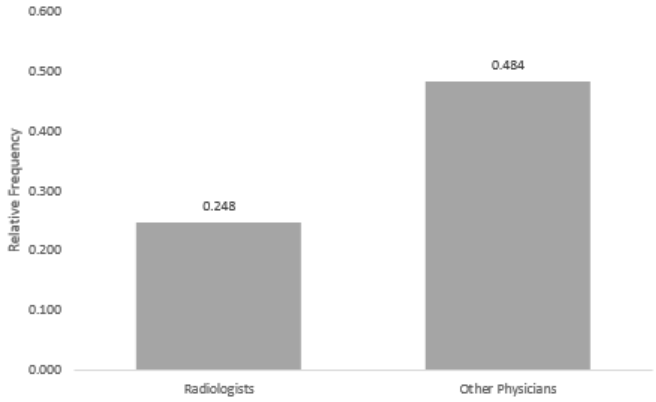


Figure 3.10: Frequency of examinations that were not accessed.

A deeper analysis was done to better understand why those situations occur, relating the modality variable with those exams. Focusing on the examinations **not accessed** by the Radiologists superset, 72% are X-Rays (CR), 7.9% Electrocardiograms, 7.4% regarding Nuclear Medicine (NM), 4.3% Ultrasounds (US) and 4.2% Bone densitometry (BMD).

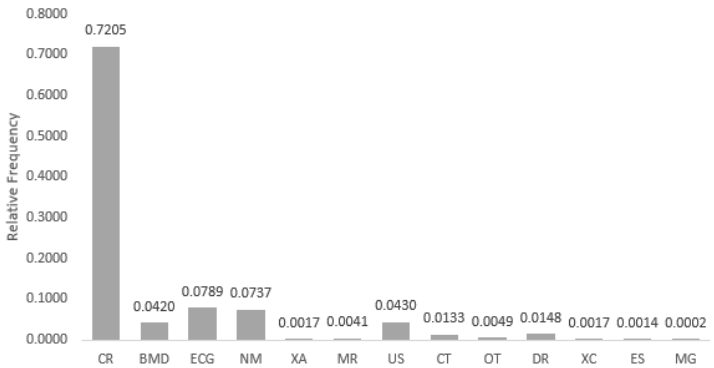


Figure 3.11: Frequency of the Modality of the exams not accessed by Radiologists superset.

For the most significant modality variables, the Body part variable was also analyzed. Looking into the CR modality, the biggest slice (around 47%) represents Thorax, the second and third

biggest (both with 5%) represents Spine and Ankle, followed by Foot with 5%, Hand with 4% and Pelvis with 2%, as shown in figure 3.12.

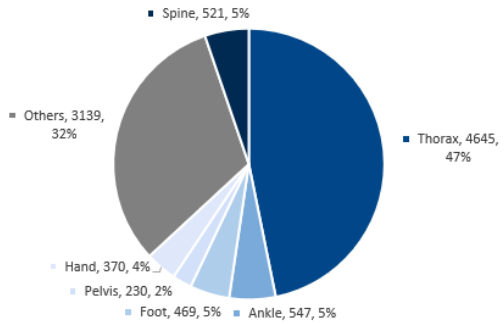


Figure 3.12: Frequency of Body Part variable in CR examinations not accessed by Radiologists superset.

In Ultrasounds, as we can observe in figure 3.13, a big percentage of the examinations is relative to Doppler ultrasounds of feet arteries (around 40%), 13% is relative to Doppler echocardiograms, 6% to Transcranial doppler ultrasound and 5% to Abdominal ultrasound.

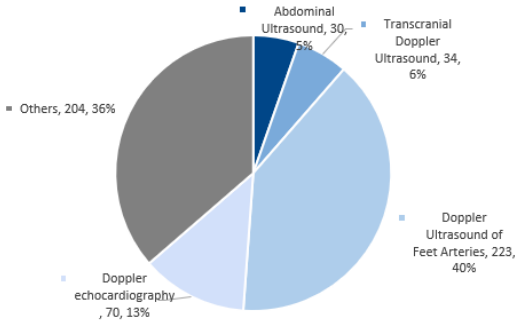


Figure 3.13: Frequency of Body Part variable in US examinations not accessed by Radiologists superset.

On the other hand, the same analysis was done to the exams which were **not accessed** by the Other physicians superset. As figure 3.14 illustrates, 48.4% of the total examinations not accessed by the Other physicians superset, 42.0% are X-Rays (CR), 25.4% are CT scans, 19.3% are Ultrasounds (US), 6.6% Magnetic Resonance (MR) and 3.4% Mammography (MG). The XA, ECG, DR, XC, NM, ES and OT modalities represent an insignificant value, sharing the remainder 3.3%.

For X-Rays, the pie chart 3.15 shows that approximately 29% of the CR examinations not seen by Other physicians superset are Thorax X-Rays, and around 19% are Spine X-Rays. Knee and Pelvis examinations represent 5% of the examinations only seen by Radiologists and Hand X-Rays 3%.

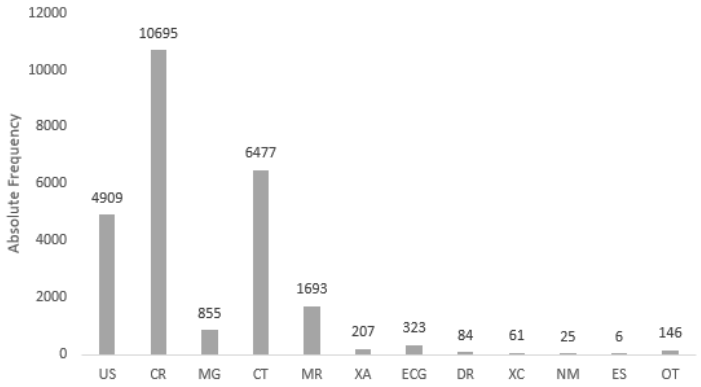


Figure 3.14: Frequency of the Modality of the exams not accessed by Others physician superset.

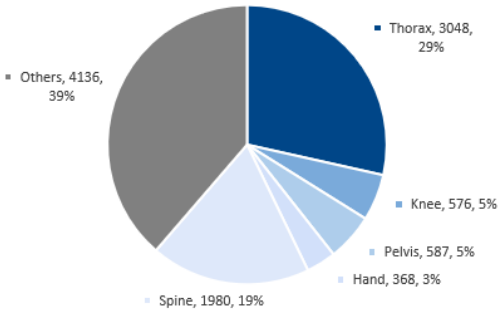


Figure 3.15: Frequency of Body Part variable in CR examinations not accessed by Other physicians superset.

