NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

# Lumbar Sciatic Pain Clinical Pathways

Internship

Filipa de Castro Fernandes

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# LUMBAR SCIATIC PAIN CLINICAL PATHWAYS

by

Filipa de Castro Fernandes

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

**Supervisor:** *Prof Doutor* Roberto Henriques

**External Supervisors***:* Paula Santos & Jéssica Zaqueu

November 2021

# DEDICATION

First, I want to dedicate this internship report to myself never to quit, be resilient, and constantly push myself to follow my dreams and projects.

To my family, who always push me to do better but to take care of myself first, for doing their best and me being able, in some way, of repaying them. A special dedication to my two sisters and my brother for looking up to me makes me want to be a better person and a good role model for them, as the oldest sister I am.

To my boyfriend, my best friend, who never doubted I could do anything I set my mind up to.

To my friends, who believe in my capacities more than I do and always encourage me to do what I love, even if we do not get to spend much time together.

To Jéssica, for being the best supervisor I could ask for in my first professional experience, and for being my friend and a big sister.

To Marta, for making sure I do not forget who I am and what I am capable of and believe more in me than I do.

To Paula and André for giving me the chance of being part of the insurance world and develop a project in the health area, which I love.

To the rest of my PAA team, also known as "the best team ever", who made me feel welcomed from the first day and from whom it was tough to say goodbye.

To professor Roberto for accepting to be my supervisor and helping me finishing this chapter.

# ABSTRACT

To provide a better experience to the clients, insurance companies are beginning to identify clinical pathways (clinical overlooks) of different non-oncological pathologies to deliver a better experience to their clients and reduce costs, by already knowing what needs to be done before, during and after a surgery, which may lead to time and money savings. This report presents a proposed approach to the lumbar sciatic pain pathology clients from the company where this internship was held. This internship had a duration of 12 months.

The creation of clinical pathways for the lumbar sciatic pain aims to help stakeholders to take a step back and see what diagnosis bring patients to the need of a surgery, which procedures are done before and after the surgery, and the main surgical procedures in this pathology. Additionally, the costs are measured, and the main providers for each procedure are analysed, in order to provide a better experience to the clients, indicating what lies ahead, where they should go to get the best treatment, and how much they will pay.

This approach was composed by a series of phases, such as business understanding, data understanding, cleaning and preparation, and modelling. There were biweekly and monthly meetings with health professionals and doctors to adjust some information that would be useful for the mentioned phases. In addition, a Predictive Model to identify which clients usually go through a lumbar sciatic surgery was built based on historical medical data, so that when clients come with specific diagnosis or already had certain pre-surgery procedures, identified in the main project of this internship, it is possible to be prepared for what will soon happen.

The most important results obtained from this approach were the main pre- and post-surgery procedures, the principal surgeries performed, the costs of each procedure and of the main pathways (pre-, day, and post-surgery), which medical providers give the best results, and the types of clients that usually suffer from lumbar sciatic pain pathology.

# KEYWORDS

Health Insurance; Lumbar Sciatic Pain; Clinical Pathways; Clustering; Association Rules; Predictive Model.

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **NHS** | National Health Service |
| **CPWs** | Clinical Pathways |
| **EPA** | European Pathway Association |
| **LBP** | Lower Back Pain |
| **CLBP** | Chronic Low Back Pain |
| **ARM** | Association Rule Mining |
| **MLP** | Multi-layer Perceptron |
| **CRISP-DM** | Cross Industry Standard Process for Data Mining |
| **RHA** | Regional Health Area |
| **PCA** | Principal Component Analysis |

# 1. INTRODUCTION

Over the years, there has been ageing of the population and an increase in the incidence of chronic diseases and comorbidities, which places healthcare services in a position where they need to create health systems that are more efficient and sustainable that deliver real value to patients, in another words, create a Value-Based Healthcare (VBHC) System. Instead of focusing on process, VBHC focus on outcomes that make the most remarkable difference to patients while making cost efficiencies within health services (European Institute of Innovation and Technology, 2021).

New healthcare delivery models highlight an approach to patient care and sharing of patient data so that care is aligned, and outcomes can be assessed easily (NEJM Catalyst, 2017). A clinical pathway, which is a clinical overlook with standard assumptions for a patient's clinical flow in disease state management or recovery (Hipp, Abel, & Weber, 2016, p. 1), is one of the tools to guide treatment decisions for a provider to build the patient's care steps based on evidence-based practice. In addition, some health insurance companies are looking at clinical pathways to limit cost while highlighting quality (Balch, 2021).

This report was done as part of an internship carried out at a health insurance company that belongs to an international group. The main goal was to construct lumbar sciatic pain clinical pathways to understand the typical path of clients in the company's portfolio and, thus, provide a better customer experience and steering, and potentially reduce costs.

## 1.1. COMPANY OVERVIEW

The Ageas Group, an insurance group headquartered in Brussels, is present in 14 countries in Europe and Asia: Belgium, United Kingdom, France, Portugal, Turkey, China, Malaysia, India, Thailand, Vietnam, Laos, Cambodia, Singapore, and the Philippines, constituting the largest share of the global insurance market. Its mission is "to provide an emotional and relevant experience in people's lives" (Médis, 2020, p. 4).

The Ageas Portugal Group is present in the Portuguese insurance ranking, comprising the following insurance brands: Ageas Seguros, Ageas Pensões, Médis, Ocidental and Seguro Direto. In addition to the insurance sector, the group has a position beyond insurance. It provides a range of services, being represented by Clínica Médis, Go Far, Kleya, Ageas Repara and Mundo Ageas (Médis, 2020, p. 5-6).

The internship was held at Médis, a Portuguese health insurance company, which currently ranks second in the global ranking in the Portuguese insurance market. It operates "a multi-channel distribution and partnerships" (Médis, Relatório sobre a Solvência e a Situação Financeira (SFCR), 2020, p. 30), from employees to individual clients, companies of all branches and tailor-made proposals.

At Médis, proposals are adapted to each person and phase of their life. An effort is made to anticipate changes and risks because the individual and health are seen as one (Médis, Sobre Nós, 2021). Médis is an integrated system available to each person, including Assistant Physicians and Nurses, available for 24-hour care through the Médis Line or Médis App (Médis, Sobre Nós, 2021).

Médis' net tax result in 2020 was 21 255 742 € (euros), an increase of 21.9% compared to 2019, with a net tax result of 17 442 585 € (Médis, Demonstrações Financeiras, 2020).

## 1.2. The team and Activities

Pricing and Advanced Analytics, the Médis department where the internship was taken, is composed of four fields: Data Science, Data Management, Solution Implementation and Actuary.

The Data Science field is responsible for developing statistical models and discovering trends or patterns in the data. Data Management is responsible for preparing the big data infrastructures for data scientists later analysis. The Solution Implementation interacts with internal stakeholders in order to understand reporting needs, gather requirements and design and build the solution created by Data Scientists and implemented by Data Managers. Finally, the Actuary is responsible for developing the pricing models, monitoring the portfolio and working closely with the underwriting and product development business.

The role I played in the internship falls within the Data Science field. The responsibilities were to carry out the project assigned, with the support and supervision of the appointed coordinator, project manager and head of the Pricing and Advanced Analytics department.

## 1.3. Internship Goals

The project proposed by the company aimed to identify the clinical pathways of a specific pathology (lumbar sciatic pain) and the creation of clinical episodes that promote a better experience and steering for the client and reduce costs.

These are procedures related to non-oncological conditions in the company's portfolio in the last five years. However, in this case, the year 2020 did not count for analysis to guarantee that the effect of the pandemic would not influence the analysis. The goal was to identify typical clinical pathways and calculate associated costs per pathology regarding the pre- and post-surgery period and the episode on the day of surgery (from the operation time until the patient is discharged).

The analysis of the clinical paths enables the company to know the behaviour of its portfolio in the face of identified pathologies. Creating these episodes will increase internal and provider efficiency; cost forecasting by pathology so that risk forecasts can be made; and the identification of typical customers, as well as customers who partially or fully complete their path with the company; and negotiation with providers.

## 1.4. INTERNSHIP REPORT OVERVIEW

This report is divided into six chapters to facilitate the understanding of the internship project.

In Chapter 1, the Introduction provides a preface of the scope of the project and a presentation of the company, and the team is made, as well as the main objectives of the internship.

Chapter 2, Literature Review, is composed by a presentation of the themes that will be discussed throughout the report. This chapter starts with a presentation of the Portuguese health system and how health insurance works in Portugal. It also provides information concerning customer steering, which allied with the clinical pathways' creation, form the aim of the project – to provide a better experience to the clients through data analysis. In addition, lumbar sciatic pain information is provided, to better understand the pathology that was examined during the internship, as well as a more technical literature review related with association rules in health insurance companies, clusters, and predictive models in health insurance companies, which were used during the internship.

Chapter 3 contains the information regarding the selected methodologies, SCRUM and CRISP-DM, and the technology and tools used to execute this project.

In Chapter 4, there are all the details about the project, regarding the Business Understanding, Data Understanding, Data Preparation, Modelling and the additional projects done during the internship.

Chapter 5 is the closing chapter with the main conclusions of the main project (Lumbar Sciatic Pain Pathways) and the internship in its whole.

# 2. LITERATURE AND TECHNICAL REVIEW

This section provides a more profound understanding of the project background, including research before the start of the project, theoretical information about Health Insurance in Portugal, customer steering, Clinical Pathways, details on Lumbar Sciatic Pain and a more technical explanation of Association Rules, Clustering and Forecasting.

## 2.1. HEALTH INSURANCE IN PORTUGAL

The Portuguese health system is composed of the National Health Service (NHS), the various health subsystems, the insurance sector and the "pure" private sector (Silva, 2009, p. 19). The NHS is universal and general; however, since health is a primary concern of citizens, there is a need to find an adequate provision of services (Silva, 2009, p. 3). Given this, due to insufficient resources to meet needs, the importance of the insurance sector has been increasing (Silva, 2009, p. 43). Galamba de Oliveira mentioned some factors for this increase, including coverage being increasingly wide-ranging, speed and accessibility to health care, and the "level of technological innovation, which allowed the availability of telemedicine and counselling remote doctor" (Rocha, 2021).

There is a strong network of hospitals and doctors all over the country when it comes to the top private insurance companies in Portugal, which means there is access to excellent healthcare services. This means people will have access to high-quality healthcare services in most cities in Portugal, even in relatively small villages (bePortugal, 2021). Portugal's most advised insurance companies are Allianz Care International Health Insurance, Cigna International Healthcare, Bupa, Médis and Multicare Fidelidade (bePortugal, 2021).

Through Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF) was possible to get information about the volume of policies and secure people with health insurance. As it is possible to see in Table 1, from 2019 to 2020, there was an increase of 4,0% when it comes to individual policies and 6,2% when it comes to group policies. In 2020 there were 1 432 143 individual "insured people" with health insurance, 12,2% more than in 2019 since in 2019 there were 1 275 987 people.

Relatively to the mean premium of the insurance, it rounded the 315,6 € in 2020, a decrease of -4,8% compared to 2019. This can also tell us that with the Covid-19 pandemic, the importance of health took a special place in the increase of health insurance policies, and with that, there was also a decrease in the mean premium.

| | 2020 | 2019 | 2018 | Δ 2020 / 2019 |
|---|---|---|---|---|
| **Number of policies** | | | | |
| Individual | 901 901 | 867 429 | 855 495 | 4,0% |
| Group | 78 810 | 74 175 | 68 477 | 6,2% |
| **Number of insured people** | | | | |
| Individual | 1 432 143 | 1 275 987 | 1 192 033 | 12,2% |
| Group | 1 792 002 | 1 614 787 | 1 479 338 | 11,0% |
| **Mean number of insured people by policy** | | | | |
| Individual | 1,6 | 1,5 | 1,4 | 7,9% |
| Group | 22,7 | 21,8 | 21,6 | 4,4% |
| **Mean premium by insured people in euros** | | | | |
| Individual | 315,6 | 331,6 | 319,8 | -4,8% |
| Group | 269,1 | 271,0 | 276,5 | -0,7% |

Table 1. General Information (direct insurance) - Health Insurance

Source: (Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF), 2020, p. 124)

## 2.2. CUSTOMER STEERING

Customer health service has grown to be very important, because customers usually do not search for healthcare companies when they are feeling good, but when they are in need for help. It is necessary to build a significant relationship with the customers to provide them a better experience.

Value Steering, in the context of this project, is a way of providing clients that suffer from lumbar sciatic pain pathology the best medical providers given the procedures most likely to happen given their medical condition and diagnosis. This provides a better experience to the client, because they already know which procedures to go through to recover from the pathology, which medical providers they should go, and how much time and money they will spend. Besides this, value steering is always getting insights about feedback, financial and implementation issues, which are present daily when attempting to improve customer service. In other words, value steering is a means to constantly develop the customer service level through the collection and anticipation of the appropriate data. Data that helps evolve according to the mission and values of the company (crmpartners, 2021). In summary, Customer Steering aims to empower customer interaction by identifying how customer alignment is employed in the company's strategy development, the different periods of the service life cycle model and decision-making and prioritisation regarding recent proposals (ICT Solutions for Brilliant Minds, 2021).

CSC – IT Center for Science is a Finnish center that has an essential role as a tool for steering and developing the Ministry of Education and Culture's education social and cultural policy. However, this can be applied to the health sector too, given that is it always adapted to the case at hand. In Figure 1 it is possible to observe an example of the CSC service lifecycle model regarding customer steering.

This example was chosen because it illustrates the reasoning behind customer steering employed in the project.

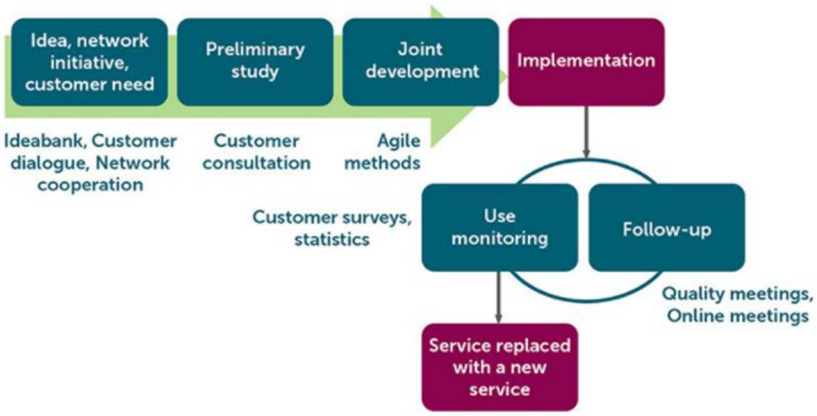



Figure 1. CSC service lifecycle model

Source: (ICT Solutions for Brilliant Minds, 2021)

The first step is for organisations to identify and comprehend segments of patients whose health and interrelated events generate a consistent set of needs (Teisberg, Wallace, & O'Hara, 2020, p. 2). In the company's case, two pathologies were already identified based on its portfolio volume (they cannot be mentioned due to professional confidentiality), and now the Lumbar Sciatic pain pathology, since it is very common for clients to have an episode of this matter. After achieving the first step, the next is to design a solution to improve health outcomes, but first, the segment must be thoroughly analysed. In the company, one way of doing it is through creating and analysing clinical pathways.

