UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Predicting post-surgical length of stay using machine learning

Patrícia Andreia Alves Falcão

Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil em Engenharia Clínica e Instrumentação Médica

Dissertação orientada por:
Professora Dra. Sara Madeira
Dr. Nuno André da Silva

2020

Acknowledgements

The development of this project was a great experience at an academic, professional and personal level. Firstly, I would like to leave my sincere acknowledgment to my supervisors: Dr. Nuno André da Silva, who has trusted and has led me during this project, and Professor Dr. Sara Madeira, who helped me whenever I needed and was always very approachable.

I must extend my deep thanks and gratitude to the collaborators of *Hospital Beatriz Ângelo* (HBA) - Rita Sousa Machado, Ana Bento, Francisco Mota, Paulo Costa and Jorge Santos – who helped me to clarify any doubts I had about the dataset given to me.

I have to express my warm thanks to Luís Venâncio and Carolina Raimundo, two of my closest friends who were passing through the same experience as me. For all the fellowship and for every time they tried to help me in my project problems.

I cannot but offer my heartfelt thanks to all the people that have helped me with this project directly, but also indirectly. I am grateful to my parents who always supported me in my every decision, without any pressure.

Thank you to other friends who were busy with their own master thesis and family. All the messages and calls were really appreciated during these months. I would not be able to complete this project the way I did without them.

Abstract

In recent years, there has been a steady increase in the number of hospitals adopting Electronic Health Records (EHR) allowing a digitalisation of patient data. In turn, the correct manipulation of these data, using Data Mining (DM) techniques, can lead to achieving solutions both related to patients' health and hospital management. Regarding hospital management problems, one of the most severe issues is related to bed management, which is associated with the Length of Stay (LOS) in the hospital.

In this way, taking advantage of the information taken from the data collected from the patients, whether of a personal or hospital nature, it is possible to solve or mitigate this complication hitherto hardly solvable.

In this follow-up, this dissertation will focus on the case study of *Hospital Beatriz Ângelo* (HBA) and proposes a Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology in order to predict the LOS of patients after surgeries. Random Forests (RF) was the technique considered to perform the classification task and F1-score was the metric selected to evaluate the results. LOS is predicted by models developed in different situations: in the postoperative period and in the preoperative period. Comparing the results between the models developed and the discharge system used in this hospital, it is possible to conclude that there are remarkable results, with an average improvement of 13.87 percentage points for the postoperative model and 12.32 for the preoperative model, in terms of F1-score.

In addition, an analysis and comparison between models that have as input merely patient-related variables and models solely containing procedure or structural-related variables was made, in order to understand the importance of each of these two types of features in the LOS. The results of this approach allowed the recognition of the importance associated with the integration of the two types of features in a Machine Learning (ML) model, adding an average improvement, in terms of F1-score, of 9.68 percentage points in relation to the exclusive use of patient-related variables and 3.83 for procedure-related variables for the post-surgical model. In turn, for the pre-surgical model, the incorporation of both variables brings an improvement of 7.67 percentage points compared to the model that uses only patient features and 5.72 for the model with only procedure-related variables.

The overall results of this work demonstrated that there was an improvement in the ML model in relation to the existing one, highlighting a better forecast of the day of discharge, which allows a better management of the beds.

Keywords: Data Mining, CRISP-DM, Bed Management, Length of Stay, Machine Learning.

Resumo

Ser saudável, em qualquer cultura, é essencialmente a condição mais importante para uma vida longa e feliz e para ela contribui toda a rede hospitalar de um país, quer seja um sistema de saúde nacional ou privado. Análogo a diferentes áreas, também a saúde deve acompanhar a evolução tecnológica para oferecer serviços avançados devido às variedades de demandas sociais. Isso acontece porque o desenvolvimento de tecnologias e metodologias em saúde permite criar novos processos aprimorados e torna os já existentes mais eficientes. A tecnologia na medicina não envolve apenas anestésicos e antibióticos ou técnicas médicas, como ressonância magnética e radioterapia. Na verdade, como os pacientes geram enormes quantidades de informações, não só médicas (como resultados de análises ao sangue), mas também relacionadas com o hospital (nomeadamente o tempo e o tipo de cirurgia), um dos avanços mais importantes dos últimos anos foi a digitalização dessas mesmas informações por meio dos registos de saúde eletrónico. Um dos maiores e mais diretos benefícios conhecidos da digitalização médica é que o atendimento ao paciente é mais fácil e eficiente. Contudo, a grande finalidade da existência destes registos vem após o tratamento e manipulação dos dados com técnicas de ciência dos dados quando, por exemplo, alguns diagnósticos, como as doenças cardíacas, podem ser previstos pelo uso dessas metodologias.

Assim, na posse dos dados em formato digital, diferentes técnicas podem ser aplicadas, conforme o caso, de modo a extrair informações que não seriam visíveis *per si*. Os resultados são tanto melhores quanto mais cógnito todo o processo por trás da coleta de dados, pois aperfeiçoa a seleção e o pré-processamento dos dados. Dentro das técnicas existentes para a previsão a partir dos bancos de dados e, consequentemente, auxiliar uma empresa a tomar as melhores decisões, está a aprendizagem automática. Esta área fornece aos sistemas a capacidade de aprender e melhorar automaticamente com a experiência, sem ser explicitamente programado, o que pode ser extremamente relevante na área da saúde.

Paralelamente à tecnologia, fatores financeiros e de gestão também devem ser considerados, pois também o hospital é uma empresa que deve ser gerida. Assim, além de contribuir para o bem-estar da população, um dos seus objetivos internos é reduzir ao máximo os custos sem prejudicar o normal funcionamento de qualquer atividade desempenhada, otimizando recursos. Neste seguimento, um dos aspetos mais problemáticos da logística hospitalar é a gestão de camas. O seu excesso, ao mesmo tempo que garante maior alocação de pacientes, leva também a um custo hospitalar excessivo. Sob outra perspetiva, um défice pode gerar situações graves para quem precisa. Em suma, a gestão profissional de camas visa uma alta taxa de ocupação, mas uma baixa taxa de cancelamentos, alcançando assim uma alocação ótima. Porém, a sua distribuição ideal é dificultada pela difícil precisão do tempo de internamento de

pacientes hospitalizados. De modo a colmatar esta adversidade, é possível a concretização de um modelo capaz de prever o tempo de estadia com maior rigor através da manipulação de um conjunto de dados composto, neste caso, por informações de pacientes.

Desta forma, esta dissertação tem como finalidade a criação e avaliação, em *Python*, de um modelo preditivo de classificação para o tempo de internamento para pacientes que sejam submetidos a cirurgia, tendo como base de comparação o adotado atualmente pelo hospital em estudo, o HBA. Por forma a alcançar este propósito, recorrendo à metodologia *Cross-Industry Standard Process for Data Mining*, este trabalho dividiu-se em três etapas: o entendimento dos dados e respetiva preparação, a sua modelação e por fim a sua avaliação e comparação com o modelo do HBA. Este estudo visa suprir as lacunas de outros estudos que não consideram simultaneamente características gerais dos pacientes e hospitalares, como a de data e hora da cirurgia. Além disso, existe ainda uma carência na literatura de estudos que utilizem aprendizagem automática no que diz respeito aos pacientes de origem exclusivamente cirúrgica.

Para o início da primeira fase, foi utilizado um *dataset* referente a 20 736 pacientes que estiveram hospitalizados no HBA entre o ano de 2017 e 2018, estando ainda asseguradas 135 características dos mesmos, quer do foro do paciente, quer do foro hospitalar. Após a receção dos dados, é necessária a sua compreensão do ponto de vista médico e comportamental, uma vez que o modo como foi preenchido está sujeito a erros de cariz humano. Estes erros podem ir desde a troca de informações no momento do preenchimento, assim como à existência de características que representam a mesma ideia, estando uma mais atualizada relativamente a outra. Assim sendo, é importante um primeiro contacto com os responsáveis pelo preenchimento do conjunto de dados por forma a garantir a sua leitura plausível e respetivo entendimento das informações fornecidas por cada uma das características. A partir desta análise é possível uma organização primordial dos dados. Ainda nesta etapa é imperativo verificar a possibilidade de formação de novas variáveis a partir de outras já existentes de forma a enriquecer o *dataset*.

O conhecimento da distribuição das variáveis torna-se essencial para a total compreensão dos dados, uma vez que permite a averiguação da repartição de categorias de cada uma das características. Nesta fase é assim necessário o conhecimento, limpeza e preparação dos dados para que estes possam ser seguidamente modelados.

A segunda etapa refere-se à modelação dos dados a um dos algoritmos de aprendizagem automática, neste caso, das *Random Forests*. Uma vez que a finalidade se prende em dois modelos diferentes – pré e pós-cirúrgico – é indispensável ter em consideração as variáveis consideradas em cada um dos modelos, tendo pleno conhecimento do momento em que cada uma delas é referenciada pela primeira vez. Tratando-se de um algoritmo de classificação com 135 *features*, é ainda imprescindível uma seleção de variáveis ideal. Esta seleção de variáveis permite um aperfeiçoamento da acuidade e uma redução do *overfitting*, face a um modelo que utilize todas as variáveis. Para além disto, o facto de haver um menor número de atributos considerados, também levará a que o tempo de treino seja menor.

Por fim, a última fase diz respeito à avaliação dos resultados. Para ambos os modelos, pré e pós cirúrgico, a métrica utilizada foi o *F1-score*, por se tratar de dados não equilibrados. Desta forma, com a elaboração destes modelos foi possível verificar-se uma melhoria notória, dependendo da especialidade, face ao modelo atualmente em vigência de, em média, 13,87 pontos percentuais para o modelo pós-operatório e 12,32 para o modelo pré-operatório.

Constrangimentos como o número restrito de pacientes considerados após a preparação do conjunto de dados para a modelação e erros comportamentais no preenchimento do *dataset* poderão ter limitado os resultados desta dissertação. No entanto, mesmo podendo beneficiar de algumas melhorias, a finalidade para o qual este projeto foi proposto, foi cumprida. Neste caso em específico, foi possível denotar melhorias face ao modelo atualmente empregue no hospital, comprovando assim o potencial de modelos que tiram proveito dos benefícios da aprendizagem automática.

Em adição ao objetivo central deste trabalho foi ainda feita uma análise e comparação entre modelos que contivessem apenas variáveis do foro do paciente e modelos que incluíam unicamente variáveis de procedimento ou estruturais. A elaboração destes modelos e posterior análise visou a comparação da influência destes dois tipos de variáveis num modelo hospitalar, com o intuito de enaltecer a importância do correto preenchimento destes atributos por parte dos profissionais. Os resultados desta abordagem permitiram reconhecer a relevância associada à integração dos dois tipos de variáveis num modelo de *Random Forests*, adicionando uma melhoria média de 9,68 pontos percentuais em relação ao uso exclusivo de variáveis relacionadas ao paciente e 3,83 para variáveis relacionadas ao procedimento para o modelo pós-cirúrgico. Por sua vez, para o modelo pré-cirúrgico, a incorporação de ambas as variáveis traz uma melhoria de 7,67 pontos percentuais em relação ao modelo que utiliza apenas características do paciente e 5,72 para o modelo apenas com variáveis relacionadas ao procedimento.

Com esta dissertação, demonstra-se que a partir da aplicação de técnicas de *Random Forests* aos registos de saúde eletrónico do hospital em estudo é possível criar um modelo preditivo para o tempo de estadia. Isto possibilita no futuro um processo de gestão de camas otimizado, permitindo assim a diminuição dos custos hospitalares.

Palavras-Chave: *Data Mining*, CRISP-DM, Gestão de Camas, Tempo de Internamento, Aprendizagem Automática.

List of Figures

List of Tables

List of Acronyms

AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
AUC	Area under Curve
BDT	Boosted Decision Tree
CART	Classification and Regression Tree
CAS	Coronary Atherosclerosis
CFS	Correlation-based Feature Selection
CM	Confusion Matrix
CRISP-DM	Cross-Industry Standard Process for Data Mining
CV	Cross-Validation
DM	Data Mining
DT	Decision Tree
EHR	Electronic Health Records
HBA	*Hospital Beatriz Ângelo*
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10-CM	International Classification of Diseases, Tenth Revision, Clinical Modification
ICD-10-PCS	International Classification of Diseases, Tenth Revision, Procedure Coding System
ICU	Intensive Care Unit
FN	False Negative
FP	False Positive
LNR	Linear Regression

LOS	Length of Stay
LR	Logistic Regression
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Splines
MI	Model Interpretability
ML	Machine Learning
MSE	Mean Square Error
NB	Naïve Bayesian
NN	Neural Networks
RAE	Relative Absolute Error
RF	Random Forests
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
SDC	Surgical Day Care
SIGLIC	*Sistema Informático de Gestão da Lista de Inscritos para Cirurgia*
SMOTE	Synthetic minority oversampling technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UO	Urgent Operation
USA	United States of America

Contents

Chapter 1

Introduction

This chapter presents the motivation and the respective context of this dissertation. The objectives outlined for this project are also described, as well as a brief explanation of how the division of this document is structured.