Concerning CT examinations, from figure 3.16, we conclude that 44% of the not accessed examinations are Brain CTs, 9% are Chest and Abdominopelvic CTs, Chest CTs and Abdominopelvic Angiography both represent 6% and 5% are Abdominopelvic CTs.

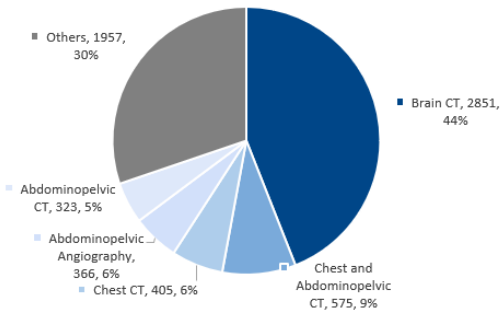


Figure 3.16: Frequency of Body Part variable in CT examinations not accessed by Other physicians superset.

Lastly, for the US examinations we can clearly distinguish ultrasonographies as the exams that are the most times not accessed by Other physicians. As the figure 3.17 shows, approximately

22% of those examinations are Abdominal Ultrasonograhies, 13% are Renal Ultrasonograhies, 9% are Abdominal and Renal Ultrasonograhies, 7% are Thyroid Ultrasonograhies, 6% are both Abdominal Pelvic Ultrasonograhies and Breast Ultrasonograhies, and with 5% Soft Tissue Ultrasonography. The remainder 32% are spread into multiple body parts.

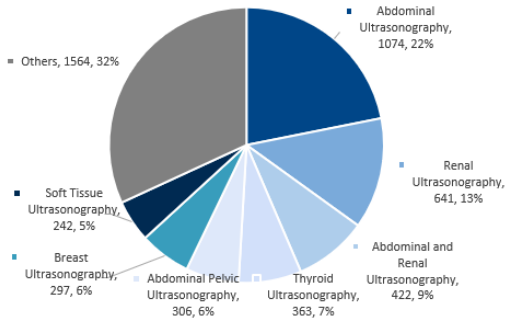


Figure 3.17: Frequency of Body Part variable in US examinations not accessed by Other physicians superset.

From this analysis we can easily observe that the examinations, which are not accessed, belong to certain types of examinations, from specific modalities. We can clearly identify which modalities are not seen the most times and which body parts are those exams referred to.

Globally for the Computed Radiographies, the examinations that were less accessed were Chest X-Rays, for CT scans were Brain CT and for Ultrasounds were Echo-doppler and Abdominal Ultrasonograhy.

3.5 Delivery

The results of the BI implementation were delivered in the form of a report using multiple charts to facilitate interpretation, as illustrated along the previous sections. Interactive dashboards and graphics, using PowerPivot[12] in Microsoft Excel, were also delivered for the entities involved in this project for an easier usage in daily operations.

Figure 3.18 illustrates the final interactive dashboard, where data is displayed regarding number of accesses, number of examinations and number of examinations not accessed. On the top layer, the distribution of the number of accesses per time interval is displayed, and the number of examinations per modality which were not accessed by either Radiologists or other physicians. On the bottom layer, the number of accesses per medical service is displayed, the number of examinations with or without report, the number of examinations per exam code, the number of examinations per medical modality and the number of examinations which were not accessed by either Radiologists or other physicians. In order to present customized information, it is possible

[12]https://support.office.com/en-us/article/get-started-with-power-pivot-in-microsoft-excel-fdfcf944-7876-424a-8437-1a6c1043a80b

to filter by month, on the bottom layer, or by service groups - i.e. Radiologists and other physicians - in the chart on the right hand size of the top layer.

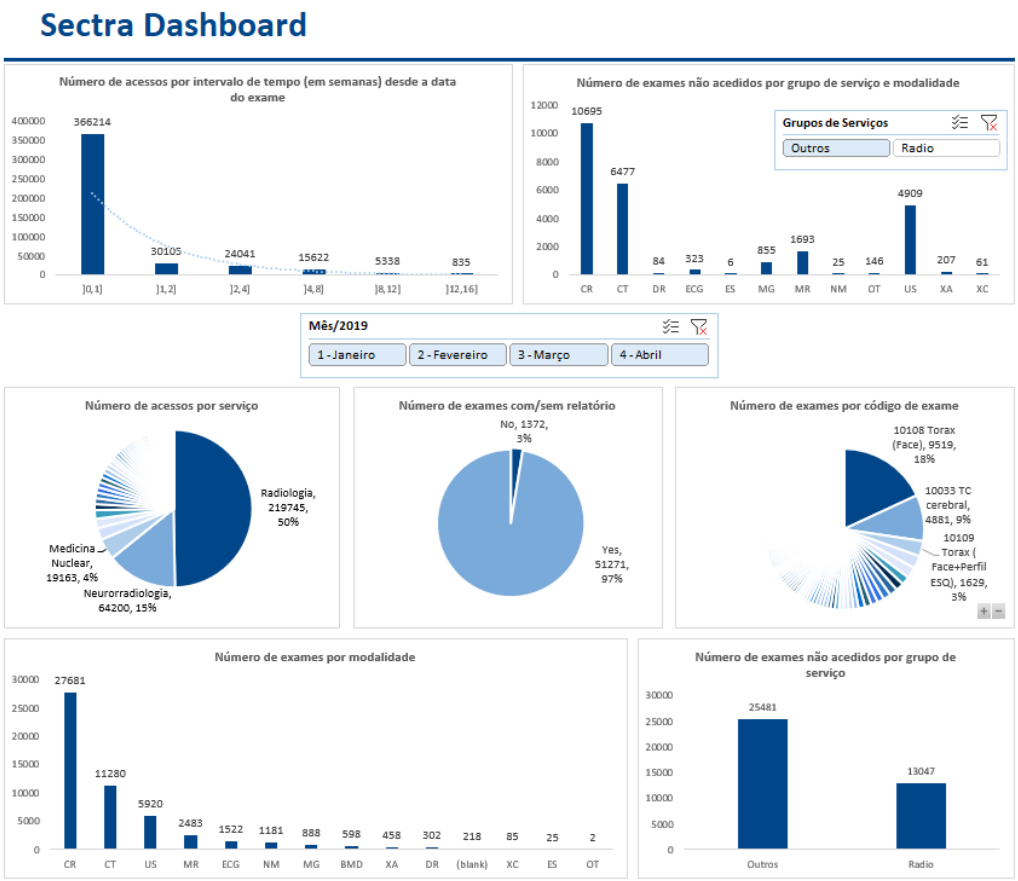


Figure 3.18: Final dashboard designed for end users.

