NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação
Master Program in Statistics and Information Management

PREDICTING LENGTH OF STAY (LOS) IN A HOSPITAL POST-SUGERY

Sara Nunes (m20201111)

Thesis Proposal for the master's degree in Information Management

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

PREDICTING LENGTH OF STAY (LOS) IN A HOSPITAL POST-SUGERY

by

Sara Nunes

Thesis Proposal presented as partial requirement for obtaining the master's degree in Information Management/ Master's degree in Statistics and Information Management, with a specialization in Business Intelligence

Advisor: Bruno Damásio

ABSTRACT

The amount of time a patient stays in the hospital after a surgery has been an issue that hospital management faces, a longer stay in the recovery room involves a high cost to the hospital and consumes a lot of hospital resources, manpower and equipment. The amount of time is difficult to predict precisely since there are many external and internal factors that account for a longer or shorter stay and it is difficult for a team to consider all these factors and make this estimation manually. With the advancement of machine learning methods and models this prediction can be made automatically. The aim of this study was to **c**reate a predicting model that look at the patient data and the procedure data and predicts the amount of time the patient will stay after the surgery to make the current prediction of the length of stay by the hospital more accurate and compliment the current surgery scheduling and discharge system. To achieve the objective, a data mining approach was implemented. Python Language was used, with particular emphasis on Scikit-Learn, pandas and Seaborn packages. Tables from a relational database were processed and extracted to build a dataset. Exploratory data analysis was performed, and several model configurations were tested. The main differences that separate the models are outlier treatment, sampling techniques, feature scalers, feature engineering and type of algorithm – Linear Regression, Decision Trees Regressor, Multilayer Perceptron Regressor, Random Forest Regressor, Light Gradient Boosting Machine Regressor and Gradient Boosting Regressor. A total of 32993 hospital episodes were observed on this study. Out of these, 2006 were eliminated due to some data anomalies, namely, values that were wrong or impossible. The data was split in training and test data. Several model configurations were tested. The main differences that separate the models are outlier treatment, feature scalers, feature engineering and the type of algorithm. The best performing model had a score of 0.73 R2 which was obtained by using the Light Gradient Boosting Machine Regressor Algorithm using outlier removal, Robust Scaling and using all the features in the dataset.

Keywords: Length of stay; Post-surgery; Recovery Room; Data Mining; Machine Learning; Portugal

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

BMI	Body Mass Index
Cm	Centimetre
DRG	Diagnosis-related group
DT	Decision Trees
ED	Emergency Department
GB	Gradient Boosting
GDPR	General Data Protection Regulation
HAI	Healthcare-Associated Infections
ICU	Intensive Care Unit
Kg	Kilogram
LGBM	Light Gradient Boosting Machine
LOS	Length of Stay
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
NDA	Non-disclosure agreement
NHS	National Health System
OR	Operating Room
R2	Coefficient of Determination
RF	Random Forest

RFE Recursive Feature Elimination

ROC Receiver Operating Characteristic

RR Recovery Room

WHO World Health Organization

XGBoost Extreme Gradient Boosting

1 INTRODUCTION

1.1 BACKGROUND AND PROBLEM IDENTIFICATION

Hospitals face huge pressure constantly. Clinical professionals need to make decisions fast and act accordingly while working under pressure. These decisions determine the patient's outcome in a hospital. The medical staff follows their experience and years of experience to diagnose and treat the patients, however, sometimes human errors happen, or some information is under looked and the patients can be misinformed or not accurately informed, this leads to patients being frustrated and unnecessary costs to the hospital.

The hospital system holds multiple service units, depending on the size of the hospital. Some medical departments interact with each other and some of these are the Operating theatre, Intensive care unit and Emergency department. The operating theatre includes the Operating Rooms (OR) and Recovery Rooms (RR). This unit is one of the costliest ones in a hospital, since they require a large capital and it is very labour intensive (Negash et al., 2022). The length of stay (LOS) is the duration of the patient's visit to the hospital, and it is measured in number of days. This paper will focus on the LOS in the RR which is the time the patient spends in the RR after going through surgery in the OR.

According to the World Health Organization report, (WHO, 2003) among the performance indicators in hospital, the average length of stay (LOS) is considered an important indicator of the hospital's performance one of the most important monitoring factors and it is often used to measure the hospital efficiency (Average Length of Stay in Hospitals | Health at a Glance 2019 : OECD Indicators | OECD ILibrary, 2019 A longer say in the hospital can reflect problems for the patients since it can lead to risk of catching multiple infections, sleep deprivation or even mental and physical deconditioning, besides these risks a too short or long stay can reduce the quality of care provided.

A patient stay in the recovery room is one of the most-resource consuming departments in a hospital. The high costs are mostly due to the number of staff being needed to run this unit and to monitor the patients, it also involves a lot of expensive monitoring equipment as well as the medicine and food that needs to be provided to the patients.

Predicting the LOS is beneficial to the hospital in multiple aspects. It can be used to inform and better prepare the patient, the doctor, the family, the hospital management, and the insurance companies. The doctors can make medical decisions and give the patient an accurate medical plan considering the length of stay so the patient can manage his budget, time, and speed of recovery. Besides, the patient can inform the family on the expected LOS to help them organize themselves. The hospital can improve the care provided by managing their resources utilization more efficiently and having better bed management.

In this context, the creation of a model able to identify and advice on an early stage the amount of time the patient will be spending at the recovery room is designed. This tool will consider the patients that have had a surgery in the hospital being studied. It starts with collecting the data from the hospital episode visit (procedure, specialty, discharge date, admission date, etc) and patient features (age, gender, BMI, etc). This data is then pre-processed and cleaned, missing data will be treated, outliers

will be handled, and the data will be normalized. This pre-processed data is then trained with different Machine Learning techniques with the collected data as input and the LOS value as the output. The dataset will be divided into test set and training set. The ML techniques were regression algorithms, namely LR, DT Regressor, MLP Regressor, RF Regressor, XGBoost, LGBM Regressor and GB Regressor. Lastly, the models will be evaluated with R2 regression score function, MAE and MSE.

The subject of hospital LOS predictions is highly researched and there are many studies about predicting the length of stay using hospital data, however there are no studies done that only look at the length of stay of patients that have been into surgery and that groups the length of stay by different surgical specialties. This is especially useful since these recoveries need to be supervised very closely due to the implications they can have in terms on mobility and even mortality. This will create a more accurate model since it will give a prediction based on the type of surgery and the patient's data instead of a general prediction.

1.2 Study Relevance and Importance

The project described in this paper is relevant for the organization where the model will be used which will be a chain of private hospitals in Lisbon and furthermore to the healthcare industry and economy as well, especially in Portugal.

From an organization perspective, the project is relevant since it will optimize the occupancy of the recovery rooms and the surgery scheduling system with the use of historical and concrete data. This will result in a higher quality of care provided and customer satisfaction.

From a healthcare industry perspective, the project will provide benefits since it can be used as a general model for other organizations and improve the current Portuguese healthcare. Besides the hospitals, it can benefit insurance companies by giving a prediction of what the costs and time after surgery usually are so they can manage their budgets.

1.3 STUDY OBJECTIVES

The goal of this project is to implement a predictive model in the hospital industry to improve the current surgery scheduling system that is performed by the hospital management. This model will be later used by the hospital management to compliment this process; this model will not replace the human side completely since there is information that the model does not take into consideration, but the surgical staff does. It will be an auxiliary tool. The creation of this model will involve the traditional five steps of a data mining process that will later be described in the Methodology section.

Taking this into consideration, the main objectives of this project were defined to answer these questions:

- Choose the variables that will be in the model by looking at the admissions and patients' data available and analysing it
- Understand what are the factors that lead to a longer stay
- Conclude on what the best algorithms are to predict the data
- Develop a model capable of predicting the length of stay of different patients once they are admitted to the hospital

2 LITERATURE REVIEW

2.1 LENGTH OF STAY AND COSTS TO THE HOSPITAL

To cut expenses, hospitals must improve the healthcare planning and structure.
Improving and reducing LOS improves financial, operational, and clinical outcomes by decreasing the costs of care for a patient. The clinical treatment methods demonstrate a substantial positive correlation between costs and LOS for both clinical and economic reasons. (Huang et al., 2013) Therefore, to keep healthcare costs down, hospitals must try to optimize and minimize the length of stay of patients. (Freeman et al., 2016)

2.2 LENGTH OF STAY AS A KEY PERFORMANCE INDICATOR

The indicator that is often used to measure the efficiency in a hospital is the average length of stay. (Nouaouri et al., 2015) According to the WHO report, the LOS is regarded as an indicator of the hospital's performance and is one of the most monitored factors in a hospital (Shaw, 2003). The LOS has been proposed as a useful outcome measure that might be used to target quality improvement efforts. (Englert et al., 2001; Guru et al., 2005).

Hospitals use the Average LOS for measuring the success of the hospital on the cost control, cost saving, service efficiency and complementary care delivery systems. (Stone et al., 2022) It is essential for hospital planning as it directly determines the number of beds to be provided. In addition, LOS is a common point of comparison between patients and hospitals, making it a suitable KPI for hospital management regardless of the healthcare setting (academic, public, or private). Therefore, it is crucial to understand the factors that influence LOS. (Kulinskaya et al., 2005)

2.3 LENGTH OF STAY AND PATIENT SATISFACTION

Patient satisfaction is a crucial indicator of healthcare quality since it provides information on the hospital's performance in satisfying clients' expectations and needs.(Xesfingi & Vozikis, 2016)

Patient satisfaction with the received medical care is essential for maintaining a positive and useful physician-patient relationship as well as patient adherence to prescribed therapies. Patient satisfaction was poorer in patients with prolonged LOS.(Parker & Marco, 2014)

2.4 LENGTH OF STAY AND PATIENT OUTCOMES

An inappropriate length of stay in the recovery room can have negative outcomes in the patient’s recovery and well-being.

Studies show that the additional LOS is associated with a HAI. In average, the risk of developing an infection is increased every day that a patient stays longer in the hospital. 37 According to a study, 25.2% of patients experienced 1 or more medical complications during hospitalization. The most common complications were urinary tract infection (15.4%), pneumonia (9.0%), and constipation (6.8%). All medical complications were associated with longer LOS (Ingeman et al., 2011). LOS is increased by surgical complications and can be used to implement discharge planning in general surgical patients. (Procter et al., 2010) It was discovered that increasing the LOS by one day increases the chance of contracting an infection by 1.37 percent and increases the average LOS by 9.32 days. 38

According to estimates, HAI increases the expense of a hospital stay by lengthening the LOS. (Hassan et al., 2010) Besides having a significant effect on risk of infection, studies results showed also a high significant association between increasing length-of-stay and mortality, at the patient and hospital levels. Patients in the upper quartile of LOS were more likely to die (odds ratio = 1.45, 95% CI) than those in the lower quartile. Long LOS was more common in hospitals with a high standardized mortality (r = 0.79, p 0.01). (Lingsma et al., 2018) Patients who stay in the ED for longer periods of time have a higher risk of morbidity and mortality than those who stay for shorter periods of time. (Englert et al., 2001) Glasgow Coma Scale, Abbreviated Injury Scores, and specific mechanisms of injury were significant predictors of the rates of death and discharge, with effects that were variable in different time intervals (Clark & Ryan, 2002). Moreover, the time after discharging a prolonged LOS patient (33 days of hospital stay) is critical as 55% of patients died within six months of being discharged (Teno et al., 2000).

2.5 LOS AND READMISSION

Hospitals with mean risk-adjusted LOS that was lower than expected had a higher readmission rate, suggesting a modest trade-off between hospital LOS and readmission (Kaboli et al., 2012)

Increasing the length of stay for some patients could help to improve treatment quality and consequently lower the readmissions during the 30-day post-discharge period.(Carey & Lin, 2014) Cases that received a short LOS were associated with a higher readmission rate within 28 days of hospitals discharge. Patients may be readmitted for further treatment if the LOS is too short. Patients readmitted with problems were discharged 41% faster than the average length of stay for their diagnosis.(Han et al., 2022)

2.6 DISCHARGE PLANNING

Discharge planning is the creation of a unique plan for each patient hospitalized in the hospital with the goal of lowering expenses and improving patient outcomes (Shepperd et al., 2013) by minimizing hospital length of stay and unexpected readmissions, and by ensuring that patients leave the hospital at the appropriate time in their treatment. (Gonçalves-Bradley et al., 2016)

Patients admitted to hospital with a clinical diagnosis and assigned to discharge planning had a statistically significant reduction in hospital LOS and readmissions. (Shepperd et al., 2013) This intervention also resulted in fewer hospital readmissions, increased time to readmission, and lower costs. (Popejoy et al., 2009)

A study on the effect of discharge planning on length of stay relation to length of stay indicates positive finding for discharge planning as an intervention, showing that discharge planning reduces hospital lengths of stay by −0.71 days (95% CI -1.05, -0.37). (Hunt-O'Connor et al., 2021)

Besides this, a personalized discharge planning can also increase the satisfaction of both patients and healthcare professionals 12, 13, 14 which is an important indicator on the hospital performance as of quality of care (Xesfingi & Vozikis, 2016). A successful discharge planning had positive outcomes in patient satisfaction and quality of life. The patient, family, nurse, doctor, hospital, and community services all benefit from the smooth and effective coordination of this process. (Carroll & Dowling, 2007)

2.7 RISKS OF FACTORS THAT INFLUENCE LENGTH OF STAY

To create a successful discharge plan, it is important to look at multiple impactful factors that correlate to the LOS.

Understanding the factors that influence patient LOS can help clinicians improve patient satisfaction and quality of care by allowing them to optimize care, rationalize their medical practice, assist administrators with budget planning and resource allocation, and potentially improve patient satisfaction and quality of care. (Lee et al., 2003)

It is important to identify the factors that contribute to the length of stay at different levels so that the hospital can make the discharge plan for the patient. The LOS is affected by factors about the hospital visit, by factors that are specific to a patient and factors that are external to these.