1.1 Motivation and Context

Currently, more and more countries are promoting the digitisation of their hospitals, namely in terms of EHR. EHR is a digital version of patient records that includes patient information such as personal, medical and procedure related [1]. In Portugal the adoption of EHR is widely spread and is beginning to verify the advantages of this medical innovation [2]. According to some studies, the implementation of EHR allows the efficient access to a large amount of information [3], enables research costs reduction, accelerates new medical investigation [4] and also, due to its efficiency, it can also be helpful in preventing medical errors [5]. As result of a study conducted in 2019, the quality of health care services, both in expectation and perception, in EHR-adopted hospitals is higher than those using paper-based records [6]. This demonstrates that the adoption of high quality EHR has a substantial impact on improving the quality of healthcare in hospitals [6]. In addition, this type of medical records can also be an excellent aid while managing a hospital.

One of the most problematic aspects in hospital logistics is the management of hospital beds. This is because an excess of beds, while ensuring greater patient allocation, leads to an unoptimised resources allocation which must be avoided in a context with limited resources. Conversely, a deficit can cause serious situations for those who may need them. Therefore, professional bed management aims a high occupancy rate, thus achieving optimal bed allocation. Thus, the management of hospital beds available for hospitalisation is a key point in improving the hospital demands, rationalising resources and avoiding complications related to poor accommodation of patients, namely overtime in Emergency Unit [7]. Thus, there is a need for an optimised control of the number of beds so that it does not harm both patients and hospital units. However, optimal bed allocation is hampered by the difficult accuracy of the LOS of hospitalised patients [8]. As a matter of fact, according to [9], hospitals that can control LOS decrease admission cost and patients' daily costs.

It is stated in many articles that LOS is an important indicator of the efficiency of hospital management and can be used for various purposes, such as the hospitalisation costs [10], since, for example, the lack of services and facilities can increase the LOS. Just like claimed by [11], LOS is one of the measures employed worldwide to measure hospital resource consumption and performance monitoring.

The LOS is influenced by multiple factors, from characteristics of the patients to operational routines of the healthcare provider. It is also influenced by the kind of hospital episode, such as urgency or planned surgeries. Consequently, ML approaches, capable to analyse a vast amount of multidimensional data, have the potential to improve LOS estimation.

According to the literature review (Chapter 3) of this report, studies have been conducted to predict LOS, however some of the results and conditions under which these models were run were not the best. It was noticed that there are a lot of models that, for instance, only consider patient or procedure-related variables neither date/time factors. Furthermore, it was noticed that the literature lacked studies utilising ML with respect to surgeries care scheme.

In the case study of this dissertation, LOS in HBA is predicted by the mean of the LOS of previous patients who had the same diagnosis. These patients, used as a base control, were admitted in 2017. Professionals at this hospital highlighted that there is a problem in the management of beds due to the fact that the accuracy in predicting hospitalisation times is challenging.

1.2 Objectives

The general objective of this dissertation is to develop a predictive model of hospitalisation days for patients undergoing surgeries at HBA and subsequent comparison with the prevailing model, highlighting the capabilities of artificial intelligence. By sectioning this general objective in stages throughout this project, we present 7 phases:

1. Understanding the current hospital discharge system under study;
2. Analysis of the literature regarding the forecast of discharge in hospitals, in order to know how to fill the gaps in existing solutions;
3. Comprehension of the entire dataset through meetings with hospital administrators, leading to understand the local context of the data;
4. Computerised analysis of the dataset to understand the distribution of the features and respective categories;
5. After the previous two steps, it is possible to analyse and perform the elimination, transformation and creation of variables;
6. Implementation of the classification algorithm;
7. Elaboration of a model containing merely patient-related variables and another solely containing procedure or structural-related variables;
8. Evaluation of results and comparison with the current model.

1.3 Document Structure

In addition to the present introductory chapter, this document is structured in six chapters as follows:

- **Chapter 2** (Theoretical Framework) introduces the basic concepts and resources that support DM techniques, namely, CRISP-DM, RF algorithm and evaluation metrics. We also include the explanation of the different steps taken in Chapter 4 without yet realising the problem itself.
- **Chapter 3** (Literature Review) presents the related work to LOS forecast in different environments and using different groups of variables.
- **Chapter 4** (Data and Methods) includes the description of the dataset used as input, all data preparation processes and the implementation of the algorithm.
- **Chapter 5** (Results) presents the results of this implementation, as well as, the outputs of the current model and the respective comparison. It also contains the variables selected in each of the models and the results for the models that only contain patient-related or procedure-related features.
- **Chapter 6** (Discussion) includes the discussion of the results and some methods, the limitations of the work and possible future improvements.
- **Chapter 7** (Conclusion) discusses the main conclusions of this work and if the final objective was achieved.

Chapter 2

Theoretical Framework

This chapter presents all the concepts and theories on which this dissertation is supported. The process through which this work went through is explained in detail, with no data and results demonstrations yet. From the analysis of the chosen methodology, CRISP-DM, to the knowledge of the metrics used to evaluate the performance of the model, everything is detailed.

2.1 Cross-Industry Standard Process for Data Mining

The aim of this work consists in the development of a ML model, through a process of DM. DM is a subdomain of artificial intelligence, being a complex process that involves several tasks and methods aimed at data exploration. Hence, it becomes crucial to select a standardised methodology at the beginning, so that the whole process is organised and structured. [12]. Currently, there are some of these well-defined and disseminated methodologies, however, according to a study published in 2007 [13], CRISP-DM was the main methodology chosen by professionals.

CRISP-DM stands for **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining. It is a framework whose purpose is to transform a business problem into knowledge and management information, splitting into well-defined mining steps so that the objective is successfully accomplished. One of its hallmarks is the focus not exclusively on technology, but also on the user's requirements [14]. This methodology is composed of 6 stages, illustrated in figure 2.1. The arrows inside the circle represent the most important dependencies between phases, however, the sequence is flexible. The result of each stage determines the next phase to be performed, with the need to sometimes return to the first steps. In turn, the outer circle indicates the cyclical nature of this methodology. When the solution to a problem is found, this does not mean that it is the end of this cycle, but rather the continuation of the discovery of a solution that can be even better and fix the flaws of the previously solutions developed [15]. Following this reasoning, the phases that make up this methodology are as follows (Figure 2.1):

- **Business Understanding**: This first stage is crucial for the rest of the process to be implemented correctly. It is necessary to understand the problem: to know the requirements and the results that the user expects to achieve and be aware of the limitations and conditions also declared by it. After this, it is essential to convert this knowledge and awareness into a DM problem and outline a plan to accomplish the objective successfully [15, 14].

- **Data Understanding**: This stage firstly refers to the collection of data and then to its familiarisation. Data should be collected from a trusted entity, so that its content can be objectively analysed. A lot of times, this is the phase that takes the longest duration, since for certain issues, data collection can be time-consuming to gather as much data as possible. In addition, sometimes data deal directly with people's identities, so it may be necessary to go through ethics committees until they can be worked on. The data collected may be under supervised or unsupervised learning [16]. In supervised learning, the data used is already labelled, that is, the data used must already be tagged with the correct answer. Conversely, in unsupervised learning, the data is unlabelled. That means that the algorithm does not have the correct answer beforehand, but from the characteristics of the unlabelled data learns to recognise it. After data collection, it becomes important to understand them in order to identify existing problems in the data, have a first perception of it or even form possible preliminary hypotheses for solving a problem. This is the stage where it is essential to check the description and perform data exploration [14].

- **Data Preparation**: The preparation of the data includes all the necessary activities in order to prepare the final dataset for its modeling. After knowledge of the data, among other tasks, data may have to be cleaned, as well as new attributes can be created and the existing ones transformed. In addition, before moving on to modelling, there may also be a need to select only some of the features, in a process called feature selection [17]. The tasks that can be performed in data preparation will be described in detail in section 2.2.

- **Modelling**: At this point, after processing the data, the next step is the choice of the model and its implementation. Which model to choose will depend on the type of the problem. In the case of supervised learning, there are essentially two different techniques – Classification and Regression. In the first one, the output is categorised into a distinct number of classes, while regression models predict a continuous value. Also, it is in this stage that a test/train method is selected. Finally, the algorithm is implemented. This phase will be detailed in sections 2.3 and 2.4.

- **Evaluation**: After building the models in the previous step, they are then evaluated with the appropriate metrics in order to understand the quality of the models. It is also at this stage that it is important to review the previous step in order to verify if any mistakes were made during the modelling. After analysing the results, it is then possible to answer the question initially presented and assess if the model built is suitable for the problem [14]. A detailed analysis of these steps will be presented in section 2.5.

- **Deployment**: When the model developed responds to the needs for which it was proposed and when other obstacles, such as costs and bureaucracy, are overcome, real application is possible. This stage is usually carried out by the user and not the responsible for the five previous steps.

2.2 Data Preparation

The data collected comes mostly in a raw form that it is not useful. As the name implies, this is the phase in which the dataset is prepared so that it can be inserted into the algorithm in the most appropriate form.

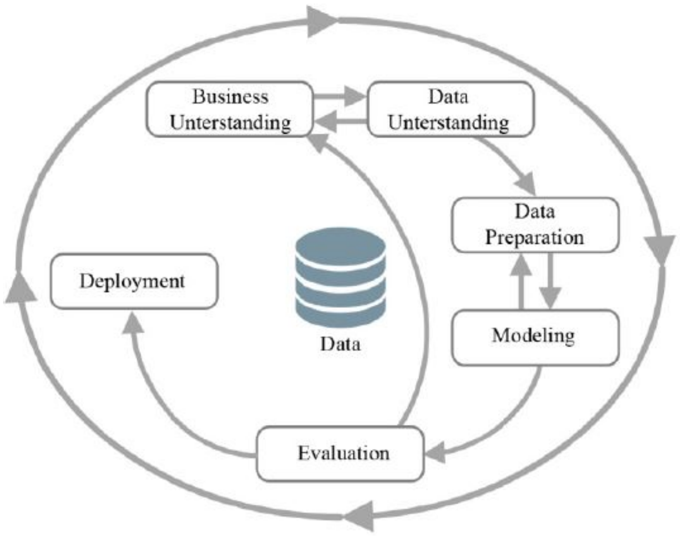

Figure 2.1: Representation of the CRISP-DM Methodology. Figure obtained from [15].

Before starting the set of tasks that prepare the data, variables must be deeply understood to avoid future errors. The size of the dataset, the type of variables and categories are some of the points that should be studied, so that if the data is inconsistent and disproportionate, it can be treated. After acquaintance with the data, it is the time where its preparation begins. It can be based on three different tasks.

2.2.1 Data Cleaning

This phase is known for eliminating inconsistencies in the dataset. This removal occurs since some data is not useful to the dataset and may even lead to inaccurate results. This step depends a lot on the type of data the user is dealing with and also the objective of the model.

Multiple points must be evaluated during this task, depending on the problem, namely:

- **Missing values**: Sometimes features/indexes contains too many missing values. One way to get rid of the problem is ignoring them. However, depending on the problem, there are other approaches that can be applied:
  - Manual filling: Although it is not mathematically complex, it is time consuming and not very effective when dealing with a large dataset containing many missing values [18];
  - Constant: It is possible to replace the missing values by constants, such as, “unknown” or “0”. However, when this term is repeated for many times it can be misunderstood by the model that may associate the constant to an interesting and important concept [18];
  - Measure of a central tendency: A missing value may be replaced by measures of central tendency, such as, mean – in case of normal distributions – or median – for skewed distributions [18];
  - Mean/Median attribute of samples belonging to the same class: The difference between this technique and the measure of a central tendency, is that in the last mentioned the average or median is made based on the samples of the class belonging to the sample with the missing value and not on all the samples [18];

- The most probable value method: This one is a prominent strategy as it takes advantage of the most information from the present data to predict missing values. This method uses different techniques such as K-Nearest Neighbors. In this specific technique, K-neighbors are chosen based on some distance measure and their average or most frequent value is used as an imputation estimate. The data analyst suggests the number of the nearest neighbors and the distance metric [19].

- **Outliers**: Outliers are identified and possibly removed or they can be replaced by the values used for missing values or with minimum and maximum percentiles, for example;

- **Duplicates**: Dropping duplicates and erroneous values is important since these data do not provide useful information.

2.2.2 Data Transforming

Sometimes the way data are presented must be modified so that they can be used and implemented in a model for better results. There are many techniques that vary from approach to approach, which includes:

- **Categorical Conversion**: Categorical features must be machine-readable. However, sometimes, these type of features have too many groups. In this case, categories with aspects in common can be converged into a single, more general category. Thus, there are two common ways to transform this type of features. When the feature is ordinal, one option is label encoding that codifies the various levels of the feature into numeric values. Nonetheless, if this attribute is not ordinal, the algorithm will misunderstand this data using label encoding. In order to overcome this problem, there is a one-hot encoding method, in which binary variables are created for each attribute in a previously existing single category. In turn, the feature that previously contained all of these categories is eliminated. Nonetheless, the number of attributes should not be very large since this leads to higher memory use and increase of dimensionality in the model.

- **Scaling**: When the algorithm is a regression or Euclidean distances related, data must be transformed since these algorithms are very sensitive to variations. In the absence of this transformation, the scale on which the variables are measured will ultimately have a negative effect on the final model due to its possible large range of values and since the algorithm misunderstands the true weight of each continuous feature. One technique for standardisation is applying the Z-score, given by:

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

  where:

  - x = Observed Feature
  - μ = Mean
  - σ = Standard Deviation

The resulting scaled variable has a mean of 0 and unit variance. In turn, if the original feature has a normal distribution, the scaled variable, achieved using equation 2.1, does too [20]. However, one limitation of variance scaling is not having a bounding range.