## 2.3. Clinical Pathways

Clinical Pathways (CPWs) appeared early in the 80s to respond to the health care reimbursement that was mainly concerned with the volume of care provided and not too worried about its quality (Hipp, Abel, & Weber, 2016, p. 2). A CPW provides clinical overlook and standard assumptions for a patient's clinical flow in disease state management or recovery (Hipp, Abel, & Weber, 2016, p. 1).

Usually, these pathways mainly focus on particular patient populations: high risk, high dollar, or high volume. However, with medical or surgical conditions that are constant and expected, these patients' data contribute to standardised health care, which provides improved results and efficiency with reduced expenses and variability (Hipp, Abel, & Weber, 2016, p. 2). In addition, from a patient point of view, CPWs deliver better assistance and awareness of what patients should be expecting during the care episode (Busse, Klazinga, Panteli, & Quentin, 2019, p. 352).

The European Pathways Association (EPA), founded in 2004, is the CPW professional leading organisation that supports the development, application, and evaluation of CPW. CPWs are being used in countries with public not-for-profit and with private for-profit healthcare providers (Busse, Klazinga, Panteli, & Quentin, 2019, p. 34).

As previously said, in the company, two pathologies were studied by identifying and analysing of clinical pathways. The choice of pathologies was made through an analysing the company's portfolio, focusing on the number of people evaluated.

The present report is about the project developed in the internship, which analysed a third pathology in the company: lumbar sciatic pain.

## 2.4. LUMBAR SCIATIC PAIN

The lumbar region of the spine, also known as the lower back, involves five vertebrae (spinal bones) called L1, L2, L3, L4 and L5, and they are the biggest in the back, giving support to the head and the body (Schnuerer, 2019).

Lower back pain (LBP) is one of the most frequent medical problems in the world (Guyer & Lewis, 2021) and the most frequent musculoskeletal condition impacting up to 84% of the adult population (Allegri, et al., 2016, p. 3).

Lower back pain can be classified as acute, subacute, or chronic. Acute lower back pain has a few days, to 4 weeks duration, while subacute lower back pain continues between 4 to 12 weeks. Nevertheless, around 20 percent of people suffering from acute back pain develop chronic low back pain (CLBP), which usually lasts 12 weeks or more, says the National Institutes of Health (https://www.spineuniverse.com/conditions/low-back-pain). These people have also demonstrated a one-year relapse rate of 24% to 80% (Casiano, Dydyk, & Varacallo, 2021).

The occurrence of acute low back pain and CLBP in adults has increased exponentially, especially in the ageing population, impacting men and women in all ethnic groups (Allegri, et al., 2016, p. 3). One Scandinavian study showed that back pain hits nearly 1% for 12-year-olds and 5% for 15-year-olds, with an accumulative rate of 50% by age 18 for females and 20 for males (Casiano, Dydyk, & Varacallo, 2021).

LBP has a meaningful impact on moving function, as pain constrains some activities and is the crucial cause of absenteeism. In addition, it has a substantial economic pressure characterised by the high costs of health care by decreased productivity. LBP symptoms may originate from anatomic and psychological factors such as nerve roots, muscle, fascial structures, bones, joints, intervertebral discs, and organs within the abdominal cavity or stress, depression, and/or anxiety (Allegri, et al., 2016, p. 3). Low back surgery intends to decrease or ease pain and repair lumbar spine stability (Schnuerer, 2019).

On another note, sciatica is a condition in which the patient suffers pain and/or paresthesia in the sciatic nerve or an associated lumbosacral nerve root (Davis, Maini, & Vasudevan, 2021). Sciatica has no distinguished gender majority and is mainly focused on people around their 40s or more, not being expected to affect people less than 20 years old (Davis, Maini, & Vasudevan, 2021).

The current project analyses the pathways related to lumbar sciatic pain because, usually, low back (lumbar) pain and sciatica, although different, are similar or consequential.

# 3. METHODOLOGY

## 3.1. METHODOLOGY

### 3.1.1. SCRUM

This chapter aims to give reasoning on the methodologies chosen to develop this project, Agile Scrum and CRISP-DM.

Agile Scrum methodology is a "sprint-based project management system" that aims to provide the highest value to stakeholders, to have more successful partnerships between teams, relying on incremental development (Business News Daily Editor, 2020). In this project, there were daily scrums, and biweekly sprint reviews, which are frameworks that transforms typical collaborations among teams into more effective ones (Business News Daily Editor, 2020), to evolve meeting after meeting, with new insights and ideas from the stakeholders.

### 3.1.1. CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) is a process model with six stages, illustrated in Figure 2, to describe the data science life cycle to help plan, organise and implement a data mining project (Data Science Process Alliance, 2021). This process model came to the public in 1999 to standardise data mining processes across industries. It became the most popular methodology for data mining, analytics, and data science projects (Data Science Process Alliance, 2021).

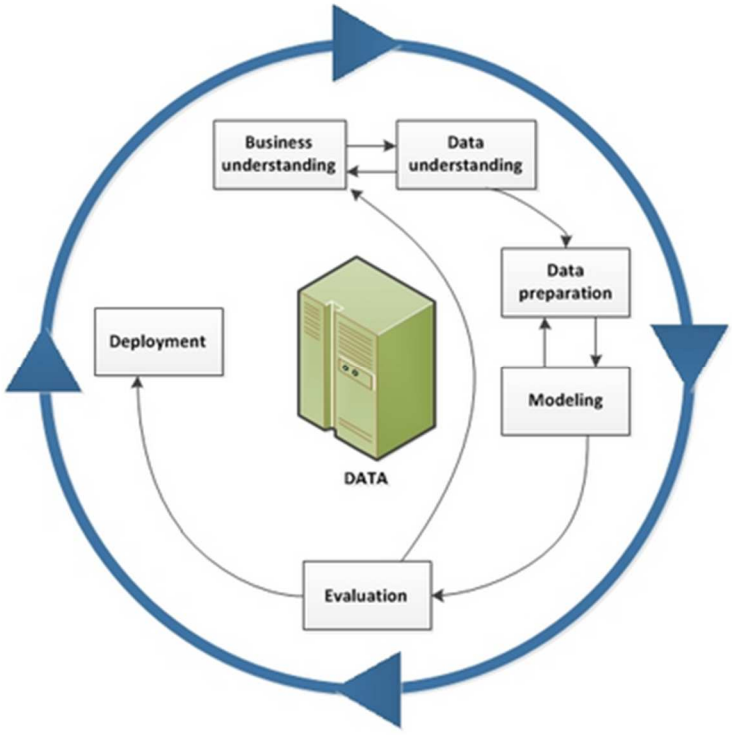



Figure 2. CRISP-DM Diagram

Source: (IBM, 2021)

Through Figure 2, it is possible to visualise the life cycle of the CRISP-DM process model. As previously stated, it is composed of six phases, and the grey arrows indicate the most common and relevant dependencies between stages. However, the order of the phases is not rigorous because, most of the time, projects move back and forth between stages, if needed (IBM, 2021).

#### 3.1.1.1. Business Understanding

The *Business Understanding* phase aims to understand the intentions and needs of the project and the customer. Four important aspects of this phase are: determine business objectives, assess the situation, determine data mining goals, and produce a project plan (Data Science Process Alliance, 2021).

*Determine Business Objectives* is helpful to understand, from a business perspective, what the customer wants to achieve, and define business success criteria (Data Science Process Alliance, 2021). This is where objectives are settled, the project plan is produced, and the business success criteria are defined.

The task *Assess the Situation* is valuable for establishing resource availability and project requirements, assessing risks and contingencies, and conducting a cost-benefit analysis (Data Science Process Alliance, 2021). Here it is supposed to be a detailed part about the inventory of resources, requirements, assumptions and constraints, risks and contingencies, terminology, and costs and benefits.

*Determine Data Mining Goals* is a part of the Business Understanding phase to present, from a technical data mining perspective, what a successful project is supposed to look like (Data Science Process Alliance, 2021). This involves defining the business success criteria and the data mining success criteria.

Finally, the task Produce Project Plan aims to select technologies and tools and define detailed plans for each project phase (Data Science Process Alliance, 2021).

#### 3.1.1.2. Data Understanding

The *Data Understanding* phase intends to identify, collect, and analyse the data sets that help achieve the project goals. It is made of four steps: collect initial data, describe, and explore the data, and verify data quality (Data Science Process Alliance, 2021).

*Collecting Initial Data* is beneficial for obtaining the essential data and, if needed, loading it into an analysis tool (Data Science Process Alliance, 2021).

*Describe the Data* is a valuable step since it analyses the data and documents its properties like data format, number of records, or field identities (Data Science Process Alliance, 2021).

*Explore the Data* is a beneficial task to dig deeper into the data, query it, visualise it, and discover relationships among the data (Data Science Process Alliance, 2021).

*Verify Data Quality* detects how clean or dirty the data is and documents any quality issues (Data Science Process Alliance, 2021).

#### 3.1.1.3. Data Preparation

The *Data Preparation* phase usually takes 80% of the project, as it involves preparing the final data set(s) for modelling (Data Science Process Alliance, 2021), and it can be a repetitive task, going back and forward. Typically, this phase implies the selection, cleaning, construction, integration, and formatting of data.

*Select Data* is where the data sets that will be used are determined and the reason behind the inclusion or exclusion (Data Science Process Alliance, 2021).

*Clean Data* is where the values are corrected, imputed or removed (Data Science Process Alliance, 2021).

*Construct Data* is an essential task because it serves to create new attributes or records that will be helpful for the modelling (Data Science Process Alliance, 2021).

*Integrate Data* may be a task or not because it creates new data sets by merging or aggregating data from multiple sources (Data Science Process Alliance, 2021).

*Format Data* might be a necessary task or not if it is necessary to convert a variable type to another (Data Science Process Alliance, 2021).

#### 3.1.1.4. Modeling

*The Modeling* phase, also known as the most exciting part of a data mining project, and the shortest, comprises four tasks: select modelling techniques, generate test design and build and assess the model (Data Science Process Alliance, 2021).

*Select Modeling Techniques* task is to decide which algorithms should be tried (Data Science Process Alliance, 2021).

*Generate Test Design* is where the data is split into training, validation, and test sets, depending on the model (Data Science Process Alliance, 2021).

*Build Model* is simply the appliance of code to fit and apply the chosen algorithm (Data Science Process Alliance, 2021), where the parameters can be adjusted.

*Assess Model* is where the models' results are translated according to domain knowledge and the business and data mining success criteria (Data Science Process Alliance, 2021), and where the parameters can be revised and tuned.

#### 3.1.1.5. Evaluation

*The Evaluation* phase is crucial because it is where the best model is chosen and what to do next. Has three critical tasks: evaluate results, review process, and determine the next steps (Data Science Process Alliance, 2021).

*Evaluate Results* tasks are observed if the model results align with the business success criteria (Data Science Process Alliance, 2021) and if they should or not be approved.

*Review Process* is a double check task to make sure all the tasks were suitably performed or make some corrections (Data Science Process Alliance, 2021).

*Determine the Next Steps* is where it is decided whether to advance to deployment, iterate further or initiate new projects (Data Science Process Alliance, 2021).

#### 3.1.1.6. Deployment

Finally, the *Deployment* phase is a big part of the data mining project because it cares for the access to the model by its customer. Therefore, plan deployment, monitoring and maintenance, producing a final report, and reviewing the project are the four tasks that compose this final phase (Data Science Process Alliance, 2021).

*Plan Deployment* is where the documentation of the plan for deploying the model is done (Data Science Process Alliance, 2021).

*Plan Monitoring and Maintenance* is to avoid issues after the model/project is complete (Data Science Process Alliance, 2021).

*The Produce Final Report* task is to document a resume of the project that may contain a final exhibition of the data mining results (Data Science Process Alliance, 2021).

*Rewrite Project* task is a retrospective about what went well, what could have been better and how to *improve in the future (Data Science Process Alliance, 2021).*

However, the data mining project may not end here, as it was previously told that the phases might be repeated and, besides, the model should be monitored and tuned often (Data Science Process Alliance, 2021).

### 3.2. TOOLS AND TECHNOLOGY

This chapter lists and explains the tools used in the internship: SAS Enterprise Guide, Jupyter Notebook and Power BI.

SAS Enterprise Guide was used to access, clean, and prepare the data for the analyses. In addition, Jupyter Notebook was employed to create Lumbar Sciatic Pain Association Rules and Clusters to

observe combinations of procedures and possible pathways and typical clients that need an intervention, either surgical or not, to treat the pain. Finally, Power BI to analyse critical information about the company's portfolio and build the final pathways.

### 3.2.1. SAS Enterprise Guide

SAS Enterprise Guide, a Windows .NET client application, has an easy-to-use graphical user point-and-click interface constructed to allow self-sufficient and guided access to the software with interactive dialogue boxes, as it is possible to see in Figure 3 and Figure 4.