Chapter 4

Segmentation by examinations and accesses

In order to obtain a segmentation of the medical services taking into account the number of accesses and the number of exams, a cluster analysis was carried out. As it was mentioned in Chapter 2, it can be helpful and/or needed to pre-process data prior to this analysis.

Along this chapter the segmentation's procedures are described and the experimental results of the process are presented. It starts with pre-processing analysis of new variables, following a brief discussion regarding the clustering methods and the number of clusters to retain. Finally, the obtained results and its analysis are presented.

4.1 Data pre-processing

At this stage, the goal is to obtain a partition of the medical services according to similar patterns of accesses. The data set was reduced to 92 objects, characterized by 11 variables. These variables were defined according to the opinion and need of the stakeholders. Each object corresponds to a medical service and the variables correspond to:

- X_1: Number of exams accessed.
- X_2: Number of accesses within the first week.
- X_3: Number of accesses within the second week.
- X_4: Number of accesses within the first month.
- X_5: Number of accesses within the second month.
- X_6: Number of accesses within the third month.
- X_7: Number of accesses within the fourth month.

- X_8: Number of exams accessed only one time.
- X_9: Number of exams accessed two times.
- X_{10}: Number of exams accessed three times.
- X_{11}: Number of exams accessed more than three times.

As observed in the Preliminary Data Analysis (section 2.2.2) the number of accesses by a medical service is, globally, proportional to the number of its associated IDs, although there is an irregular distribution. Thus, the medical services with more associated IDs tend to have more accesses, which in turn tend to access more exams. The number of accesses within the first week (X_2) also represent the majority of the Access Interval variables, which means that the value of the variable X_2 is going to be considerably higher than the value of the variables X_3, X_4, X_5, X_6 and X_7. Regarding the variables X_8,X_9,X_{10} and X_{11} it is evident that on average a medical service is more frequent to access twice an examination, and less frequent to access three times an examination.

In order to overcome the significant variance between values, normalization was done for each variable. Hence, the value in each variable corresponds to the percentage (instead of absolute number) of examinations or accesses. For X_1 variable, the number of exams was divided by the total number of examinations. For the variables X_2 to X_7, the number of accesses within a specific period was divided by total number of accesses of its respective medical service. For the variables X_8 to X_{11} the number of exams accessed a certain number of times, was divided by the total number of exams accessed (initial X_1 variable) by the respective medical service.

4.2 Cluster Formation Techniques

Regarding the clustering technique, we tested both K-means and DBSCAN. By performing the DBSCAN we obtained only one cluster with 60 objects and 32 objects identified as noise points. Since the results of the DBSCAN algorithm were unsatisfactory, only the results obtained from the K-means will be presented and discussed.

We used Jupyter Notebook[1] as IDE, to run the algorithms because it offers an interactive approach. Moreover, the main programming language is Python 2.7 and the core libraries used are NumPy[2] and pandas[3] to work with multidimensional data efficiently, Scikit-learn[4] for clustering analysis, and Matplotlib[5] to visualize results.

Prior to applying the algorithm, the number of clusters (K) needs to be determined, which is the number of groups (clusters) the algorithm is going to generate. Less significant parameters were also specified: the method for initialization, the number of times the algorithm will be run with different centroid seeds (*n_init*) and the maximum number of iterations (*max_iter*). This

[1] https://jupyter.org/
[2] https://www.numpy.org/
[3] https://pandas.pydata.org/
[4] https://scikit-learn.org/stable/
[5] https://matplotlib.org/

way, we randomly chose K objects for the initial centroids (method of initialization), defined the value 500 to *n_init* and 100 to *max_iter*. In order to determine the number of clusters, we used the Elbow Method and the Davies-Bouldin Index. Figure 4.1 illustrates a case where the Elbow Method is ambiguous. According to the Elbow Method, typically the plot resembles an "arm", and the “elbow” (the point of inflection on the curve) will serve as the reference for the determination of K. In this case there is clearly no point of inflection so another scoring method, Davies-Bouldin Index, was used.

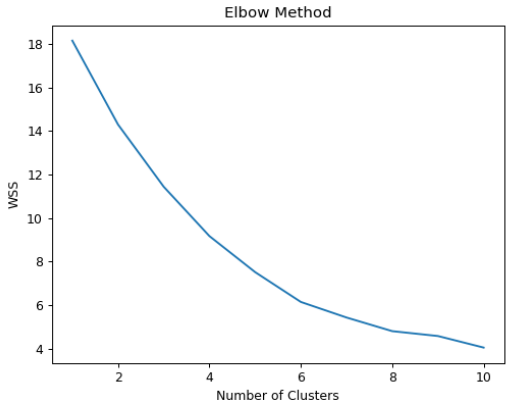


Figure 4.1: Elbow Method Graph.

The second method is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Clusters which are further apart and less dispersed will result in a better (lower) score. This way, we are looking for lower values which mean better clusters. We infer from figure 4.2 that the number k of clusters to use is 6.

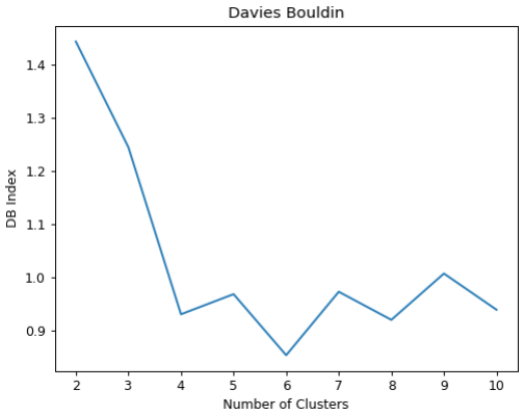


Figure 4.2: Davies-Bouldin Index Graph.

After defining the parameters, the K-means algorithm was applied. In the following section, we proceed to the analysis of the obtained clusters.

4.3 Cluster Analysis

As stated previously, the K-means algorithm divided the data set into 6 clusters. The respective distribution through the multiple clusters can be seen in figure 4.3. Cluster 5 is the biggest with 67 medical services, followed by cluster 3 with 10, cluster 4 with 7, cluster 6 with 6, and cluster 5 and cluster 1 with only 1 medical service. Since clusters 1 and 5 only have one client each, they do not have sufficient data sample. Due to their less importance, they can not represent a pattern or a segment on their own.