2.7.1 Hospital Level Factors

In the hospital, LOS was longer on days with a higher percentage of daily admissions, more elopements, longer periods of ambulance diversion, and on weekdays, but shorter on weekends and days with a higher number of discharges. (Wiler et al., 2012)

Patients' hospital LOS is reduced because of increased bed pressure caused by increasing demand for surgical care. LOS is also reduced for those patients that were admitted via a 24h emergency department, receiving surgery on the same day of admission. (Castelli et al., 2015)

Los varies geographically as well since LOS varies across different hospitals. Routine data showed that there were variations in LOS between countries, regions, and hospitals.(Clarke & Rosen, 2001)

The type of specialty had a big influence on LOS. Specialty consultation was similarly linked to longer LOS and the effect varied greatly depending on the service sought. (Yoon et al., 2003) Moreover, within the same Diagnostic group, length of stay also differed when episodes were treated by different specialty of doctor. (Liu et al., 2001). Patients within the same Diagnosis-related group differed in length of stay when they were admitted from different referral sources to the hospital. (Liu et al., 2001)

The most significant influence on LOS was discharge destination according to Kulinskaya. Length of stay is at least 25% longer for patients transferred from other hospitals and not admitted as emergencies, and LOS for patients discharged to private facilities is more than twice that of patients discharged to NHS facilities or to their own homes. (Kulinskaya et al., 2005). Patients that have been transferred between hospitals or readmitted within 28 days had significantly longer LOS. (Castelli et al., 2015)

Furthermore, patients discharged to a nursing home (14.2 days LOS), or a rehabilitation institution (11.5 days LOS) had a greater length of stay than those discharged to any other facility (9.6 days LOS). (Brasel et al., 2007). LOS was lower for those discharged to their own homes. (Castelli et al., 2015).

2.7.2 Patient Level Factors

In multivariate analysis of patient data, factors significantly associated with extended LOS and higher costs included age, sex, race/ethnicity, insurance status and Revised Trauma Score (Brasel et al., 2007), at danger or undernutrition patients, and BMI 32 as well as patients that were coming from more deprived areas (Castelli et al., 2015) and different payment classifications (Liu et al., 2001)

Although the impact of many clinical characteristics on LOS are intuitive and data-backed, studies of medical and voluntary surgery patients demonstrate the significance of non - clinical factors. Insurance or payer type are one of these variables. (Brasel et al., 2007; Kagan et al., 2002; KHALIQ et al., 2003)

Insurance's positive impact on hospital admission and length of stay varies by income quintile, area, and type of health facility. (Yoon et al., 2003) Patients with Medicaid had a considerably longer mean LOS (11.3 days) than patients with commercial insurance, uninsured patients (each 9.3 days), and Medicare patients (8.8 days). (Brasel et al., 2007)

A study found that patients with clinically significant depression (N.=296; median: 5 days, interquartile range: 3-8 days) had a longer LOS than patients without a clinical depression (N.=2328; median: 4 days, interquartile range: 2-6 days). (Kerper et al., 2014)

Patients with specific co-diseases have higher LOS.(Castelli et al., 2015) Patients with cardiovascular illnesses, numerous diseases, nervous system disorders, and cerebrovascular diseases had a significantly longer length of stay, according to a study. Furthermore, we discovered that as urea, creatinine, and salt levels rise, so does the length of stay. (Toptas et al., 2018)

2.8 DATA MINING IN LOS PREDICTIONS

Data mining has been used to predict LOS in multiple studies. To predict the LOS of patients, several works propose statistical approaches or Artificial Neuronal Networks (ANN) as well as ways to deal with outliers and missing data.

(Ng et al., 2006) constructed an ANN to predict the length of stay in the ICU. (Wrenn et al., 2005) developed and validated an ANN model to predict LOS for an ED. (Hachesu et al., 2013) use the techniques of machine learning modelling with three algorithms (Decision tree (DT), Support Vector Machines (SVM), and ANN) to predict LOS with the models performing quite well with various high degrees of accuracy. (Azari et al., 2012) proposed an approach for predicting LOS using a multi-tiered data mining approach. They utilized clustering to create training sets to test different algorithms. (Marie & Davis, 2010) used four Machine Learning algorithms (logistic regression, neural network, decision tree, and ensemble model) to analyse the patient discharge data for average LOS based on input variables.

It was established that identifying LOS outliers, which are data points that have been eliminated from most data, can lead to a better understanding of hospital expenses, and assist hospital management in controlling those expenditures. (Lingsma et al., 2018) The current tendency to exclude such outlier stays in data reporting due to assumed rare occurrence may need to be revisited or else the typical statistical values (for example, means and deviations) will be unreliable. (Hughes et al., 2021)

3 METHODOLOGY

This study was conducted using the Python coding language supported by Jupyter Notebook technology. Python is the world's third most popular programming language and is described as an interpreted, high-level, and general-purpose programming language. Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations, and narrative text (*Project Jupyter | Home*, 2022)

Python was chosen to achieve the goal on this project since it contains a lot of packages that data visualization and the modelling more efficient and easier since several Data Science packages are available to complement this language and provide functions for machine learning without having to code them ourselves. Below are the key packages that were used in this project:

- Pandas - Pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series. One of those structures is called Data frame, which is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns like a spreadsheet or SQL table. (*Pandas - Python Data Analysis Library*, 2022)
- Scikit-Learn - Scikit-Learn is a software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forest, gradient boosting, k-means and DBSCAN. (*Scikit-Learn*, 2022)
- Seaborn - Seaborn is a Python data visualization library. It provides a high-level interface for drawing attractive and informative statistical graphics. (Waskom, 2021)

3.1 SEMMA OVERVIEW

The methodology chosen for this project was one of the most used Data Mining frameworks, called Sample, Explore, Modify, Model, and Assess (SEMMA), which was developed by the SAD institute. Figure 1 shows the overview of the workflow of this methodology and the tasks involved in each of these phases.

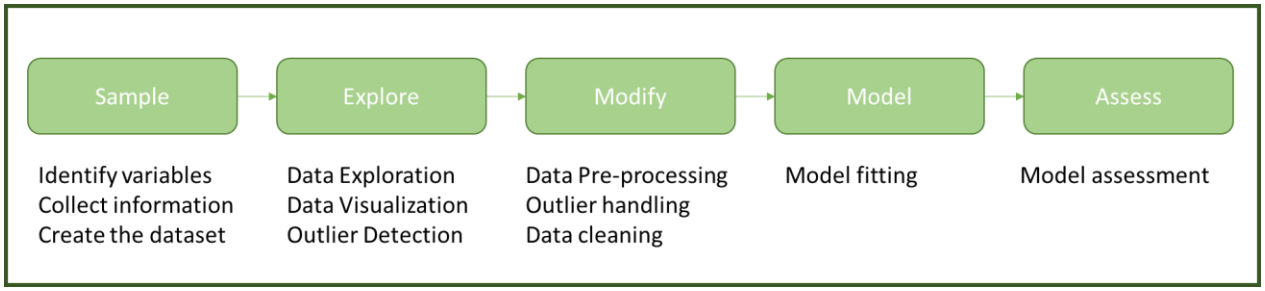

Figure 1: SEMMA Overview

In the following sections, there will be a more detailed explanation for each phase and all the tasks completed in each one.

3.2 SAMPLE

The study is based on sample data collected by a private hospital in Lisbon, which is one of the biggest hospitals in Lisbon. The hospital has eight surgical wards in total and comprises 24 different types of surgery. This health centre collects information of all hospitalizations, diagnostics, treatments, and some individual characteristics of the patients according to the national standards of Diagnostic Related Groups (DRG) records.

The data was firstly extracted from multiple databases and joined with an SQL Query. The data was then extracted into an excel file to allow for fast analysis. A SQL query was used to extract the data needed from the hospital's databases, which in this case is all the patients' data and the corresponding hospital visit data from the patients that have had a surgery in the past 3 years. Since the data is from multiple database tables and in different formats the data will be joined and transformed to group all the different sources into one.

The variables were chosen based on the information from the literature about the most important variables and matched with the data that was available from the patients and hospital visit so that we could include every significant available variable. Since this data is classified as sensitive and protected by GDPR, authorization was needed to access the information and the NDA Agreement was signed on 1st of July of 2022.

The time frame observed in the sample is between 2019 and 2022. The sample contains 32993 observations, and each row represents a unique visit to the hospital. Each observation in the sample is identified with two main ids: the patient id which is a unique identifier for the patient and the visit number, which is associated to each patient, and it's generated a new visit record every time the patient returns to the hospital.

The personal characteristics contained in the dataset are gender, weight, height, BMI and district and city of the patient; there is no information regarding the employment status, income, or the civil status (married, cohabitation, unmarried or divorced).

The visit number associated with each patient makes it possible to count the number of times a patient returned to the Hospital in the last 3 years. (Identified in the dataset as Number of Admissions). It also allows us to determine if the visit has been a readmission, an episode is considered a readmission when a patient who has been discharged from a hospital is admitted again within 30 days with the same diagnosis group (Identified in the dataset as Readmission).

The length of stay is measured in hours, and it is calculated starting from the days in which the patient is admitted to the hospital for the surgery procedure until the patient is given a discharge note from the doctor. This variable will be the dependent variable, the one that will be predicted.

The table below (Table 1: Dataset Description), represent all the columns in the dataset and their descriptions:

Column	Type of Data	Example Values	Description
Unit Area	Categorical	Hospital Lisboa	Hospital Unit
Visit Number	Categorical	AD1FB1BBF3DB8CCD8F0F9B0765D1E718	Identifier of the visit (Visit ID)
Patient Id	Categorical	1E06B868ABB03C6063DDD1A70D135F6C	Identifier of the patient (Patient ID)
Age	Numerical	99	Age of the client at the data of the admission
Gender	Categorical	F	Gender of the patient **Possible Values** F - Feminine M - Male I - Undetermined
Number of Admissions	Numerical	99	Number of visits to the hospital of that patients since 2019
Weight	Numerical	99	Weight of the patient (Kgs)
Height	Numerical	199	Height of the patient (Cms)
BMI	Numerical	99.99	BMI of the patient
District	Categorical	Lisboa	District where the patient lives
City	Categorical	Lisboa	City where the patient lives
Insurance Type	Number	1	Type of insurance **Possible Values:** 1-No Insurance 2-Insurance
Specialty of Admissions	Categorical	Cirurgia Geral	Specialty of the episode
Admission Cause	Categorical	Cirurgia Electiva	Reason for the visit
Urgency Type	Numerical	Avaliação Urgência Geral	Type of Urgency, it is filled only if the patient was admitted to the Emergency room **Possible Values:** 1-Not Urgent 2-Urgent

			3-Very Urgent 4-Critical
Admission Diagnosis	Categorical	366- CATARATA;	Diagnosis on the admission to the patient
Admission Data	Date	1/1/2020 0:00	Date when the patient visited the hospital
Discharge Date	Date	1/1/2020 0:00	Date when the patient was discharged
Admission Priority	Categorical	3 - URGENTE	If the patient has been before
Surgery Date	Date	1/1/2020 0:00	Date of the surgery
Procedure Code	Categorical	39.15.00.13	Code of the surgery
Procedure	Categorical	Cir.Arterial Directa - Desobst. BilateralS/Desobstruçã o Aórtica Via Inguinal (Aneis)	Description of the surgery
Visit Price	Numerical	1	Price paid by the patient (Ranges of values) **Possible Values:** 1- <= 500€ 2- >500 e <=1000 3- >1000 e <=10000 4- >10000 e <=50000 5- >50000
Exemption	Numerical	1	Payment Type (By who was the procedure paid) **Possible Values:** 1- Insurance 2- Patient 3- Shared
Voluntary	Numerical	1	Type of Surgery **Possible values:** 1- Emergent 2- Optional 3- Urgent
Discharge Destination	Categorical	Domicílio	Discharge Destination

Length of Stay	Numerical	99	Total duration of the hospital stay (in hours)
Procedure Duration	Numerical	99	Surgery duration (in minutes)
Readmission	Numerical	1	Indicates if the visit is a readmission **Possible values:** 0- Not Readmission 1- Readmission

Table 1: Dataset Description

3.3 EXPLORE

Now that the variables are defined and the dataset is built, this phase consists of visualizing the data and exploring. The dataset built has a total of 32993 records and 31 columns. To better understand the data, multiple plots and visualizations have been used for aid. First an overview of the variables was analysed, and they were separated on the fact if they were numerical or categorical, since different statistical functions are available for each of them. Lastly, a correlation matrix was plotted to see what variables are related and which ones are important to our dependent variable.

3.3.1 Numerical Columns

When exploring the data, a few anomalies were found in the data, as seen in Table 2. The columns Weight, Height and BMI which are all related seem to have wrong numbers, the Weight variable had a Maximum value of 995 which is physically impossible for someone to weigh that much, it had a minimum of value 0 is also an impossible value. The Height has a maximum value of 1800 which is impossible since no human being can be that tall and the minimum height of 0 is also wrong because it is impossible. Lastly, the BMI maximum of 10000 and minimum of 0 is an impossible value since the BMI usually ranges from 19 to 50. These errors are due to input mistakes, since these values were filled by the hospital staff when admitting the patients, it is possible that sometimes the values are filled as 0 when they are unknown and when the values are too large, for example “995” is it very likely a typing error, and the staff probably wanted to type “99.5”. From the minimum value we can that other variables contain wrong values, for example the Procedure Duration has a minimum of -1365.0, which is impossible since it’s a negative value, the Length of Stay has a minimum value of 0, which will not be considered since for the studies only patients who stay in the hospital after surgery are considered. These values will be removed since they are data anomalies. After removing these rows, the total number of rows were 30987.

Variable	Number of Rows	Average	Standard Deviation	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Weight	32774.0	68.67	25.41	0.0	60.00	70.0	80.00	995.0
Height	32786.0	160.29	30.18	0.0	160.00	165.0	173.00	1000.0
BMI	32049.0	29.98	185.94	0.0	22.49	25.1	28.37	10000.0
Procedure Duration	32189.0	110.38	87.48	-1365.0	55.00	87.0	143.00	1387.0
Length of Stay	32908	67.76	186.36	0.0	7.75	27.0	54.0	3977.0

Table 2: Data anomalies

Table 3: Table showing statistical values for the numerical variables shows a summary of the statistical values of the numerical columns after deleting the data anomalies, it is showing information on each numerical variable such as the number of rows, the average, the standard deviation, the minimum value, the lower quartile (25% of the data), median (50% of the data), upper quartile (75% of the data) and the maximum value. We can see from the number of rows, which is the total number of the records per variable, taking into consideration that the total number of records is 30987, it is possible to see that some of the columns had missing values which will be dealt with later, such as the Exemption and Voluntary. Lastly, the maximum values reveals that a few variables contain outliers, such as the Number of Admissions, Length of Stay, and Procedure Duration. All this will be analysed closer in the Data Transformation step.