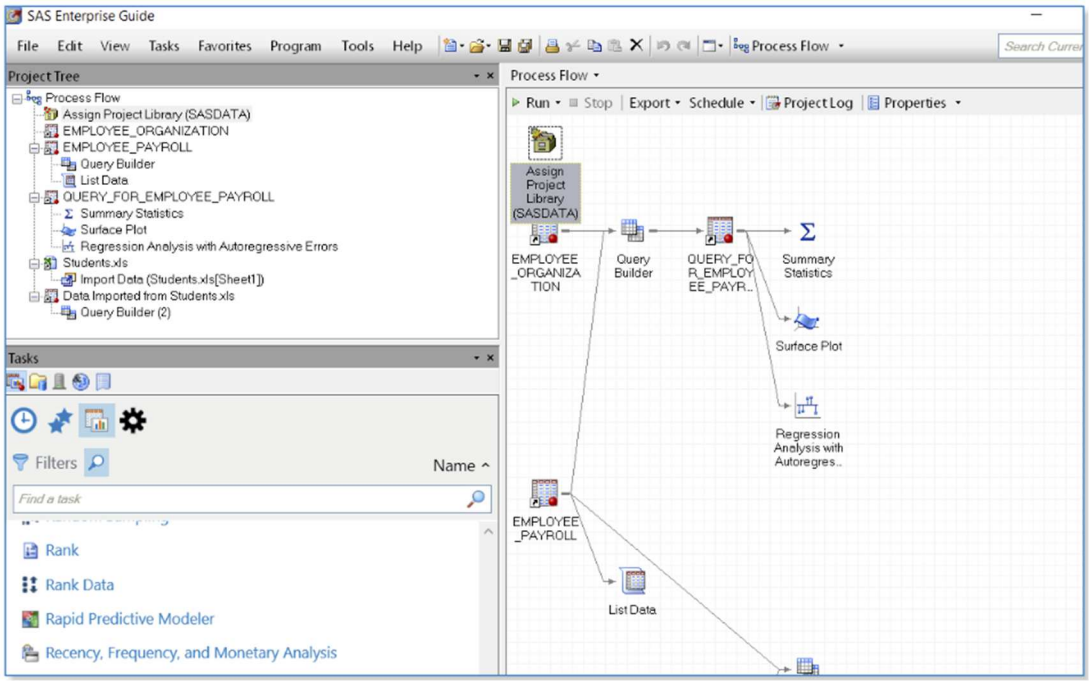



Figure 3. Process flows help users visually organise and maintain their projects projects

Source: (SAS, 2021, p. 2)

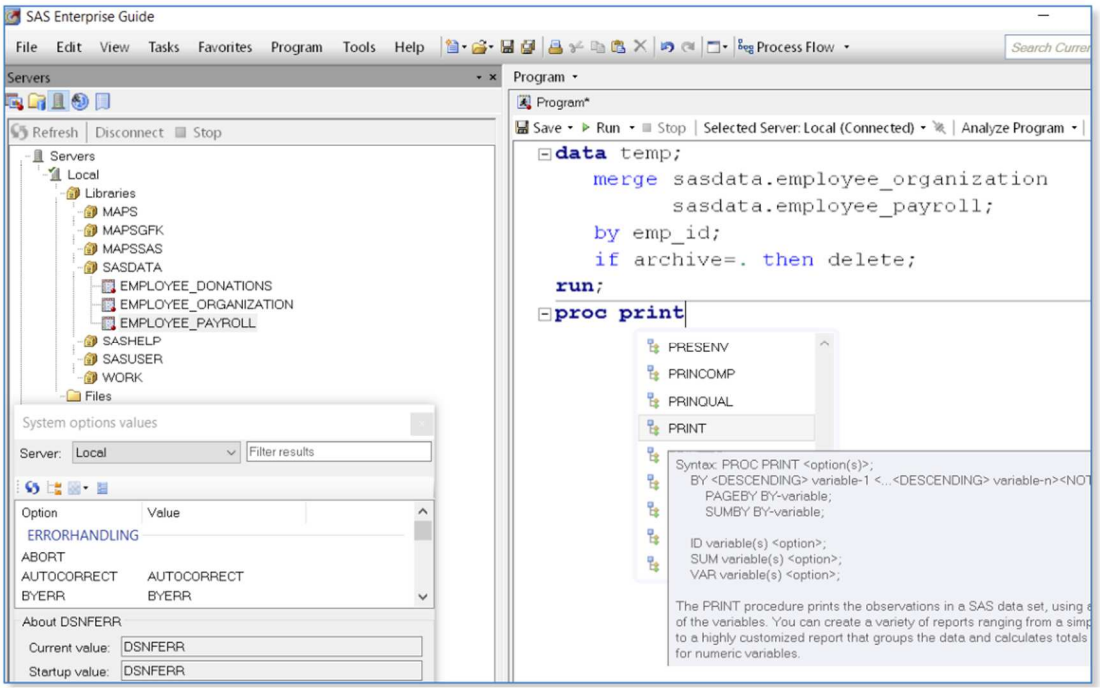



Figure 4. Pop-up tips and an autocomplete feature make programming faster and easier

Source: (SAS, 2021, p. 2)

It is a valuable tool to help business analysts, programmers, and statisticians access data, manipulate it, perform basic reporting, and develop dense and straightforward analyses. It also allows access to external data and export results, facilitating the usage for others (SAS, 2021, p. 1-2).

When it comes to the access and management of data, users can visually access any data type. The graphical query builder enables users to create, update, subset and join tables. They can easily manipulate data visually without using any programming language unless they want to – in this case, SQL code is the one to use. Nevertheless, the code behind the query builder is automatically generated and can be viewed and authenticated (SAS, 2021, p. 2)

Some of the benefits of using SAS Enterprise Guide are: the self-service and friendly user environment it provides to analysts and statisticians for workflow-based projects, where they can create analyses and reports; the centralised system for handling access to enterprise data guaranteeing appropriate access privileges to the users; the capacity to develop and deploy tailored tasks; results can be delivered through a framework with the ability to "publish dynamic, interactive content to Microsoft Office and web users" (SAS, 2021, p. 2).

### 3.2.2. Jupyter Notebook

The Jupyter Notebook is a free, interactive, open-source web application to produce and share documents with live code, equations, visualisations, and narrative text. It is helpful to clean, transform and visualise data, perform statistical modelling, and use machine learning. Jupyter allows using more than 40 programming languages, including Python (Jupyter, 2021), which was used in this internship.

#### 3.2.2.1. Association Rules

The usage of association rules is occasionally mentioned as "association rule mining" or "mining associations" (Lutkevich, 2020). Association Rule Mining (ARM), a crucial research method when it comes to data mining (Zhan, 2019, p. 2), is referred to as the detection of connections among two or more variables (Goh & Ang, 2007, p. 1), an "if-then" statement (Lutkevich, 2020). It is widely used in medicine, but it originated from the basket analysis problem to discover relationships among different commodity transaction databases to detect purchase patterns (Zhan, 2019, p. 2).

An association rule has two elements: an *antecedent* ("if") and a *consequent* ("then"). An *antecedent* is an item inside the dataset. A *consequent* is an item found in a pattern with the *antecedent*. Association rules are built through seeking frequent if-then patterns and using the support and confidence measure to detect the most important relationships (Lutkevich, 2020). In ARM, rules are only chosen if they fulfil both a minimum support and a minimum confidence threshold (Goh & Ang, 2007, p. 2).

Support gives the information of how frequently an item or an itemset appears in the data. It measures how often that association rule occurs in the entire set of transactions (Goh & Ang, 2007, p. 2). It can be represented by the following ratio, A and B being items or an itemset, and N the total number of transactions (Ansari, 2019, p. 21):

$$\boldsymbol{Support = P(A \cap B)/N}$$

Confidence specifies how often the "if-then" statements happen in the total transactions and measures the strength/reliability of a rule (Goh & Ang, 2007, p. 2). It can be characterised by the following fraction (Ansari, 2019, p. 21):

$$\boldsymbol{Confidence = \frac{P(A \cap B)}{P(A)}}$$

Another metric, named Lift, is used to calculate how many times an "if-then" statement is expected to be found faithful (Lutkevich, 2020). If the Lift result is more than 1, it indicates that the occurrence of the items on the LHS (Left-Hand-Side, or "A", or antecedent) has increased the probability that the items on the RHS (Right-Hand-Side, or "B", or consequent) will take place on this transaction. If the Lift is less than 1, it is the opposite. Finally, if the lift result is equal to 1, it suggests that items on the LHS and RHS are independent, which means that the items on the LHS do not impact the probability of RHS items occurring. The Lift can be stated by the following equation (Ansari, 2019, p. 21):

$$\boldsymbol{Lift = \frac{P(A \cap B)}{P(A) * P(B)}}$$

Furthermore, the Apriori algorithm was the one carefully chosen since it uses a strategy to calculate the support of itemsets. It performs the following structure of computations: 1. Determine support for itemsets of size 1; 2. Apply the minimum support threshold and prune itemsets that do not match the threshold; 3. Move to size 2 itemsets and duplicate phases one and two; 4. Keep up with the process until no further itemsets pleasing the minimum threshold can be obtained (Dobilas, 2021). To better understand this sequence, here is Figure 5 that illustrates it:

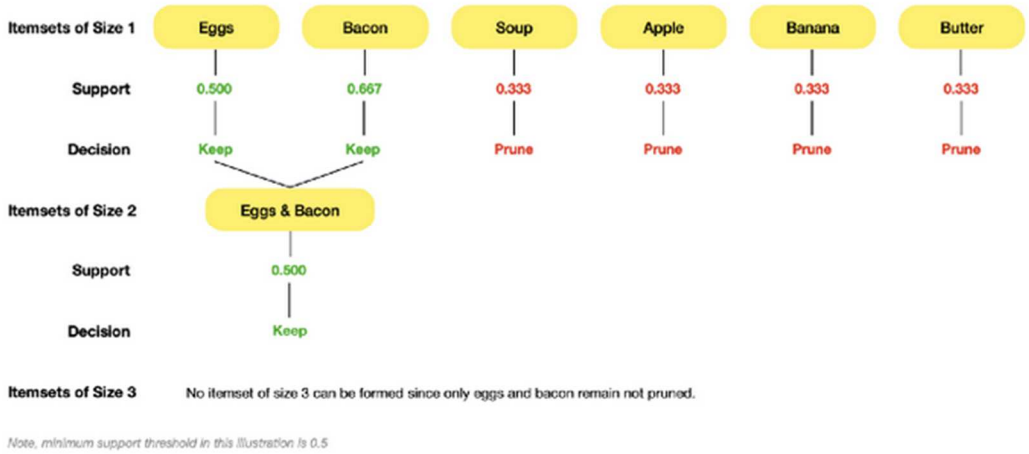



Figure 5. Process tree of Apriori algorithm

Source: (Dobilas, 2021)

#### 3.2.2.2. Clustering

Clustering is an unsupervised machine learning method of finding and gathering similar data points in greater datasets with no concern for the exact result (Explorium Data Science Team, 2020). In other words, it is an unsupervised learning method because there is no target, so it is impossible to compare the outputs with the proper labels to assess its performance (Dabbura, 2018). In this context, during the internship, data clustering, which is the process of finding and determining groups or "clusters" based on their similarity (Omran, Engelbrecht, & Salman, 2007, p. 2), was implemented to find typical clients who were going through lumbar sciatic pain, for example, their age or gender, their origin, and what do they usually go through.

The algorithm applied was K-means due to its simplicity, since it was preferred to have a simpler and already well-known algorithm. K-means is an iterative algorithm that partitions the dataset into k pre-specified subgroups, where each data point fits only in one group (Dabbura, 2018). These groups are heterogeneous, but each group is composed of homogeneous data points. This algorithm allocates data points to a cluster such that "the sum of the squared distance between the data points and the cluster's centroid – arithmetic mean of all data points that belong to the cluster – is at the minimum" (Dabbura, 2018). The less the variation, the more similar the data points inside each cluster. K-means works in the following way: 1. Indicates the number of clusters (k); 2. Prepares the centroids of each cluster by shuffling the dataset and selecting random k data points for the centroids; 3. Keep repeating the 2nd step until the allocation of the data points is no longer changing, i.e., until the closest data points to a cluster are aggregated (Dabbura, 2018). K-means approach to solve this problem is known as Expectation-Maximization. The Expectation step designates the data points to the nearest cluster, and the Maximization step computes the centroid of each cluster (Dabbura, 2018).

#### 3.2.2.3. Predictive Model using MLP Classifier

Predictive analytics analyses present and past facts to predict upcoming events by employing a range of statistical, data mining, and game theories (visionedge, 2021). To build a predictive model to forecast if a client will have a surgery or not, Multi-layer Perceptron (MLP) was the supervised learning algorithm chosen. MLP Classifier depends on a fundamental Neural Network to perform the task of classification (Nair, A Beginner's Guide To Scikit - Learn's MLPClassifier, 2019). This classifier applies an MLP algorithm that trains to use Backpropagation (learn, 2021), which is a training algorithm that accepts the adjustment of the weights in a multi-layer feedforward neural network, no matter how complex.

The training algorithm adjusts the synaptic weights to achieve the requested results. First, it randomly initialises the weights. Then, it stops if the result is good, but a slight adjustment is made if the results are wrong. The MLP works, as it is possible to see from Figure 6, through Input, Hidden and Output Layers.

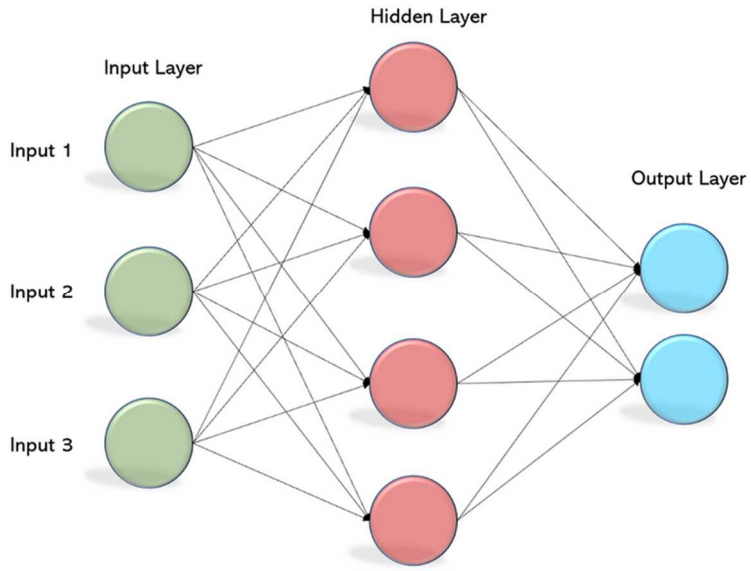



Figure 6. How the Multi-layer Perceptron works

Source: (Mohanty, 2019)

In the Input Layers, each neuron gets only one input directly from outside, and there is no activation function or any kind of processing. The Hidden Layers connect Input and Output Layers, where the classifications of the characteristics are done. Finally, the Output Layers are the output of each neuron that directly goes outside, and they have the same functionality as the Hidden layers.

### 3.2.3. Power BI

Another helpful tool used during the internship project was Microsoft Power BI to analyse data and visualise some tendencies, transforming it into knowledge. It also made it easier to interact with the stakeholders visually.