- **New Features**: Sometimes some features do not have usable content in their original state as they do not provide readable information. As a result, they can sometimes be converted into another type of feature or their content can be transformed into more than one new feature. For example, in dates related features, it may only make sense to know the day and not the complete date (Feature Extraction). Conversely, occasionally the relation between two or more features (resulting from sum or fraction, for instance) can result in a very useful attribute.

2.2.3 Feature Selection

Feature Selection represents the process of selecting a subset of the most relevant attributes in the dataset. In a nutshell, feature selection allows the separation between the most and least important features of the dataset so that a limited number of them are included in the model. This selection helps training the model faster, since the number of features selected by feature selection is smaller than originally improves model performance because it only includes the relevant attributes for prediction. It can also contribute to reduce the overfitting since the noise that comes from irrelevant features is removed after excluding them from the model. There are 3 main different methods for feature selection (Figure 2.2):

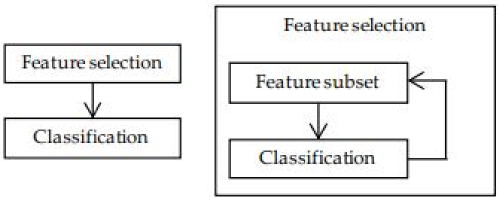

Figure 2.2: Overview of the filter (left scheme), wrapper (middle scheme) and embedded methods (right scheme). Figure obtained from [21].

2.2.3.1 Filter

This technique ranks each attribute based on univariate or multivariate metric by selecting the highest-ranking ones. It only takes into account the intrinsic properties of the feature. Filter techniques are computationally simple, fast (univariate is faster than multivariate) and independent of the algorithm, so it must be performed only once. However, it ignores features dependencies, since each feature is evaluated individually, which may lead to worse performance of models. Nonetheless, multivariate filter techniques already incorporate features dependencies in some level [22, 23]. Some examples of filter methods are:

- χ^2: It tests the independence of two variables, in this case, the predictor and the target. So, the higher the χ^2 value, the higher the dependence is between the independent and dependent variables.

$$\chi_c{}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{2.2}$$

where:

- O = Observed Values
- E = Expected Values
- c = Degrees of Freedom

$$\text{Degrees of Freedom for the contingency table} = (Columns - 1)\,(Rows - 1) \tag{2.3}$$

In order to perfom the χ^2 test, the hypotheses must be defined:

$$\begin{aligned} H_0 :&\quad \text{Two variables are independent.} \\ H_1 :&\quad \text{Two variables are not independent.} \end{aligned}$$

Choosing the desired confidence interval (95%, for example), we will then check whether the χ^2 value (calculated using equation 2.2) for these features is in the acceptance or rejection region. The critical χ^2 for the α in question (in the case of 95% of confidence interval, α = 5%) is calculated using the χ^2 table, taking also into account the degrees of freedom (equation 2.3) of the contingency table (frequency distribution table). The null hypothesis is not rejected when the χ^2 value is smaller than the critical χ^2 value. In addition to this, the $p-value$ is another measure of significance which the greater it is (the larger α), the greater the evidence that the null hypothesis must not be rejected. The $p-value$ represents the probability of getting the actual, or more extreme, results when the null hypothesis is assumed to be true.

- **Analysis of Variance (ANOVA)**: ANOVA may be considered as the extension of the t-test, a statistical test that evaluates whether the means of two populations greatly differ from one another. In turn, ANOVA stands for Analysis of Variance and is applied when the comparison to be made is between more than two populations. So by using a T-test it is possible to verify with some margin if a single variable is statistically significant, while a F-test verifies, with the same margin, if a group of variables are jointly significant. ANOVA assumes that the populations have the same variance, are normally distributed and each sample is independent from each other. The main goal of ANOVA is to compare the means from more than two groups:

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}} = \frac{MS_{Between}}{MS_{Within}} = \frac{\frac{SS_{Between}}{df_{Between}}}{\frac{SS_{Within}}{df_{Within}}} \tag{2.4}$$

where:

- MS = Mean Squares
- SS = Sum of Squares
- df = Degrees of Freedom

$$SS_{Between} = \sum \frac{(\sum x^2)}{n} - \frac{(\sum\sum x)^2}{n_t} \tag{2.5}$$

$$df_{Between} = k - 1 \tag{2.6}$$

$$SS_{Within} = \sum \sum (x^2) - \sum \frac{(\sum x)^2}{n} \tag{2.7}$$

$$df_{Within} = n_t - k \tag{2.8}$$

where:

- n = Sample Size
- n_t = Total Sample Size across all groups
- k = Number of Groups

And, similarly to χ^2, it is important to state the hypotheses:

$$\begin{aligned} H_0 &: \quad \mu_0 = \mu_1 = ... = \mu_k \\ H_1 &: \quad \mu_i \neq \mu_j, \text{ some } i \neq j \end{aligned}$$

When using the F-test, the outputs will be the F-value (calculated from data, using equation 2.4) and F-critical value or F statistic (from the F-Distribution table). In general terms, whether an F-value is larger than the F statistic, the null hypothesis can be rejected. Nonetheless, the F-statistic must be used in combination with the $p - value$ because if the overall results are significant, it does not explicitly mean that all the variables are. So, if the $p - value$ is less than the α value, the null hypothesis can be rejected and the $p - value$ of each feature should be analysed in order to find out which of them are statistically significant.

- **Mutual Information**: Mutual Information is a dependence measure between two random variables (X,Y). This method determines the ability of the independent feature to predict the target variable. It evaluates the gain of each variable in the context of the target variable. Mutual information for discrete distributions is given by equation 2.9 [24]:

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(X,Y)(x,y) \log \frac{p(X,Y)(x,y)}{p_X(x)\, p_Y(y)} \tag{2.9}$$

where:

- $p(X,Y)$ = Joint probability mass function of X and Y
- p_X , p_Y = Marginal probability mass functions of X and Y, respectively

The equation above demonstrates that the mutual information method determines the similarity between the joint distribution and the products of the factored marginal distributions. In this context, if variables X and Y are independent, then $p(x,y) = p(x)p(y)$, so the mutual information would be 0.

2.2.3.2 Wrapper

This type of feature selection selects the best subset of input features to predict the target variable. It searches for the attributes that provide the best model by evaluating a specific ML algorithm. This method uses the result from the previous model developed to add or remove a new feature. This technique allows the recognition of the importance of features when combined, even if some of these variables are useless when analysed individually. Some of the wrapper methods are:

- **Permutation Importance**: Permutation feature importance measures the increase in the prediction performance of the model after feature values are shuffled. When a feature is highly important, the permutation of its values will cause a drastic change to the results of the prediction, whereas the permutation of the values of a less important feature will not damage the outcome. This happens since in the first case, the model relied on that feature for predicting, while in the second one, the model ignored it [25, 26].

- **Recursive Feature Elimination (RFE)**: It aims to find the best features for the optimal performance. It creates various models, saving the best and the worst performing feature at each iteration. With the remaining features, it builds the follow-up model, until there are no more features. Based on the order they are eliminated, the features are ranked [22, 27].

2.2.3.3 Embedded

These methods perform feature selection during the model training process. However they are not as powerful as the wrapper methods, they are much cheaper and select features specific to the model, which is a major advantage compared to filtering methods [20]. There are some algorithms that perform this kind of feature selection, however RF will be the only one addressed in the scope of this work. It will also be described in more detail in the section 2.3.

- **RF**: RF consists of a construction of decision trees. In every tree, at each node there is a condition on a single feature, splitting the dataset into two sets. For classification, Gini impurity is one of the measures widely used to choose the locally optimal condition [28]. Gini impurity is the probability of incorrectly classifying a randomly chosen element and is given by the following expression [22]:

$$G = \sum_{i=1}^{C} p(i) \times (1 - p(i)) \tag{2.10}$$

  Where:

  - C = Total Classes
  - $p(i)$= Probability of picking a datapoint of class i

  Thus, a Gini Impurity of 0 is the lowest and consequently represents the classification purity. Conversely, 1 indicates maximal inequality among values. When a RF is being trained, it is viable to compute for every tree how much each feature decreases the impurity [29]. The closer the Gini Index is to 0, the more important the feature is. Since RF is an association of trees, the impurity decreases from each feature can ben averaged across the combination of trees, resulting in the final importance for the predictor. By following this line of reasoning, in general, the most important features are selected at the top of the trees, while the less important are at the end nodes [22].

2.3 Model Selection

After processing the data, the next step is based on the choice of the model. In the case of the problem under study, we will use the supervised learning algorithm. In particular, classification, where the output is categorised into a distinct number of classes [16]. In this section, we will describe two types of algorithms. In the present work, only RF was used:

- **RF**: RF algorithm is one of the most popular supervised ML algorithm that is capable of performing both regression and classification tasks. As the name implies and as already mentioned, RF is a combination of decision trees where, in order to classify a new object based on features, each tree gives a classification and vote for each class. As result, in classification, the forest selects the class with the most votes over all trees and in regression the forest takes the average of the outputs by the different trees [30]. The low correlation between the trees is a great advantage in this kind of algorithm. The reason for this is that trees protect each other from errors, by allowing each tree to randomly sample from the dataset with replacement, in a process called *Bootstrap Aggregation* or *Bagging*. In this process, each tree has a training set of size N, however, instead of the original data, a random sample of size N with replacement is taken. Conversely, feature randomness is also crucial. While in a decision tree algorithm every feature is considered and only the one that causes the better separation is chosen when the decision to split a node comes, in a RF, each tree only has access to a random subset of features. So, having M input variables, m of these (where $m < M$) are selected at random out of the M at each node. The best feature of the m is selected to split the node. These two characteristics when combined result in trees with more variation and lower correlation between them, preventing each other from their individual errors [30]. Nevertheless, moving downwards, the tree needs a splitting measure in order to calculate the level of impurity and uncertainty so that the best feature of the subset is chosen. Information Gain and, as mentioned before, Gini Index are some measures construct the trees. In Information Gain, in every node there is a reduction in entropy, which is called information gain. Then, the greater the decrease in entropy, the greater the information gain value. This way, the construction of a decision tree is based on finding the attribute that returns the highest information gain [22]. So first, the entropy of the target feature, called the entropy before the split, should be calculated following the expression:

$$E = -\sum_{i}^{C} p_i * \log_2(p_i) \tag{2.11}$$

Where:

  - C = Total Classes
  - $p\,(i)$ = Probability of getting a datapoint of class i

Then the entropy of each node based on a specific attribute should be calculated. The total entropy for the split is the result of the sum of entropy of every branch, resulting in the entropy after split, according to equation 2.12:

$$E(T, X) = \sum_{c \in X} P(c)\, E(c) \tag{2.12}$$

Where:

- $E(T, X)$ = Weighted average entropy of the split
- $P(c)$ = Probability of getting a sample from branch c
- $E(c)$ = Entropy of branch c

Finally, the attribute with the largest information gain value is selected as the decision node, represented by the following equation. This process is repeated on every branch, since a branch with no null entropy needs further splitting [22].

$$gain(split) = Entropy\,(\text{prior to split}) - Entropy\,(\text{after split}) \tag{2.13}$$

- **Simple and Multiple Linear Regression**: Linear Regression (LNR) establishes the linear relationship between the predictor and the target, however, in simple LNR there is only one independent variable. Equation 2.14 represents the simple LNR.

$$y = \beta_0 + \beta_1 x \tag{2.14}$$

Where:

- β_0, β_1 = Coefficients (abscissa and slope)
- y = Target Variable
- x = Predictor Variable

After the values of the coefficients have been estimated, it is important to know how relevant they are to predict the response. One way to perform this task is by using the $p - value$, where the null hypothesis supports that there is no correlation between the independent and the dependent variables, while the alternative hypothesis supports the existence of it. In general, it is considered that when the $p - value$ for each coefficient is less than 0.05, then the null hypothesis can be rejected. Finally, in order to evaluate the performance of the model it is necessary to apply metrics such as R^2, addressed in detailed in section 2.5. Conversely, multiple LNR uses more than two independent variables to predict the target variable by adapting the previous equations into equation 2.15:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \tag{2.15}$$

It is important to perform a hypothesis test so it is possible to access the relevancy of a feature and comprehend if at least one of the features is useful in predicting the output, . The two hypotheses are demonstrated below:

$$\begin{aligned} H_0 &: \quad \beta_0 = \beta_1 = \beta_2 = \beta_k = 0 \\ H_1 &: \quad \text{At least one } \beta_i \neq 0 \end{aligned}$$

The F-statistical test must be performed to accept or reject the null hypotheses. If the $p - value$ for the F-test of overall significance test is less than the significance level chosen, the null-hypothesis can be rejected, concluding that at least one $\beta_i \neq 0$ [31]. The final step is about the evaluation of the performance of the model that will be explained in detail in section 2.5.

2.3.1 Model Interpretability

According to [32], interpretability is the ability to explain or to present in simple terms to a human. Model Interpretability (MI) becomes relevant when there are implications involved in a model's prediction that affects the real world, in particular, the management of a hospital. Moreover, when a model achieves a good performance, that does not always mean that it is doing in the right way, so it becomes important to analyse if the model is trustable and how it is making its predictions. One of the principals is that if the complexity of a model is increased, it will get harder to interpret it.