Variables	Number of Rows	Average	Standard Deviation	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Age	30987.0	52.70	20.32	0.0	40.00	55.00	69.00	103.00
Number of Admissions	30987.0	47.74	61.82	1.0	12.00	26.00	58.00	763.00
Weight	30987.0	69.73	19.74	0.0	60.00	70.00	80.00	197.00
Height	30987.0	164.34	16.65	0.0	160.00	166.00	173.00	205.00
BMI	30987.0	25.37	5.61	0.0	22.49	25.05	28.30	59.25
Insurance Type	30987.0	1.97	0.16	1.0	2.00	2.00	2.00	2.00
Visit Price	30987.0	1.65	0.92	1.0	1.00	1.00	2.00	5.00
Exemption	30694.0	2.65	0.67	1.0	3.00	3.00	3.00	3.00

Voluntary	29651.0	2.10	0.317	1.0	2.00	2.00	2.00	3.00
Length of Stay	30987.0	66.68	177.810	1.0	8.00	27.00	54.00	3977.00
Procedure Duration	30987.0	111.96	82.23	6.0	55.00	88.00	145.00	1387.00
Readmission	30987.0	0.03	0.17	0.0	0.00	0.00	0.00	1.00

Table 3: Table showing statistical values for the numerical variables

3.3.2 Categorical Variables

The table below (Table 4: Table showing statistical values for the categorical variables) represents a summary of the categorical variables and its values. In the table is showing information on each of the variables, such as the count, which is the number of rows, the unique which is how many unique values exist in the dataset, the most frequent which is the value that is the most frequent and the frequency which is how frequent the most frequent values appear.

On the Unit Area there is only distinct value, which is expected since the data belongs all to same hospital. The visit number is repeated a few times, which is expected since some visits to the hospital require multiple procedures. The patient Id is repeated a few times and that is because the same patient can have multiple visits to the hospital. The gender has two values which means that the only values were feminine and masculine in the dataset, even though there is an unidentified gender value. The district value has 32 different values and since there are only 20 districts in Portugal (including the island), this means that some patients do not live in Portugal, which is the same with the city since there are only 308 different cities in Portugal (*Distritos/Concelhos - GEE*, n.d.), besides they contain some missing values (653 for District and 655 for City) and the most common value is Lisboa which is expected since the hospital is located in Lisbon. The specialty of Admission contains 24 unique values, and the most common value is *Ortopedia* (Orthopaedics) which occured 6913 times. The admission cause has 14677 unique values being the most common one *Cirurgia Electiva* (Elective Surgery) occurring 891 times and it has one missing value. The urgency type has 28487 missing values and the value *Avaliação Urgência Geral* (General Urgency Evaluation) is the most common value happening most of the time (2251 out of 2500). The admission diagnosis has 30970 unique values and 17 missing values. The Admission Date has 29589 unique values and no missing values. The discharge Date has 28405unique values and 85 missing values. The surgery date, the procedure Code and the procedure variables have no missing values. Lastly, Discharge Destination has 143 missing values and 7 unique values. The admission priority has 28692 missing values.

Variables	Number of Rows	Unique	Most Frequent Value	Frequency
Unit Area	30987	1	Hospital Lisboa	30987
Visit Number	30987	30319	D6797397DC3EDA00226FF90D03C19A45	19
Patient Id	30987	26399	7EF2F91613BCABD0CBB41C612800B283	20

Gender	30987	2	F	17089
District	30334	32	Lisboa	22151
City	30332	928	Lisboa	8184
Specialty of Admission	30987	24	Ortopedia	6613
Admission Cause	30986	14677	cirurgia electiva	891
Urgency Type	2500	3	Avaliação Urgência Geral	2251
Admission Diagnosis	30970	3179	621.0- POLIPO DO CORPO DO UTERO;	1019
Admission Date	30987	29589	2022-01-17 07:17:00	19
Discharge Date	30987	28405	2022-03-25 18:00:00	19
Admission Priority	2295	4	3 - URGENTE	1215
Surgery Date	30987	29574	2022-08-10 08:00:00	9
Procedure Code	30987	6554	46.05.00.06	2164
Procedure	30987	6858	Facoemulsificação Cristalino C/Implantação de ...	1128
Discharge Destination	30844	7	Domicílio	22515

Table 4: Table showing statistical values for the categorical variables

3.3.3 Data Visualization of Categorical Variables

After analysing the summary of the variables, a closer evaluation was done by visualising the variables and their distribution, the visualizations will be done the categorical variables since these types of variables are a lot harder to analyse and modify than the numerical variables because of the different levels they may have. Visualizations of the Categorical variables containing the least unique values, such as the gender, specialty of admission, urgency type, admission priority and discharge destination will be created since the ones that contain a lot of unique values are very difficult to visualize in a graph.

Figure 2: Gender shows the number of patients per gender for the 30987. It is visible that the sample contains more female and male.

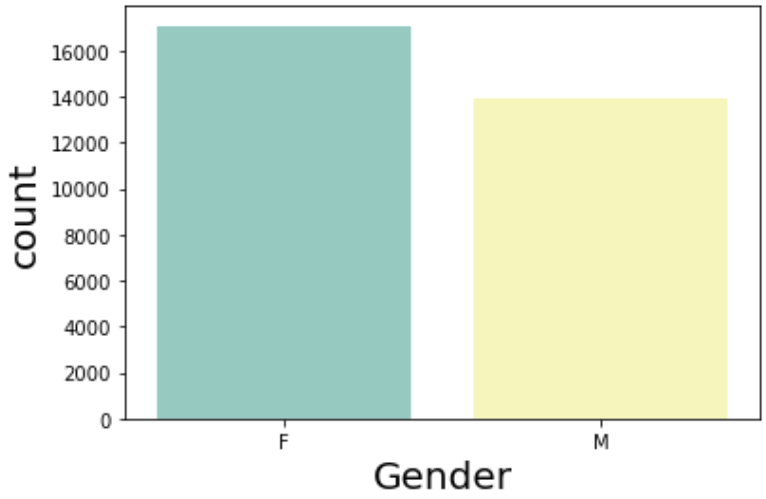

Figure 2: Gender

Figure 3: Urgency Type shows the different Urgency Types. We can see that the most common Urgency is *Avaliação Urgencia Geral*, followed by *Avaliação Urgência Pediátrica* and *Avaliação Urgência Ginecologia-Obstreticista.* In Figure 4: Urgency Type per Gender, we can see that *Avaliação Urgencia Geral* has the same number of Female and Males patients, the Avaliação *Urgência Pediátrica* has more Male patients than Female patients and the *Avaliação Urgência Ginecologia-Obstreticista* has exclusively female patients which is expected since this is exclusively for urgencies in the gynecologist and obstetrics department, and it is related with issues in women.

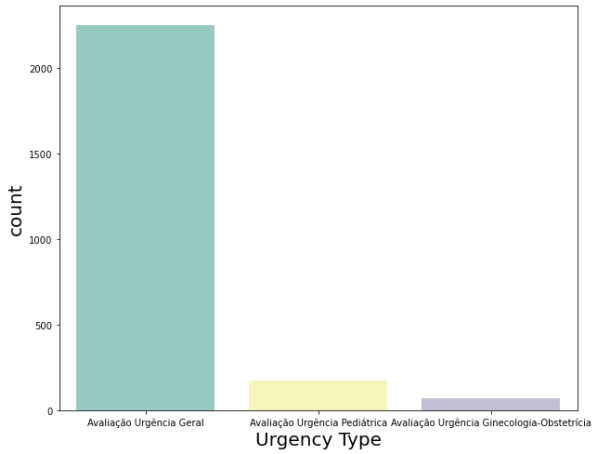

Figure 3: Urgency Type

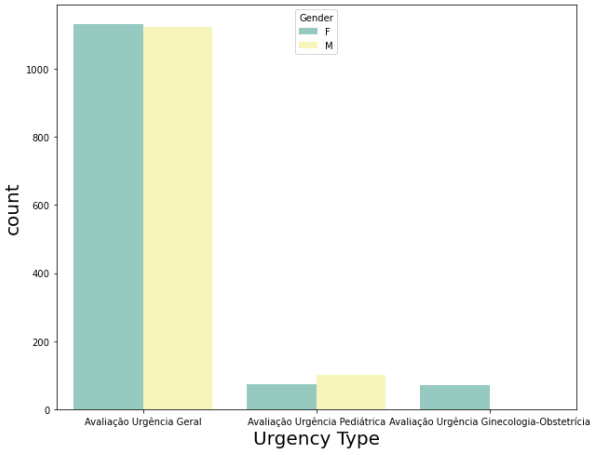

Figure 4: Urgency Type per Gender

In Figure 5: Count of Patient per Admission Priority it is visible that the most common Admission Priorities are 3 – Urgente, followed by 4-Não Urgente, 2- Muito Urgente and 1-Emergente. This column has a lot of missing values, which indicates that most of the patients are not getting surgery because of an emergency, which explains also why most values in the Urgency Type are missing, since these columns are only filled if the patient has been admitted to the emergency room.

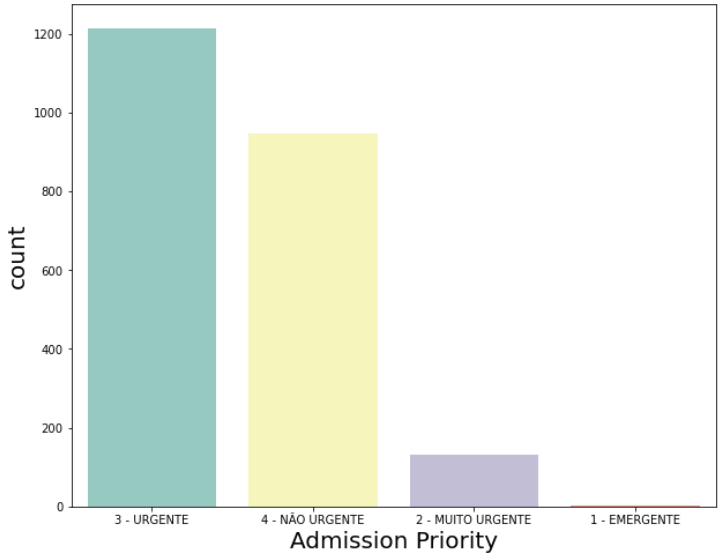

Figure 5: Count of Patient per Admission Priority

Figure 6: Count of Patients per Discharge Destination is showing the count of records per Discharged Destination, it is seen that *Domicilio* is the most common one meaning that most patients go home after being discharged, followed by *ConsultaExternaDoHospital,* meaning that patients have another hospital appointment before being discharged, followed by *Óbito* which means the patient has died and closely followed by *Transferidoparaoutrohospital* which means that the patient has been transferred to another hospital for continuing treatment. The least common discharge destinations are *Saídacontraparecermédico,* meaning that the patient decided to abandon the hospital without a discharge date. *Lastly, AltaporAbandono,* which means the customer left without having the appointment.

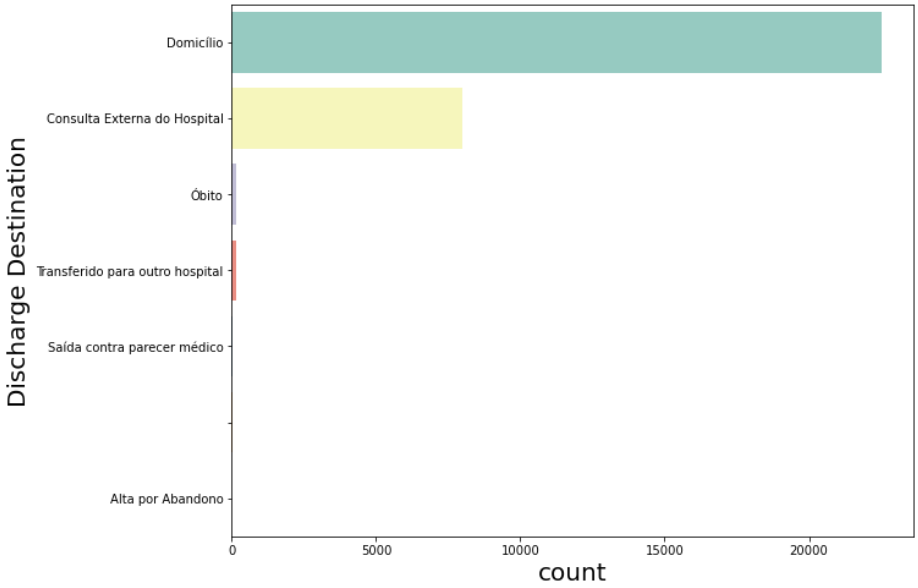

Figure 6: Count of Patients per Discharge Destination

In Figure 7: Count of Patients per District it is visible that Lisbon and Setubal are the most common districts, this is expected since the Hospital is in the Lisbon district, and Setúbal is located close to Lisbon. It is also possible to observe that some districts do not belong to Portugal, for example “São

Paulo" and "Madrid", which explains why there are more than 20 districts in our dataset, which is the number of districts in Portugal.