Microsoft Power BI is a Business Intelligence and Analytics solution where users can connect, transform, and model data, build charts, graphs, reports, and dashboards with visuals, and share those reports with other clients that use Power BI service (Stitch, 2021).

Power BI uses a user-friendly interface that allows its users to see, through charts and graphs, for example, what happened in the past, what is happening in the present, and what may happen in the future (Wright, 2019), as it is possible to observe in Figure 7.

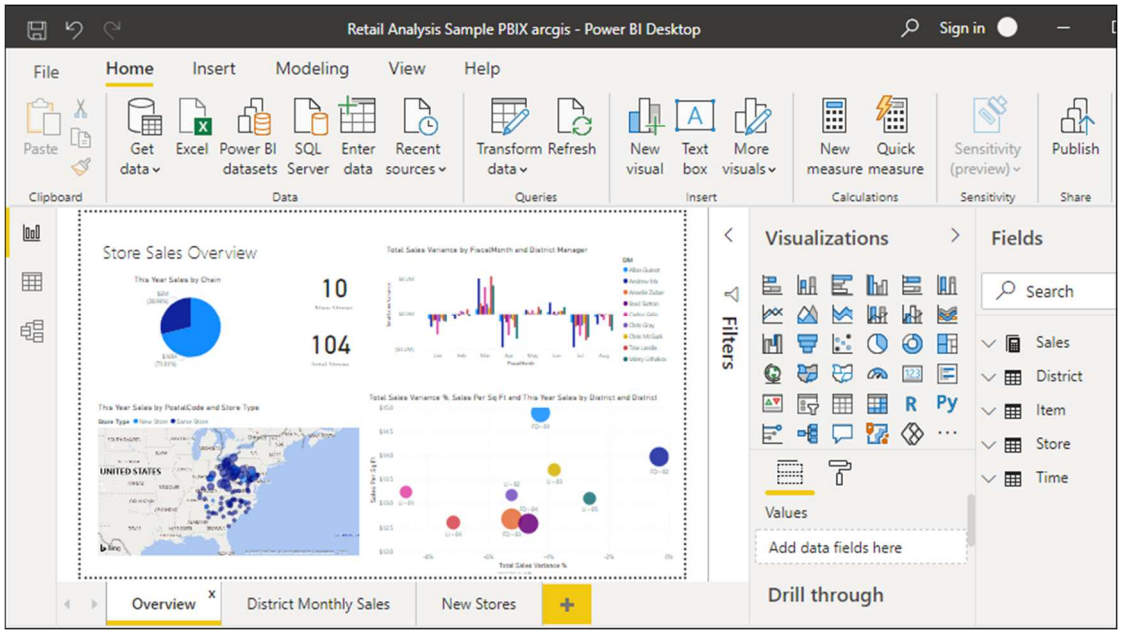



Figure 7. Power BI Desktop interface

Source: (Iseminger, 2021)

There are many benefits of using Power BI. For example, it is possible to input large amounts of data into the user-friendly and easy to navigate platform and to build machine learning features to identify trends and build predictions; through plenty of templates, users can develop dashboards where they can get instant access to the data and have a better perception of it; if users need to be aware of metrics and measurements, Power BI provides KPIs that with the option of alerting users (Wright, 2019); and dashboards can be updated in real-time, saving time and allowing users to work out their difficulties and find opportunities quickly.

During the project, Power BI was used to provide biweekly results to the stakeholders, by examining policies, company's portfolio, lumbar sciatic pain data, and, therefore, getting powerful insights about history data and essential information to make better decisions for the final project.

# 4. PROJECT

The main project on this internship was the development of clinical episodes through pathways for the lumbar sciatic pain pathology. It was done in nine months, the initial period scheduled for the internship.

However, the internship was extended to 12 months, and, in the remaining three months, two more projects were built – an automation of the data preparation used in the main project and a predictive model to forecast if people with lumbar sciatic pain history or diagnosis will need surgery or not.

## 4.1. LUMBAR SCIATIC PAIN CLINICAL EPISODES

### 4.1.1. Business Understanding

From a business perspective, the company wants to create lumbar sciatic pain clinical episodes to understand the typical path of clients in its portfolio and, thus, provide a better customer experience and steering, and potentially reduce costs. The business success criteria here is to get valuable insights about clients personal and medical characteristics through the company's specialists' judgement.

Now, assessing the situation, it is essential to point out the project's resources, either the people, the data, the computing resources, or the software. When it comes to the people that contributed to the project, me and my external supervisor are the project owners; the Clinical and Providers Network department was the project owners as well as the main stakeholders. Additional stakeholders of the project included the Claims Management and Marketing departments. The project sponsors were in the Pricing and Advanced Analytics department; the project manager is the solution implementer, and a clinical consultant of orthopaedics gave the clinical support. The data used for the project is part of the company's libraries. Regarding the computing resources or the software, the primary tool was SAS Enterprise Guide for the data preparation phase, where the data sources were allocated. For creating models, python was used on the Jupyter Notebook, and for data analysis, Power BI was the critical tool.

The requirements of this project were for the project to be delivered and presented at the end of the nine months of internship, May 27 2021, and that every 15 days, scrum meetings existed with the business owners. The constraints found on this project were concerned with the data quality because data was not always correct, for example, the gender or the date of birth information, and the data set was lacking a critical variable, a unique customer identifier. The creation of this variable, which will be explained in the Data Preparation phase, was useful because sometimes people end their insurance policy. When they come back, sometimes they return with only a few names, for example, the first time they come as Filipa Fernandes, but after they return as Filipa de Castro Fernandes or even with errors, such as Felipa or Flipa. Other times they have the birth date incorrect, for example, instead of 31/12/2000, they come with 21/12/2000, which becomes problematic to aggregate this data as one person and may affect the final results. Given this, the data preparation took a significant part in the project's construction (about 80%), and an ID variable was built.

This project may cost a lot to the company as it takes a great deal of time to understand the pathology and know what is essential to consider when preparing the data. However, the project may have better results, as there is time to reflect on what is essential to understand what else can be done. In addition, as the internship was extended, automatisation of the data preparation phase was built after completing the main project so that the company could perform clinical episodes for other pathologies faster. Finally, a predictive model to predict which clients with lumbar sciatic pain and the company's historical data would go through an intervention to treat the pain, surgical or not, so that the company may use that to provide a better experience to its customers, enabling information about the possibility of them needing surgery, based on historical data.

From a technical analytics perspective, the goal is to build clinical pathways for the lumbar sciatic pain pathology based on Médis clients that performed a low back intervention. This, therefore, looked at surgeries or treatments for pain relief, which procedures were associated to these treatments, and in which providers they took place. Additionally, it was important to consider the costs associated with the pathway. In order to do this, it was crucial to look at the episode in three distinct phases, composed of the pre-surgery, the day of the surgery, and the recovery process, post-surgery.

The plan to achieve the data mining business goals are, first, to understand the lumbar sciatic pain pathology and see what might be important to take into consideration, which involves talking to Dr António Rodriguez, orthopaedics specialist, regarding what usually happens before and after the day of surgery, and which procedures of the day of surgery are to be analysed. After this, analysis and understanding of the data sets to be used is made, and the data preparation phase starts with the SAS Enterprise Guide tool by creating a client identifier to not lose information about the clients that leave during a certain period, and the selection of the years and procedures of analysis chosen. After the data preparation phase is fully completed, data is analysed on Power BI, which facilitates depicting essential aspects of the final data set used to create the clinical pathways. After this, two models are also applied to consolidate the creation of the pathways, Association Rules and Clustering, using python, a programming language. The timeline of the project may be seen in Figure 8.

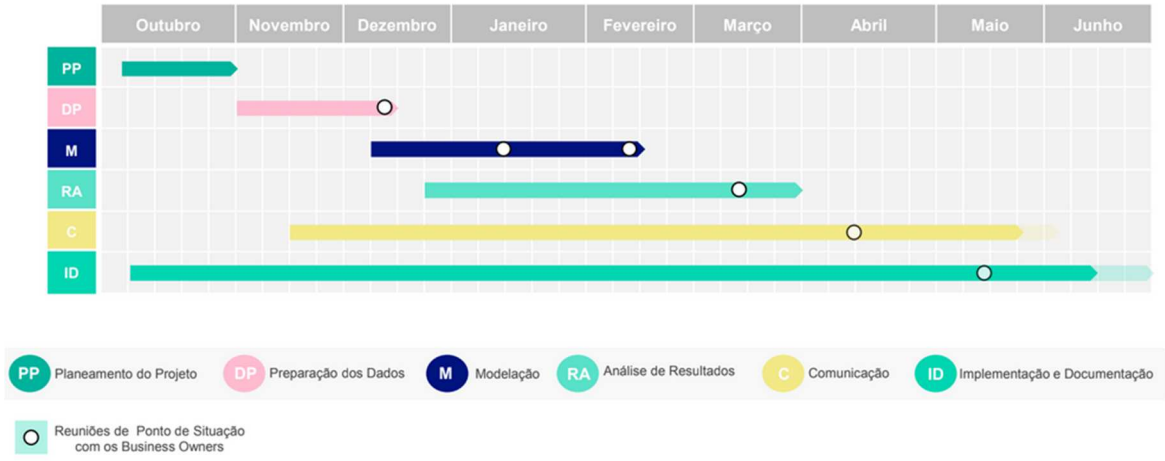



Figure 8. Project Timeline

### 4.1.2. Data Understanding

It was essential to understand what the lumbar sciatic pain consisted of, which implied deep research and multiple consulting meetings with the company's clinical consultants, during the data preparation of the project. One of these meetings was extremely important because the project was initially going to be about lumbar pain, but then the scope changed to lumbar sciatic pain. Also, the project was previously supposed to be about 60 procedures, but later it became about 44.

Afterwards, as it was necessary to create an ID variable, a data set containing data regarding the clients were observed, and variables that were considered to help build the ID variable were analysed to see how the aggregation could be built, as it is possible to observe in the example present in Table 2. The original data set is composed of millions of rows and about one hundred columns.

| **Policy Number** | **Birth Date** | **…** | **Name** | **Client Identifier** |
|---|---|---|---|---|
| 111111 | 01/01/1990 | … | A B C D | 123456789 |
| 222222 | 31/12/1976 | … | E F G H | - |

Table 2. Data Set 1 Example

Upon the completion of the creation of a unique ID per client, multiple data sets were chosen to prepare the data for the final pathways.

After choosing the type of clients for the project in a conversation with the clinical consultants, two data sets that contained data regarding the clinical aspects about the clients were used. As in Table 3, this data involves procedures, diagnosis, dates, values, and providers. Here, only a particular type of clients were selected for the project. The company's portfolio used in this project must have had at least 1 of 44 procedures present in their historical data. Furthermore, it should be clients with one of these procedures in the last five years, without counting with the year 2020 given the pandemic caused by Covid-19, which means the years being analysed were 2015, 2016, 2017, 2018 and 2019. The original data sets, one more recent and another before that, are composed of millions of rows and around one hundred columns.

| **Client ID** | **Claim ID** | **Date** | **Diagnosis** | **…** | **Procedure** | **Provider** | **Provider Value** |
|---|---|---|---|---|---|---|---|
| 1 | 12345 | 01/01/2015 | Back Pain | … | Lumbar Back Surgery | Hospital X | 3 000 € |
| 2 | 67890 | 31/12/2019 | Sciatic Pain | … | Magnetic Resonance | Clinic Y | 30 € |

Table 3. Data Set 2 Example

Also, an analysis file containing information about updates on the main Procedures IDs regarding surgery procedures was used, as it is possible to see in the example in Table 4. The original data set is composed of 3 600 rows and 3 columns. It was created by the person who did the first clinical episode.

| Old Id Procedure | New Id Procedure | Procedure |
|---|---|---|
| 1 | 12 | Lumbar Back Surgery |
| 2 | 34 | Sciatic Surgery |

Table 4. Data Set 3 Example

Moreover, finally, a Médis data set containing data regarding the number of units and the name of the doctors that performed the procedures, as it is possible to see in the example of Table 5. The original data set is composed of millions of rows and near one hundred columns.

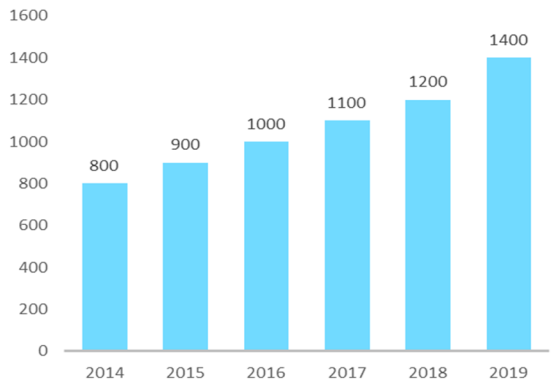
| Claim ID | Date | Procedure | … | Provider | Doctor | No. of Units |
|---|---|---|---|---|---|---|
| 12345 | 01/01/2015 | Lumbar Back Surgery | … | Hospital X | A B C D | 3 |
| 67890 | 31/12/2019 | Magnetic Resonance | … | Clinic Y | E F G H | 0 |

Table 5. Data Set 4 Example

After defining and exploring the data sets that were going to be used, an analysis of relevant data for the construction of the clinical pathways was done in Power BI, before the data preparation, regarding providers, procedures, the volume of claims, among other things.

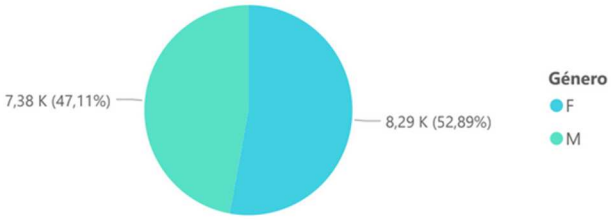
When analysing data, it was possible to observe 6 725 people with lumbar pain procedures from 2014 to 2019. Here the year 2014 was being analysed to see any significant differences from the period chosen to be analysed. From Figure 9, it is possible to notice an increase in claims, related to lumbar pain surgery procedures, where y axis is the number of claims, and x axis is the year. An insurance claim is an application to an insurance company requesting a payment where the insurance company evaluates and then pays or not based on the contract agreements (General Insurance Council, 2021).