Generally, linear models and tree-based models are easy to interpret but also to underfit or overfit. In general, the advantage of these models being more easily interpretable as they are simpler, brings the disadvantage of not being able to adapt frequently to complex datasets leading to an inadequate fit. In particular, linear models can be easily underfitted when subjected to non-linear data and overfitted when there are many features compared to the number of observations in the training set, since a single sample is used to estimate the coefficients for all of the terms in the model. In turn, tree-based models can be overfitted especially if the tree is deep, leading to a greater specificity. Whereas the shallower the depth of the tree is, the more chances of a biased tree.

It is crucial to know what features drive predictions, as well as the features that are not effective. By using suitable methods for this purpose, it becomes possible to explain what and how features are important as well as the way they interact with each other. Using the citation in [25], "A feature's importance is the increase in the model's prediction error after we permuted the feature's values (breaks the relationship between the feature and the outcome)". Therefore, a feature is important if shuffling its values greatly increases the model error, meaning that the model relies on that feature for the prediction. Otherwise, that feature may be unimportant.

2.4 Model Validation

Data cleaning and preparation are essential to later evaluate the model performance by using strategies of training and testing the dataset. There are some methods that may be applied, namely the common approaches: Hold-out and Cross-Validation (CV).

- **Hold-out**: Hold-out (Figure 2.3) aims to use data for testing that was not used for training nor validation. Thus, the dataset is split into three subsets: Training set – contains data that when trained build predictive models; Validation set – subset of data used do assess the performance of the model obtained from the training phase. This phase is used for fine-tuning the model's parameters and select the best model; Test set – subset used to assess the probable performance of a model. However, overfitting may occur if a model fits a training set better than the validation and test sets.

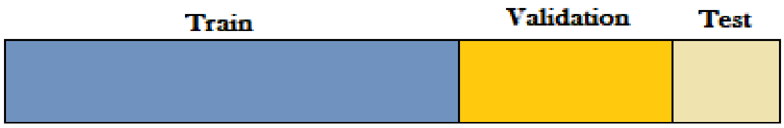

Figure 2.3: Schematic of a split of a dataset using the hold-out method.

- Cross-Validation: CV (Figure 2.4) is a technique that splits the original dataset into a training set to train the model and an independent set to validate the analysis. The most common technique is the k-fold CV, where the original dataset is divided into k equal subsets (folds). One of the folds is used for testing, while the remaining are used for training. This is repeated k times, so that the same fold is not used more than once as the testing set. The estimation of the defined metric is averaged over all k trials, getting the global result of the model [33].

  Nevertheless, in CV, the results may be erroneous due to the random partitioning of the data into k subsets. This happens because, on certain occasions, these subsets may not include many or even any samples of the minority class, more noticeable if the dataset is either imbalanced or small. In such cases, there is another type of CV that ensures the fairness of instances of each class in each subset, involving bootstrapping [34] - Stratified CV. This method allows the splitting of data into folds in such a way the number of observations of each class is preserved throughout all folds.

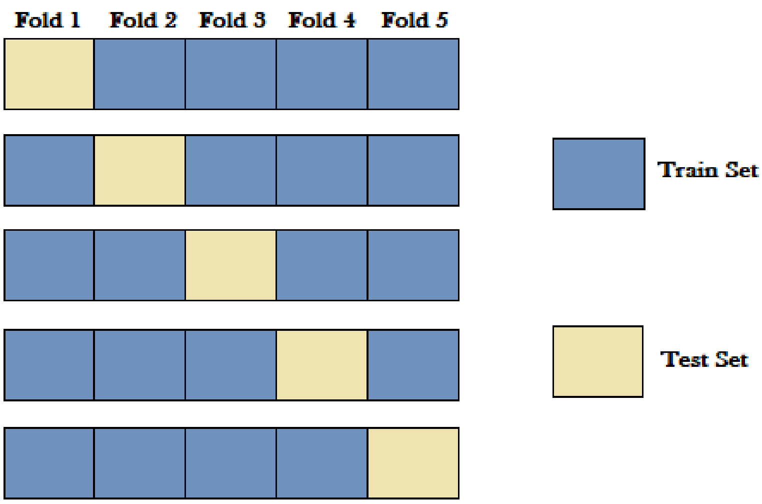

Figure 2.4: Schematic of a 5-fold CV.

2.4.1 Imbalanced Data

In many situations in real life, the datasets can be extremely imbalanced and algorithms will be biased resulting in a poor performance when classifying the minority class. The class imbalance may either occur due to absolute rarity, meaning that the absolute number of samples associated with the minority class is small, or relative rarity, when the minority class is smaller than other classes but not overall [35].

Sampling is one of the most common and simplest way to overcome this constraint. Sampling is a process where the training set is modified in such a way as to balance the classes distribution. It can be split into two different types: undersampling and oversampling. Undersampling focuses on reducing samples from the majority class, while oversampling is concerned with the increase in cases in the minority class. In addition, there are some cases where both undersampling and oversampling must be applied, leading to an ensemble method, so better results can be achieved [36].

Random oversampling relies on duplicate random examples already existing in the minority class, balancing the number of samples of other classes. Since this process performs in a random way, it is complicated for the decision function to find out a well-defined borderline between classes. Thus, this technique may be inefficient at improving the predicting capability of a model by a large margin, due to the possible overfitting on account of replication of samples of a minority class.

Synthetic minority oversampling technique (SMOTE) is a good alternative to *Random oversampling*. SMOTE is another oversampling method which adds new samples by synthesising them from the existing ones, and not replicating them. These artificial examples are extrapolated and created using k-Nearest Neighbor algorithm, the neighbours are randomly selected. In contrast to *Random oversampling*, the application of SMOTE forces a more general bias, but affecting the minority class [37].

2.5 Evaluation Metrics

Evaluating learning algorithms is crucial so that it is possible to have feedback from metrics and, consequently, make improvements to get the best result. After the models are trained, the unseen data is classified and evaluated with the most appropriate metric. These metrics depend on either the type of model (like classification or regression) or the type of data (such as imbalanced and balanced datasets). Therefore, there are metrics exclusively to classification and to regression.

Confusion Matrix (CM) is an evaluation metric for classification problems. It provides a detailed analysis of correct and incorrect classifications for each class. In Figure 2.5, True Positive (TP) means all the positives samples that were predicted correctly, while the False Positive (FP) represents all the positive samples that were predicted wrong. Conversely, False Negative (FN) means that the positive samples were predicted as negative, while True Negative (TN) means all the negative samples that were predicted correctly. Some other metrics arise from CM.

- **Recall**: Recall (equation 2.16) represents how many positive samples were predicted correctly.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.16}$$

- **Precision**: Precision (equation 2.17) represents the ratio between the true positives samples and all the positives samples.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.17}$$

- **Specificity**: Specificity (equation 2.18) or False Positive Rate corresponds to the ratio of negative data observations predicted as positive, with respect to the negative observations.

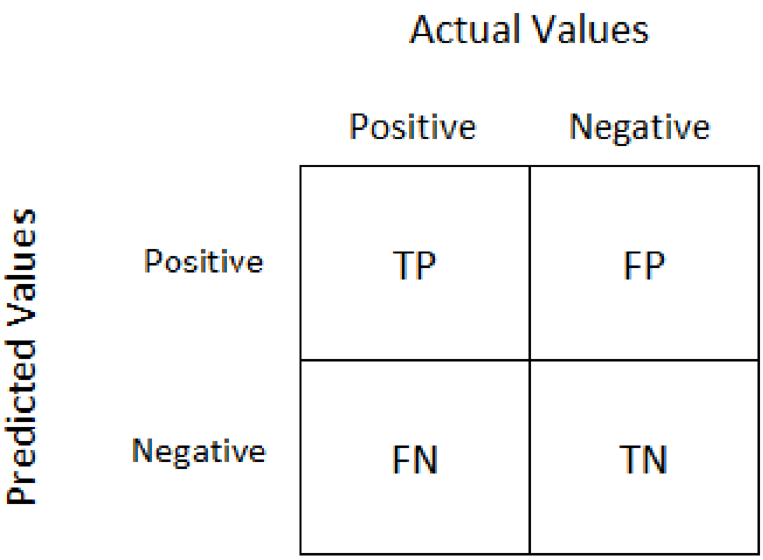

Figure 2.5: Representation of a CM for 2-class classification.

$$Specificity = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.18}$$

- **Sensitivity**: Sensitivity (equation 2.19) or True positive Rate corresponds to the ratio of positive data observations predicted as positive, with respect to all positive data observations.

$$Sensitivity = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{2.19}$$

- **F1-Score**: F1-score (equation 2.20) helps to measure recall and precision at the same time, so it can be considered the harmonic mean between recall and precision. Thus, when it is difficult to compare two models having low precision and high recall (or vice versa), F-score is used.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{2.20}$$

- **Accuracy**: Accuracy (equation 2.21) represents the ratio between the number of correct predictions and all predictions.

$$Accuracy = \frac{\text{TP} + \text{TN}}{Allsamples} \tag{2.21}$$

- **Weighted Average**: While in macro average the equal weight is given to each class (which might not be a realistic metric), in weighted average, the Recall/Precision /F1-score of each class is weighted by the number of samples from that specific class. This is important, for instance, when there is a large amount of class imbalance.

In turn, the Area under Curve (AUC) measures the 2-dimensional area underneath the Receiver Operating Characteristic Curve (plot of true positive and false positive rate), i.e. it measures the ability of a binary classifier to discriminate between positive and negative classes. AUC has a range from 0 to 1. The greater the value, the better the performance of the model.

Conversely, there are also metrics exclusively for regression algorithms. For equations 2.22, 2.23 and 2.24:

- y_i = Original Values
- $\hat{y}_i$ = Predicted Values
- N = Total of Samples

Some of these metrics include:

- **Mean Absolute Error (MAE)**: It represents the average of the difference between the true values and the predicted ones. Thus, this metric measures how far the prediction was from the original output.

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N}|y_i - \hat{y}_i| \tag{2.22}$$

- **Mean Square Error (MSE)**: Similar to the previous metric, MSE represents the average of the square of the difference between the true values and the predicted ones. It has the advantage to highlight the larger errors comparing to the smallest ones, since it takes the square of the error,

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.23}$$

- **Root Mean Square Error (RMSE)**: It is the square root of MSE. Thus, while MSE represents the variance of the error value, RMSE represents the standard deviation of errors.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{N}} \tag{2.24}$$

- $\boldsymbol{R^2}$: R^2 or coefficient of determination represents how strong the relationship between the model and the dependent variable is. It ranges from 0 to 1, where 1 represents the best model.

2.5.1 Parameter Tuning

This is an experimental process, where once the first evaluation is done, the parameters of the model may be tuned in order to further improve the training. At the beginning of a train, there are few parameters that have default values. However, some other values should replace the originally set parameters to evaluate the model's new performance. It is important to note that after parameter tuning, the evaluation of the results using metrics must always follow.

Chapter 3

Literature Review

Over the last decade, the number of research studies that have been developed to study the LOS in a hospital has increased. These models are applied to a particular type of disease or patient, using different algorithms and metrics.

3.1 Statistical Analysis

There were studies with the aim to determine factors that affects LOS, using only statistical analysis [38, 39]. One of those studies [38], containing 640 patients, applied to teaching hospitals in Iran, demonstrated that factors such as age, employment, marital status, history of previous admission, patient condition at discharge, method of payment, and type of treatment influenced on LOS. However, variables like gender, place of residence, and type of admission, did not influence it. To analyse data, since the dependent variable of LOS did not follow normal distribution, Mann–Whitney U test (nonparametric statistical test comparing the medians of two independent random samples) and Kruskal–Wallis test (nonparametric test equivalent to ANOVA) at the significant level of 0.05 were used [38]. Along the same lines, on other study with the same aim but with data provided from public hospitals in Lorestan Province (Iran) [39] used T-test and one-way ANOVA. Moreover, it also used an embedded feature selection method (multifactor regression) to determine the most important features related to LOS. With 662 patients, it demonstrated that factors such as gender, age, marital status, residence, job, referral type, type of insurance, type of disease, and discharge status affect LOS. None of these studies provide prediction model, since it only focuses on descriptive analysis [39]. Conversely, in [40], using data from 145 patients, in order to identify the factors related to LOS in the Intensive Care Unit (ICU) stay in patients after cardiac surgery, it was performed univariate analysis using Mann-Whitney U test or the independent sample t-test, as appropriate, for continuous variable. For categorical variables, the χ^2 test was performed. In that study, where the variables have sequential evaluations in the first 24 hours postoperatively, it was applied repeated measures of ANOVA in order to evaluate its time course, where a new variable was created for that purpose. However, in addition to statistical inference, since the number of characteristics that influenced LOS could be large, the RF model (a feature selection embedded model) was used to identify those features influenced the LOS the most. From this RF analysis, preoperative hemoglobin concentration, aortic cross clamp time, PaO_2/FiO_2 ratio and blood glucose measured during the first 1 to 4 postoperative hours were the four more relevant variables. In this paper a classification model was built using Logistic Regression (LR), obtaining AUC = 0.79 [40]. Lastly, for statistical analysis, in [41], one of the aims was determine which factors were associated to LOS, the following major features were identified. These statistical analysis techniques were applied: Student's T-test and ANOVA, as well as regression analysis. Nonetheless, this article states that analyse the absolute LOS for all diagnoses was not a good technique since there are huge differences in the LOS for different diagnoses. Z-scores were used to analyse any differences in LOS by diagnosis. The last purpose of this research was to use the resulting variables to build a predictive model of patient's LOS. While LOS was the dependent variable, frequency of surgery, frequency of diagnosis, frequency of patient transfer, severity and insurance type were the independent variables. For both the regression (Multiple Regression) and classification model (RF), 80% and 20% of data became the training and test data, respectively. The R^2 for the regression model from the training dataset was equal to 0.267 and the MAE of the test dataset equals to 4.68. The classification model had as aim to classify long-term hospitalised patients (more than 30 days), having an accuracy of 0.973 [41]. Data coming from about 45 000 patients admitted to a tertiary general university hospital in South Korea were used. However, this article has some limitations since it does not consider general characteristics of the patients as well as their environmental and seasonal factors and, finally, date/time factors.