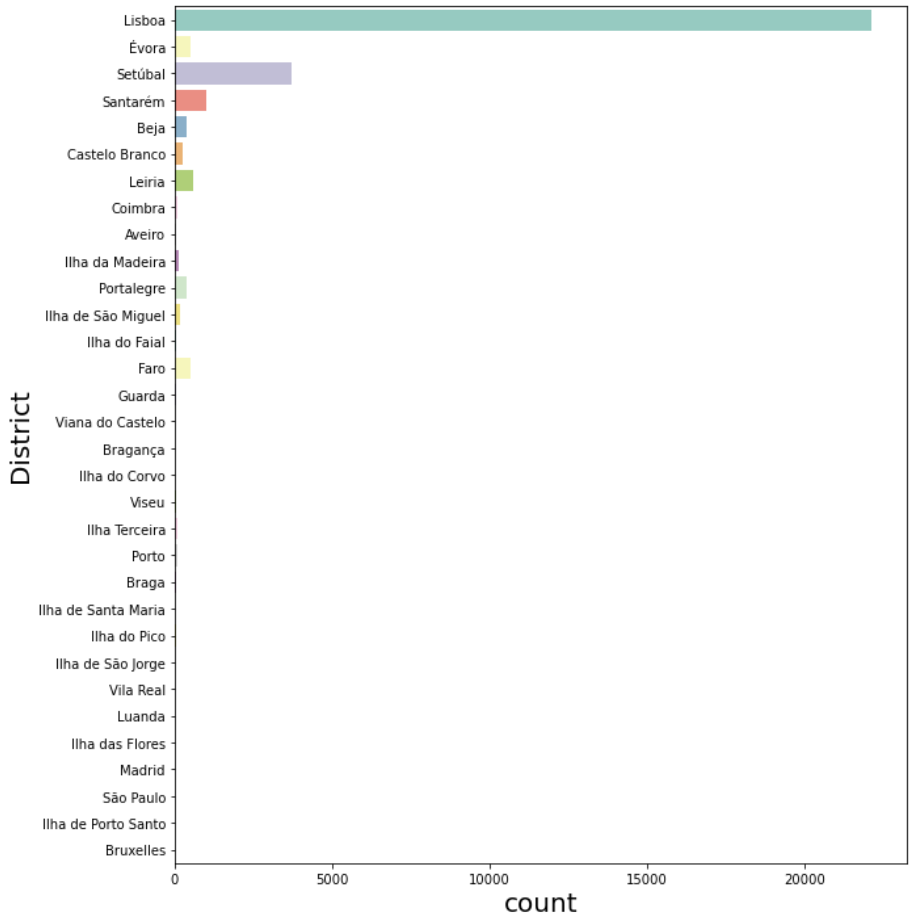

Figure 7: Count of Patients per District

3.3.4 Variable correlation

In this section, Pearson´s correlation was calculated and a Heatmap was plotted to visualize the correlations better. The Pearson´s correlation is the most common method to use for numerical variables, it varies from -1 to 1, where -1 is total negative correlation, 0 means no correlation and 1 total positive correlation. (Nettleton, 2014)

In Figure 8: Spearman correlation matrix, there are a few highly correlated variables such as the BMI and Weight are positively correlated with a value of 0.83, The Height and Weight are also positively correlated with a value 0.59. This indicates a few details that need to be examined, the BMI and Height should be correlated as well since the BMI is the division of the weight by their height. (*What Is the Body Mass Index (BMI)? - NHS*, 2019). This could be due to anomalies in our Height, Weight and BMI variables correlation matrix be plotted again after removing outliers.

The length of stay, which is our dependent variable and the one that we want to predict by looking at the most important factors. Looking at the correlation between the dependent variable and the independent variables is a good indicator since correlated variables can be used to predict one another. The length of stay is positively correlated with the Procedure duration with a value of 0.79. Lastly, the length of stay is positively correlated with the Visit price, with a value of 0.39.

3.3.5

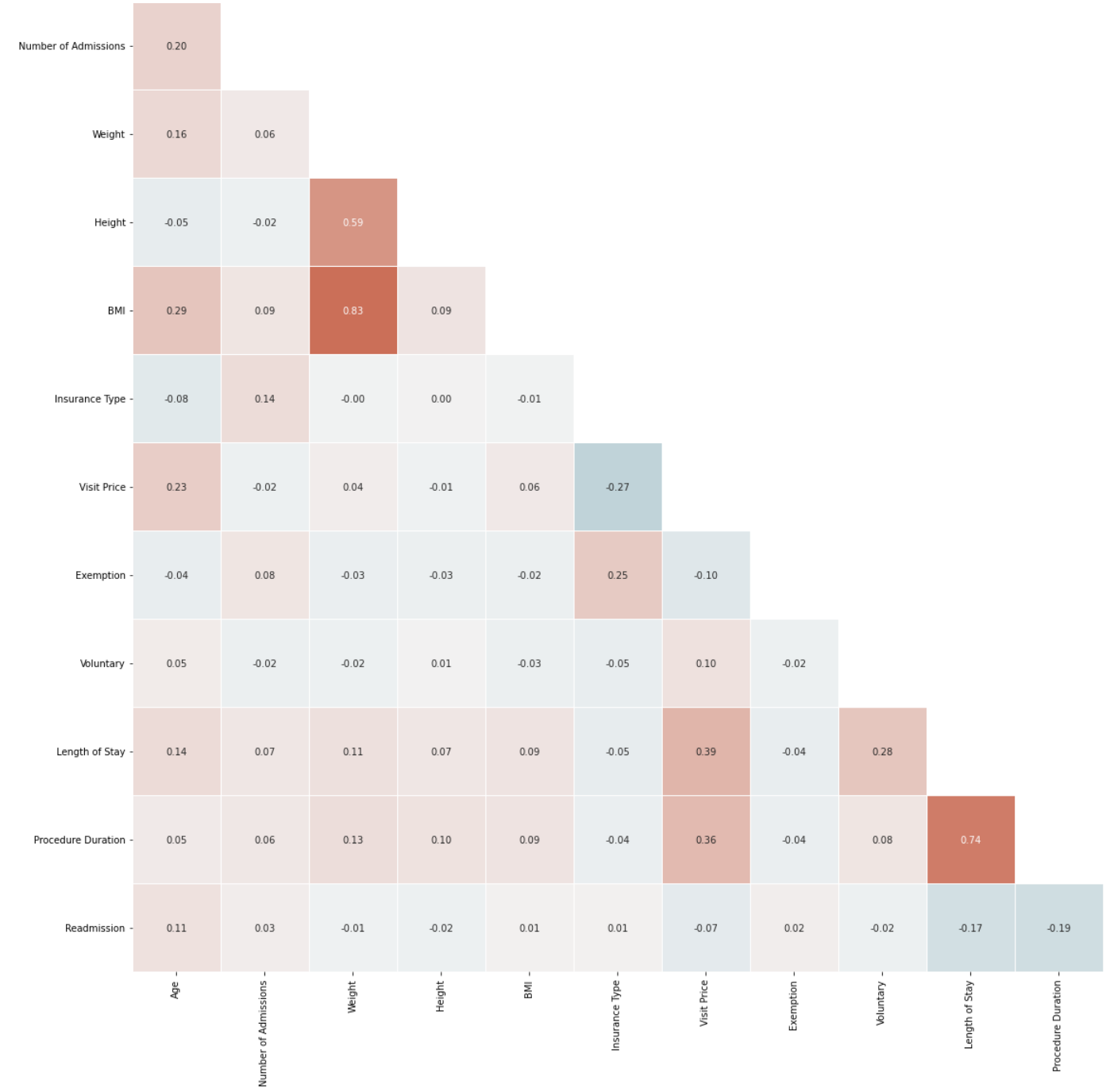

Figure 8: Spearman correlation matrix

3.4 MODIFY

For the modify part of our Data mining process, a lot of data transformations were done, such as changing data types, analysing variable correlation handling missing values, creating new variables, and removing outliers.

3.4.1 Missing Data

In this section, missing data was handled. It is important to handle missing data because any statistical results based on a dataset with non-random missing values could be biased. Additionally, many ML algorithms do not support data with missing values.

Table 5: Missing values shows the percentage of missing values on the variables, there are 9 columns that have missing values. The total number of missing values is 60277. The way these are handled depends on the Data Type of the column and on the percentage of missing values. In the next sections, the different handlings will be explained as well as what variables the handling was applied to.

Variable	Missing Values	Percentage of Missing Values
Admission Priority	28692	92.6
Urgency Type	28487	91.9
Voluntary	1336	4.3
City	655	2.1
District	653	2.1
Exemption	293	0.9
Discharge Destination	143	0.5
Admission Diagnosis	17	0.1
Admission Cause	1	0.0

Table 5: Missing values

3.4.1.1 Drop columns

For the columns that have more than 5% of missing values, the choice was to delete the variables. In this case Admission Priority (92.6% missing values) and Urgency Type (91.9% missing values), they will be deleted and will not be a part of the model. After deleting these columns, the dataset had in total 27 columns.

3.4.1.2 KNNImputer

In the case of the variables that are numerical and had less than 5% missing values, the K-Nearest Neighbour Imputation algorithm was used to estimate and replace missing data. The reason was to not use a big representation of the dataset. The k-neighbours are chosen using some distance measure and their average is used as an imputation estimate. This could be used for estimating both qualitative attributes (the most frequent value among the k nearest neighbours) and quantitative attributes (the mean of the k nearest neighbours). For this, the package from sklearn, KNNImputer was used (Sklearn.Impute.KNNImputer, 2022). The variables where the algorithm was applied were the

Voluntary (4.3% missing values), the City (2.1% missing values), the district (2.1% missing values) and the Exemption (0.9% missing values).

3.4.1.3 Imputation by Mode

In the case of variables that are categorical, the missing values were replaced by the mode of the column. These variables were the city (2.1% of missing values), the district (2.1% of missing values), the discharge destination (0.5% of missing values), admission diagnosis (0.1% of missing values) and the admission cause (0.0% of missing values).

3.4.2 Variable Transformation

In Table 6: Variable Transformations it is showing the transformation that were done to the variables in the dataset. The Surgery Date, Admission Date and Discharge Date were converted to Date Time since they were of the Type String in the original dataset. Lastly, the Discharge Destination had empty strings which were replaced with null values since an empty string would not be showing as null even though there is no value. The columns Procedure and Procedure Code were split in 4 since they contained all the procedures in one column separated by a semicolon, after they were dropped since the new 4 columns contain all the procedure.

Column	Transformation
Surgery Date	Convert from String to Date Time
Admission Date	Convert from String to Date Time
Discharge Date	Convert from String to Date Time
Discharge Destination	Replace empty string with null value
Procedure	Split the column in 4 by the separator ";" and create new columns (Proc 1, Proc 2, Proc 3, Proc 4)
Procedure Code	Split the column in 4 by the separator ";" and create new columns (Proc Code 1, Proc Code2, Proc Code 3, Proc Code 4)
Procedure	Drop the column
Procedure Code	Drop the column

Table 6: Variable Transformations

3.4.3 Outlier Removal

In this phase, the outliers were removed. This task is essential since most machine learning algorithms do not work well with outliers, since they are sensitive to the range and distribution of attribute values. Extreme values can mislead the training process resulting in longer training times, less accurate models, and ultimately poorer results. Besides these, outliers in the data can be a sign of anomalies in the data since it can have improbable values due to, for example data insertion mistakes by humans.

In Table 3: Table showing statistical values for the numerical variables in the section above, the outliers were identified based on their max and min value comparatively with the lower and upper quartile. In this section, these variables will be examined closer to determine how many outliers should be

removed. The identified variables were the Number of Admissions, Weight, Height, BMI, Length of Stay and Procedure Duration.

3.4.3.1 Number of Admissions

In Figure 9: Distribution of the Number of Admissions is represented the distribution of the number of admissions. It is visible that there are outliers in the data, from 120 number of admissions it is marked as an outlier in the boxplot, it was decided to remove the number of admissions higher than 550, which were 31.

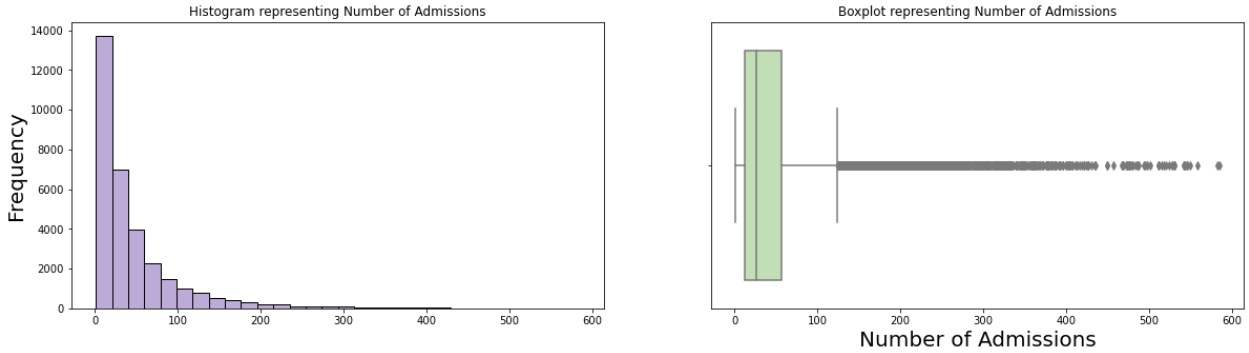

Figure 9: Distribution of the Number of Admissions

3.4.3.2 Weight

In Figure 10: Distribution of the Weight it is showing the distribution of the weights. Below around 20 and above around 100 it is visible in the boxplot as being an outlier. It was decided to remove the weights that are above 150 and below 10, this was 460 rows.

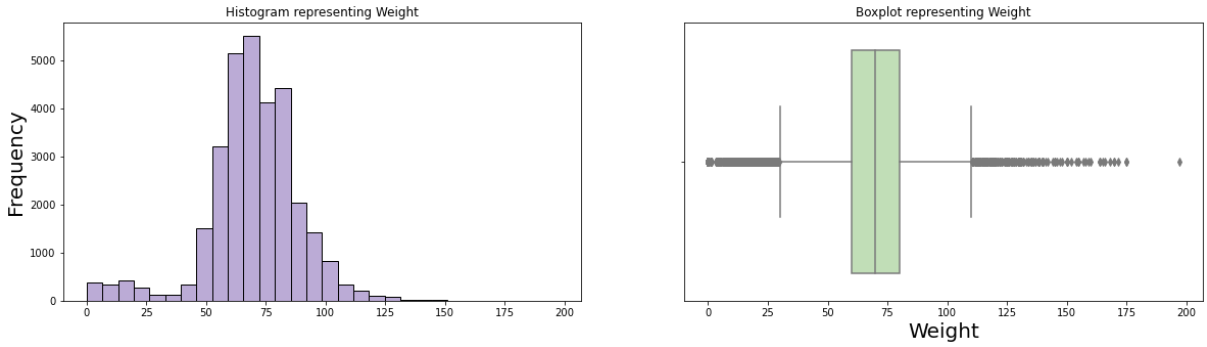

Figure 10: Distribution of the Weight

3.4.3.3 Height

In Figure 11: Distribution of the Height, it is represented the distribution of the height variable. In the boxplot, it is visible that heights under 160 and heights above 200 are outliers. It was decided to remove heights above 200 and heights under 50, which were 19 rows.