Figure 9. Claims volume of lumbar pain surgery claims by year



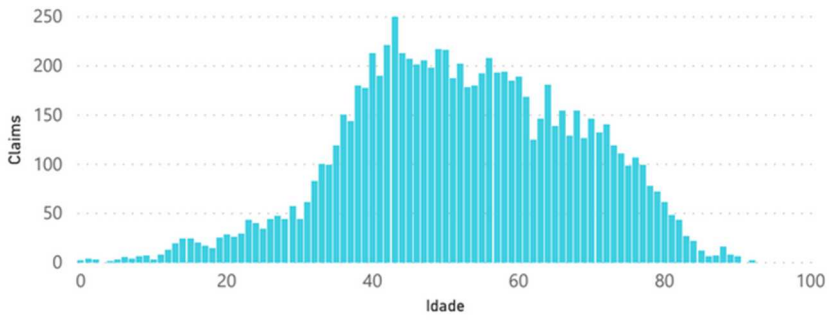

From Figure 10, it is acceptable to say that there is not much difference in gender distribution, although the feminine gender reveals a few more claims.

Figure 10. Claims distribution by gender



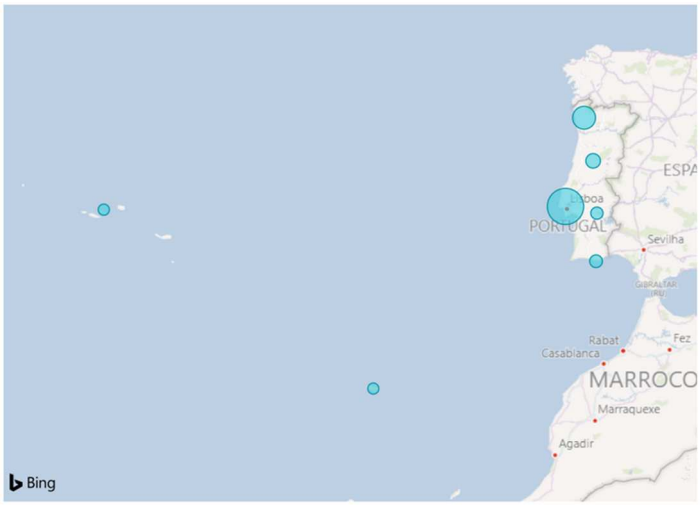

A conclusion extracted from observing Figure 11 is the focus of claims from the late thirties through the late seventies.

Figure 11. Claims distribution by age



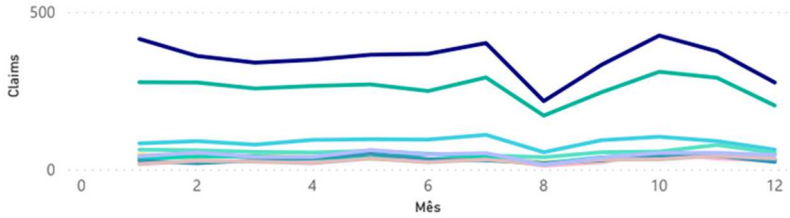

Through Figure 12, it is possible to tell there is a higher volume of claims in the Regional Health Area (RHA) of Lisbon, followed by the RHA of Porto.

Figure 12. Claims distribution by Regional Health Area




Finally, one aspect obtained from the analysis present in Figure 13, where each colour represents one procedure, was that looking at the top 5 procedures claims through the year, there is a decrease of claims in the Summer. This happens because in the Summer months, people go less to the doctors.

Figure 13. Top 5 procedure claim distribution by month




After ending the analysis of the data, there are a few characteristics that need to be reviewed. First, there are a few clients with more than one gender and age-associated. Besides this, claims were analysed separately but also aggregated. Most of the time, a combination of procedures is done on the same day or complementary, so they need to be aggregated when preparing the data. Second, there is missing data, but on variables we will not need or cannot get more information to input the missing information.

### 4.1.3. Data Preparation

The data sets mentioned in the Data Understanding phase were chosen to prepare the data for the final goal. The Data Preparation was done using SAS Enterprise Guide and consisted of 8 process flows, which are "pages" on one SAS E-Guide project.

The first process flow consisted of the creation of a unique ID for each client. Based on the Data Set 1, four variables were used to help build this variable: a Client Identifier, which is a unique number that identifies a client, but it is not always filled, which is the reason for the need to build this variable, because there are a lot of missing values; the name of the client, which also brought some issues given the fact it is not mandatory for the client to fill in the full name, the birth date, which as it is so common, is not always filled correctly, and the Policy ID, which indicated the insurance ID of a client, and that changes if the client cancels their insurance, even if they come back. To create the unique ID, a few tasks were done: 1 - if the person had the same name and same birth date, despite having the Client Identifier variable filled or not, it would be given the same ID; 2 - if the person had the same name and two of the birth date fields were equal (day, month, year), despite having the variable Client Identifier filled, the same ID would be given, because sometimes people write one of the fields in an incorrect way; 3 – for the rest of the people an ID starting from 0 was allocated. However, sometimes the name is different because people fill the full name one time, another time they abbreviate it, so it is expected that in the future, a technique can be used to detect similar names and, with the help of the birth date, detect as happened in task 1 and 2 if they are the same client or not. This process flow made it possible to go through more than 2 million IDs to 1,7 million unique IDs.

The second process flow is the creation of the base table where the chosen clients for the project are present, and there is a cleansing of the data according to some insurance information regarding these surgeries. As previously mentioned, only a particular type of clients were selected for the project. The company's portfolio used in this project must have had at least 1 of 44 procedures present in their historical data. It should be clients with one of these procedures in the last five years, without counting with the year 2020 given the pandemic caused by Covid-19, which means the years being analysed were 2015, 2016, 2017, 2018 and 2019. The original data sets (Data Set 2 Example), one more recent

and another before that, are composed of millions of rows and about one hundred columns, respectively. Furthermore, after choosing the clients based on their unique ID on the most recent dataset, it was possible to extract historical data from the older dataset and merged it to the most recent. This was the process flow where Data Set 3 was used to correct the primary procedures IDs.

The third process flow was also to clean the data according to some insurance information regarding these surgeries and to filter the procedures and diagnoses related to lumbar sciatic pain, as indicated by the clinical consulting team. Furthermore, an important variable was created, the date of the first primary surgery, which will be used in the following process flows. This variable was useful to rectify cases where more than one gender and age were assigned on the day of the first primary surgery. Additionally, the variable diagnosis was corrected because some rows had general diagnosis written, and in Data Set 4, there was more detailed information.

In the fourth process flow, a few rows were removed from the data set due to a few clients with the same surgery in a short amount of time, 30 days, defined based on a business rule.

The fifth process flow was the one where the day and the inpatient days of the first surgery table was built. The data about the day where the first lumbar sciatic pain surgeries were performed was identified through the variable created in the previous process flow. Then a deep data preparation was done because it was crucial to have the real initial and final date of this episode, or in another words, the date when the person entered for the surgery and the date when the client was discharged. For this, it was necessary to resort to Data Set 4 due to the units, if not missing, and sum it to the initial data variable from Data Set 2, resulting in the final official date. When missing, the already final date that the dataset contained was accepted. Once the initial and final dates were defined, all these clients' data between the initial and final data were extracted to create the final data set. Through Data Set 4, information regarding the doctor was also joined to the leading data set. Once this table was completed, through a state point meeting, it was decided that there were primary surgeries and others that were interventions for pain treatment without surgery intervention, which resulted in all the analysis being segmented for surgeries or pain treatment. Given this, a variable was created: the procedure of the day, which was an aggregation of the surgeries that happened on the same day in the first intervention. As there were many combinations, it was decided via medical advice that it was better to have a top 9 + "Others". Besides this, another variable was created: the days that the person had the day episode.

The sixth process flow was built for the creation of the pre-surgery table. With the final data set from the fourth process flow, the data kept was all before the first surgery date. Based on medical consulting, the main procedures before the surgeries were defined: C, X, M, T, F, and the rest aggregated into "Others", where each letter represents a procedure, which cannot be shared due to professional secrecy and data protection. Based on these procedures, they were also aggregated. It resulted in so many aggregations that based on the percentage of representation of the top 9 aggregated procedures of the day, the top pre-surgery aggregated procedures were selected, and the remaining ones aggregated in "Other Combinations". Besides this, the doctor and unit data were merged to the existing one, and another variable was created: the days that the person had until the day of the first surgery.

The seventh process flow was conceived to create the post-surgery table. With the final data set from process flow 3, the data kept was all after the final date of the first surgery. Based on medical

consulting with Dr Rodriguez, the main procedures that occur before the surgeries were defined: C, X, M, T and R, which stands for Relapse or Reoperation, being the same surgery or not, or in the same bone or not. The rest was aggregated into "Others". Based on these procedures, they were also aggregated. It resulted in so many aggregations that based on the percentage of representation of the top 9 aggregated procedures of the day, the top pre-surgery aggregated procedures were selected, and the remaining ones aggregated in "Others". Besides this, data regarding the doctor and units was merged to the existing one. Other variables were created: the days the person had since the day of the first surgery; if the relapse is the same as the first, different, or both.

The eighth process flow was an additional process flow to build the table to model the clusters in Jupyter Notebook. There is no distinction between the segments, surgery intervention, or pain treatment to see if the model will do it. Based on the tables from process flows 3, 4 and 5, before distinguishing between segments, the variables were selected based on the importance of personal and medical data. Besides this, the providers' data was also analysed to understand who had done the entire path with Médis and who switched before or/and after the intervention, surgical or not. The variables present in the final table of this process flow are personal, such as age, gender, and RHA, and historical, such as the procedures of the pre-, day, and post-surgery, the costs of the provider, days that the person stays at the pre-, day, or post-moment, the number of units if found necessary for the cluster analysis, the number of main pre- and post-procedures and some data related with reoperations.

The ninth process flow was the last process flow to be done because it created an extra table to provide information for the final presentation when it comes to "personas". Although clusters were created, it is also essential to analyse the typical clients who go through surgery or pain treatment, distinctively. The variables chosen were the names of the day's procedures, before and after main procedures, age, days in each stage (pre-, day, and post-), and the mean provider cost of each stage.

### 4.1.4. Modeling

As modelling techniques, Association Rules and Clusters were built, to see if they would help provide better data insights regarding the types of clients who suffer from this pathology and their pathways.

#### 4.1.4.1. Association Rules

The data used for Association Rules modelling was simply the ID, the day's procedure, and the pre- and post-surgery procedures before dividing into surgery intervention or pain intervention.

Two approaches were performed: one with the three tables together, where the algorithm is applied at the same time; another with two stages, where the first involved the before and day tables and the other the day and after tables, to see which approach would provide better results. In both approaches, the *Apriori* algorithm was used, as previously mentioned, with minimum support of 0,05, a minimum threshold of 0,25 for the confidence metric and a minimum threshold of 1 for the lift metric.

In Approach 1, the top itemsets above the minimum support were the ones in Table 6.

| Support | Itemsets |
|---|---|
| 0,329 | After - C |
| 0,328 | Day - E+FO |
| 0,234 | Before - C |
| 0,195 | Before - C+M |
| 0,171 | Day - Other Combinations |
| 0,164 | Day - Treatment of Pain |
| 0,16 | Before - Others |

Table 6. Itemsets Support - Approach 1

After, the Association Rules were generated, first by confidence, as it is possible to observe in Table 7, and then by lift, as in Table 8.

| Antecedents | Consequents | Antecedent Support | Consequent Support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| Before - C+M | Day - E+Fo | 0,195 | 0,329 | 0,085 | 0,435 | 1,327 |
| Day - E+Fo | After - C | 0,328 | 0,328 | 0,143 | 0,435 | 1,321 |
| After - C | Day - E+Fo | 0,329 | 0,328 | 0,143 | 0,434 | 1,321 |
| Before - C | After - C | 0,234 | 0,329 | 0,094 | 0,403 | 1,224 |
| Before - C+M | After - C | 0,195 | 0,329 | 0,077 | 0,4 | 1,203 |
| Before - C | Day - E+Fo | 0,234 | 0,328 | 0,078 | 0,333 | 1,014 |
| Day - Treatment of Pain | After - C | 0,164 | 0,329 | 0,052 | 0,316 | 0,961 |
| After - C | Before - C | 0,329 | 0,234 | 0,094 | 0,287 | 1,224 |
| Day - E+FO | Before - C+M | 0,328 | 0,195 | 0,085 | 0,259 | 1,327 |

Table 7. Association Rules by Confidence - Approach 1

| Antecedents | Consequents | Antecedent Support | Consequent Support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| Day - E+Fo | Before - C+M | 0,328 | 0,195 | 0,085 | 0,259 | 1,327 |
| Before - C+M | Day - E+Fo | 0,195 | 0,328 | 0,085 | 0,435 | 1,327 |
| Day - E+FO | After - C | 0,328 | 0,329 | 0,143 | 0,435 | 1,321 |
| After - C | Day - E+Fo | 0,329 | 0,328 | 0,143 | 0,34 | 1,321 |
| After - C | Before - C | 0,329 | 0,234 | 0,094 | 0,287 | 1,224 |
| Before - C | After - C | 0,234 | 0,329 | 0,094 | 0,403 | 1,224 |
| After - C | Before - C+M | 0,329 | 0,195 | 0,077 | 0,235 | 1,203 |
| Before - C+M | After - C | 0,195 | 0,329 | 0,077 | 0,4 | 1,202 |
| Day - E+Fo | Before - C | 0,328 | 0,234 | 0,078 | 0,238 | 1,014 |
| Before - C | Day - E+Fo | 0,234 | 0,328 | 0,078 | 0,333 | 1,014 |

Table 8. Association Rules by Lift - Approach 1

In Conclusion, by analysing the most frequent itemsets, the results were:

1. Day - E+Fo + Before - C;
2. Before - C + After - C;
3. Before - C+M + Day - E+Fo;
4. Day - E+Fo + After - C;
5. Before - C+M + After - C;
6. Day - Treatment of Pain + After - C.