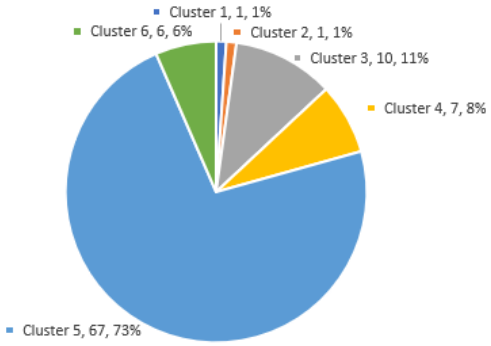


Figure 4.3: Cluster distribution dashboard.

In chart 4.4 it is shown the average value of each of the normalized variables (X_n), which represent the number of accesses and the number of visualizations. We can not compare values between all the variables, we can only compare values between variables regarding periods of access, or the frequency of accesses, or between values in the same variable. So, from the chart's observation we can clearly distinguish that for the medical services in the clusters 3, 5 and 6, the majority of the accesses are done within the first week, decreasing continuously for the next five variables. Regarding cluster 4, the accesses within the second week decrease significantly, increasing afterwards for the accesses within the first and second months. We can also identify a similar pattern in clusters 3, 4, 5, and 6. All of them increase from the value of exams accessed once to exams accessed twice, decreasing in the normalized number of exams accessed three times and increasing for exams accessed more than three times. The medical services in cluster 6 are the ones who access more examinations. Clusters 2 and 1 have no accesses in all the time periods except in the first month and in the third month, respectively. They also exclusively accessed examinations more than three times and exactly three times. We can conclude that clusters 1 and 2 represent completely different patterns comparing to the other clusters. Thus, these two clusters are considered outliers.

In table 4.1, it is represented the average value of the normalized variables for each cluster, which are illustrated in chart 4.4, when comparing to the average of each medical service, ignoring any cluster partition. From there we can confirm the interpretation of the chart 4.4, and we can also compare to the average of all medical services. As can been seen from the table data, cluster 6 has a really high volume of examinations accessed when comparing to the other clusters and

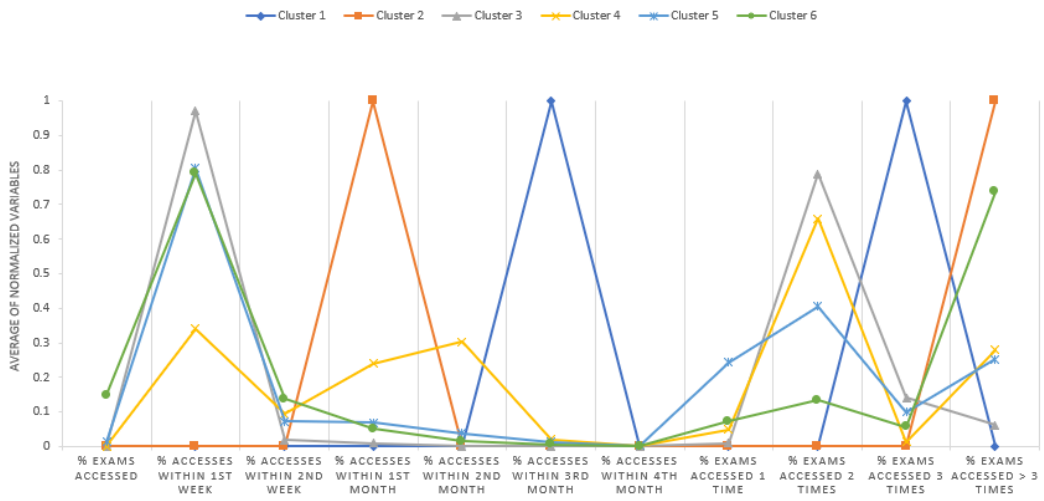


Figure 4.4: Average value of the normalized variables for each cluster.

to the global average. Medical services in cluster 5 have a similar behaviour to the average. In conclusion, both clusters 1 and 2 are not representative patterns nor do they have the same behaviour as the other clusters.

To complement the information of table 4.1, table 4.2 shows the average number of exams accessed regarding each variable (excluding time interval variables) and the global average.

In summary, from the K-means algorithm six different groups were identified. Clusters 1 and 2 are groups with one medical service each, with very distinct behaviour patterns from the global average. Cluster 3 aggregates 10 medical services with values far below average. However we can see that in "Accesses within the first week" and "Exams accessed twice" variables, its values are actually significantly above the average. This means that the medical services of cluster 3 generally access the examinations within the first week and access two times per examination. Cluster 4 represents medical services that access examinations more times after the first week and tend to access twice or more than three times to the examinations. Cluster 5 includes the majority of the medical services, approximately 73%, although usually with values very similar to the global average. Cluster 6 includes the medical services, such as *Radiologia, Neurradiologia* that access more number of times to the examinations and a higher number of exams. This cluster has a more homogeneous distribution of the frequency accesses variables, being the one that accesses more exams.

	% Exams Accessed	% Accesses within 1st week	% Accesses within 2nd week
Cluster 1	0.00004	0	0
Cluster 2	0.00004	0	0
Cluster 3	0.00012	0.97143	0.01905
Cluster 4	0.00044	0.34069	0.09374
Cluster 5	0.01227	0.80444	0.07262
Cluster 6	0.14866	0.79025	0.13685
Average	0.01881	0.77393	0.07078
	% Accesses within 1st month	% Accesses within 2nd month	% Accesses within 3rd month
Cluster 1	0	0	1
Cluster 2	1	0	0
Cluster 3	0.00952	0	3.47E-18
Cluster 4	0.24065	0.30296	0.01992
Cluster 5	0.06890	0.03908	0.01330
Cluster 6	0.05129	0.01562	0.00534
Average	0.08187	0.04966	0.02234
	% Accesses within 4th month	% Exams Accessed once	% Exams Accessed twice
Cluster 1	0	0	0
Cluster 2	0	0	0
Cluster 3	2.17E-19	0.00909	0.78924
Cluster 4	0.002045	0.04888	0.65964
Cluster 5	0.001674	0.24358	0.40586
Cluster 6	0.000656	0.07217	0.13397
Average	0.00141	0.18892	0.43753
	% Exams Accessed 3 times	% Exams Accessed >3 times	
Cluster 1	1	0	
Cluster 2	0	1	
Cluster 3	0.14000	0.06167	
Cluster 4	0.01235	0.27913	
Cluster 5	0.09856	0.25199	
Cluster 6	0.05655	0.73731	
Average	0.10343	0.27012	

Table 4.1: Value of the normalized variables by cluster and global average.