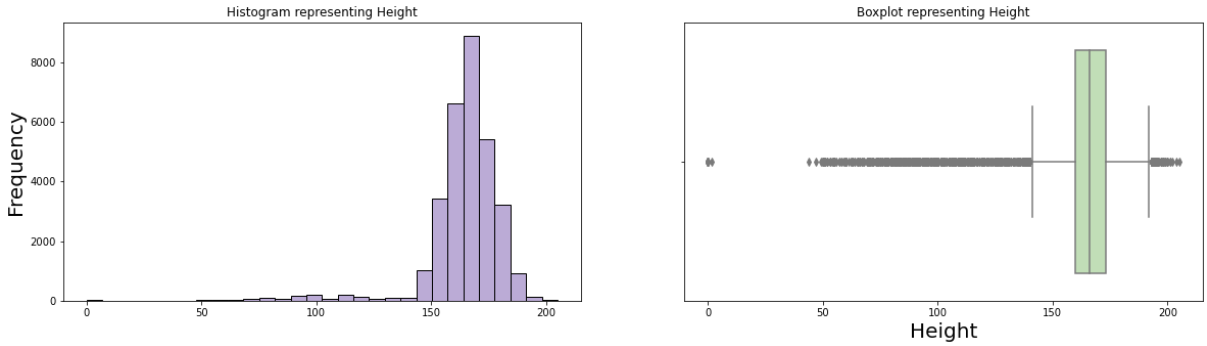

Figure 11: Distribution of the Height

3.4.3.4 BMI

In Figure 12: Distribution of the BMI, it is represented the distribution of the BMI variable. In the boxplot, it is visible that BMI values of above 35 were outliers. It was decided to remove BMIs above 50 and below 10, which were 440 rows.

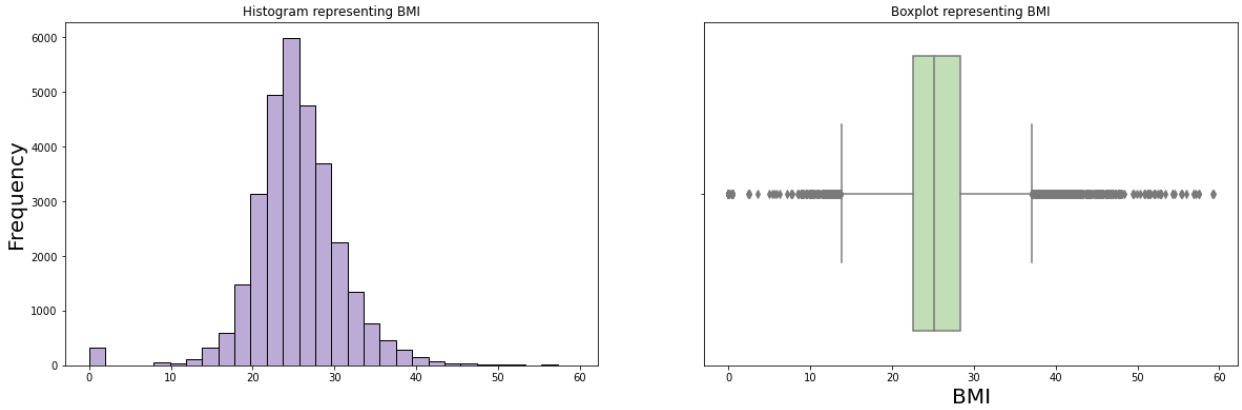

Figure 12: Distribution of the BMI

3.4.3.5 Length of Stay

In Figure 13: Distribution of the Length of Stay, it is represented the distribution of the Length of Stay variable. In the boxplot, it is visible that Length of Stay values of above 100 were outliers. It was decided to not remove any Length of Stay since these values are rare, but they happen in a hospital context in some situations, and it should be part of the model.

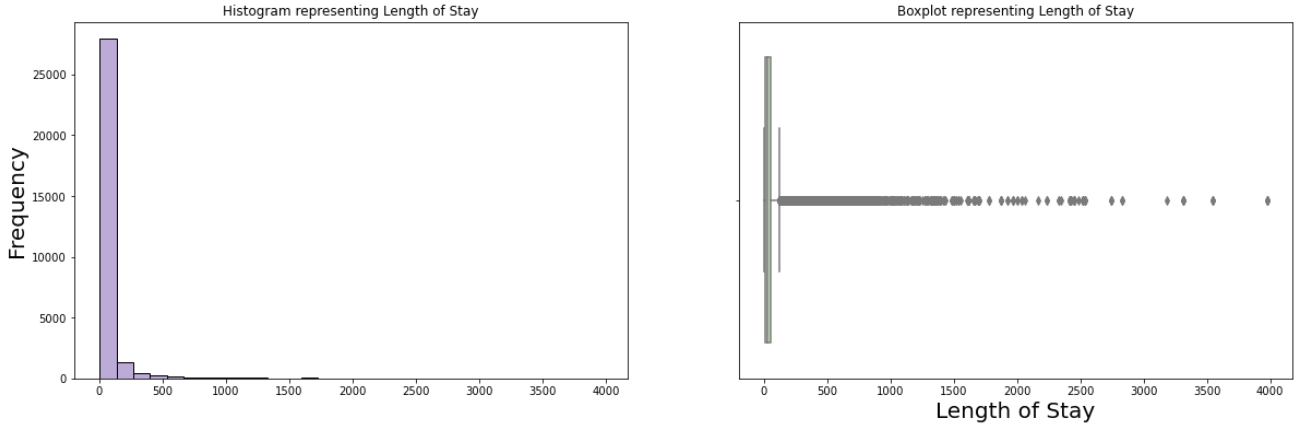

Figure 13: Distribution of the Length of Stay

3.4.3.6 Procedure Duration

In Figure 14: Distribution of the Procedure Duration it is represented the distribution of the Procedure Duration. It is visible in the Boxplot that there were some negative values for the procedure, this is impossible since duration cannot hold negative values. Therefore, the procedures under 0 hours were removed, which were 16.

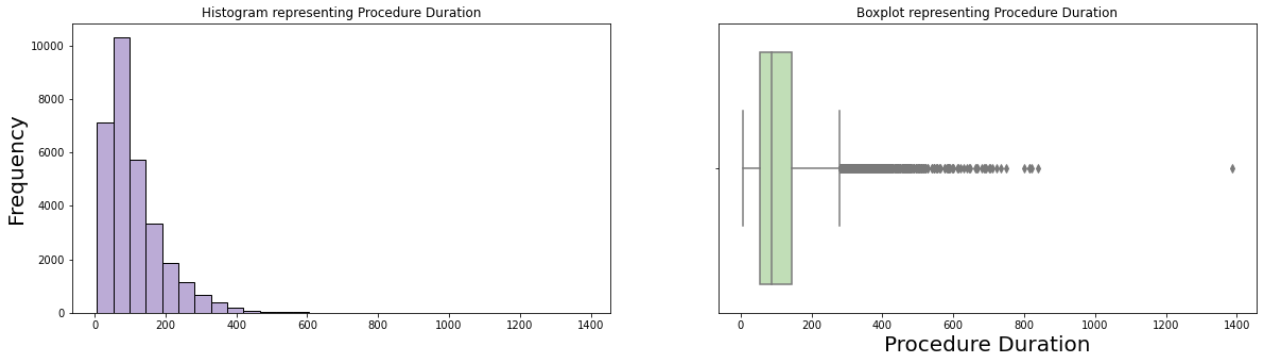

Figure 14: Distribution of the Procedure Duration

Before removing the outliers, the number of rows in the dataset was 30987. After removing some of the outliers that were defined in this section, the number of rows in the Sample was 30532 rows, which means that the number of deleted rows was 455.

3.4.4 Encoding Categorical Data

Many machine learning algorithms cannot handle categorical variables as they require all inputs to be numerical. For this reason, it was needed to transform the values of categorical variable to numbers. (Victor Popov, 2019)

For this, a label encoder was used. In label encoding every distinctive category is mapped to some arbitrary number. The sklearn.preprocessing package was used for this encoding. The label encoder was applied to the categorical variables in our dataset, such as the Gender, the district, the city, the specialty of admission, the proc code 1, the proc code 2, the proc code 3, the discharge destination, the admission cause, and the admission diagnosis. In Table 7: Encoding for the Gender variableis an example of this encoding for the variable gender, where the value F for female is replaced with 0 and the value M is replaced with 1, which means that the variable is now numerical and can be used in the model.

Original Value	Encoded Value
F	0
M	1

Table 7: Encoding for the Gender variable

3.4.5 Scaling

Standardization of datasets is a common requirement for many machine learning algorithms since the models might behave badly if the individual features do not roughly look like standard normally distributed data: Gaussian with zero mean and unit variance. (Sebastian Raschka, 2014)

The sklearn.preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. Below is the one that was tested in this project:

3.4.6 Robust Scaler

This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). This scaler is not affected by outliers as their values are not considered for scaling. (*Scikit-Learn*, 2022)

The Figure 15: Robust Scaler Formula shows the formula used by the Robust Scaler:

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Figure 15: Robust Scaler Formula

The scaler chosen for this study was the Robust Scaler since it is not affected by outliers. Not scaling the data will also be tested.

3.4.7 Feature Selection

Feature Selection is the process of selecting the number of variables used to fit the model by selecting the most relevant features from the dataset. The feature selection is an importance step for reducing the storage capacity and execution time, preventing the curse of dimensionality problem, minimizing the over-fitting issue, resulting in improved model generalization, and increasing the performance attainability (Gu et al., 2011)

Before selecting the most important features, some columns were deleted since they were not important for the study. In Table 7 is the list of features that were deleted and the reason they were deleted.

Feature	**Reason**
Patient ID	Only an identifier column
Visit Number	Only an identifier column
Unit Area	Unit area is always the same since the data is all from the same hospital
Discharge Date	The discharged date will be dropped since the discharge date will only be determined after the surgery
Proc 1	Deleted because the procedure code 1 is the same but number
Proc 2	Deleted because the procedure code 2 is the same but a number
Proc 3	Deleted because the procedure code 3 is the same but a number

Table 8: Dropped Columns

The classes in the sklearn.feature_selection module can be used for in-sample feature selection/dimensionality reduction to improve the accuracy rating of estimators or to improve their performance on very high dimensional datasets. (*Scikit-Learn*, 2022).

3.4.7.1 RFE

RFE is a wrapper class in sklearn that works by iteratively training a model, computing the ranking criterion for all features in a way that the best ranking is the top contributor to the model and eliminating the lowest scoring feature from the subset. This process occurs until all features are eliminated. The subset of features that gives the best overall assessment score is the one selected. The model used was a Random Forest Regressor with default parameters.

3.4.7.2 SelectKBest

SelectKBest method selects the features according to the k (number of features) highest scores. The function parameter takes a machine learning algorithm. The function was the sklearn.feature_selection.f_regression, and the number of features was 9.

3.4.7.3 Pearson Correlation

Lastly, the Pearson correlation which is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.(Shaun Turney, 2022) This was used to evaluate the correlation between the Length of Stay and the independent variables. In table 8 there is the pearson correlation for the length of stay and the independent variables, the correlation was viewed for the dataset before removing the outliers, scaling, and feature selection.

Voluntary	0.29
Visit Price	0.28
Procedure Duration	0.24
Discharge Destination	0.14
Number of Admissions	0.08
BMI	0.05
Weight	0.04
Gender	0.03
Height	0.02
Admission Diagnosis	0.01
City	0.01
Surgery Date	-0.01

Exemption	-0.01
Admission Date	-0.01
District	-0.01
Insurance Type	-0.02
Readmission	-0.02
Proc Code 2	-0.03
Admission Cause	-0.05
Proc Code 1	-0.06
Proc Code 3	-0.06
Specialty of Admission	-0.07

Table 9: Pearson Correlation

In the end, these three estimators were compared to determine the feature importance, by evaluating it through different measures. This is done for every dataset since feature importance varies depending on the data transformation done.

At the end of the modify phase, a total of 8 datasets were created to test different methods in data mining. Some datasets will be tested with the features selected based on the feature selection for that dataset and the datasets that will not be tested with feature selected will have all the variables such as Visit Price, Specialty of Admission, Procedure Duration, Discharge Destination, Age, Voluntary, BMI, Admission Diagnosis, Proc Code 3, Proc Code 1, Number of Admissions, Exemption, Weight, Readmission, Proc Code 2, Insurance Type, Height, Gender, District, City and Admission Cause. The datasets will be tested with and without outlier removal as well as no scaling applied. Below are the different datasets created:

1. Dataset 1 -no outlier removal, no scaler applied and no feature selection
2. Dataset 2 - outlier removal, no scaler applied and no feature selection
3. Dataset 3 - no outlier removal, scaler applied and no feature selection
4. Dataset 4 - no outlier removal, no scaler applied and feature selection
5. Dataset 5 - outlier removal, scaler applied and no feature selection
6. Dataset 6 - no outlier removal, scaler applied and feature selection
7. Dataset 7 - outlier removal, no scaler applied and feature selection
8. Dataset 8 - outlier removal, scaler applied and feature selection

For the datasets that have feature selection, the estimators were applied to them, such as the Pearson correlation, the RFE and the SelectKBest with f regression. 8 features were chosen from each estimator. In the end, the chosen features were the features that were selected by 1 or more estimators.

Below is the visualization used to determine the features. This visualization was used for every dataset; hence the datasets will have different variables. In Table 9 is the example for the dataset 1.

Feature	Pearson	RFE	SelectKBest	Total
Visit Price	True	True	True	3
Specialty of Admission	True	True	True	3
Procedure Duration	True	True	True	3
Discharge Destination	True	True	True	3
Age	True	True	True	3
Weight	True	False	True	2
Voluntary	True	False	True	2
Number of Admissions	True	False	True	2
Proc Code 1	False	True	False	1
BMI	False	True	False	1
Admission Diagnosis	False	True	False	1
Surgery Date	False	False	False	0
Readmission	False	False	False	0
Proc Code 3	False	False	False	0
Proc Code 2	False	False	False	0
Insurance Type	False	False	False	0
Height	False	False	False	0
Gender	False	False	False	0
Exemption	False	False	False	0
District	False	False	False	0
City	False	False	False	0
Admission Date	False	False	False	0

Admission Cause	False	False	False	0

Table 10: Feature Selection for dataset 1

Below how the first rows of the 8 datasets that were created look. It is visible looking at the data they have all been randomized and it is visible that some datasets were scaled. The first 4 datasets have less features since they have feature selection, and the rest of the datasets contain all features.