In Approach 2, Stage 1 (Before + Day) the top itemsets above the minimum support were the ones in Table 9.

| Support | Itemsets |
|---|---|
| 0,329 | Day - E+Fo |
| 0,234 | Before - C |
| 0,195 | Before - C+M |
| 0,171 | Day - Other Combinations |
| 0,164 | Day - Treatment of Pain |
| 0,16 | Before - Others |

Table 9. Itemsets Support - Approach 2 – Stage 1

After, the Association Rules were generated, first by confidence, as it is possible to observe in Table 10, and then by lift, as in Table 11.

| Antecedents | Consequents | Antecedent Support | Consequent Support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| Before - C+M | Day - E+Fo | 0,195 | 0,328 | 0,085 | 0,435 | 1,327 |
| Before - C | Day - E+Fo | 0,234 | 0,328 | 0,078 | 0,333 | 1,014 |
| Day - E+Fo | Before - C+M | 0,328 | 0,195 | 0,085 | 0,259 | 1,326 |

Table 10. Association Rules by Confidence - Approach 2 - Stage 1

| Antecedents | Consequents | Antecedent Support | Consequent Support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| Before - C+M | Day - E+Fo | 0,195 | 0,328 | 0,085 | 0,435 | 1,327 |
| Day - E+Fo | Before - C+M | 0,328 | 0,195 | 0,085 | 0,259 | 1,327 |
| Day - E+Fo | Before - C | 0,328 | 0,234 | 0,078 | 0,238 | 1,014 |
| Before - C | Day - E+Fo | 0,234 | 0,328 | 0,078 | 0,333 | 1,014 |

Table 11. Association Rules by Lift - Approach 2 - Stage 1

In Conclusion, by analysing the most frequent itemsets, the results were:

1. Day - E+Fo + Before - C;

2. Before - C+M + Day - E+Fo.

In Approach 2, Stage 2 (Day + After) the top itemsets above the minimum support were the ones on Table 12.

| Support | Itemsets |
|---|---|
| 0,329 | After - C |
| 0,328 | Day - E+Fo |
| 0,171 | Day - Other Combinations |
| 0,164 | Day - Treatment of Pain |

Table 12. Itemsets Support - Approach 2 - Stage 2

After, the Association Rules were generated, first by confidence, as it is possible to observe in Table 13, and then by lift, as in Table 14.

| Antecedents | Consequents | Antecedent Support | Consequent Support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| Day - E+Fo | After - C | 0,328 | 0,329 | 0,143 | 0,435 | 1,321 |
| After - C | Day - E+Fo | 0,329 | 0,328 | 0,143 | 0,434 | 1,321 |
| Day - Treatment of Pain | After - C | 0,164 | 0,39 | 0,052 | 0,316 | 0,961 |

Table 13. Association Rules by Confidence - Approach 2 - Stage 2

| Antecedents | Consequents | Antecedent Support | Consequent Support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| Day - E+Fo | Day - E+Fo | 0,328 | 0,329 | 0,143 | 0,435 | 1,321 |
| After - C | Day - E+Fo | 0,329 | 0,328 | 0,143 | 0,434 | 1,321 |

Table 14. Association Rules by Lift - Approach 2 - Stage 2

In Conclusion, by analysing the most frequent itemsets, the results were:

1. Day - E+Fo + After - C;

2. Day - Treatment of Pain + After - C.

After comparing the two approaches, it is safe to say the results are very similar, being the most frequent items: Before - C, Before - C+M, Day - E+Fo, Day - Treatment of Pain, and After - C.

Since the Day Combination E+Fo is the most frequent procedure, it was not easy to get other combinations of procedures. However, these results, aligned with the Power BI and the following Cluster analysis, provide sound and well sustained final results.

#### 4.1.4.1. Clusters

The data used for Cluster modelling was the one that resulted from the eighth process flow. When importing the data into Jupyter Notebook, another Data Understanding and Data Preparation phases took place to see what else could be done. After the missing values were filled, the variables related to the units were removed. A few variables were created: one related with the past or future historical data was created to understand if the client began or/and continued with the company; and others with bins, such as the age, provider costs, days passed on each stage (pre-, day or post-), and the number of main before and after procedures. Additionally, the HRAs were aggregated into three groups, that will be detailed after the creation of clusters.

After this, the variables were encoded using the One Hot Encoder method because it lets the illustration of categorical data be more expressive. Next, they were normalised using the Min-Max Scaler preprocessing method because it conserves the shape of the initial distribution. After this, the dimension reduction was applied through the dataset, through Principal Components Analysis (PCA), using 41 components, as it is possible to see from Figure 14.

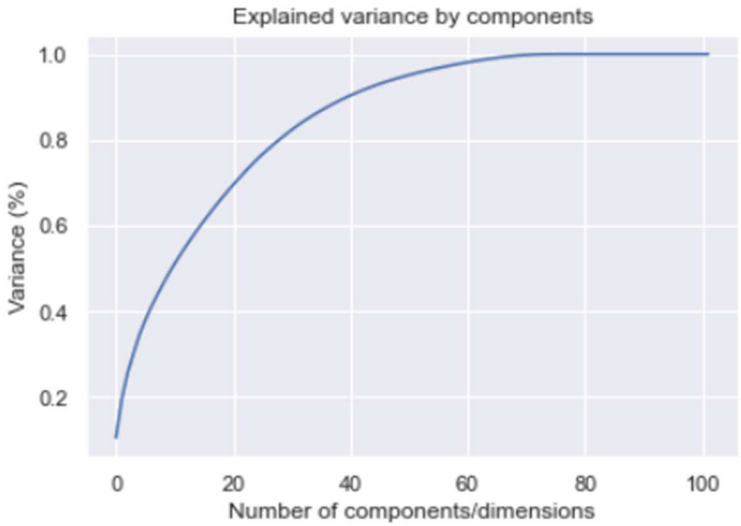



Figure 14. Cumulative Explained Variance

Afterwards, the number of clusters (k) was selected through the analysis of the elbow graph and the silhouette plot, as it is possible to observe from Figure 15 and Figure 16.

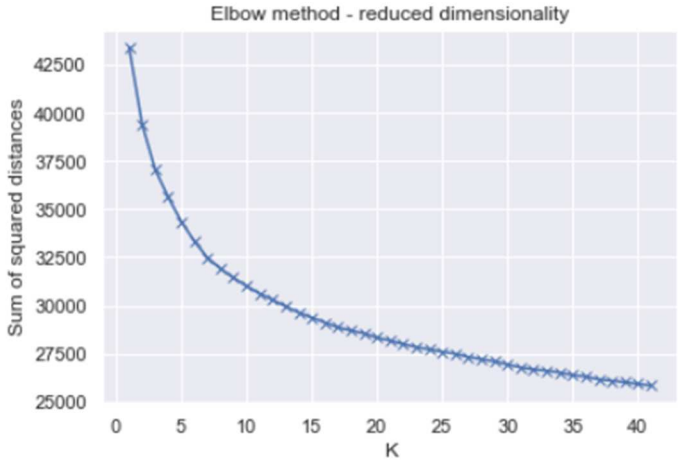
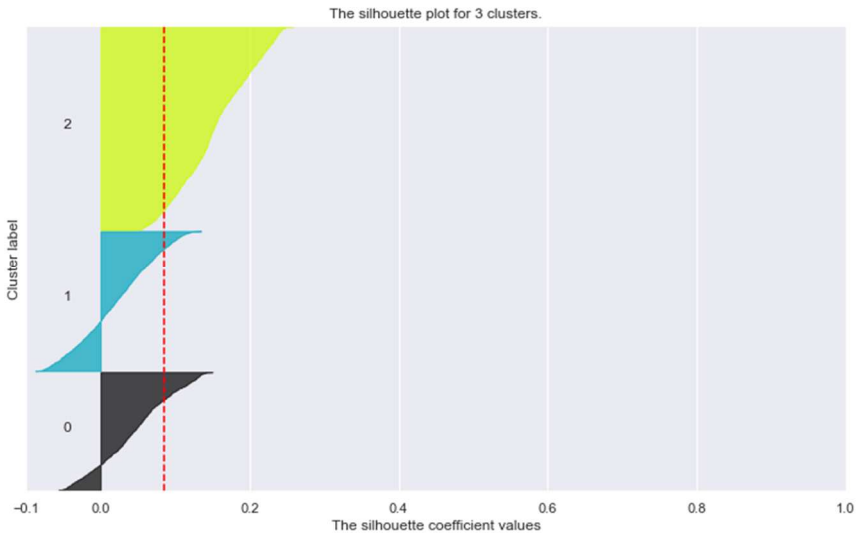
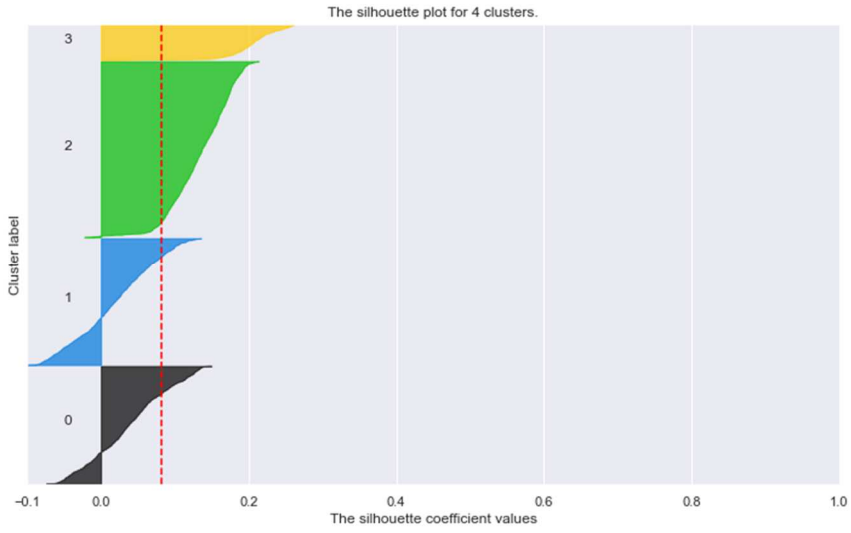



Figure 15. Elbow Method



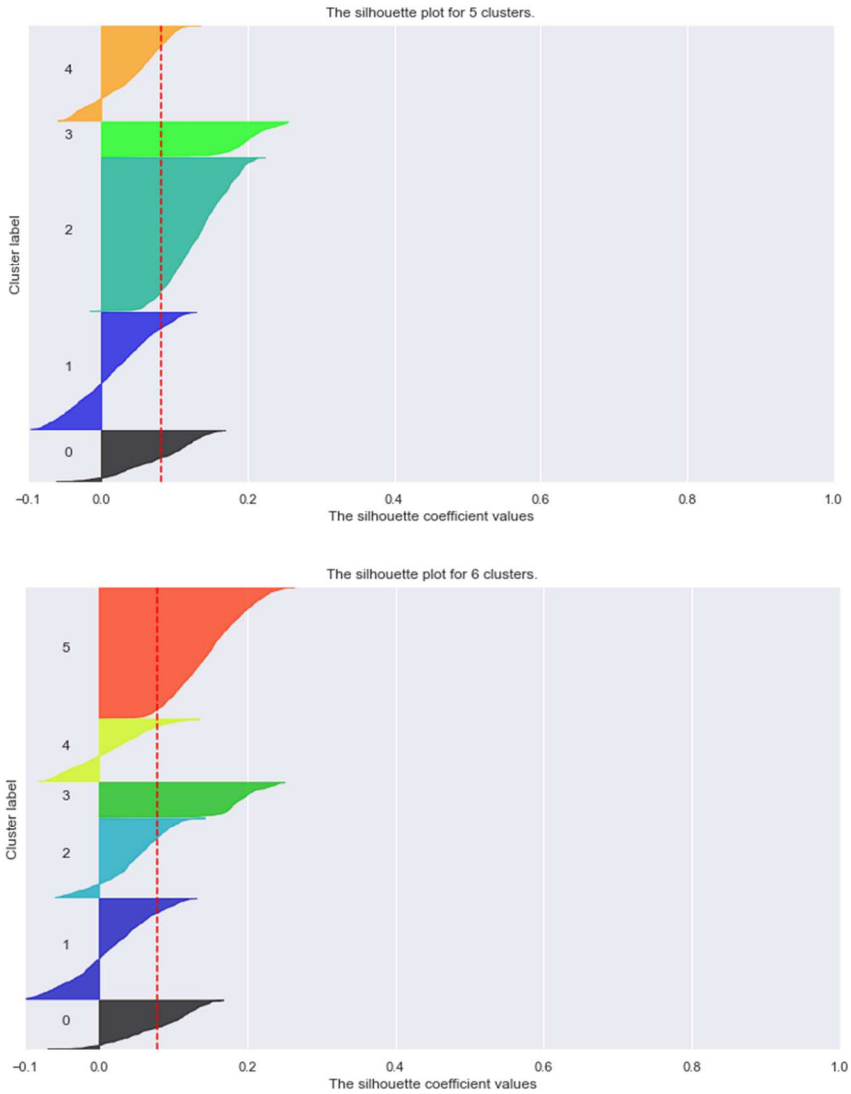



Figure 16. Silhouette Method for K=3, K=4, K=5 and K=6

From Figure 16 it may not be clear, but for k=3, the average silhouette score is 0,086; for k=4 it is 0,082; for k=5 it is 0,0814; and for k=6 it is 0,783.

Given all this, the number of clusters chosen was 3, which resulted in the following distribution: Cluster 0 with 1280 clients, Cluster 1 with 1082 clients, and Cluster 2 with 1858 clients, resulting in the two principal components plot present in Figure 17.

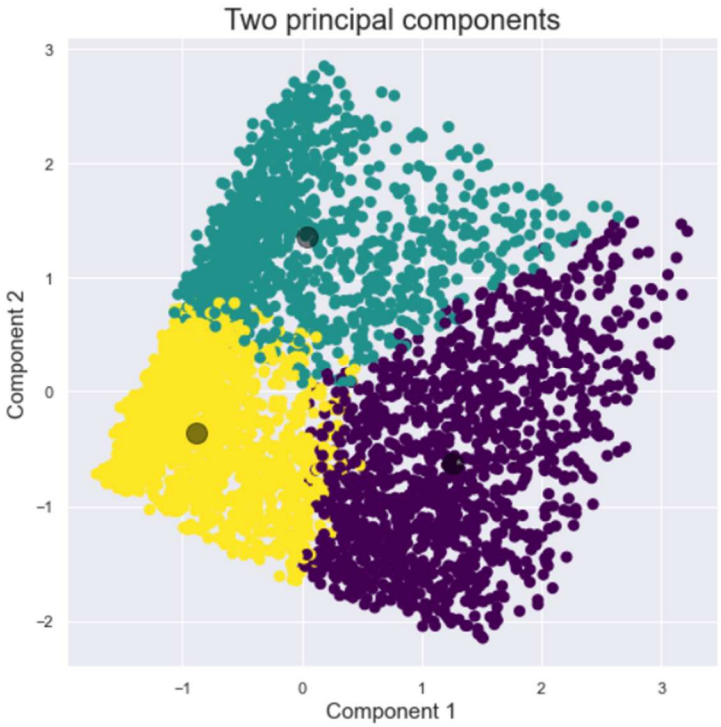



Figure 17. Two Principal Components Plot

The results were beneficial; however, as the data cannot be public, some codes will be used, for example, genders AA and AB, age interval BA, BB and BC, and Regional Health Area CA, CB and CC, and not all characteristics will be shared.