	Exams Accessed	Exams Accessed 1 time	Exams Accessed 2 times	Exams Accessed 3 times	Exams Accessed >3 times
Cluster 1	2.0004	0	0	2.0004	0
Cluster 2	2.0004	0	0	0	2.0004
Cluster 3	6.3572	0.0578	5.0173	0.8900	0.3920
Cluster 4	22.9997	1.1243	15.1715	0.2839	6.4200
Cluster 5	646.0438	157.3626	262.2059	63.6757	162.7996
Cluster 6	7826.1666	564.8570	1048.4540	442.5578	5770.2979
Average	990.0978	185.8804	309.3696	89.4891	404.3587

Table 4.2: Number of the examinations by cluster and global average.

Chapter 5

Mining Association Rules

In the following sections we explain how the association rules were obtained. Similar to chapter 5 pre-processing the data is also required prior to the analysis, then the association rules mining and its respective results' analysis are presented.

5.1 Data pre-processing

At this stage, the goal is to obtain associations between multiple variables. To do so, a binary representation was done similarly to the market basket problem (vide Chapter 2 section 2.2.2.2). The data set consists of 52 643 objects, whereby each object represents an examination, characterized by 875 variables. The variables consist of medical services (92), medical modalities (14), modules (6), exam codes (758) and visualization patterns (5) - Visualized more than three times, Visualized more than eight times (above average), Visualized after two weeks, Visualized after one month, Visualized after two months.

Variables on medical services, indicate whether a certain medical service accessed an exam. We can have multiple medical service variables in the same object. Modalities, modules and exam codes variables represent the modality, the module and the exam code of an examination, respectively. These three sets of variables are mutually exclusive regarding each set of variables. If an exam corresponds to CT modality, it can not correspond to the MR modality. Visualization pattern variables represent the number of times an exam was accessed by any medical service - Visualized more than three times, Visualized more than eight times - and the longevity of the accesses - Visualized after two weeks, Visualized after one month, Visualized after two months. These two type of variables are not mutual exclusive. For instance, if an examination was accessed ten times, both variables "Visualized more than three times" and "Visualized more than eight times" are represented.

5.2 Association Rules Mining and Analysis

For rule generation we used the Apriori algorithm (vide Chapter 2 section 2.2.2.2). Once again Jupyter Notebook was used with Python 2.7, NumPy, pandas and apyori[1], which is a library for the implementation of the Apriori algorithm.

In this approach the minimum support (*minSup*) value chosen was 1%, the minimum confidence 50% and minimum lift 1, which generated 855 rules. As previously mentioned, support indicates how often a certain rule is applicable to a given data set, confidence determines how frequently items in the consequent (Y in $X \rightarrow Y$) appear in transactions that contain the antecedent (X) and lift expresses how X and Y are correlated.

Due to a Apriori theorem - if an itemset is frequent, then all of its subsets must also be frequent - many rules were discarded for being redundant and not adding valuable information. For instance, the rule {10008}→{CT} is discarded because the exam code 10008 refers to a Chest Thorax, which can only be performed in a CT modality. Other subsets that were generated due to this theorem, were also discarded, except specific cases. In table 5.1 we can observe a set of rules, chosen randomly, obtained from the Apriori algorithm, with the respective value of the three measures - lift, confidence and support. The number of generated rules from the algorithm was considerably high, so it was impractical to present them all. From the table we can obtain the following conclusions:

- Brain CTs (10033) done regarding inpatient care that are visualized by the medical service *Radiologia*, then are likely to also be visualized by the medical service *Neurorradiologia*.
- Computed Radiographies that are visualized by the medical service *Cirurgia Cardiotorácica*, are likely to be regarding inpatient care.
- Thorax X-Rays (10108) that are visualized by the medical service *Neurologia*, are likely to concern inpatient care.
- Examinations that are visualized by the medical service *Cirurgia Cardiotorácica* are likely to be Thorax X-Rays (10108).
- Examinations regarding inpatient care, that are visualized by the medical services *Radiologia* and *Neurorradiologia*, and are visualized after weeks and more than three times, are likely to be visualized more than eight times.
- CT examinations regarding inpatient care that are visualized by the medical service *Radiologia*, and are visualized more than three times, are likely to be visualized more than eight times.
- Thorax X-Rays that are visualized by the medical service *Anestesiologia*, are likely to be regarding inpatient care.

[1] https://pypi.org/project/apyori/

- Examinations that are visualized by the medical services *Psiquiatria* and *Neurorradiologia* and are visualized more than three times, are likely to be visualized more than eight times.
- Examinations regarding inpatient care that are visualized by the medical service *Pediatria*, are likely to be Thorax X-Rays.
- Examinations visualized more than three times and visualized by the medical services *Medicina B Internamento* and *Neurorradiologia*, are likely to be visualized more than eight times.
- CT scans that are visualized after one month and are visualized by the medical service *Neurorradiologia*, are likely to be visualized more than eight times.
- Examinations visualized by the medical services *Neurocirurgia* and *Radiologia*, are likely to be visualized more than eight times.
- Examinations visualized by the medical service *Neurofisiologia/EEG* and visualized more than three times, are likely to be visualized more than eight times
- Nuclear medicine examinations regarding medical appointments, that are visualized more than three times, are likely to be visualized more than eight times.
- Magnetic resonance imaging examinations visualized by the medical services *Radiologia* and *Neurorradiologia* performed due to medical appointments, are likely to be visualized more than eight times.
- Electrocardiograms (13099) that are visualized more than three times, are likely to be regarding emergency (URG).

All things considered, globally the support values can be perceived low, however this is due to the large amount of data. Observing the confidence and the lift values, it is possible to evaluate the rules as reliable, uncovering strong relations between the multiple variables.

Rule	Lift	Confidence	Support
{INT, Radiologia, 10033, CT}→{Neurorradiologia}	5.7587	0.9680	0.0138
{CR, Cirurgia Cardiotoracica}→{INT}	3.6655	0.8584	0.0113
{CR, 10108, Neurologia}→{INT}	3.5777	0.8378	0.0100
{Cirurgia Cardiotoracica}→{10108}	3.2134	0.5811	0.0113
{INT, Visualized after 2 weeks, Radiologia, Neurorradiologia, Visualized > 3 times}→{Visualized > 8 times}	3.1342	0.9579	0.0117
{INT, Visualized after 2 weeks, Visualized > 3 times, Radiologia, CT}→{Visualized > 8 times}	3.1173	0.9528	0.0130
{CR, Anestesiologia, 10108}→{INT}	3.0831	0.7220	0.0145
{Psiquiatria, Neurorradiologia, Visualized > 3 times}→{Visualized > 8 times}	3.0144	0.9213	0.0100
{INT, Pediatria}→{10108}	3.0100	0.5443	0.0103
{Medicina B Internamento, Neurorradiologia, Visualized > 3 times}→{Visualized > 8 times}	2.9214	0.8929	0.0103
{Visualized after 1 month, Neurorradiologia, CT}→{Visualized > 8 times}	2.8088	0.8585	0.0120
{Neurocirurgia, Radiologia}→{Visualized > 8 times}	2.6936	0.8233	0.0106
{Neurofisiologia/EEG, Visualized > 3 times}→{Visualized > 8 times}	2.6777	0.8184	0.0122
{Medicina Nuclear, Visualized > 3 times, NM, CON}→{Visualized > 8 times}	2.6687	0.8157	0.0140
{MR, Radiologia, Neurorradiologia, CON}→{Visualized > 8 times}	2.6369	0.8059	0.0103
{ECG, 13099, Visualized > 3 times}→{URG}	2.6045	1.0	0.0119