3.4.8 Dataset 1

	Visit Price	Specialty of Admission	Procedure Duration	Discharge Destination	Age	Weight	Voluntary	Number of Admissions	Proc Code 1	BMI	Admission Diagnosis
4172	0.0	-0.916667	0.270588	0.0	-0.413793	1.10	0.0	-0.066298	0.091232	1.533913	-0.753989
7698	0.0	0.333333	-0.035294	-1.0	-1.172414	-3.50	0.0	1.259669	-0.528436	-0.130435	0.894282
28777	0.0	-0.916667	-0.411765	-1.0	-1.241379	0.05	0.0	0.022099	-1.100711	-0.488696	0.357713
29906	0.0	0.333333	-0.352941	0.0	0.068966	0.00	0.0	-0.265193	-0.610190	0.074783	0.555851
8464	0.0	0.000000	0.964706	0.0	-0.172414	0.35	1.0	0.817680	0.539100	1.377391	-1.009309

Table 11: Dataset 1 First Rows

3.4.9 Dataset 2

	Voluntary	Visit Price	Specialty of Admission	Procedure Duration	Discharge Destination	Age	Number of Admissions	Admission Cause	BMI
30682	0.0	0.0	-0.916667	-0.379310	0.0	0.448276	1.369565	-0.794551	1.084922
12509	0.0	0.0	0.000000	-0.505747	0.0	0.551724	1.260870	0.303121	0.559792
3874	1.0	0.5	0.333333	0.988506	0.0	0.448276	1.869565	-0.319975	0.239168
8230	0.0	0.0	0.416667	-0.494253	0.0	-0.931034	-0.347826	0.519563	0.474870
15238	0.0	0.0	-0.916667	-0.413793	-1.0	-0.793103	1.826087	-0.351148	0.558059

Table 12: Dataset 2 First Rows

3.4.10 Dataset 3

	Visit Price	Specialty of Admission	Procedure Duration	Discharge Destination	Age	Weight	Voluntary	Number of Admissions	Proc Code 1	BMI	Admission Diagnosis
4172	1.0	4	108.0	3	42.0	92.0	2.0	23.0	1060	33.790001	532
7698	1.0	19	82.0	2	20.0	0.0	2.0	83.0	537	24.219999	3011
28777	1.0	4	50.0	2	18.0	71.0	2.0	27.0	54	22.160000	2204
29906	1.0	19	55.0	3	56.0	70.0	2.0	14.0	468	25.400000	2502
8464	1.0	15	167.0	3	49.0	77.0	3.0	63.0	1438	32.889999	148

Table 13: Dataset 3 First Rows

3.4.11 Dataset 4

	Visit Price	Specialty of Admission	Procedure Duration	Discharge Destination	Age	Voluntary	Number of Admissions	Admission Cause	Proc Code 1	BMI	Admission Diagnosis
30682	1.0	4	53.0	3	67.0	2.0	89.0	646	809	31.410000	58
12509	1.0	15	42.0	3	70.0	2.0	84.0	9308	1433	28.379999	719
3874	2.0	19	172.0	3	67.0	3.0	112.0	4391	264	26.530001	2818
8230	1.0	20	43.0	3	27.0	2.0	10.0	11016	650	27.889999	1314
15238	1.0	4	50.0	2	31.0	2.0	110.0	4145	961	28.370001	1593

Table 14: Dataset 4 First Rows

3.4.12 Dataset 5

	Age	Gender	Number of Admissions	Weight	Height	BMI	District	City	Insurance Type	Specialty of Admission	Admission Cause	Admission Diagnosis	Admission Date
4172	-0.413793	0.0	-0.066298	1.10	0.000000	1.533913	0.0	-0.301639	0.0	-0.916667	0.167508	-0.753989	0.128463
7698	-1.172414	1.0	1.259669	-3.50	-12.692308	-0.130435	0.0	0.000000	0.0	0.333333	-0.338824	0.894282	-0.785842
28777	-1.241379	1.0	0.022099	0.05	1.076923	-0.488696	6.0	1.118033	0.0	-0.916667	-0.640713	0.357713	0.578357
29906	0.068966	0.0	-0.265193	0.00	0.076923	0.074783	0.0	0.127869	0.0	0.333333	0.206922	0.555851	-0.137599
8464	-0.172414	0.0	0.817680	0.35	-0.923077	1.377391	0.0	0.000000	0.0	0.000000	0.361554	-1.009309	0.237194

	Surgery Date	Visit Price	Exemption	Voluntary	Discharge Destination	Procedure Duration	Readmission	Proc Code 1	Proc Code 2	Proc Code 3
4172	0.127686	0.0	0.0	0.0	0.0	0.270588	0.0	0.091232	0.000000	0.0
7698	-0.784500	0.0	-2.0	0.0	-1.0	-0.035294	0.0	-0.528436	0.000000	0.0
28777	0.578748	0.0	-2.0	0.0	-1.0	-0.411765	0.0	-1.100711	-1.447263	0.0
29906	-0.137875	0.0	0.0	0.0	0.0	-0.352941	0.0	-0.610190	0.000000	0.0
8464	0.237556	0.0	0.0	1.0	0.0	0.964706	0.0	0.539100	0.000000	0.0

Table 15: Dataset 5 First Rows

3.4.13 Dataset 6

	Age	Gender	Number of Admissions	Weight	Height	BMI	District	City	Insurance Type	Specialty of Admission	Admission Cause	Admission Diagnosis	Admission Date
4172	42.0	0	23.0	92.0	165.0	33.790001	20	313	2.0	4	8582	532	17536
7698	20.0	1	83.0	0.0	0.0	24.219999	20	405	2.0	19	4394	3011	3426
28777	18.0	1	27.0	71.0	179.0	22.160000	26	746	2.0	4	1897	2204	24479
29906	56.0	0	14.0	70.0	166.0	25.400000	20	444	2.0	19	8908	2502	13430
8464	49.0	0	63.0	77.0	153.0	32.889999	20	405	2.0	15	10187	148	19214

	Surgery Date	Visit Price	Exemption	Voluntary	Discharge Destination	Procedure Duration	Readmission	Proc Code 1	Proc Code 2	Proc Code 3
4172	17920	1.0	3.0	2.0	3	108.0	0	1060	1501	820
7698	3507	1.0	1.0	2.0	2	82.0	0	537	1501	820
28777	25047	1.0	1.0	2.0	2	50.0	0	54	146	820
29906	13724	1.0	3.0	2.0	3	55.0	0	468	1501	820
8464	19656	1.0	3.0	3.0	3	167.0	0	1438	1501	820

Table 16: Dataset 6 First rows

3.4.14 Dataset 7

	Age	Gender	Number of Admissions	Weight	Height	BMI	District	City	Insurance Type	Specialty of Admission	Admission Cause	Admission Diagnosis	Admission Date
30682	0.448276	1.0	1.369565	1.20	0.538462	1.084922	0.0	-0.58	0.0	-0.916667	-0.794551	-1.085383	-0.211082
12509	0.551724	1.0	1.260870	0.65	0.384615	0.559792	0.0	0.82	0.0	0.000000	0.303121	-0.633880	0.752351
3874	0.448276	1.0	1.869565	0.75	1.000000	0.239168	0.0	0.00	0.0	0.333333	-0.319975	0.799863	-0.040845
8230	-0.931034	0.0	-0.347826	0.25	-0.153846	0.474870	0.0	0.00	0.0	0.416667	0.519563	-0.227459	0.271857
15238	-0.793103	0.0	1.826087	0.60	0.307692	0.558059	0.0	0.00	0.0	-0.916667	-0.351148	-0.036885	-0.271178

	Surgery Date	Visit Price	Exemption	Voluntary	Discharge Destination	Procedure Duration	Readmission	Proc Code 1	Proc Code 2	Proc Code 3
30682	-0.211189	0.0	0.0	0.0	0.0	-0.379310	0.0	-0.188249	0.000000	0.0
12509	0.754759	0.0	0.0	0.0	0.0	-0.505747	0.0	0.559952	0.000000	0.0
3874	-0.041627	0.5	0.0	1.0	0.0	0.988506	1.0	-0.841727	-1.257855	-676.0
8230	0.273247	0.0	0.0	0.0	0.0	-0.494253	0.0	-0.378897	-0.893824	0.0
15238	-0.271320	0.0	0.0	0.0	-1.0	-0.413793	0.0	-0.005995	0.000000	0.0

Table 17: Dataset 7 First Rows

3.4.15 Dataset 8

	Age	Gender	Number of Admissions	Weight	Height	BMI	District	City	Insurance Type	Specialty of Admission	Admission Cause	Admission Diagnosis	Admission Date
30682	67.0	1	89.0	94.0	173.0	31.410000	20	221	2.0	4	646	58	11734
12509	70.0	1	84.0	83.0	171.0	28.379999	20	641	2.0	15	9308	719	25922
3874	67.0	1	112.0	85.0	179.0	26.530001	20	395	2.0	19	4391	2818	14241
8230	27.0	0	10.0	75.0	164.0	27.889999	20	395	2.0	20	11016	1314	18846
15238	31.0	0	110.0	82.0	170.0	28.370001	20	395	2.0	4	4145	1593	10849

	Surgery Date	Visit Price	Exemption	Voluntary	Discharge Destination	Procedure Duration	Readmission	Proc Code 1	Proc Code 2	Proc Code 3
30682	11984	1.0	3.0	2.0	3	53.0	0	809	1459	796
12509	26522	1.0	3.0	2.0	3	42.0	0	1433	1459	796
3874	14536	2.0	3.0	3.0	3	172.0	1	264	298	120
8230	19275	1.0	3.0	2.0	3	43.0	0	650	634	796
15238	11079	1.0	3.0	2.0	2	50.0	0	961	1459	796

Table 18: Dataset 8 First Rows

3.5 MODEL

Simple Regression is the investigation of the relationship between variables, it investigates the casual effect of one variable upon another. It estimates the quantitative effect of the independent variables upon the dependant variable. Multiple regression is a technique that allows multiple variables to be analysed separately so that the effect of each can be estimated on the variable to be predicted. (Sykes, 1993)

The main objective of this study is to build a predictive model capable of predicting new lengths of stay of a certain patient, given multiple variables as an input. The model will learn a function to make predictions of a defined variable (LOS) based on the input data, which means it is a regression problem and since multiple variables it is a multiple regression problem, therefore it will make use of regression algorithms. The models chosen were Linear Regression (LR), Decision Tree Regressor (DT) Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB) and Artificial Neural Networks (ANN).

3.5.1 Data Splitting

Before the data was used to train the models, it was split in two, training and test data. Data splitting is commonly used in machine learning to split data into a train, test, or validation set. This approach allows us to find the model hyper-parameter and estimate the generalization performance. (Vijaya et al., 2018)

The dataset with the selected variables, Visit Price, Specialty of Admission, Procedure Duration, Discharge Destination, Age, Voluntary, Proc Code 1, BMI and Admission Diagnosis will be divided into training and test dataset. The training dataset is used to train and develop models. The package of sklearn, sklearn.model_selection.train_test_split was used for splitting the data. The parameters used were the train_size which was set to 0.75, meaning that 75% of the data was used to train the model and test_size which was set to 0.25, meaning that 25% of the data was used to test and assess the model on unknown data and the random state was set to 42, meaning that the data was shuffled.

3.5.2 Model Training

During the model training, the training dataset will be fitted into the different models. All the algorithms used in this study were implemented using the Scikit-Learn library. Below each of these algorithms will be explained.

3.5.2.1 Decision Tree (DT)

Decision Trees map the possible outcomes of a series of related choices. A decision tree typically starts with a node (root node), which branches into more nodes with possible choices (Decision node) and outcomes (leaf node). This will lead to a tree-like shape as seen in figure 15.

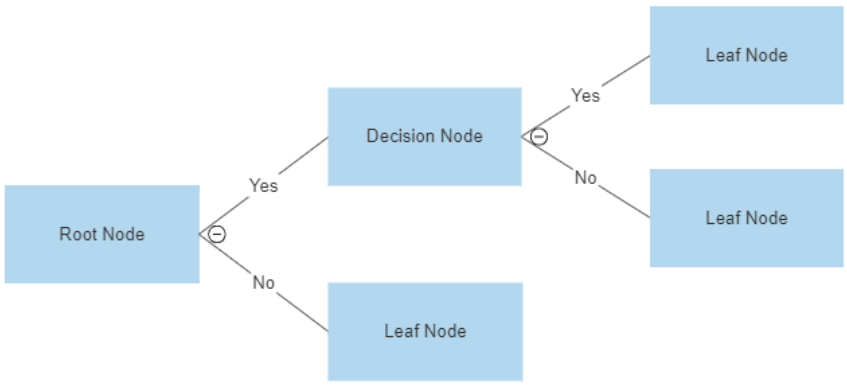

Figure 16: Decision Tree Ilustration

3.5.2.2 Gradient Boosting (GB)

Gradient boosting is an ensemble model comprising of a set of weak learners, which are typically decision trees with only one split, obtained in a stage-wise fashion through the minimization of some differentiable prediction loss using functional gradient descent. (*Scikit-Learn*, 2022). In this study, a few different variations of the gradient boosting decision trees were tested since they have different implementations. The algorithms that were tested were the XGBoost and the LightGBM. The main difference between them is that in the XGBoost the trees grow depth-wise while in LightGBM, trees grow leaf-wise. For XGBoost the sklearn.ensemble. XGBRFRegressor package was used and for the LightGBM the sklearn.ensemble. LGBMRegressor package was used. As seen in the pictures below:

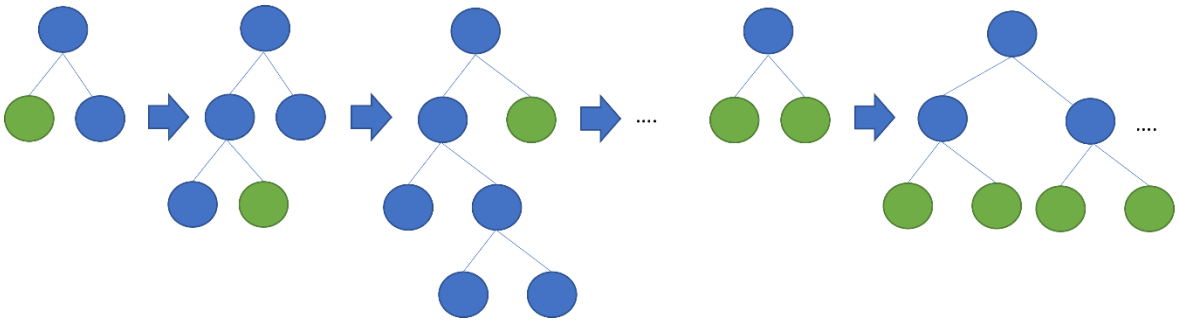

Figure 17: LightGBM (left) and XGBoost (right)