- Cluster 0: mostly gender AA, from the Regional Health Areas CA and CC, focusing on age interval BA. One particular characteristic of these clients is that they have the same number of main procedure C, before and after surgery.
- Cluster 1: mainly gender AB, from the Regional Health Area CA, emphasising BB age interval. Two interesting aspects about these clients are that they have a more significant number of procedures C after than before surgery, and the day's value is around Y €.
- Cluster 2: predominantly gender AB, from the RHAs CA and CB, inserted in the age interval CC. This cluster differs from the other two mostly because it spends a considerable amount of money after the procedure, including surgeries and pain treatment, having an enormus number of main procedures realised before the first intervention, mostly procedures F.

These results help the company segment their clients when concerned with this pathology. However, when combined with the Association Rules and the Power BI analysis results, the final results provide Médis with the so wanted clinical pathways, provider and cost information needed for the improvement of the client experience and steering.

The project resulted in the analysis of 4 388 clients, about 84% with surgery as the first intervention and approximately 16% with a procedure to treat the pain, with a representation in the gender of 49% and 51%, male and female, respectively. There is a focus on surgeries as people get old, especially between ages 36 and 65.

The top 5 pathways for each segment were identified through a Sankey plot in Power Bi, as it is possible to see in Figure 18 for surgeries, and Figure 19, pain treatment. The thickness of the lines depends on the number of clients that do the pathway.

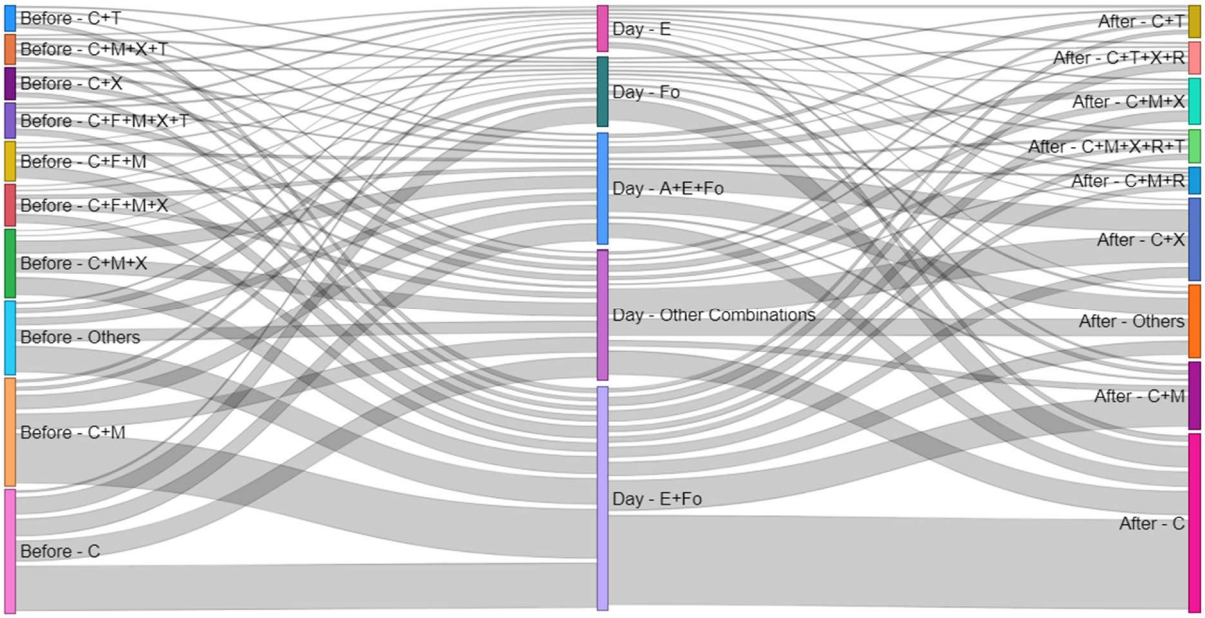



Figure 18. Surgery Segment Top 5 Pathways

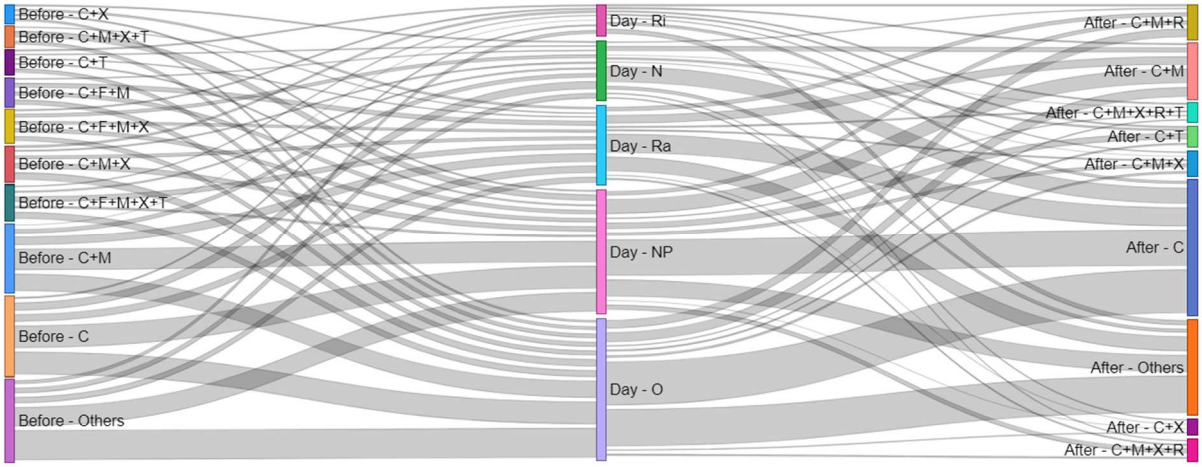



Figure 19. Treatment of Pain Top 5 Pathways

Thus, it is possible to say that the most common surgical pathway is composed by the pre-surgery procedures C or C+M, the main procedure E+Fo, and the post-surgery procedure C. As for the most

frequent pain treatment pathway, it is composed by the pre-procedures C, or Others, or C+M, followed by the main pain treatment procedures O or NP, ending with the post-procedures C or Others.

Besides this, the top 3 procedures for each segment of the pre-day were analysed when it comes to its volume, mean provider cost and mean days on this stage, representing 60% of the clients in each segment. C was the main procedure for the surgery segment, with the lowest average cost but with the most significant amount of time in this segment. However, nothing compared with what is happening on the segment of pain treatment, since Others was the primary procedure, and it was the most expensive one, taking three times more time of procedure C in the surgery segment. The top providers for both segments were providers 1, 2 and 3, representing approximately 86 % for the surgery segment and 90% for the pain treatment. The average number of main procedures was also calculated, although not being shared, as it is only possible to say that the results are similar when comparing both segments.

Concluding the pre-intervention analyses next is the analysis of the day of the first intervention. The top 3 procedures and diagnoses of the day were observed. For the surgery segment, the top 3 procedures, E+Fo, Other Combinations, and A+R+F, represent around 75% of the clients. The average provider cost of the E+Fo procedure is almost three times less than the Other Combinations and two times less than the A+R+F procedure. The inpatient days increase as the top 3 decrease. The top 3 surgery diagnoses represent almost 38% of the diagnosis. As for the treatment of pain segment, the top 3 procedures represent 77% of the clients, there are no inpatient days, and the average provider costs are very similar. When it comes to the top 3 diagnoses, they represent about 56%.

Finally, the analysis of the post-intervention was a little denser as there might be some reoperations. The top 3 procedures were C, C+X and Others in the surgery segment, which represents around 66% of the clients. The average cost and days spent in the after stage are very similar, except for Others which is three times more costly and almost two times longer. As for the pain treatment segment, the top 3 procedures are C, Others and C+M, almost 69% of the clients. As for the costs and time, it happens the same as in the surgery segment. When it comes to the providers, the top 3 for the surgery segment is the same as in the before stage. The top 3 for the pain treatment stage is the same except for one of the providers, B, being replaced by provider D. When analysing the reoperations, about 14,5% of the surgery segment client's relapse or perform a new operation, almost half of the clients in the pain treatment segment.

As for the personas, for the surgery segment, the typical client is male, has over 50 years old, the main procedure of the day is E+Fo, the before stage is C+M, and the after stage is C. The client usually stays in the hospital three days, and z days in the before stage, two times more than the after stage. The mean value of the three stages is very similar, although the after stage is a bit more expensive. Furthermore, when it comes to the pain treatment segment, the typical client is female or AB, has one more year than the typical surgery client. The main procedure of the day is O, anticipated by C or Others, and followed by C. As for the days spent in each stage, as previously referred, the pain treatment patient does not have any inpatient day, but the before and after average days is very close to the ones in the surgery segment. Additionally, the final provider cost value is very similar to the surgery segment. However, there are more differences between the stages, being the after stage the most expensive, the before stage half of it, and the day stage also half of the before stage.

To end the analysis, it was also discovered that k% of the company's clients with a specialty in neurosurgery and orthopedics performed a lumbar sciatic intervention, surgical or not. More than g% do the entire course with the company. In the case of the surgeries, about h% of the clients goes to the same provider when needing a reoperation, x% more clients than in the pain treatment segment.

### 4.1.5. Evaluation

After nearly nine months of deep data analysis and models, the final results were delightful and valuable for the following steps: clinical pathway validation, episode systematisation and use cases exploration, such as client steering. This because it was possible for the stakeholders to understand which medical providers were more famous, which were the lumbar sciatic pain pathways, the main procedures and its monetary costs, and provide information about typical personas that usually suffer from this pathology. With this, it is possible to beware of the needed procedures when patients get to the company's medical providers with specific diagnosis ou historical medical data, and also to guide them to the best medical providers when it comes to the needed surgery of pain treatment. In addition, this may reduce costs and time, because it is known which procedures need to be done to the patient.

It also decided that the analyses performed in this pathology that was not performed on the other pathologies should be added. These analyses covered essential aspects of the Médis information, which means the project's purpose was achieved and overreached.

### 4.1.6. Deployment

The deployment phase of this project is the responsibility of the Network and Clinic Department, as they are torn to improve the customer experience and steering, considering the possibility of creating clinical packages.

## 4.2. Additional Projects

The internship was extended to 12 months, so two additional projects were done after presenting the main project during the remaining three months.

### 4.2.1. Data Preparation Automatization

The creation of lumbar sciatic pain pathways took approximately eight months, which may sound like it takes too much time to develop a pathology pathway. Since the data preparation on SAS took most of the time, it was found necessary to automate the data preparation process. This was done using python on Jupyter Notebook, according to the process flows of SAS Enterprise Guide. Only a few adjustments need to be performed based on the next pathologies, such as the main procedures to be selected in SAS E-Guide first, because there is still no connection to Médis libraries and the years of

analysis. In Jupyter Notebook, adjustments are related to filtering diagnosis or selecting the main procedures of the pre- and post-surgery.

### 4.2.2. Predictive Model

The purpose of building a predictive model was to predict which clients usually end up needing a surgery or pain treatment based on the historical data, the after stage.

Given this need, a table was created on SAS. The base table used was one like in Dataset Example 2. With the prior knowledge from the lumbar sciatic pain pathways project, only clients with historical data in the same years as the main project were selected, which means, data from year 2015 until year 2019, and the following filters were performed, to select only people that had, in fact, lumbar sciatic pain history:

- Having realised procedures that are inserted in the following aggregators: Physical Medicine and Rehabilitation, Neurosurgery, Orthopaedics, C, F, Rheumatology, X, M, and T.
- Had history with the most common diagnosis detected in the lumbar sciatic pain pathways project, although here only the ones related with the lumbar sciatic region were selected, not including the broadest ones.

The first thing to be done was a slice of the data to have train and test datasets, bearing in mind that the training dataset would also be divided into train and validation datasets in the future. This split was done in the following way: 80% train dataset and 20% test dataset.

After filtering the database, a new variable was created, the target, containing the information if the client went through an intervention, surgical or not. This was done with the help of the main project, since the dataset only contained clients that had had an intervention. Additionally, as the dataset was composed with a lot more clients that did not had an intervention, a SAS code was applied to have the same number of clients with target 0 and target 1. After this creation, data prior to the date of the first surgery was removed.

In the Data Preparation phase, the variable age was corrected, since through the historical data, people age was changing, and it did not make sense to have multiple ages associated with the clients. The median age of each client was selected as the final age.

Besides this, the following variables were created: the number of years the person had historical data with the company; the number of days the person got the insurance until it left, or its due date; the total amount of units and the total costs, the main pre- and post- procedures, and how many times they occur, the number of procedures realised inside and outside the provider network, and the number of providers the person has had realised procedures on. In addition, two other variables were created, based on the previous knowledge from the pathways project: a variable that says if the person has had main procedures and main diagnoses with a certain word on it, in this case, frequent words found on the main project, and how many times they occur.

Moreover, the detection and removal of outliers was executed. Feature Engineering was performed as well, to bin the numeric variables with lots of possible values, based on the minimum, mean, median and maximum values.

After this, Feature Selection was performed using the Random Forest Classifier because it delivers higher accuracy through cross validation and handles missing values while preserving the accuracy of a large proportion of data (Anurag, 2018). The data was split into 70 % train and 30 % test, which resulted in 49 optimal number of features.

After selecting those 49 features, named by the Random Forest Classifier, the Modelling phase started. The dataset was divided into train and validation datasets, with a representation of 70 % and 30 %, respectively. A great number of classifiers were fitted to see which one would perform better. The results were the ones present in Table 15.

| Classifier | Train | Validation |
|---|---|---|
| LR | 0,888 | 0,892 |
| LSVC | 0,886 | 0,877 |
| Voting | 0,917 | 0,874 |
| GradientBoosting | 0,924 | 0,856 |
| AdaBoost | 0,881 | 0,853 |
| MLPClassifier | 0,973 | 0,85 |
| ExtraTrees | 0,993 | 0,847 |
| RandomForest | 0,994 | 0,844 |
| XGB | 0,994 | 0,844 |
| SDG | 0,881 | 0,841 |
| SVC | 0,889 | 0,841 |
| NSVC | 0,87 | 0,835 |
| Bagging | 0,982 | 0,823 |
| DecisionTree | 0,994 | 0,802 |
| GaussianNB | 0,813 | 0,79 |
| BNB | 0,751 | 0,745 |
| KNN | 0,994 | 0,7 |

Table 15. Accuracy performance of each classifier on the train and validation sets

Although it seems there are excellent results, a part of them resulted in overfitting, which resulted in the selection of the Multi-layer Perceptron (MLP) Classifier, which is a classifier that connects to a

Neural Network (Nair, 2019), and may seem like one of the cases. However, it did, indeed, result in great results.