Table 5.1: Sample of association rules obtained from Apriori.

Chapter 6

Conclusions

In this work, we have explored the visualizations of medical examinations by the medical services of *Centro Hospitalar Universitário São João*. With the goal of determining the usage patterns of the medical services, a cluster analysis was performed based on the number of exams accessed, the time interval of the accesses and their frequency. Cluster analysis is one of the most used techniques for segmentation, however usually underlying choices in this analysis are based on previous works and not particularly on the research problem. This leads to the fact that sometimes the outcome might not be accurate [TSK06].

The cluster formation techniques used were K-means, one of the most used and simple clustering algorithms, and DBSCAN, a density-based clustering algorithm [Ber02]. This way we could obtain clusters based on the behaviour of its medical services.

We obtained six clusters with the K-means algorithm, in which we could identify clusters 1 and 2 for having an anomalous behaviour. The medical services in cluster 3 access more exams within the first week, and are likely to access two times or three times an examination. Cluster 4 has a more homogeneous distribution in the time interval variables. It aggregates the medical services that access more times an examination within the first and second months and are likely to access two times. In cluster 5, the medical services access more examinations within the first week and access once or twice per examination. Lastly, cluster 6 represents the medical services who access more examinations within the first and second weeks and are likely to access an examination more than three times. In general, clusters 3, 5 and 6 all have similar behaviours, although having slight differences in the proportion of the number of examinations. Although the medical services in cluster 4 access much more examinations within the first and second months comparing to the other clusters, we can generalize that cluster 4 also follows a similar pattern of visualizations (vide chapter 5 section 4.3).

In this work we also explored frequent patterns between the different variables. In order to accomplish such goal, the Apriori algorithm was used resulting in multiple rules which correspond to relationships between the multiple variables - medical services, medical modalities, examinations'

modules, examinations' exam code and visualization variables. Apriori algorithm implementation result in strong, relevant and informative rules, meeting expectations of both parties - Sectra and *Centro Hospitalar Universitário São João*.

6.1 Future Work

Although a consistent and complete analysis was done to our data set, the data collection period may be considered quite small. However, due to the need of researching, coding and collecting daily data, it turned impractical for extending the data collection period. Nevertheless, collecting data for a year may uncover more relationships between the different characterization variables, and may also result in a different and more precise segmentation.

The automation of processes, resulting in an analytical tool where all data analysis is presented and illustrated for the multiple end-users in a hospital, would also be a major advance for this work.

References

[AbhKS17] Karim Abouelmehdi, Abderrahim beni hssane, Hayat Khaloufi, and Mostafa Saadi. Big data security and privacy in healthcare: A review. *Procedia Computer Science*, 113:73–80, 12 2017.

[ACJ16] Osama Ali, Pete Crvenkovski, and Helen Johnson. Using a business intelligence data analytics solution in healthcare. In *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1–6. IEEE, oct 2016.

[AFHC13] Tim Aubry, Susan Farrell, Stephen W. Hwang, and Melissa Calhoun. Identifying the Patterns of Emergency Shelter Stays of Single Individuals in Canadian Cities of Different Sizes. *Housing Studies*, 28(6):910–927, sep 2013.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[AÖNC13] Osama Ali Özkan, Ali Nassif, and Luiz Capretz. Business intelligence solutions in healthcare a case study: Transforming oltp system to bi solution. June 2013.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[Ber02] Pavel Berkhin. Survey of clustering data mining techniques. *A Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data: Recent Advances in Clustering.*, 10, October 2002.

[BVB+13] Michelle A. Borkin, Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, December 2013.

[Che95] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, August 1995.

[COV+10] Alin Cordos, Bogdan Orza, Aurel Vlaicu, Serban Meza, Carmen Avram, and Bogdan Petrovan. Hospital information system using HL7 and DICOM standards. *WSEAS Transactions on Information Science and Applications*, 7(10):1295–1304, 2010.

[CR06] Aaron Ceglar and John F. Roddick. Association mining. *ACM Comput. Surv.*, 38(2), July 2006.

REFERENCES

[CV10] Christi E. Carter and Beth L. Vealé. *Digital radiography and PACS*. Mosby/Elsevier, 2010.

[DM92] William Delone and Ephraim McLean. Information systems success: The quest for the dependent variable. *Information Systems Research*, 3:60–95, 03 1992.

[DS76] E. Diday and J. C. Simon. Clustering Analysis. pages 47–94. Springer, Berlin, Heidelberg, 1976.

[dS18] Serviço Nacional de Saúde. MCDT Realizados em Entidades Convencionadas. Available at https://www.sns.gov.pt/monitorizacao-do-sns/mcdts/, 2018. Accessed: 2019-02-05.

[Dub87] Richard C. Dubes. How many clusters are best? - An experiment. *Pattern Recognition*, 20(6):645–663, jan 1987.

[Eck09] Wayne Eckerson. Using Predictions to Power the Business. Technical report, The Data Warehousing Institute, 2009. Accessed: 2019-02-18.

[EFKS00] Martin Ester, Alexander Frommelt, Hans-Peter Kriegel, and Jöorg Sander. Spatial data mining: Database primitives, algorithms and efficient dbms support. *Data Min. Knowl. Discov.*, 4(2-3):193–216, July 2000.

[ENBJ05] Jason Ernst, Gerard Nau, and Ziv Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics (Oxford, England)*, 21 Suppl 1:i159–68, 07 2005.

[Fer12] Ana Luísa Rodrigues Ferreira. Identificação de Perfis de Hábitos de Consumo Através da Utilização de Cartões Bancários. 2012. Accessed: 2019-05-19.