3.5.2.3 Random forest Regressor

The sklearn.ensemble.RandomForestRegressor package was used for training the data. RF is an ensemble learning method, which constructs multiple decision trees (each on a randomly sampled feature set) at the training stage. Their outputs will be aggregated in the prediction stage (usually through majority voting) as the result. This algorithm can also be used for classification problems when the output is categorical. (*Scikit-Learn*, 2022)

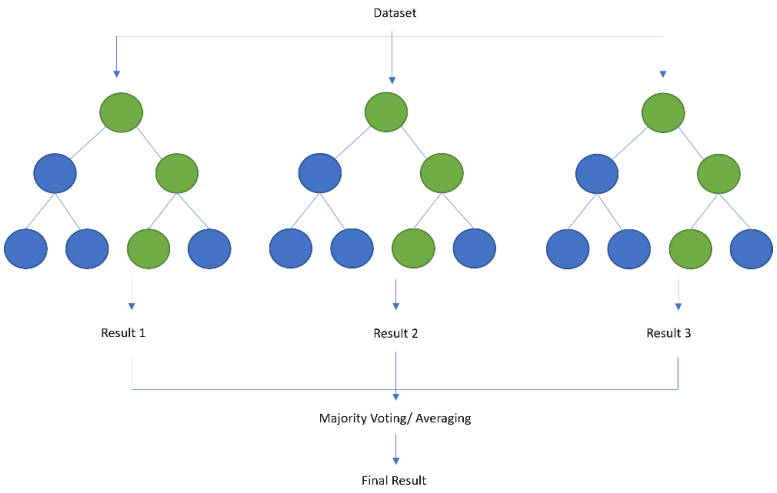

Figure 18: Random Forest

3.5.2.4 Multi-Layer Perceptron Regressor

The sklearn.neural_network. MLPRegressor class was used for training the data. A multilayer perceptron is a class of feedforward artificial neural network. It consists of an arrangement of layers. Each layer is composed of nodes, which one with a nonlinear activation function, that connect between each subsequent layer, with weights attributed to which connection. Moreover, some nodes receive additional information through what is called a bias. To train such algorithm, an optimization method like the gradient descent is used. In its minimal state, the algorithm is composed of 3 layers, the input, hidden and output layer, as seen on Figure 19. Figure is from (Meng et al., 2021)

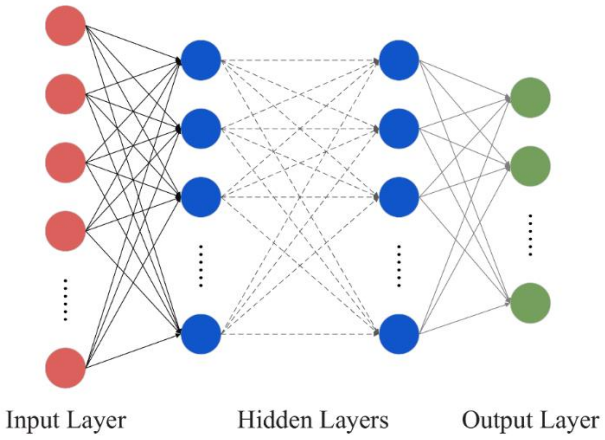

Figure 19: MLP Architecture

3.6 ASSESS

The final phase of the SEMMA methodology is the assess phase. The objective of this phase is to evaluate and compare the several models' performance with the use of metrics. The metrics chosen were the Mean Absolute Error, Mean Squared Error and R Squared.

3.6.1 R Squared (R2)

R2 is a measure that determines the amount of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. If the R^2 of a model is 0.50, then approximately half of the observed variation in the independent variable can be explained by the dependent variables.(*Scikit-Learn*, 2022)

$$R^2 = 1 - \frac{RSS}{TSS}$$

Figure 20: R Squared Formula

3.6.2 Mean Absolute Error (MAE)

MAE is the mean of the absolute errors; the absolute error is the absolute value of the different between the predicted value and the actual value. This metric is often used in regression models and helps predict the accuracy of a model. When a model has no error, the MAE equals zero. (Yang & Sun, 2021)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Figure 21: Mean Absolute Error Formula

3.6.3 Mean Squared Error (MSE)

MSE is the mean of the squared errors, the squared error is the value of the different between the predicted value and the actual value. This measure is used to tell how close a regression line is to a set of points. When a model has no error, the MSE equals zero. (Yang & Sun, 2021)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure 22: Mean Squared Error Formula

4 RESULTS AND DISCUSSION

After creating the datasets and apply all the transformations to them including the feature selection, the eight training datasets were all trained with the five different models, creating 40 different model configurations, and producing different results. In this section, these results will be analysed, and the best performing models will be discussed.

4.1 MODELS COMPARISON

4.1.1 Decision Tree

Table 18 shows the scores of the decision tree model for each dataset. The decision tree best performed in the dataset 1 and dataset 2, with a R2 of 0.48. The results indicate that this model performs better with the presence of outliers since the datasets had both outliers. It is not affected by scaling the model since the scores did not change when scaling was applied. It also indicates that the model performed better with less features, since the datasets had feature selection.

Dataset	Outliers	Scaled	Feature Selection	R2	MAE	MS2
1	**Yes**	**Yes**	**Yes**	**0.48**	**35.19**	**15023.26**
2	No	Yes	Yes	0.26	40.89	20237.67
3	**Yes**	**No**	**Yes**	**0.48**	**35.38**	**15006.00**
4	No	No	Yes	0.14	41.85	23694.83
5	Yes	Yes	No	0.35	38.07	18709.20
6	Yes	No	No	0.40	37.44	17260.79
7	No	Yes	No	0.27	39.65	19911.29
8	No	No	No	0.22	40.14	21292.72

Table 19: Decision Tree Assessment

4.1.2 MLP Regressor

Table 19 shows the scores of the MLP Regressor for each of the datasets. The MLP Regressor performed better in dataset 6 and 7, with a R2 of 0.50 and 0.52, respectively. The results indicates that this model performs better without the presence of outliers since both datasets have no outliers. Scaling the models did not improve the performance significantly. It also indicates that the model performed better with more features, since the datasets with no feature selection had higher scores than the ones without.

Dataset	Outliers	Scaled	Feature Selection	R2	MAE	MS2
1	Yes	Yes	Yes	0.41	47.35	17101.48

2	No	Yes	Yes	0.45	45.52	15188.29
3	Yes	No	Yes	0.21	65.60	22884.38
4	No	No	Yes	0.18	59.62	22392.07
5	Yes	Yes	No	0.39	45.62	17487.41
6	Yes	No	No	0.46	61.09	15525.14
7	**No**	**Yes**	**No**	**0.52**	**48.50**	**13135.46**
8	**No**	**No**	**No**	**0.50**	**53.41**	**13642.43**

Table 20: MLP Regressor Assessment

4.1.3 Random Forest Regressor

Table 20 shows the scores of the Random Forest Regressor for each of the datasets. The Random Forest Regressor performed better in the dataset 3 and 5, with a R2 of 0.60. The results indicates that this model performs better with the presence of outliers since the outliers were not removed from the datasets. It is not affected by scaling the model since the scores did not change when the dataset was scaled. Feature selection did not affect the score either.

Dataset	Outliers	Scaled	Feature Selection	R2	MAE	MS2
1	Yes	Yes	Yes	0.59	34.23	11864.32
2	No	Yes	Yes	0.55	36.14	12372.02
3	**Yes**	**No**	**Yes**	**0.60**	**34.13**	**11596.44**
4	No	No	Yes	0.56	35.61	12353.81
5	**Yes**	**Yes**	**No**	**0.60**	**34.64**	**11575.92**
6	Yes	No	No	0.63	34.04	10746.75
7	No	Yes	No	0.57	35.83	11711.43
8	No	No	No	0.58	35.19	11440.17

Table 21: Random Forest Assessment

4.1.4 XGBoost Regressor

Table 20 shows the scores of the XGBoost Regressor for each of the datasets. The Random Forest Regressor performed better in the dataset 5,6,7 and 8, with a R2 of 0.69. The results indicates that the model's performance is not affected by the outliers or the scaling. It performs slightly better with more features, since there was no feature selection involved.

Dataset	Outliers	Scaled	Feature Selection	R2	MAE	MS2
1	Yes	Yes	Yes	0.66	33.23	9860.62

2	No	Yes	Yes	0.62	35.09	10259.58
3	Yes	No	Yes	0.66	33.23	9860.62
4	No	No	Yes	0.68	31.37	8199.25
5	**Yes**	**Yes**	**No**	**0.69**	**32.51**	**8817.63**
6	**Yes**	**No**	**No**	**0.69**	**32.51**	**8817.44**
7	**No**	**Yes**	**No**	**0.69**	**32.02**	**8404.73**
8	**No**	**No**	**No**	**0.69**	**32.00**	**8079.72**

Table 22: XGBoost Regressor Assessment

4.1.5 Light Gradient Boosting Regressor

Table 22 shows the scores of the LightGBM for each of the datasets. The Random Forest Regressor performed better in the dataset 7,8. The results indicates that the model's performance is slightly improved with the presence of outliers. The scaling does not affect the model. The model also works better when there has been feature selection.

Dataset	**Outliers**	**Scaled**	**Feature Selection**	**R2**	**MAE**	**MS2**
1	Yes	Yes	Yes	0.68	32.22	9144.30
2	No	Yes	Yes	0.65	34.47	9666.17
3	Yes	No	Yes	0.68	32.03	9063.86
4	No	No	Yes	0.69	31.37	8199.25
5	Yes	Yes	No	0.70	31.59	8483.17
6	Yes	No	No	0.71	31.58	8422.33
7	**No**	**Yes**	**No**	**0.73**	**30.90**	**7418.70**
8	**No**	**No**	**No**	**0.72**	**30.01**	**8079.72**

Table 23: LGBM Regressor

In Table 23 is the best scores for each of the models. The Light GBM was the best performer out of the models, with a R2 of 0.73.

Decision Tree			**MLP Regressor**			**Random Forest**			**XGBoost**			**Light GBM**		
R2	**MAE**	**MS2**	**R2**	**MAE**	**MS2**	**R2**	**MAE**	**MS2**	**R2**	**MAE**	**MS2**	**R2**	**MAE**	**MS2**
0.48	35.19	15023.26	0.52	48.50	13135.46	0.60	34.13	11596.44	0.69	32.00	8079.72	**0.73**	**30.90**	**7418.70**

Table 24: Best Scores for each model

Looking at the R2 value, the best performing score had an R2 of 0.73 which means that 73% of the variance of the length of the stay can be explained by the model. This means that the variables included in the dataset were good predictors of the changes in the length of stay and it indicates that the variables selected from the database were significant in the study.

In terms of accuracy, the model performs poorly. Since the MAE was 30.90, the model has an average absolute error of 30.90, this means that the model will predict the length of error with an average error of 30,90 hours, for example if the model predicts the length of stay to be 30.90, the actual length of stay could be either 60 or 0.90, which is around 1 day delay or forward, it would predict the patient to leave the hospital 1 day earlier or 1 day later and this value would be excessively high to use in the hospital since it represents more than 1 day in length of stay is a great inconsistency between the actual value and the predicted values. The value of length of stay is on average 67,78 for this dataset and 30,90 is a big part of this value.

Looking at the 40 models' configurations some conclusions can be drawn:

- Outlier removal does not have a significant impact on the performance of the models
- Scaling the data with the Robust Scaler does not have a significant impact
- The models performed better with no FTE to perform feature selection
- The Boosting algorithms work the best with the data

5 CONCLUSION

In this study, the predictive power was examined by using previously known data such as the patient data and hospital data when the patient is admitted, to calculate and determine the length of stay. This was done with data from more than 30,000 surgeries in the hospital from 2019 until 2022. Five different regression models were used to develop this prediction and 23 variables were used to build the model and different settings were tested with the models.

For each of the models, eight datasets were created meaning that 40 predictors were created and 20 used distinct features' subsets according to the feature selection. Decision Trees, Random Forests, Multilayer Perceptron Neural Network and Gradient Boosting Algorithms. The data was split into training and testing data which was used to train and then test the accuracy of the model, respectively.

This study's findings suggest that clinical data such as patient data and admission data are good at predicting length of stay, however the accuracy cannot be considered high enough. Even though it did not calculate the exact length of stay instances, the created models can classify longer lengths of stays that would demand better healthcare and should be primary targets for preventive care programs. This study also contributes to improving the patients' quality of life and medical outcome since these kinds of predictions may be important to improve clinical outcomes while controlling healthcare costs' increase.

The patient level factors proved to be of high importance, which is aligned with the literature, the age, insurance type and BMI were good determinants of the length of stay since they were positively correlated with the dependent variable, and they were selected during the feature selection phase. The higher the age and the higher the BMI generally mean higher lengths of stay.

Similarly, the specialty of admission was an important feature in the model, as it was always selected in the feature selection phase. However, since there are 3278 different specialties of admission, the model would need to be trained with more data for each of the specialties to have a higher representation and therefore a more accurate prediction.

Even though the initial dataset had more than 30,000 rows and all the surgery data from the last 3 years, this number is still not very impressive, especially when compared to other studies, for example the study by (Brasel et al., 2007) which comprised over 300.000 medical records.

In the literature, it mentions that the outliers' removal in data analysis should be reconsidered since removing these outliers can lead to inconsistencies and unreliable results in the data since it in these studies there is outliers in length of stay and they should be tested in the analysis. This is aligned with the results in this study, since the model configuration that included outliers had a better result.

There were some limitations in this study, since it was not possible to acquire some of the patient data that the literature mentioned as important. The reason for this was because most of these variables were confidential or they were not available. Some of these variables are the patient data such as the income quintile of the patient, if the patient has co-diseases, if the patient is clinically depressed. In the future, it would be important to include these factors to reach a more accurate prediction on the length of stay.