3.2 Predictive Models for length of stay

Conversely, there are articles that elaborate predictive models for LOS in the most diverse situations, divided into two major groups: classification and regression. Regarding the first groups there is already some research applied to LOS that concerns a specific disease [42, 43]. A recent study conducted in 2019 [42], with a dataset of 12 000 patients, had as purpose to develop a ML-based model for predicting in-hospital LOS for cardiac patients, using 4 different algorithms in order to build a classification model with three classes: Short (< 3 days), Intermediate (3-5 days) and Long Stay (>5 days). Unlike the other articles mentioned so far, the information gain was used as the method to select the most relevant attributes of the dataset. The attributes with information gain of more than zero were the only attributes used as features of interest in the model. In this article four different models were evaluated: RF, Neural Networks (NN), Support Vector Machine (SVM) and Bayesian Network (probabilistic model that takes advantage of Bayesian inference for probability computations). In the model evaluation the 10-fold CV method was used. An accuracy of 0.97, 0.80, 0.50 and 0.67 for the RF, NN, SVM and Bayesian Network algorithm were obtained, respectively [42]. By the same token, in 2013 an article was published in Healthcare Informatics Research [43] whose objective was to use DM techniques to determine and predict LOS of coronary artery patients, but this time using a much smaller dataset: 2064 patients. During data preparation, if a feature contained more than 50% of missing values in the records, that characteristic was determined not to be an effective feature in the analysis. Then, conversely, if a feature was found in less than 12% of records with missing values, the mean values of records replaced the missing values with numeric features. Finally, for those features with more than 10% of missing values, the C5.0 Decision Tree (DT) was applied. The missing values of these features were filled using this algorithm. Conversely, the technique used to resolve the outliers issue was to replace its value by the nearest acceptable non-outlier. For the classification methods to be applied, 3 classes were created: '1', if $0 \leq$ LOS ≤ 5, '2' if $6 \leq$ LOS ≤ 9 and '3': if LOS >10. Then 4 different algorithms were adopted: DT (C5.0), SVM, NN and

Ensemble Algorithm that created a new model combining the other three algorithms used. The accuracy for DT was 0.835, and 0.539 for NN. The best results (overall accuracy) were 0.964 and 0.959 for SVM and ensemble algorithm, respectively [43]. One of the most significant limitations was that this model did not consider individual characteristics such as weight or disease status.

Regarding the implementation of classification models, there is still some research to perform in the surgical area [44, 10]. In a paper containing a dataset of 896 patients, which aimed at developing predictive models for determining whether patient LOS is within the standard LOS after surgery, divided the cases into Urgent Operation (UO) and non-UO [44], resulting in two different classification models. 7 different supervised learning techniques were applied: C4.5, C5.0 (successor of C4.5, it generates fewer rules and more accurate results and automatically removes redundant attributes, however it has low memory usage), Classification and Regression Tree (CART) (similar to C4.5, but the construction of the tree is based on a numerical splitting criterion recursively applied to the data), LR, SVM, RF and Multivariate Adaptive Regression Splines (MARS) – a non-parametric regression technique that models nonlinearities and interactions between variables. Feature selection involved three steps. All the variables were filtered out if there were more than 99% of missing values. Secondly, by using the *GainRatioAttributeEval* module of Weka (an open source DM software), less important variables were excluded. Lastly, based on two surgeons and two physician opinions, the final subset of features was defined by consensus. While for the UO model, comorbidity, body temperature, blood sugar, and creatinine were the most influential features, for the non-UO model were blood transfusion, blood pressure, comorbidity and the number of ICU admissions. The metrics presented in this article were: accuracy, specificity, sensitivity and AUC. Regarding the UO group model, the accuracy goes from 0.719 for LR to 0.857 for both RF and MARS. Whereas for the Non-UO group model, the worst algorithm presents an overall accuracy of 0.747 for CART and the best accuracy value of 0.894 for RF [44]. Another article in 2016 also studies LOS in surgery [10]. It aimed to determine the factors influencing LOS and also to build a predictive model of LOS in the general surgery department and it comprehends a dataset of 327 patients. The features were selected if four specialists in general surgery believed to be related to LOS. Using 70% of the data for training and 30% for testing and defining 3 output classes ('1' if $1 \leq$ LOS ≤ 3, '2' if $4 \leq$ LOS ≤ 5 and '3' if LOS ≥ 6), DT was the shortlisted algorithm. Only the pre-operation information was used to predict post-operation LOS. Finally, the accuracy of the DT model was equal to 0.8469 [10].

A different study [45], which aims to develop a pre-surgical classification model to determine whether the LOS of a patient is within the standard number of days is performed, they proved that supervised learning techniques can serve as a way to analyse EHR to accurately predict a prolonged LOS. In this research, by using a dataset of 913 patients, the final subset of features was selected by two general surgery surgeons and two senior physician assistant clinicians. Three different algorithms were used to construct the LOS model: DT, SVM and RF. Also, in order to evaluate the performance of the LOS model 4 different metrics were considered: accuracy, sensitivity, specificity and AUC. The sample group was divided into 2 groups, UO and non-UO, where UO was defined as an acute aortic syndrome needing invasive monitoring in an ICU. The results demonstrated that in both groups the RF algorithm constituted the most accurate prediction model, presenting a value in terms of accuracy of 0.853 for UO and 0.877 for Non-UO [45].

From another perspective, there are still studies equivalent to [41], which contain both regression and classification models [46]. In a research from 2013, DM techniques for predicting incubator LOS were studied in Egypt (302 patients) and United States of America (USA) (5000 patients). Two working strategies were defined: 1 – Predict LOS as a categorical variable, where the classification models were built using Naïve Bayesian (NB) algorithm another one using SVM and the third one using the LR algorithm; 2 – Predict LOS as a continuous variable, using two different models: SVM and LNR. During the data processing phase, attribute importance (which takes advantage from the Minimum Description Length algorithm - Given a limited set of data, it states that the best explanation is the one that allows the greatest compression of the data [47] - and is provided by Oracle DM) was used as principle to rank the attributes by significance. When evaluating the classification models, the highest overall accuracy for Egypt is equal to 0.89 with SVM, while the lowest is 0.82 (NB). Conversely, for USA, LR had the best performance, achieving an overall accuracy of 0.98, while the worst was for SVM (0.79). The metric used for evaluating the regression models were the MAE and the RMSE. In the Egypt case, the MAE for SVM was 5.65, while for LNR was 10.44. Regarding USA, also the highest value corresponds to LNR (3.56) and the lowest to SVM (2.49). Regarding RMSE in Egypt, SVM presents a value of 11.76 and LNR 14.9, whereas in USA case, SVM leads to 7.38 and LNR to 8.21 [46]. This study thus demonstrates the prevailing difference in the evaluation of results when using classification or regression algorithms, with the classification more likely to be more successful even though with potentially less precision of the exact day depending on the classes considered.

Although this dissertation focuses on the elaboration of a classification model, it is also important to be aware of the state-of-the-art regarding the regression approaches in order to understand how the problem is addressed, how the data are processed and what the outcome means. As such, just as there are articles devoted to purely classification algorithms, there is also research that is dedicated to regression algorithms only. Such as [42] and [43], [48], [49] and [50] are also focused in LOS for a specific disease, while using regression methods. The first [48] concerns hip-fracture patients, containing a dataset of 2000 patients. The feature selection on this research was performed by permutation importance method. 4 different regression models were built: RF, Boosted Decision Tree (BDT) – type of decision tree where the boosting technique was implemented. In boosting, at each iteration, the misclassified data points, increase their importance, so the learner can improve accuracy - NN and LNR. Also using the 10-fold CV for training, the method used to evaluate the performance of the regression models was the Relative Absolute Error (RAE) - it measures a performance of a predictive model comparing a mean error to errors produced by a trivial model. RF obtained a RAE equal to 0.26, BDT to 0.34, NN to 0.55 and LNR to 0.93 [48].

Other publication [49], a regression model was proposed to predict LOS for inpatients with one of the three diagnoses, in a cardiovascular unit: Coronary Atherosclerosis (CAS), heart failure and acute myocardial infarction. However, in this research two stages in LOS prediction were presented. The first - the predischarge stage - uses all clinical factors, while the second, the preadmission stage, uses only factors available before admission. For both models the same two algorithms were applied: LNR and NN. In order to evaluate the relative effectiveness in predicting LOS at the preadmission stage, the prediction results were used at the predischarge stage. In the statistical analysis the Pearson's correlation coefficient was applied in order to study the relationships between LOS and inpatients' characteristics. However, since the distribution of LOS for CAS patients was significantly different to patients with the other two conditions, two different prediction models were built: one for CAS patients and another one for non-CAS patients. In order to avoid overfitting, the training data was separated into two sets. A training set – in order to update the weights and biases – and a validation set – to stop training as soon as the NN was overfitting. The authors used, for both models, 30% of the data for test and 70% for training. The MAE for the predischarge model for CAS using LNR algorithm was 1.09 and when using NN it was between 1.06 and 1.11. In turn, 1.00 and 1.03~1.07 were the values for the MAE for the LNR algorithm and NN, respectively, in the preadmission model. For Non-CAS patients, the MAE for the predischarge model using the LNR was 3.76 and for NN, it was between 3.83 and 3.91. Finally, for the preadmission model, the MAE using the LNR was 3.76 and for NN it was between 3.87 and 3.97 [49].

A different type of comparison was also performed in a research article from 2010, with a dataset of 1080 patients [50]. It aimed to compare prediction results among different clinical stages (admission, acute and post-treatment) of burnt patients. The independent variables used for the first stage consisted only in patient demographics, possible medical conditions and burn injury characteristics. For the second stage, the independent variables associated were also used for the acute stage in addition to the surgical operation that was performed and the interval between admission and the first surgery. For the post-treatment stage, it was considered three additional variables: number of escharotomy treatments, the number of wound excisions and skin grafts and the total number of surgical operations. The evaluation procedure used to estimate the effectiveness of the LOS prediction techniques for each clinical stage was the 10-fold CV strategy. This strategy was performed three times so potential biases could be minimised, being the overall performance the average of the performance estimates. In turn, among others, MAE was the shortlisted metric. LNR, M5 (model-tree-based regression) and SVM regression were the selected algorithms. Regarding the MAE of the admission stage, the lowest value was equal to 8.992 and corresponded to the SVM regression, while the highest was equal to 9.532 for the LNR. For the acute stage, it goes from 8.994 to 9.503, for SVM and LNR, respectively. Finally, for post-treatment it ranges between 6.074 for SVM and 6.331 for M5. However, in order to improve this error, the Correlation-based Feature Selection (CFS) method was implemented for feature selection purpose. This method has as hypothesis that good subsets of variables should contain independent variables highly correlated with the target dependent variable, however uncorrelated with each other. By using the variables selected by CFS, the highest MAE for admission stage was 9.625 for LNR and the lowest 9.189 for SVM regression, while in the acute stage, the lowest values were 9.237 for SVM and the highest 9.665 for LNR. In the post-treatment case, also SVM yielded the lowest values (similarly when the CFS method had not been applied): 6.543 for SVM and 9.648 for LNR [50].

LOS regression models applied to surgery were also developed [51, 52]. Since there are many algorithms that can be applied to this case, there are also studies that compare the performance of different techniques used when estimating the LOS. Comparing the performance of NN and adaptive neuro-fuzzy system algorithms (a type of NN) to predict patients LOS on ICU after cardiac surgery, a research article published in 2018 [51] described that an adaptive neuro-fuzzy algorithm (MSE = 7 and R = 0.88, where R represents the correlations between the predicted value and the real value) was more precise than a NN (MSE = 21 and R = 0.6). This article was based on data from 311 patients. During the cleaning phase, some nominal variables were subdivided into fewer categories, while independent quantitative variables were changed into qualitative ones. For the feature selection, in this case, CART DT method was used, where 23 variables were identified as related to the LOS [51].

Similarly, a research article published in 2015 [52] compared a NN model to several other models in predicting LOS in the cardiac surgical ICU based on pre-incision characteristics of 185 patients [52].

By using automatic linear modelling (a method that enables researchers to select the best subset automatically) 8 pre-incision factors statistically associated with ICU's LOS were identified. Using 90% of the data for training and 10% for testing, the selected data was implemented in four different algorithms: Automatic Linear Modelling (achieving a R^2 = 0.36), an NN (R^2 = 0.54), a DT (R^2 = 0.50) and, the optimal, a RF (R^2 = 0.84) . The dataset of this article contained data from 185 patients [52].