After selecting the most appropriate classifier, the train dataset was normalised using the Min Max Scaler, although Standard Scaler and Robust Scaler were also validated, but did not provide better results. Finally, the classifier was fitted to the data with some adjusted parameters, with the help of GridSearchCV, such as *activation, hidden_layer_sizes, learning_rate, learning_rate_init,* and *solver*, which resulted in the confusion matrix present in Table 16, and in the following scores:

| | **0** | **1** |
|---|---|---|
| **0** | TN = 109 | FN = 43 |
| **1** | FP = 15 | TP = 166 |

Table 16. Train/Validation Confusion Matrix

- A Precision score of 0,917, which is used to count the true positives out of all positive predictions made. As high as possible, the better;
- Recall score of 0,794, which is used to measure the true positives of all actual positives. As high as possible, the better;
- The accuracy score of 0,826, which measure the model in terms of true positives and true negatives out of all the predictions performed. As high as possible, the better;
- F1 score of 0,851, which is a mean of precision and recall scores, to not compromise results. As high as possible, the better.

When testing the model with the test data, it resulted in the confusion matrix present in Table 16, and the following scores:

| | **0** | **1** |
|---|---|---|
| **0** | TN = 94 | FN = 13 |
| **1** | FP = 33 | TP = 151 |

Table 17. Train/Test Confusion Matrix

- Precision Score = 0,821;
- Recall Score = 0,921;
- Accuracy Score = 0,842;
- F1 Score = 0,868.

As it is possible to observe, the results were outstanding, which means the predictive model is appropriate to predict who needs surgery or not. However, it is now vital to update the train and test datasets over the years because patterns may change, and some tasks may need to be modified.

# 5. CONCLUSIONS

The main goal of this project was to build lumbar sciatic pain pathways to understand the procedures the company's clients usually do before and after a surgery or pain treatment and the most frequent surgeries of pain treatments. This was also an approach to estimate the costs and analyse the most common providers to deliver a better experience to the company's clients.

The project was helpful because it answered these questions and exceeded the analyses expectations, as it was possible to withdraw helpful information from the data. Besides, during the internship period, as it was extended, two additional projects were done, automatisation of the leading project data preparation phase and a predictive model to predict the clients needing a lumbar sciatic pain intervention, surgical or not.

The conclusions that can be extracted from the main project, lumbar sciatic pain pathways, is that the principal pathways are, for the surgery segment, the pre-surgery procedures C or C+M, the main procedure E+Fo, and the post-surgery procedure C, and for the pain treatment segment the pre-procedures C, or Others, or C+M, followed by the main pain treatment procedures O or NP, ending with the post-procedures C or Others. In addition, there are a few more conclusions related to the provider and providers' costs, the number of main pre- and post-procedures, the number of clients that do the whole path with the company, and the personal characteristics of these clients. Unfortunately, these results cannot be shared due to professional secrecy.

The only conclusion obtained from the data preparation phase automatisation additional project is that this will help other company employees build the clinical pathways for other pathologies faster.

The conclusion achieved with the predictive model was that it is possible to predict around 83% of the cases who need and do not need surgery.

## 5.1. CONNECTION TO THE MASTER PROGRAM

The master program contributed to the fulfilment of the internship and its projects since it was necessary to know data science and analytic tools, such as prior understanding of the python programming language, algorithms for the prediction model, SQL code for the SAS Enterprise Guide tool, and Power BI to analyse the data to its fullest.

## 5.2. INTERNSHIP EVALUATION

When evaluating the internship, the results were achieved, and they exceeded the expectations of the governance of the main project. Besides, the internship was extended, even after finishing the proposed project, which means good results and performance were delivered. This contributed to the development of two more projects, which also provide time efficiency for future pathologies and the capability to know if clients will need surgeries or pain treatments in the future.

## 5.3. LIMITATIONS

There are a few limitations that were found during the execution of the project. The fact that the company datasets were not clean having errors due to data being entered as the mood fits; not having a unique ID for the clients, which made the analysis a bit difficult; and that when there is only one project to do for such a long time, it is difficult to become creative and think outside the box, as I would like to have learned more or be present in other projects so that when returning to the main project, new ideas or other ways of looking at the data could arise. However, I feel like so many more analyses could have been performed, aligned with the Department of Marketing and the Department of Health and Wellbeing.

## 5.4. LESSONS LEARNED

Regarding the lessons learned in the internship, I would say there is always something new to learn and that at first, it may seem complicated, but it only takes a few weeks to be entirely inside the area of analysis. For example, I learned how to work with SAS Enterprise Guide from the ground, and now I can say I am an active user. Also, the health area and the specific pathology were unfamiliar, and I had to learn everything about it. However, after two or three months, I was able to say the names of the procedures front to back and almost help my family with their back pain diagnosis.

I would recommend this internship to any Advanced Analytics master program student. It puts into practice lots of things learned during the master's degree and is an excellent way to enter the market, accompanied by an incredible team.

## 5.5. FUTURE WORK

As for the future, I believe the company will benefit from the delivered projects and the automatization of the data preparation phase, as it allows to analyze more pathologies in a shorter amount of time.

As for me, I leave my internship with a happy face and a feeling of accomplishment, as I was able to achieve one more goal, and that opened the doors for my future, as one colleague invited me to be a part of the consulting company where she belongs. I believe this will be an essential part of my career in the data science industry, as I will try different things and conclude which I like the most.

Finally, I wish never to stop learning and never stop trying to be better every day.

## 6. BIBLIOGRAPHY


Allegri, M., Montella, S., Salici, F., Valente, A., Marchesini, M., Compagnone, C., . . . Fanelli, G. (2016). Mechanism of low back pain: a guide for diagnosis and therapy. *F1000Research*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4926733/pdf/f1000research-5-10546.pdf

Ansari, S. Z. (2019). *Market Basket Analysis - Trend Analysis of Association rules in different time periods.* NOVA Information Management School. Retrieved from https://run.unl.pt/bitstream/10362/80955/1/TEGI0458.pdf

Anurag. (2018, August 17). *Random Forest Analysis in ML and when to use it*. Retrieved from newgenapps: https://www.newgenapps.com/blogs/random-forest-analysis-in-ml-and-when-to-use-it-2/

Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF). (2020). *Estatísticas de Seguros.* Retrieved from https://www.asf.com.pt/ISP/Estatisticas/seguros/estatisticas_anuais/historico/ES2020/EstatSeguros2020.pdf

Balch, A. (2021, May 28). *Clinical Pathways: What are they and how do they impact patient care?* Retrieved from METAVIVOR: https://www.metavivor.org/blog/clinical-pathways-what-are-they-and-how-do-they-impact-patient-care/

bePortugal. (2021, June 11). *Health Insurance in Portugal: What You Need to Know*. Retrieved from bePortugal: https://beportugal.com/health-insurance-in-portugal/

Business News Daily Editor. (2020, february 24). *What Is Agile Scrum Methodology?* Retrieved from Business News Daily: https://www.businessnewsdaily.com/4987-what-is-agile-scrum-methodology.html

Busse, R., Klazinga, N., Panteli, D., & Quentin, W. (2019). *Improving healthcare quality in Europe.* Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK549276/pdf/Bookshelf_NBK549276.pdf

Casiano, V., Dydyk, A., & Varacallo, M. (2021, July 18). *Back Pain*. Retrieved from STATPEARLS: https://www.statpearls.com/ArticleLibrary/viewarticle/18089

crmpartners. (2021, June 18). *How vallue steering contributes to a better customer relationship*. Retrieved from crmpartners: https://crmpartners.com/en/insights/how-value-steering-contributes-to-a-better-customer-relationship/

Dabbura, I. (2018, September 17). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Retrieved from towards data science: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Data Science Process Alliance. (2021). *What is CRISP-DM?* Retrieved from Data Science Process Alliance: https://www.datascience-pm.com/crisp-dm-2/

Davis, D., Maini, K., & Vasudevan, A. (2021, September 2). *Sciatica*. Retrieved from STATPEARLS: https://www.statpearls.com/ArticleLibrary/viewarticle/28772

Dobilas, S. (2021, July 11). *Apriori Algorithm for Association Rule Learning - How To Find Clear Links Between Transactions*. Retrieved from towards data science: https://towardsdatascience.com/apriori-algorithm-for-association-rule-learning-how-to-find-clear-links-between-transactions-bf7ebc22cf0a

European Institute of Innovation and Technology. (2021, May 28). *Value-based Healthcare*. Retrieved from EIT Health: https://eithealth.eu/think-tank-topic/value-based-healthcare/

Explorium Data Science Team. (2020, january 3). *Clustering - When You Sould Use it and Avoid it*. Retrieved from Explorium: https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/

General Insurance Council. (2021, August 20). *Insurance Claims*. Retrieved from General Insurance Council: https://www.gicouncil.in/insurance-education/insurance-claims/

Goh, D. H., & Ang, R. P. (2007). An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents. *Behavior Research Methods*, 259-266. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03193156.pdf

Guyer, R. D., & Lewis, D. (2021, March 30). *Lower Back Pain Causes, Symptoms, Diagnosis and Treatment*. Retrieved from spineuniverse: https://www.spineuniverse.com/conditions/low-back-pain

Hipp, R., Abel, E., & Weber, R. J. (2016). A Primer on Clinical Pathways. *Director's Forum*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896352/pdf/i0018-5787-51-5-416.pdf

IBM. (2021). *CRISP-DM Help Overview*. Retrieved from IBM: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

ICT Solutions for Brilliant Minds. (2021, June 25). *Customer steering*. Retrieved from ICT Solutions for Brilliant Minds: https://www.csc.fi/en/web/guest/customer-steering-model

Iseminger, D. (2021, september 23). *O que é o Power BI Desktop?* Retrieved from Microsoft: https://docs.microsoft.com/pt-pt/power-bi/fundamentals/desktop-what-is-desktop

Jupyter. (2021, July 16). *Jupyter*. Retrieved from Jupyter: https://jupyter.org/

Lutkevich, B. (2020, September). *association rules*. Retrieved from SearchBusinessAnalytics: https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

Médis. (2020). *Demonstrações Financeiras.* Retrieved from https://www.grupoageas.pt/media/4080/demonstracoes-financeiras_medis_2020.pdf

Médis. (2020). *Relatório de Gestão.* Grupo Ageas Portugal. Retrieved from https://www.grupoageas.pt/media/4081/af_rg-medis-2020-high.pdf

Médis. (2020). *Relatório sobre a Solvência e a Situação Financeira (SFCR).* Retrieved from https://www.grupoageas.pt/media/3899/sfcr_2020_medis.pdf

Médis. (2021, June 5). *Sobre Nós*. Retrieved from Médis: https://www.medis.pt/sobre-nos/#quem-somos

Mohanty, A. (2019, May 15). *Multi layer Perceptron (MLP) Models on Real World Banking Data*. Retrieved from Medium: https://becominghuman.ai/multi-layer-perceptron-mlp-models-on-real-world-banking-data-f6dd3d7e998f

Nair, A. (2019, June 20). *A Beginner's Guide To Scikit - Learn's MLPClassifier*. Retrieved from Analytics India Magazine: https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/

Nair, A. (2019, June 20). *A Beginner's Guide To Scikit - Learn's MLPClassifier*. Retrieved from Analytics India Magazine: https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/

NEJM Catalyst. (2017). What Is Value-Based Healthcare? *NEJM Catalyst*. Retrieved from https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0558

Omran, M., Engelbrecht, A., & Salman, A. A. (2007, November). *An overview of clustering methods.* Retrieved from ResearchGate: https://www.researchgate.net/publication/220571682_An_overview_of_clustering_methods

Rocha, C. (2021, February 5). *Seguros de saúde continuam a crescer. Prémios aumentaram 8,2% em 2020*. Retrieved from Diário de Notícias: https://www.dn.pt/dinheiro/seguros-de-saude-continuam-a-crescer-premios-aumentaram-82-em-2020-13317893.html

SAS. (2021, July 16). *Delivering the Power of SAS Analytics and reporting from an easy-to-use, point-and-click Windows interface.* Retrieved from SAS: https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-guide-101431.pdf

Schnuerer, T. (2019, July 23). *Lumbar Spine Surgery: Will You Need Surgery for Your Lower Back Pain?* Retrieved from spineuniverse: https://www.spineuniverse.com/conditions/back-pain/low-back-pain/lumbar-spine-surgery-will-you-need-surgery-your-lower-back-pain

scikit learn. (2021, July 30). *1.17. Neural network models (supervised)*. Retrieved from scikit learn: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Silva, J., Varela, N., López, L. A., & Millán, R. H. (2019). Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm. *Procedia Computer Science, 151*, 1207-1212. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050919306404

Silva, S. N. (2009). *Os Seguros de Saúde Privados no contexto do sistema de saúde português.* Associação Portuguesa de Seguradores. Retrieved from https://pns.dgs.pt/files/2010/08/aps1-ss.pdf

Stitch. (2021). *7 reasons to use Microsoft Power BI*. Retrieved from Stictch - A Talend Product: https://www.stitchdata.com/resources/7-reasons-power-bi/

Teisberg, E., Wallace, S., & O'Hara, S. (2020, May). Defining and Implementating Value-Based Health Care: A Strategic Framework. *Academic Medicine, 95*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7185050/pdf/acm-95-682.pdf

visionedge. (2021, July 30). *Six Steps for Building Predictive Models*. Retrieved from visionedge: https://visionedgemarketing.com/six-steps-for-building-predictive-models/

Wright, N. (2019). *Everything you ever wanted to know about Microsoft Power BI*. Retrieved from Nigel Frank International: https://www.nigelfrank.com/blog/everything-you-ever-wanted-to-know-about-microsoft-power-bi/

Zhan, F. (2019). Summary of Association Rules. *IOP Conference Series: Earth and Environmental Science*. Retrieved from https://iopscience.iop.org/article/10.1088/1755-1315/252/3/032219/pdf