[FK14] Neil Foshay and Craig Kuziemsky. Towards an implementation framework for business intelligence in healthcare. *International Journal of Information Management*, 34(1):20–27, 2014.

[GL86] John Gower and P Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48, 02 1986.

[GMVJ18] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi, and Meiko Jensen. Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. pages 1–7, 2018.

[GNS17] Rikke Gaardboe, Tom Nyvang, and Niels Sandalgaard. Business Intelligence Success applied to Healthcare Information Systems. *Procedia Computer Science*, 121:483–490, 2017.

[Gos12] A. Ardeshir Goshtasby. *Image Registration: Principles, Tools and Methods*. Springer Publishing Company, Incorporated, 2012.

[GT15] Junhao Gan and Yufei Tao. Dbscan revisited: Mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 519–530, New York, NY, USA, 2015. ACM.

[HGN00] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining - a general survey and comparison. *SIGKDD Explorations*, 2:58–64, 06 2000.

REFERENCES

[HK06] Jiawei Han and M Kamber. *Data Mining : Concepts and Technique*. Morgan Kaufmann, 01 2006.

[HRP+16] Raymond Heatherly, Luke V. Rasmussen, Peggy L. Peissig, Jennifer A. Pacheco, Paul Harris, Joshua C. Denny, and Bradley A. Malin. A multi-institution evaluation of clinical profile anonymization. *Journal of the American Medical Informatics Association*, 23(e1):e131–e137, 2016.

[HSM01] David J. Hand, Padhraic Smyth, and Heikki Mannila. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001.

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[KJR90] Leonard Kaufman and Peter J. Rousseeuw. *Partitioning Around Medoids (Program PAM)*, pages 68 – 125. January 1990.

[KKS17] Ayman Khedr, Sherif Kholeif, and Fifi Saad. An Integrated Business Intelligence Framework for Healthcare Analytics. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(5):263–270, 2017.

[Kol01] Erica Kolatch. Clustering algorithms for spatial databases: A survey. April 2001.

[Liu06] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[MC05] Francois Meyer and Jatuporn Chinrungrueng. Spatiotemporal clustering of fmri time series in the spectral domain. *Medical image analysis*, 9:51–68, March 2005.

[Min17] Lloyd B Minor. Harnessing the Power of Data in Health. *Stanford Medicine Health Trends Report*, 2017. Accessed: 2019-02-03.

[MP87] László Manczinger and Gábor Polner. Cluster analysis of carbon source utilization patterns of trichoderma isolates. *Systematic and Applied Microbiology*, 9(3):214 – 217, 1987.

[MS03] Ron Martin and Peter Sunley. Deconstructing clusters: Chaotic concept or policy panacea? *Journal of Economic Geography*, 3:5–35, February 2003.

[Mur12] Ishola Dada Muraina. Healthcare Business Intelligence: The Case of University's Health Center. *International Conference on E-CASE & E-TECH*, 2012.

[NCH11] Ali Nassif, Luiz Capretz, and Danny Ho. Estimating software effort based on use case point model using sugeno fuzzy inference system. pages 393–398, nov 2011.

[NCH12] Ali Bou Nassif, Luiz Fernando Capretz, and Danny Ho. Software Effort Estimation in the Early Stages of the Software Life Cycle Using a Cascade Correlation Neural Network Model. In *2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 589–594. IEEE, aug 2012.

[NHC13] Ali Bou Nassif, Danny Ho, and Luiz Fernando Capretz. Towards an early software estimation using log-linear regression and a multilayer perceptron model. *Journal of Systems and Software*, 86(1):144–160, jan 2013.

[NMN13] John W. Nance Jr, Christopher Meenan, and Paul G. Nagy. The Future of the Radiology Information System. *American Journal of Roentgenology*, 200(5):1064–1070, may 2013.

[O'C17] Stephen O'Connor. What Are the Differences Between PACS, RIS, CIS, and DICOM? Available at https://www.adsc.com/blog/what-are-the-differences-between-pacs-ris-cis-and-dicom, 2017. Accessed: 2019-02-08.

[Pag17] Ugo Pagallo. The Legal Challenges of Big Data: Putting Secondary Rules First in the Field of EU Data Protection. *European Data Protection Law Review*, 2017.

[Pec11] Mark Peco. Tdwi business intelligence fundamentals: From data warehousing to business impact. 2011.

[PWSTE03] Mark P. Wachowiak, Renata Smolíková, Georgia Tourassi, and Adel Elmaghraby. Similarity metrics based on nonadditive entropies for 2d-3d multimodal biomedical image registration. *Proc SPIE*, 5032, may 2003.

[RA15] Chandan K. Reddy and Charu C. Aggarwal. *Healthcare Data Analytics*. CRC Press, 2015.

[RP17] John Mark Michael Rumbold and Barbara Pierscionek. The effect of the general data protection regulation on medical research. *Journal of Medical Internet Research*, 19(2):1–6, 2017.

[RSJM10] Anant Ram, Jalal Sunita, Anand Jalal, and Kumar Manoj. A density based algorithm for discovering density varied clusters in large spatial databases. *International Journal of Computer Applications*, 3, 06 2010.

[SAS+17] Muneeb Ahmed Sahi, Haider Abbas, Kashif Saleem, Xiaodong Yang, Abdelouahid Derhab, Mehmet A. Orgun, Waseem Iqbal, Imran Rashid, and Asif Yaseen. Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions. *IEEE Access*, 6:464–478, 2017.

[SAW15] Ali Seyed Shirkhorshidi, Sr Aghabozorgi, and Teh Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE*, 10:e0144059, dec 2015.

[SAWH14] Ali Seyed Shirkhorshidi, Sr Aghabozorgi, Teh Wah, and Tutut Herawan. Big data clustering: A review. volume 8583, june 2014.

[Shn96] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, Sep. 1996.

[SSS14] Marc J Schniederjans, Dara Schniederjans, and Christopher M Starkey. *Business Analytics Principles*. Pearson Education, Inc., 2014.

REFERENCES

[Teg99] David P. Tegarden. Business information visualization. *Commun. AIS*, 1(1es), January 1999.

[tHE] Integrating the Healthcare Enterprise. Ihe wiki scheduled workflow. Available at https://wiki.ihe.net/index.php/Scheduled_Workflow. Accessed: 2019-02-07.

[TSK06] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[Ves91] Iris Vessey. Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature. *Decision Sciences*, 22(2):219–240, mar 1991.

[VP06] Herna Viktor and Eric Paquet. *Visualization Techniques for Data Mining*. 01 2006.