3.3 Predictive Model: Portuguese Case

Not only have LOS related studies been performed abroad in hospitals, but also in Portugal. A study dating from 2014 it is an example, whose aim was predicting inpatient LOS, using a regression model, in a Portuguese hospital, applying the CRISP-DM methodology [53]. This project comprised data from 26431 patients. In the data preparation phase, the missing values were replaced using the hot deck technique, which consists of looking for the most similar example and replace the missing value by it. The main procedure and main diagnosis attributes had too many categories, so the categories were grouped in order to have fewer levels. 28 attributes were selected based on other literature and later confirmed and validated by a panel of 9 specialists from 7 hospitals. To analyse model validity, a 5-fold CV method was applied and, for a better robustness of the results, it was performed 20 times. In order to evaluate the models, 3 regression metrics were used: R^2, MAE and RMSE. The best model was the RF with an average R^2 = 0.813, MAE = 0.224 and RMSE = 0.469. Globally, the best results refer to NN, RF and SVM, which present higher coefficients of determination and lower values for the MAE and RMSE. The simple method of forecasting based on the mean, multiple regression and DT presented the worst results. Taking only these algorithms into consideration, DT yielded the best results for the R^2 (6.22) and MAE (0.415). In turn, multiple regression showed the lowest value among the 3 for the RMSE of 0.650.

Although the good results, *Hospital das Forças Armadas* chose not to implement the models because it plans to develop further research in this area, such as, investigation on specialised modelling for some types of services, such as orthopedics [53].

3.4 Influence of risk factors as predictors of length of stay

An important analysis carried out by [54], with a cohort of 4509 patients advised to undergo total knee arthroplasty, is the comparison between the influence on LOS caused by patient-related and procedure or structural-related risk factors. According to this study, risk factors can be categorised as patient-related (such as demographic characteristics) and as procedure or structural-related risk factors (for instance, day of the week of the surgery and surgery time). The performance between a model containing patient-related risk factors only, model A, and another one with both patient and procedure or structural-related risk factors, model B, was evaluated. The data included information from multiple hospitals from the same network. This analysis resorted to the use of the Akaike Information Criterion (AIC), which measures the quality of a model – the lower the AIC, the better the fitting. Although model A has demonstrated that the patient-related risk factors are significant predictors of LOS , model B proved, by a decrease of 1670 units in AIC, that procedures or structural-related risk factors are important variables that influence the outcome of the LOS . As such, this work organises an idea of the importance of both features, early discarded by many articles, which can help to more accurately control the LOS.

3.5 State of the art challenges summary

As described throughout this chapter, there are several studies around the world that provide models for predicting LOS for a specific disease, surgery, and more general cases and also determine the most LOS-related attributes. However, despite some of the good results in these articles, most of them used few data (hundreds) and often focused solely on either patient or procedure characteristics, as well as the feature selection stage was only performed mostly for one technique. Furthermore, it was noticed that the literature lacked studies utilising ML with respect to surgery. Thus, the purpose of this study is, based on EHR from HBA, to predict the LOS after surgeries of patients in this particular institution by applying a classification algorithm.

Chapter 4

Data and Methods

This chapter aims to outline the exploreddata and methods to ultimately develop the ML models used to predict the LOS after surgery. Firstly,a contextualisation of the current hospital dynamics is granted in order to understand the whole process behind the data. Next, an analysis about the current model adopted is presented. In the third part of this chapter, a brief description of the data is present, where it is included the definition of the models. Then, all the preparation process, which the data went through until reaching the final dataset used as input in the models, is described. Last but not least, how the final subset was modelled and evaluated is specified.

4.1 Contextualisation

A patient may undergo surgery for several health reasons, however, regardless of that reason, the surgery only assumes to have one of three dynamics:

(A) The vast majority of patients enter the Surgical Day Care (SDC), which is where patients who will undergo surgery will register and be prepared for it, on the day of the surgery and are immediately referred for surgery, without going through hospitalisation. Only after recovery, which can last up to 48 hours after surgery, patients are hospitalised.

(B) Very few patients (5 to 10%) enter the day before surgery. These are 'Enhanced Recovery After Surgery' patients or those with a specific self-preparation. In this case, these patients are hospitalised before and after surgery. It is important to note that the bed occupied by these patients after surgery is not the same as the one assigned to them before surgery.

(C) The remaining percentage refers to patients undergoing urgent surgery. They can directly enter in the operating theatre or be primarily hospitalised and then operated.

With the exception of the third case mentioned, before surgery doctors define the type of surgery: outpatient, inpatient care and one-day surgery. Outpatient surgeries are surgeries in which patients are discharged in less than 24 hours after the time they entered, being even hospitalised after surgery. Inpatient Care corresponds to the type of surgery that leads to the patient being discharged in more than 24 hours. One-day surgery is identical to outpatient, however patients are discharged from recovery, and do not go through hospitalisation.

4.1.1 Surgical Process

For an effective treatment of data, it is also necessary to become more aware of the reality in hospitals so that a more critical understanding of the data occurs. Due to the visit to HBA, it was possible to know the reality for patients who require surgery, concerning the most common dynamic of surgeries.

At the beginning of the process, a patient has an appointment where he/she is told about the need for the surgery. The patient goes to the surgery waiting list, taking into account a range of factors, such as, the type of surgery and anaesthesia. Then, as soon as possible, the day of surgery is communicated to the patient, as well as the surgery plan, including the procedure. At this time, generally, all the process on the day of surgery are known, except for the surgery room where it will take place.

On the day of the surgery, the patient must go to the SDC where it is admitted and given the suitable wristband. The patient is then called to the room in SDC and a box is assigned a to her/him, where the preparation for surgery takes place. In this context, a box is a compartment with a bed for the patient, separated by curtains from other boxes.

At the time of surgery, the patient is taken to the operating theatre. If something unexpected happens during the surgery, it is registered either by the surgeon responsible or by the nurse. As soon as the surgery is finished, the patient enters the post-anaesthesia care unit (also called recovery). The patient remains in this unit until necessary or until a bed in the inpatient area leaks. After being assigned a bed, and a room, and there is medical consensus on his/her healthy stability, the patient is discharged.

4.1.2 Surgery List

Sistema Informático de Gestão da Lista de Inscritos para Cirurgia (SIGLIC) is a software system that aims to organise surgical care services in the hospital from the National Health System. Its purpose is to maintain the balance between demand and response, in an articulated manner, focusing on the citizens' access needs. In this way, SIGLIC regulates all information related to the scheduled surgical activity and that is performed by the emergency services of the establishments of the National Health System [55]. Taking into account the patient's clinical situation, a level of clinical priority is assigned, among the following [56]:

- Level 1 - if the patient can wait up to 270 days for the surgery, or 60 days if they have an oncological disease.
- Level 2 - if the patient cannot wait more than 60 days for the surgery, or 45 days if it is an oncological disease;
- Level 3 - if the patient cannot wait more than 15 days for surgery;
- Level 4 - if the surgery has to be carried out within a maximum period of 3 days or during hospitalisation;

4.2 Current Model

In order to predict the day of discharge and estimate the LOS, HBA has its own model. Before going into detail, it should be noted that this model does not comprise some specialties, such as oncology.

Currently HBA calculates the expected LOS for a given patient taking into account their main diagnosis. Based on a dataset from the previous year (here 2017), the average number of days of hospitalisation for each main diagnosis is calculated and thus, when a new patient enters in 2018, according to the average of that diagnosis in 2017, this suggestion of LOS is assigned to the patient and, consequently, the expected day of discharge is provided. This value must be validated by the responsible team that can still modify this day, in consideration of possible complications.

Furthermore, it is also important to emphasise that the discharge dates are always communicated to the bed management department. An email is also sent to the doctors notifying how many discharges are foreseen for a certain day. Finally, contextualisation is also necessary for some special cases. Although some patients are able to leave the hospital on the planned date, sometimes there are some who may not leave on the day of discharge as they may be elderly or have a social impediment and may have to stay in the hospital for a longer period.

4.3 Data Description

The dataset used in the study comes from the database of HBA. The cohort comprises 20736 cases of patients undergoing surgery at HBA, between 2017 and 2018, and contains 135 features, including patient and procedure-related features.

The dataset contained patients of 10 medical specialties: General Surgery, Ophthalmology, Orthopaedics, Urology, Otorhinolaryngology, Gynaecology-Obstetrics, Angiology and Vascular Surgery, Reconstructive and Aesthetic Plastic Surgery, Dermatology and Cardio-Thoracic Surgery.

It should also be mentioned that there is a single variable for the main diagnosis and a single variable for the main procedure, however, there are 20 different variables for secondary diagnoses and 6 for associated procedures. This is because a patient may have more than one diagnosis, such as having diabetes and hypertension, and be subjected to more than one surgical procedure in the same surgery.

4.3.1 Definition of Models

For this work, LOS is defined as the time difference, in days, between the day of surgery and the day of discharge, regardless of whether the patient was admitted days before surgery. With the available data, two models were developed: pre-surgical and post-surgical, using only pre-surgical or post-surgical information, respectively. In the post-surgical model, besides the pre-surgical information, for instance, the time of anaesthesia and time of surgery are also already known variables.

For comparison purposes, it is also necessary to replicate the model in use at the hospital, here named as HBA Model (details in section 4.2), so its performance can be known.

4.3.2 Diseases and Procedures Coding

In Portugal, since the 80s, clinical coding has been used, carried out by doctors who codify episodes of hospitalisation, outpatient surgery and part of the medical outpatient clinic. This coding was first done by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), which consists of 3 volumes [57]:

- Volume 1: A tabular list of diseases;

- Volume 2: An alphabetical index to the disease and injuries entries;
- Volume 3: A classification system for procedures (alphabetic index and tabular list).

The list of diseases was divided into 19 large groups, differentiating, for example, respiratory diseases from circulatory diseases. In turn, the classification system for procedures has 18 different groups, splitting surgeries, such as, of the digestive system and of the urinary system [58].

Although over the years ICD-9-CM has undergone updates, in order to adequately portray the spectrum of pathologies and procedures existing in hospitals, as well as the technological innovations that exist every year, the ICD-9-CM has become obsolete. Accordingly, World Health Organization authorised the USA government to convert ICD-9-CM into a more robust and suitable classification. Thus, International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) was created to replace ICD-9-CM (Volumes 1 and 2) for the purposes of classifying diagnoses and International Classification of Diseases, Tenth Revision, Procedure Coding System (ICD-10-PCS) was created to replace ICD-9-CM (Volumes 3) for the classification of procedures [59]. Similarly to ICD-9-CM, ICD-10-CM is also divided into groups. This presents 22 different sets, each of which presents diseases that derive from a common scope. In turn, the ICD-10-PCS is divided into 17 different groups [60]. Although the ICD-10-CM system has also been adopted in Portugal, the present dataset still presents the main diagnosis and the associated procedures in ICD-9-CM. Conversely, the secondary diagnoses are already in ICD-10-CM, as well as the codes associated with the main procedure are in ICD-10-PCS.

4.4 Data Preparation

To achieve the objective of this work, data was prepared and cleaned, following the CRISP-DM methodology. Data preparation process and the evaluation phase were performed using Python.

4.4.1 First Data Cleaning

First, the features were analysed in order to understand which columns in the first instance could be excluded, or, taking into account their content, could serve as a basis for excluding some of the cases who might have some error. This includes evaluating data consistency, for instance, if dates of the process are not consistent.

Regarding cases (i.e. the rows of this dataset), 13997 of the initials 20736 were eliminated, taking into account the following detailed exclusion criterion for each elimination:

- All the 50 patients who died before hospital discharge were not considered.
- In order to make a comparison between the actual LOS and the estimated LOS all patients who did not have the *Expected Discharge Date* were excluded, resulting in 13644 eliminations.
- In order to get the LOS, two features were necessary: *Surgery Date* and *Real Discharge Date*. However, despite every patient having the surgery date, 13132 did not have the real discharge date.
- 1 patient had not defined his *ICD9 Diagnosis*.

- 1 patient did not have his *ICD10 Main Procedure Code*. Also, 161 patients did not have codes that corresponded to the used classification system and were also eliminated.

- All 402 patients who did not have a category associated with the main surgeon were deleted.

- Since all patients must follow the normal hospital dynamics - admission, surgery, discharge - All patients whose *Surgery Date* was later than the *Real Discharge Date* were eliminated (there were no patients under this condition). Similarly, patients with *Admission Date* later than *Surgery Date* were removed, resulting in 2 eliminations.

- 7 patients who do not have anaesthesia details and have an start and end date and time of anaesthesia were eliminated. In turn, 100 and 202 patients who have anaesthesia and do not have an anaesthesia start time and end time, respectively, were excluded.

- 544 patients whose height was less or equal to 45 cm were eliminated. In turn, 569 patients measuring less or equal to 65 cm and older than 0 years were also excluded.

- All 584 patients weighing 0 kg were excluded. Likewise, 2 patients who weighed more than 300 kg were also eliminated.

Since some of these eliminations coincided in the same cases, the total number of rows eliminated is 13997, resulting in a total of suitable 6739 cases.

Conversely, regarding the features, some of them were also removed. Firstly, there are 12 features that are duplicates of other features coexisting in the dataset and so were removed. According to HBA hospital administrators, 36 variables have a very low degree of reliability, which is why they have been deleted from the model. This low reliability is due to the fact that this dataset contains automatic features that are sometimes not controlled or supervised, resulting in errors. In addition, this dataset results from a set of variables under the responsibility of different people, some of whom were not present during the development of this work and, consequently, it was not possible to fully perceive these features.

Lastly, 24 other features were also discarded due to disparate reasons, which are described below:

- *ID* represents the number for each case. Being different for each situation, the variable does not introduce any useful information.