In the literature it mentioned that the length of stay varies geographically across different hospitals and between countries and regions. In this study the geographic data was not varied enough to affirm this since the data was all from the same hospital and therefore, the data is all from the same country and region. In the future, it would be important to gather data from different hospitals across different

countries or regions to be able to expand this prediction tool across various hospitals. Another limitation in the study was the fact that the data collected was all from a private hospital and according to the literature, the type of facility is significant in predicting the length of stay.

It is important to point the fact that the data included some wrong data, such as the Weight, Height, BMI and Procedure duration. These columns were wrong due to input mistakes, since the data was typed manually and there is no way to correct these values, unless there was some data validation done directly from the data source to possibly identify the error. Another possibility to correct these values would be to get the Weight and Height from the patients' identity cards and calculate the BMI, for example.

Time was the biggest restriction to this project and not all available techniques were hypothesized. It is proposed as future work that this paper serves as a building block for performance prediction researchers in hopes to build upon the study presented and build a new process capable of overcoming the restrictions and present a solution with a more powerful prediction capability. There are a few modifications that can be made to the data collected and the model:

- Acquire more data from more surgeries
- Acquire data from different hospitals from different locations and different types
- Test different parameters on the models to achieve a better score by implementing a grid search on the models
- Different methods should be tested, not only different algorithms but also different sampling techniques, feature engineering technique and variables

Average length of stay in hospitals | Health at a Glance 2019 : OECD Indicators | OECD iLibrary. (n.d.). Retrieved January 29, 2022, from https://www.oecd-ilibrary.org/sites/0d8bb30a-en/index.html?itemId=/content/component/0d8bb30a-en

Azari, A., Janeja, V. P., & Mohseni, A. (2012). Predicting Hospital Length of Stay (PHLOS) : AAA multi-tiered data mining approach. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 17–24. https://doi.org/10.1109/ICDMW.2012.69

Brasel, K. J., Lim, H. J., Nirula, R., & Weigelt, J. A. (2007). Length of Stay: An Appropriate Quality Measure? *Archives of Surgery*, *142*(5), 461–466. https://doi.org/10.1001/archsurg.142.5.461

Carey, K., & Lin, M. Y. (2014). Hospital length of stay and readmission: An early investigation. *Medical Care Research and Review*, *71*(1), 99–111. https://doi.org/10.1177/1077558713504998

Carroll, A., & Dowling, M. (2007). Discharge planning: Communication, education and patient participation. *British Journal of Nursing (Mark Allen Publishing)*, *16*, 882–886. https://doi.org/10.12968/bjon.2007.16.14.24328

Castelli, A., Daidone, S., Jacobs, R., Kasteridis, P., & Street, A. D. (2015). The determinants of costs and length of stay for hip fracture patients. *PLoS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0133545

Clark, D. E., & Ryan, L. M. (2002). Concurrent prediction of hospital mortality and length of stay from risk factors on admission. *Health Services Research*, *37*(3), 631–645. https://doi.org/10.1111/1475-6773.00041

CLARKE, A., & ROSEN, R. (2001). Length of stay: How short should hospital care be? *European Journal of Public Health*, *11*(2), 166–170. https://doi.org/10.1093/eurpub/11.2.166

Distritos/Concelhos - GEE. (n.d.). Retrieved October 25, 2022, from https://www.gee.gov.pt/pt/documentos/publicacoes/estatisticas-regionais/distritos-concelhos

Englert, J., Davis, K. M., & Koch, K. E. (2001). Using Clinical Practice Analysis to Improve Care. *The Joint Commission Journal on Quality Improvement*, *27*(6), 291–301. https://doi.org/https://doi.org/10.1016/S1070-3241(01)27025-4

Freeman, W. J., Weiss, A. J., & Heslin, K. C. (2016). *Overview of U.S. Hospital Stays in 2016: Variation by Geographic Region*. https://www.kff.org/other/state-indicator/admissions-by-

Gonçalves-Bradley, D. C., Lannin, N. A., Clemson, L. M., Cameron, I. D., & Shepperd, S. (2016a). Discharge planning from hospital. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org/10.1002/14651858.CD000313.pub5

Gu, Q., Li, Z., & Han, J. (2011). Generalized fisher score for feature selection. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, 266–273.

Guru, V., Anderson, G. M., Fremes, S. E., O'Connor, G. T., Grover, F. L., & Tu, J. v. (2005). The identification and development of Canadian coronary artery bypass graft surgery quality indicators. *The Journal of Thoracic and Cardiovascular Surgery*, *130*(5), 1257.e1-1257.e11. https://doi.org/https://doi.org/10.1016/j.jtcvs.2005.07.041

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare Informatics Research*, *19*(2), 121–129. https://doi.org/10.4258/HIR.2013.19.2.121

Han, T. S., Murray, P., Robin, J., Wilkinson, P., Fluck, D., & Fry, C. H. (2022). Evaluation of the association of length of stay in hospital and outcomes. *International Journal for Quality in Health Care*, *34*(2), 1–9. https://doi.org/10.1093/INTQHC/MZAB160

Huang, Z., Juarez, J. M., Duan, H., & Li, H. (2013). Length of stay prediction for clinical treatment process using temporal similarity. *Expert Systems with Applications*, *40*(16), 6330–6339. https://doi.org/10.1016/J.ESWA.2013.05.066

Hughes, A. H., Horrocks, D., Leung, C., Richardson, M. B., Sheehy, A. M., & Locke, C. F. S. (2021). The increasing impact of length of stay "outliers" on length of stay at an urban academic hospital. *BMC Health Services Research*, *21*(1), 1–7. https://doi.org/10.1186/S12913-021-06972-6/FIGURES/4

Hunt-O'Connor, C., Moore, Z., Patton, D., Nugent, L., Avsar, P., & O'Connor, T. (2021). The effect of discharge planning on length of stay and readmission rates of older adults in acute hospitals: A systematic review and meta-analysis of systematic reviews. *Journal of Nursing Management*, *29*(8), 2697–2706. https://doi.org/10.1111/jonm.13409

Kagan, S. H., Chalian, A. A., Goldberg, A. N., Rontal, M. L., Weinstein, G. S., Prior, B., Wolf, P. F., & Weber, R. S. (2002). Impact of age on clinical care pathway length of stay after complex head and neck resection. *Head and Neck*, *24*(6), 545–548. https://doi.org/10.1002/hed.10090

Kerper, L., Spies, C., Buspavanich, P., Balzer, F., Salz, A.-L., Tafelski, S., Lau, A., Weiss-Gerlach, E., Neumann, T., Glaesmer, H., Wernecke, K.-D., Brähler, E., & Krampe, H. (2014). Preoperative depression and hospital length of stay in surgical patients. *Minerva Anestesiologica*, *80*.

KHALIQ, A. A., BROYLES, R. W., & ROBERTON, M. (2003). THE USE OF HOSPITAL CARE: DO INSURANCE STATUS, PROSPECTIVE PAYMENT, AND THE UNIT OF PAYMENTS MAKE A DIFFERENCE? *Journal of Health and Human Services Administration*, *25*(4), 471–496. http://www.jstor.org/stable/25790654

Kulinskaya, E., Kornbrot, D., & Gao, H. (2005). Length of stay as a performance indicator: robust statistical methodology. *IMA Journal of Management Mathematics*, *16*(4), 369–381. https://doi.org/10.1093/IMAMAN/DPI015

Lee, A. H., Fung, W. K., & Fu, B. (2003). Analyzing Hospital Length of Stay: Mean or Median Regression? *Medical Care*, *41*(5), 681–686. http://www.jstor.org/stable/3768028

Lingsma, H. F., Bottle, A., Middleton, S., Kievit, J., Steyerberg, E. W., & Marang-Van De Mheen, P. J. (2018). Evaluation of hospital outcomes: The relation between length-of-stay, readmission, and mortality in a large international administrative database. *BMC Health Services Research*, *18*(1), 1–10. https://doi.org/10.1186/S12913-018-2916-1/FIGURES/4

Liu, Y., Phillips, M., & Codde, J. (n.d.). *Factors influencing patients' length of stay*.

MacKenzie, E. J., Morris, J. A., & Edelstein, S. L. (1989). Effect of pre-existing disease on length of hospital stay in trauma patients. *The Journal of Trauma*, *29*(6), 757–764; discussion 764-5. https://doi.org/10.1097/00005373-198906000-00011

Marie, L., & Davis, B. (2010). *Using Data Mining to Analyze Patient Discharge Data for an Urban Hospital. Predictive Mathematical Modeling: Traditional Versus Client-Choice Food Pantries View project Center for Advanced and Transportation Mobility View project*. https://www.researchgate.net/publication/220704956

Meng, Z., Zhao, F., & Liang, M. (2021). SS-MLP: A Novel Spectral-Spatial MLP Architecture for Hyperspectral Image Classification. *Remote Sensing 2021, Vol. 13, Page 4060*, *13*(20), 4060. https://doi.org/10.3390/RS13204060

Negash, S., Anberber, E., Ayele, B., Ashebir, Z., Abate, A., Bitew, S., Derbew, M., Weiser, T. G., Starr, N., & Mammo, T. N. (2022). Operating room efficiency in a low resource setting: a pilot study from a large tertiary referral center in Ethiopia. *Patient Safety in Surgery*, *16*(1). https://doi.org/10.1186/S13037-021-00314-5

Nettleton, D. (2014). Selection of Variables and Factor Derivation. *Commercial Data Mining*, 79–104. https://doi.org/10.1016/B978-0-12-416602-8.00006-6

Ng, S. K., McLachlan, G. J., & Lee, A. H. (2006). An incremental EM-based learning approach for on-line prediction of hospital resource utilization. *Artificial Intelligence in Medicine*, *36*(3), 257–267. https://doi.org/10.1016/J.ARTMED.2005.07.003

Nouaouri, I., Samet, A., & Allaoui, H. (2015). Evidential data mining for length of stay (LOS) prediction problem. *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, 1415–1420. https://doi.org/10.1109/CoASE.2015.7294296

pandas - Python Data Analysis Library. (2022). https://pandas.pydata.org/about/

Parker, B. T., & Marco, C. (2014). Emergency department length of stay: accuracy of patient estimates. *The Western Journal of Emergency Medicine*, *15*(2), 170–175. https://doi.org/10.5811/westjem.2013.9.15816

Popejoy, L. L., Moylan, K., & Galambos, C. (2009). A Review of Discharge Planning Research of Older Adults 1990-2008. *Western Journal of Nursing Research*, *31*(7), 923–947. https://doi.org/10.1177/0193945909334855

Procter, L. D., Davenport, D. L., Bernard, A. C., & Zwischenberger, J. B. (2010). General surgical operative duration is associated with increased risk-adjusted infectious complication rates and length of hospital stay. *Journal of the American College of Surgeons*, *210*(1). https://doi.org/10.1016/J.JAMCOLLSURG.2009.09.034

Project Jupyter | Home. (2022). https://jupyter.org/index.html

scikit-learn. (2022). https://scikit-learn.org/stable/

Sebastian Raschka. (2014). *About Feature Scaling and Normalization*. https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Shaun Turney. (2022). *Pearson Correlation Coefficient (r) | Guide & Examples*. https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Shaw, C. (2003). *How can hospital performance be measured and monitored?* World Health Organization Regional Office for Europe.

Shepperd, S., Lannin, N. A., Clemson, L. M., McCluskey, A., Cameron, I. D., & Barras, S. L. (2013). Discharge planning from hospital to home. In S. Shepperd (Ed.), *Cochrane Database of Systematic Reviews* (Issue 1). John Wiley & Sons, Ltd. https://doi.org/10.1002/14651858.CD000313.pub4

sklearn.impute.KNNImputer — scikit-learn 1.1.3 documentation. (n.d.). Retrieved October 27, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

Stone, K. I., Zwiggelaar, R. I., Jones, P., & mac Parthalá in, N. (2022). A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, *1*(4), e0000017. https://doi.org/10.1371/JOURNAL.PDIG.0000017

Sykes, A. O. (1993). *An Introduction to Regression Analysis*. https://chicagounbound.uchicago.edu/law_and_economics

Toptas, M., Sengul Samanci, N., Akkoc, İ., Yucetas, E., Cebeci, E., Sen, O., Can, M. M., & Ozturk, S. (2018). Factors Affecting the Length of Stay in the Intensive Care Unit: Our Clinical Experience. *BioMed Research International*, *2018*, 9438046. https://doi.org/10.1155/2018/9438046

Victor Popov. (2019). *Dealing with Categorical Data. Categorical data can screw up you ML… | by Victor Popov | machine_learning_eli5 | Medium*. https://medium.com/machine-learning-eli5/dealing-with-categorical-data-f4c8556cbda0

Vijaya, R., Reddy, K., & Ravi Babu, U. (2018). *A Review on Classification Techniques in Machine Learning*.

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

What is the body mass index (BMI)? - NHS. (2019). https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/

WHO. (2003). *How can hospital performance be measured and monitored?* https://apps.who.int/iris/handle/10665/363763

Wiler, J. L., Handel, D. A., Ginde, A. A., Aronsky, D., Genes, N. G., Hackman, J. L., Hilton, J. A., Hwang, U., Kamali, M., Pines, J. M., Powell, E., Sattarian, M., & Fu, R. (2012). Predictors of patient length of stay in 9 emergency departments. *The American Journal of Emergency Medicine*, *30*(9), 1860–1864. https://doi.org/https://doi.org/10.1016/j.ajem.2012.03.028

Wrenn, J., Jones, I., Lanaghan, K., Congdon, C. B., & Aronsky, D. (2005). Estimating Patient's Length of Stay in the Emergency Department with an Artificial Neural Network. *AMIA Annual Symposium Proceedings*, *2005*, 1155. /pmc/articles/PMC1560706/

Xesfingi, S., & Vozikis, A. (2016). Patient satisfaction with the healthcare system: Assessing the impact of socio-economic and healthcare provision factors. *BMC Health Services Research*, *16*(1), 94. https://doi.org/10.1186/s12913-016-1327-4

Yang, C., & Sun, B. (2021). Additive requirement ratio estimation using trend distribution features. *Modeling, Optimization, and Control of Zinc Hydrometallurgical Purification Process*, 63–82. https://doi.org/10.1016/B978-0-12-819592-5.00014-4

Yoon, P., Steiner, I., & Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department. *Canadian Journal of Emergency Medicine*, *5*(3), 155–161. https://doi.org/DOI: 10.1017/S1481803500006539