[VW07] Fernanda B. Viégas and Martin Wattenberg. Artistic data visualization: Beyond visual analytics. In *Proceedings of the 2Nd International Conference on Online Communities and Social Computing*, OCSC'07, pages 182–191, Berlin, Heidelberg, 2007. Springer-Verlag.

[WKB18] Yichuan Wang, LeeAnn Kung, and Terry Anthony Byrd. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126:3–13, jan 2018.

[WQK+17] D. D. Wang, T. X. Quan, A. Khoo, S. Sridharan, S. Ramachandran, S. H. Ng, and S. N. A. Rahman. Cluster analysis on utilization patterns of patients with chronic diseases based on flattened electronic medical records. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 293–298, June 2017.

[You18] Sang Young Lee. Architecture for Business Intelligence in the Healthcare Sector. *IOP Conference Series: Materials Science and Engineering*, 317(1):012033, mar 2018.

[Zha18] Dongpo Zhang. Big data security and privacy protection. *8th International Conference on Management and Computer Science (ICMCS 2018)*, 77(Icmcs):275–278, 2018.

[Zhe17] Jack G. Zheng. Data visualization in business intelligence. *Global Business Intelligence*, pages 67–81, 2017.

[ZK04] Ying Zhao and George Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331, 06 2004.

REFERENCES

Appendix A

Additional information

In the following section, it is presented additional information regarding the medical services in study.

A.1 Health professional services

In this study, we consider 92 different medical services: *Anatomia Patológica, Anestesiologia, Anestesiologia/Reanimação, Bloco Operatório Central, Bloco Operatório Serviço Urgência, C. Estomatologia, Cardiologia, Cardiologia Pediátrica, Cirurgia A, Cirurgia B, Cirurgia Cardiotorácica, Cirurgia Geral, Cirurgia Pediátrica, Cirurgia Plástica Reconstrução Estética e Maxilo Facial, Cirurgia Vascular, Cuidados Paliativos, Dermatovenereologia, Doenças Infeciosas, Endocrinologia, Estomatologia, Estomatologia-Valongo, Gastroenterologia, Ginecologia, Hospital Dia/Quimioterapia, Imagiologia, Imunoalergologia, Imunohemoterapia, Internato Medico Geral, Interno Ano Comum, Medicina 1, Medicina A Internamento, Medicina B Internamento, Medicina do Trabalho, Medicina Fisica e Realibilitação, Medicina Fisica e Realibilitação-Valongo, Medicina Intensiva, Medicina Interna, Medicina Nuclear, Medicina Outros Serviços, Nefrologia, Neonatalogia, Neurocirurgia, Neurofisiologia/EEG, Neurologia, Neurologia Pediátrica, Neurologia UCI, Neurorradiologia, Obstetrícia Unidade Cuidados Intermédios, Obstetrícia, Obstetrícia Piso 4, Obstetrícia Piso 5, Obstetrícia/Ginecologia, Oftalmologia, Oncologia Médica, Ortopedia, Ortopedia-Valongo, Otorrinolaringologia, Patologia Clínica, Pediatria, Pediatria Médica, Pneumologia, Psiquiatria, Psiquiatria-Valongo, Radiologia, Radiologia-Valongo, Radioterapia, Reumatologia, Unidade Hemato-Oncologia, UCI Cardiologia, UCI Cirurgia Programada, UCI Pediatria, UCI de Neurocriticos, UCI Polivalente Geral, UCI Polivalente da Urgência, Unidade Exploração Funcional Respiratória, Unidade Convalescença-Valongo, Unidade Cuidado Intermédios, Unidade AVC, Unidade de Cuidados Paliativos, Unidade de Queimados, Unidade Doentes Neutropénicos, Unidade Integrada Processos, Unidade Oncologia, Unidade Reumatologia, UPC*

Intermédios Geral, UPC Intermédios Urgência, Urgência-Adultos, Urgência-Valongo, Urgência Geral/SO, Urgência Pediátrica, Urgência/Outras, Urologia.

A.2 Medical Services in Clusters

Each cluster contains the following medical services:

- Cluster 1: *Neurologia Pediátrica.*
- Cluster 2: *Medicina Fisica e Realibilitação-Valongo.*
- Cluster 3: *Bloco Operatório Serviço Urgência, Cirurgia B, Estomatologia-Valongo, Medicina 1, Obstetrícia Piso 5, Ortopedia-Valongo, UCI Cirurgia Programada, Unidade de Cuidados Paliativos, Unidade Doentes Neutropénicos, Urgência Valongo.*
- Cluster 4: *Medicina do Trabalho, Neurologia UCI, Oncologia Médica, Pediatria Médica, Unidade AVC, Unidade Integrada Processos, Unidade Oncologia.*
- Cluster 5: *Anatomia Patológica, Anestesiologia, Anestesiologia/Reanimação, Bloco Operatório Central, Cardiologia, Cardiologia Pediátrica, Cirurgia A, Cirurgia Cardiotorácica, Cirurgia Geral, Cirurgia Pediátrica, Cirurgia Plástica Reconstrução Estética e Maxilo Facial, Cirurgia Vascular, Cuidados Paliativos, Dermatovenereologia, Doenças Infeciosas, Endocrinologia, Estomatologia, Gastroenterologia, Ginecologia, Hospital Dia/Quimioterapia, Imagiologia, Imunoalergologia, Imunohemoterapia, Internato Medico Geral, Interno Ano Comum, Medicina A Internamento, Medicina B Internamento, Medicina Fisica e Realibilitação, Medicina Intensiva, Medicina Interna, Medicina Outros Serviços, Nefrologia, Neonatalogia, Neurocirurgia, Neurofisiologia/EEG, Neurologia, Obstetrícia Unidade Cuidados Intermédios, Obstetrícia, Obstetrícia Piso 4, Obstetrícia/Ginecologia, Oftalmologia, Ortopedia, Otorrinolaringologia, Patologia Clínica, Pediatria, Pneumologia, Psiquiatria, Radioterapia, Reumatologia, Unidade Hemato-Oncologia, UCI Cardiologia, UCI Pediatria, UCI de Neurocriticos, UCI Polivalente Geral, UCI Polivalente da Urgência, Unidade Exploração Funcional Respiratória, Unidade Convalescença-Valongo, Unidade Cuidado Intermédios, Unidade de Queimados, Unidade Reumatologia, UPC Intermédios Geral, Urgência-Adultos, Urgência Geral/SO, Urgência Pediátrica, Urgência/Outras, Urologia.*
- Cluster 6: *C.Estomatologia, Medicina Nuclear, Neurorradiologia, Psiquiatria-Valongo, Radiologia, UPC Intermédios Urgência.*