- *Parish*, *Municipality* and *Health Centre Registration* has been withdrawn since it contains too many categories and because it is assumed that their information was already considered in the variable *District*.

- *Anaesthesiologist* and *Instrumentalist* were also removed due to the existence of an abundant number ($>$ 100) of categories.

- Since the ICD-10-PCS code is the most suitable for contemporary times and there is a procedure variable with this classification in the dataset, *Procedure*, *ICD9 Procedure* and *Main Procedure Description ICD9* were eliminated because it is the procedure variable in ICD-9-CM.

- *Main Diagnosis* and *Diagnosis Description* were discarded because they contain the description of the variable that contains the ICD-9-CM code for diagnosis.

- *SIGLIC Patient* refers to whether the patient is enrolled in the SIGLIC or not. In turn, *Delivery Room Hospitalisation* corresponds to the patients who were admitted in the delivery room. However, since the categories for these two variables were always the same, these variables do not introduce additional information to the model.

- *Main Surgeon*, *Assistant Surgeon 1* and *Assistant Surgeon 2* represent the surgeon responsible for the surgery and, if any, the first and second assistant surgeon, respectively. However, as all these features have a high number (> 150) of categories, it is not feasible that they remain in the dataset.

- *Discharge Hour* and *Physician Responsible for Discharge* were not considered since they correspond to the moment when the patient is already leaving the hospital.

- *Expected Discharge Date* was removed because it is associated to the current HBA model.

- Once used to remove patients who died before being discharged, feature *Date and Time of Death* were eliminated.

- *Floor* and *Bed* were eliminated because these variables are sometimes only acquainted with way after the end of the surgery, so it cannot be placed on any of the models (pre-surgical or post-surgical).

- *Procedure Description* and *ICD10 Main Procedure Code Description* are not included in the final model since they refer to the description of the variable in ICD-10-PCS code of the procedure.

- All columns containing more than 85% missing values were removed. However, no column was under this condition. In fact, after the first data cleaning, no missing values were considered in the dataset.

The removal of the aforementioned features resulted in a decrease from 135 to 63 features.

4.4.2 Data Transforming

Although the original dataset has many features, sometimes the combination of some of them can give rise to new attributes that enrich the dataset with important information. Thus, in order to extract relevant knowledge, the following features were created. In addition, some of the currently existing features can also be deployed in a greater number of attributes in order to provide greater information fluency.

- Since the variable *Days of Hospitalisation* proved not to correspond sometimes to the subtraction between *Real Discharge Date* and *Surgery Date*, a new variable was created: *LOS*, corresponding to the mathematical operation mentioned.

- A variable referring to the number of days a patient had been in the hospital until surgery (*difAdmissionDaySurgery*) was also created.

- A variable was also included representing the time, in days, between the *Surgery Date* and the *Suggested Date of Surgery*, named *difSuggestRealSurgery*.

- A new feature, referring to the time that a patient spent in the operating theatre, was also implemented (*timeOt*).
- The time under which a patient remained anaesthetised (*timeAn*) was also associated with a new attribute.
- A feature corresponding to the difference, in hours, between the time of the last suture and the time of the first incision was additionally created, *timeSurgery*.
- A variable corresponding to the hour and minute of leaving the operating theatre (*exitOperating*) was also implemented.
- A variable associated with the number of secondary diagnoses (*nrDiagnsec*) and another corresponding to the number of associated procedures (*nrProcedassoc*) were established.
- Since one of the features that the dataset provided corresponded to the ID number of the patient, it was possible to create a variable that would count the number of readmissions in the last six months (*Readmission*), and another that would quantify the number of readmissions in total (*ReadmissionEver*).
- When *Weight* and *Height* features were given, the variable corresponding to the body mass index (*BMI*) was calculated.
- The month of surgery (*MonthSurgery*) was also associated with a new variable, previously converted from *Surgery Date*.
- At last, the feature that is used in the current HBA model to define the LOS was also included. This attribute (*MeanDiagnosis*) corresponds to the average LOS for diagnosis. Nonetheless, all 248 patients who underwent a surgery in 2018 with diagnoses that were not part of the 2017 diagnosis list were eliminated, as the prediction of the average LOS for these diagnoses did not exist.

After new features were created, the size of the dataset comprised 6491 cases and 77 features.

Categorical variables were encoded using the one-hot-encoding technique. However, since some categorical variables have many categories, some of these variables were grouped into larger sets.

The variable *District* initially had 32 categories: 18 associated with the districts of mainland Portugal, 1 corresponding to the island of Madeira, 5 corresponding to Azores, 7 relating to different countries and 1 unknown. However, in order to reduce the number of categories, the 5 districts of Azores were grouped into a single category defined as *Azores* and all the 7 countries were brought together into a single category named *Other Countries*. Thus, the district was finally made up of 22 categories instead of 32.

Similarly, also in *Origin*, the 10 categories were grouped into 2 different sets: *Internal Origin*, which comprised the categories External Consultation, Emergency and SDC, and *External Origin* which contained the categories INEM, Health Centre, External and Health 24 Line.

In turn, since initially there were 1410 categories corresponding to the main diagnosis, these were gathered according to the existing groups according to the ICD-9-CM, resulting in 19 different groups. 2 of the 19 groups were also merged due to the fact that they are supplementary classifications, resulting in 18 groups. In addition, all codes that were part of at least 250 cases were also considered as unique categories, adding 2 categories to the existing ones, making 20 categories.

Analogously, the categories of the main surgical procedure were also reduced from 2200 to 17 categories, using the groups in the ICD-10-PCS.

Finally, the 20 secondary diagnoses features were also coded, this time, following the ICD-10-CM, which comprises 22 large groups. Notwithstanding the existence of patients who may not have all 20 variables filled, led to the creation of a new category defined as *noSecundaryDiagnoses*. In addition, 7 new categories were assigned to codes that were diagnosed in at least 100 cases, resulting in a total of 30 groups.

The 6 associated procedural variables were organised into 18 groups resulting from the ICD-9-CM. A variable defined as *noAssociatedProcedure* was also created, thus resulting in 19 groups.

Finally, all the remaining categorical variables (30 + 18 = 48 features) that presented the final categories, were converted into dummy variables, via one-hot-encoding. However, since variables *noSecundaryDiagnoses* and *noAssociatedProcedure* provided no extra information, they were not considered.

However, as each secondary diagnosis variable had 30 different categories, which would represent 600 (30*20 = 600) new features when applied one-hot-encoding technique, a count was made of the total number of secondary diagnoses for each group, resulting in 29 features corresponding to the groups already mentioned. The same was applied for the associated procedures that resulted not in 114 new features (6 associated procedures features x 19 categories), but in 18 (according to ICD-9-CM categories for procedures).

It is necessary to mention that whenever there were no cases for any category, this column was disconsidered. This means that in the final dataset, not all categories/columns mentioned here were actually created.

After the application of one-hot-encoding technique, the consequent removal of the features that resulted in the binary variables, the final number of features was 197.

4.4.3 Last Data Cleaning

At last, after some of the features were used to generate new attributes, since they no longer provided any extra information, they were removed. In addition to these eliminations, some other cases were also not considered due to the existence of possible errors in the veracity of their data, or because they where outside the scope of this dissertation:

- Only patients who stayed at least one day in the hospital were considered, following the purposes of this work. Thus, all 12 patients with zero *LOS* were not considered.
- Due to its low probability, it was assumed that the 10 cases whose *BMI* is greater than 70 and the 10 cases whose *BMI* is less than 9, are errors. In order to reduce the likelihood of unreliable data, these 20 cases were removed from the dataset.
- *ID NP* corresponds to the exclusive number of patient, so it was not considered.
- *Suggested Date of Surgery*, *Surgery Date*, *Date and Time Operating Theatre Entry*, *Date and Time Anaesthesia Induction*, *Date and Time First Incision*, *Date and Time Last Suture*, *Date and Time Anaesthesia End*, *Date and Time of Operating Theatre Exit*, *Real Discharge Date* and *Admission Date* were eliminated because they had already served the purpose for which they were used, namely, the creation of new variables.

- Although it was generated from existing features in order to provide information, the evaluation of *difAdmissionDaySurgery* demonstrated its inconsistency with that reported by hospital administrators. According to HBA, a large majority of surgeries take place on the same day or the day after the patient is admitted, however, the variable reflects that 82% of cases were admitted at least 2 days before undergoing surgery. Consequently, *difAdmissionDaySurgery* was not considered.

- Finally, it is necessary to underline that, since this dissertation also consists of a pre-surgical model, the following variables were not considered in this model, since they are only known after the end of the surgery: *timeOt, timeAn, timeSurgery, Room* and *exitOperating*.

After data preparation, the dataset that will be used as input in the development of the ML model, comprises a total of 185 and 180 features for the post and pre-surgical models, respectively, and 6459 cases. For further information about the description of each feature used in the final dataset, table A.1, in *Appendix* section, should be consulted.

4.4.4 LOS Intervals

Once data preparation is complete, categorising the target variable, *LOS*, is crucial. This task took into account the testimonies of HBA managers regarding the major needs in the hospital. Thus, 3 different classes were considered:

- Class 0: $LOS = 1$
- Class 1: $1 < LOS \leq 3$
- Class 2: $LOS > 3$

However, for the purpose of comparison and the possibility of better performance, the split of the previous classes into 2 complementary models (B and C) of 2 classes each was considered. Table 4.1 shows the intervals of that models and its designation from now on.

Table 4.1: Designation of each model regarding its classes. Model B and C are complementary.

Class	Model/Interval A	Model/Interval B	Model/Interval C
0	LOS = 1	LOS = 1	$1 <$ LOS ≤ 3
1	$1 <$ LOS ≤ 3	LOS > 1	LOS > 3
2	LOS > 3	-	-

It is necessary to highlight that, only patients who are predicted in model B as class 1 proceed to model C so that it can be predicted if patients remain between 2 to 3 days or more than 3 days in hospital. Thus, only patients with a LOS higher than 1 day were considered for model C, making a total of 3566 cases for this model.

4.4.5 Feature Selection

Initially the ambition of this dissertation focused on the development of a unique model including all the medical specialties presented in the dataset. However, as hospital administrators detected an imbalance in the success of LOS prediction depending on the clinical area, the development of exclusive models by specialty was also considered.

For the specialty case, due to the very limited number of samples for some of them, only specialties containing more than 400 cases were considered to specific models: General Surgery, Gynaecology-Obstetrics, Orthopaedics, Urology and Otorhinolaryngology.

Since the prepared dataset contains many variables, there is still one last task to be performed, with the possibility of decreasing the number of features and increasing the performance of the model - the selection of the most important features for each model. Before applying the feature selection methods, correlated (in at least 85%, using Pearson's correlation) and constant variables for each model were eliminated. Next, in this dissertation, four different feature selection techniques were considered: Mutual Information, Permutation Importance, RFE and RF. The selection of all variables that were considered in at least 3 of the 4 methods was adopted as criterion.

The estimator parameter used in RFE and Permutation Importance was the RF Classifier, with 100 trees.

The criteria for selecting the variables for each technique were as follows:

- **Mutual Information**: All variables with a mutual info value higher than the mean of mutual info values of all variables were considered.

- **Permutation Importance**: Applying CV along with RFE, the features with permutation importance positives values were the only ones selected.

- **RFE**: Resorting to the use of CV along with RFE, the optimal number of features is found. In each loop, one feature is recursively eliminated. This system scores different feature subsets and select the best scoring collection of features. For this work, the scoring parameter selected was F1-score weighted.

- **RF**: Similarly to Mutual Information, this embedded technique will select all the features whose importance is greater than the mean importance of all variables.

4.5 Data Modelling

Data Modelling, as the name implies, is the phase where a pre-determined algorithm is trained so it can predict the classes from the features when applied to unseen data. The dataset prepared and described in the previous sections served as input for the chosen algorithm: RF Classifier. Regarding this specific work, all models rely exclusively on this algorithm with defined parameters: 100 trees and the criteria adopted to measure the quality of a split was Gini.

In order to develop the models a stratified 10-fold CV was applied.

Nonetheless, due to the discrepancy between the quantity of samples of each class, during the training SMOTE was applied so that the number of samples in each label was identical and the model was not biased. However, there is an exception in the case of otorhinolaryngology, as this specialty has a much higher value of samples in class 0 compared to other classes, both in model A or model B. Thus, an undersampling technique was applied to the majority class, which comprised 300 samples, and later, SMOTE was implemented to the rest of the classes.

4.6 Evaluation Metrics

Lastly, in order to make possible the evaluation of the models, for both ML and HBA models, F1-score was the metric adopted during model evaluation. The lack of balance between classes, led to this decision. For an equitable assessment between all models, not only the F1-Score Weighted but also the F1-score for each class were considered. So that there is also a comparable analysis between the ML models and the HBA model, the results presentedin chapter 5 refer only to the assessment of the 2018 patient subset.

Chapter 5

Results

This chapter aims to present the major results of this work. Firstly, the features selected for each model are provided. Then, the results for every classification model, whether post-surgical and pre-surgical model, are introduced. Afterwards, the results of the best classification models are compared with HBA model. Finally, a final table also includes the results for models that only use procedure or patient variables.

5.1 Selected Features

The selected features for each model are presented in Tables 5.1, 5.2, 5.3, 5.4, 5.5 and 5.6. For reasons unrelated to the focus of this work, the specialties in the referenced tables will be replaced by:

- General Model: G
- General Surgery: S
- Gynaecology-Obstetrics: W
- Orthopaedics: O
- Urology: U
- Otorhinolaryngology: E

5.2 Developed Models

This section presents the models developed throughout this work. The general model is introduced, without specification by specialty. The second section presents the results of another approach: the development of models *per* specialty.

5.2.1 General Model

Table 5.7 presents the results of the general model, taking into account the application or the absence of SMOTE, and it also allows to identify which models are more suitable: 2-class or 3-class.

Table 5.1: Selected Features according to specialties and general model for Model A, Post-Surgical.

Model A: Post-Surgical						
Features	**G**	**S**	**W**	**O**	**U**	**E**
timeOt, MeanDiagnosis	x	x	x	x	x	x
Maximum Blood Pressure	x	x	x	x	x	
difSuggestRealSurgery	x	x	x	x		x
neoplasmsDiagSec	x	x	x		x	x
Age	x	x		x	x	x
nrDiagnsec, healthstatus, Weight	x	x		x	x	
Height	x	x		x		x
exitOperating	x	x		x		x
Medical device placement proposal	x			x	x	
neoplasms	x	x				x
SDC, hospitalisation	x	x				
genitourinaryDiagSec, noLatProcMain	x		x		x	
Room4, Room1, Specific needs for perioperative support or special techniques	x				x	
respiratoryDiagSec	x					x
BMI		x		x		
tuesday, thursday, Minimum Blood Pressure			x	x		
noLatDiagMain			x		x	
AssistantSurgeon1Spec				x	x	
Room3, Otorhinolaryngology, Gynaecology-Obstetrics	x					
endocrine, Gender, Cholelithiasis, digestiveDiagSec		x				
timeAn, leftLatProcMain, femaleGenitalProcSec, nrProcedassoc			x			
muscular, injury, Room5, ImprovementOrmitigation, hypertension, injuryDiagSec, musculoDiagSec, monday, friday				x		
wednesday, AssistantSurgeon2Res, InpatientCare, Main Surgeon Category, infectiousDiagSec					x	
cardiovascularProcSec						x

In order to select the best approach (with or without SMOTE), F1-Score Weighted and the standard deviation of F1-score between classes were taken into account. The model with the highest F1-Score Weighted and the lowest standard deviation is the chosen one. In case both values are higher, the difference between the standard deviations must be less than the difference between the metrics. Otherwise, the model chosen will be the one with the lowest F1-score weighted.

Regarding the standard deviation and the F1-Score Weighted, it is possible to observe that the application of SMOTE always enhances the outcome.

Considering only the results with the application of SMOTE, Table 5.7 shows that 2-class models

Table 5.2: Selected Features according to specialties and general model for Model A, Pre-Surgical.

Model A: Pre-Surgical						
Features	**G**	**S**	**W**	**O**	**U**	**E**
MeanDiagnosis, difSuggestRealSurgery	x	x	x	x	x	x
nrDiagnsec, Age	x		x	x	x	x
neoplasmsDiagSec	x		x		x	x
Maximum Blood Pressure	x		x	x		x
BMI	x			x	x	x
Height			x	x	x	x
healthstatus, Medical device placement proposal, Minimum Blood Pressure	x		x	x	x	
genitourinaryDiagSec, Specific needs for perioperative support or special techniques	x		x		x	
thursday	x		x	x		
Weight	x		x			x
neoplasms	x				x	x
Cholelithiasis, hospitalisation, digestiveDiagSec	x	x				
circulatoryDiagSec, genitourinary	x		x			
musculoDiagSec	x			x		
respiratoryDiagSec	x					x
Tuesday			x	x		
friday			x		x	
nrProcedassoc, cardiovascularProcSec			x			x
AssistantSurgeon2Spec, endocrine, RegionalCentral, Otorhinolaryngology, SDC, Gynaecology-Obstetrics, DeviatedNasalSeptum, conventionalSurg	x					
noLatProcMain, noLatDiagMain, AssistantSurgeon2Res, october, july, leftLatProcMain, rightLatProcMain, rightLatDiagMain, femaleGenitalProcSec			x			
monday, wednesday, muscular, injury, General, ImprovementOrmitigation				x		
AssistantSurgeon1Spec, ResolutionOrHealing, InpatientCare, Gender, leftLatDiagMain, urinaryProcSec, infectiousDiagSec, Readmission_ever					x	
Readmission						x

always have better results for the corresponding classes in the 3-class model. Regarding the blue-coloured class, model B is on average 0.69 percentage points higher than model A. As for the orange-coloured class, model C is on average 18.72 percentage points better than model A and, finally, model C is, on average, 4.68 percentage points higher than model A for the green-coloured class.

Table 5.3: Selected Features according to specialties and general model for Model B, Post-Surgical.

Model B: Post-Surgical						
Features	**G**	**S**	**W**	**O**	**U**	**E**
timeOt, MeanDiagnosis, Height, exitOperating	x	x	x	x	x	x
difSuggestRealSurgery	x	x	x	x		x
Age	x	x	x	x	x	
neoplasmsDiagSec, neoplasms	x	x			x	x
Minimum Blood Pressure	x		x	x	x	
nrDiagnsec		x	x	x	x	
Medical device placement proposal	x			x	x	
genitourinaryDiagSec	x		x		x	
noLatDiagMain, noLatProcMain		x	x		x	
Weight, healthstatus		x			x	x
Cholelithiasis, hospitalisation	x	x				
respiratoryDiagSec	x					x
BMI		x				x
Maximum Blood Pressure			x	x		
AssistantSurgeon2Res, Room3, Otorhinolaryngology, hypertension	x					
SDC, digestiveDiagSec, Main Surgeon Category		x				
timeAn, tuesday, nrProcedassoc, caesareanRoom11, Room4, cardiovascularProcSec			x			
friday, monday, muscular, injury, AssistantSurgeon2Spec, Room6, ResolutionOrHealing, RegionalCentral, General, injuryDiagSec, musculoDiagSec, BilateralLatProcMain, thursday				x		
genitourinary, InpatientCare, Gender, leftLatProcMain, Readmission_ever					x	

5.2.2 Medical Specialties

The results for models that evaluate *per* specialty are presented in this section.

5.2.2.1 Synthetic minority oversampling technique

In order to compare the influence of SMOTE on the application of the model, models were developed without and with its application. Table 5.8 represents these same results in the case of the general model and table 5.9 for specific models. It is important to mention that here the general model is the same as in section 5.2.1, however, the results are evaluated *per* specialty and not as a whole. Analogously to the section on the general model, the criterion of standard deviation and F1-score weighted was also adopted here to choose the best approach. Based on these conditions, the best models (With or without (No) SMOTE) for each case are represented in yellow in table 5.8 and 5.9.

Table 5.4: Selected Features according to specialties and general model for Model B, Pre-Surgical.

Model B: Pre-Surgical						
Features	**G**	**S**	**W**	**O**	**U**	**E**
MeanDiagnosis, difSuggestRealSurgery	x	x	x	x	x	x
Weight, Age, nrDiagnsec	x	x	x	x	x	
neoplasmsDiagSec	x	x	x		x	x
Maximum Blood Pressure	x	x	x	x		
neoplasms	x	x			x	x
BMI	x	x		x		x
Medical device placement proposal	x		x	x	x	
Height, Minimum Blood Pressure	x		x	x	x	
genitourinaryDiagSec, nrProcedassoc	x		x		x	
tuesday	x		x			x
hospitalisation	x	x			x	
SIGLIC_2		x	x	x		
digestiveDiagSec, Cholelithiasis	x	x				
noLatDiagMain,Specific needs for perioperative support or special techniques	x		x			
hypertension	x			x		
healthstatus	x				x	
respiratoryDiagSec	x					x
SDC, Main Surgeon Category		x			x	
thursday, AssistantSurgeon1Res			x	x		
noLatProcMain, rightLatProcMain, june, rightLatDiagMain			x		x	
Readmission			x			x
bloodDiagSec				x	x	
AssistantSurgeon2Res, endocrine, Urology, SIGLIC_3, Otorhinolaryngology, january, nose&mouthProcSec, earProcSec, BilateralLatProcMain	x					
E_V_codes, circulatoryDiagSec, endocrineDiagSec, historyChemoterapy		x				
leftLatProcMain, leftLatDiagMain, femaleGenitalProcSec, maleGenitalProcSec, mentalDiagSec			x			
friday, monday, muscular, injury, AssistantSurgeon2Spec, september, ResolutionOrHealing, RegionalPeripheral, hyperlipidaemia, musculoDiagSec				x		
genitourinary, InpatientCare, Readmission_ever, october, february, infectiousDiagSec, conventionalSurg, august, april					x	
cardiovascularProcSec, earDiagSec						x

Table 5.5: Selected Features according to specialties and general model for Model C, Post-Surgical.

Model C: Post-Surgical						
Features	**G**	**S**	**W**	**O**	**U**	**E**
nrDiagnsec, MeanDiagnosis, difSuggestRealSurgery	x	x	x	x	x	x
Height, exitOperating, timeOt	x	x		x	x	x
Maximum Blood Pressure	x	x	x	x	x	
genitourinaryDiagSec	x	x	x		x	
Age	x	x		x	x	
SDC, hospitalisation	x	x				x
healthstatus	x	x			x	
thursday	x			x		x
digestive, historyChemoterapy	x	x				
Minimum Blood Pressure	x		x			
Gender, musculoDiagSec, monday, tuesday, BMI	x			x		
genitourinary, Room4, Specific needs for perioperative support or special techniques	x				x	
Weight	x					x
hypertension		x		x		
neoplasmsDiagSec, neoplasms		x				x
Main Surgeon Category			x		x	
SIGLIC_2, SIGLIC_1			x			x
wednesday				x		x
overweight					x	x
noLatProcMain, noLatDiagMain, Setúbal, Room5, Medical device placement proposal, may, Gynaecology-Obstetrics, leftLatProcMain, rightLatProcMain, circulatoryDiagSec, infectiousDiagSec, AngiologyAndVascularSurg, Orthopaedics	x					
AssistantSurgeon1Res, endocrine		x				
timeAn, friday, cardiovascularProcSec, nrProcedassoc			x			
AssistantSurgeon2Res, AssistantSurgeon1Spec, Room1, ResolutionOrHealing, november, bloodDiagSec, digestiveDiagSec, april					x	
nose&mouthProcSec, mentalDiagSec, respiratoryDiagSec						x

5.2.2.2 Best Intervals

In this section, the results corresponding to 2 models (post and pre-surgical) are presented, for each of the specialties, either as a general model or as a specific model, in 3 different intervals (Model A, B and C), in a total of 36 models. The objective of this approach is, not only to verify which is the best model for each specialty (whether within a general or a specific model), but also to discover the most advantageous intervals for the results of the models.

Table 5.6: Selected Features according to specialties and general model for Model C, Pre-Surgical.

Model C: Pre-Surgical						
Features	**G**	**S**	**W**	**O**	**U**	**E**
MeanDiagnosis, difSuggestRealSurgery	x	x	x	x	x	x
Maximum Blood Pressure	x	x	x	x	x	
Height	x	x	x	x		x
neoplasmsDiagSec	x	x			x	x
nrDiagnsec, Age	x	x		x		x
healthstatus	x	x		x	x	
neoplasms	x	x	x			x
thursday	x		x	x		x
SDC	x	x			x	
hospitalisation	x	x				x
genitourinaryDiagSec	x		x		x	
musculoDiagSec	x			x		x
mentalDiagSec		x			x	x
Weight			x	x	x	
nrProcedassoc	x		x			
tuesday, Gender, Minimum Blood Pressure	x			x		
infectiousDiagSec, Specific needs for perioperative support or special techniques	x				x	
Main Surgeon Category			x		x	
SIGLIC_2			x			x
BMI, wednesday				x		x
genitourinary, Medical device placement proposal, Orthopaedics, november, Gynaecology-Obstetrics, General Surgery	x					
endocrine, digestiveDiagSec, AssistantSurgeon1Res		x				
friday, august, cardiovascularProcSec			x			
AssistantSurgeon2Spec, AssistantSurgeon1Spec, ResolutionOrHealing, RegionalCentral, hypertension, monday				x		
AssistantSurgeon2Res, Readmission_ever, General, circulatoryDiagSec					x	
SIGLIC_1, march, nose&mouthProcSec, earProcSec						x

In Table 5.10 all the results from the classification models developed are presented. The metric adopted to evaluate the models is the F1-Score, the F1-weighted and the F1-score for classes 0, 1 and 2 and represented in Table 5.10

Before comparing the results with the HBA model, it is necessary to analyse Table 5.10 for a selection of the best model in each specialty, in pre-surgical or post-surgical mode.

For greater ease in comparing the values of model A with models B and C, colours were used.

Table 5.7: General Model results taking into account 3 different scopes: With/Without(No) Smote, post/pre-surgical and intervals. Some rows are coloured to facilitate comparison between the equivalent intervals of the 3-class model (Model A - blue, orange and green) with those of the 2-class model (Model B - blue and Model C - orange and green). W means F1-Weighted. 0, 1 and 2 represent the F1-Score values for class 0, 1 and 2, respectively. STD stands for Standard Deviation and it represents the standard deviation of F1-score between classes.

	General Model											
	Model A				Model B				Model C			
	Post		Pre		Post		Pre		Post		Pre	
	No	With	No	With	No	With	No	With	No	With	No	With
W	74.62	74.89	71.01	71.28	84.21	86.01	82.33	82.74	76.29	77.80	75.90	76.62
0	85.58	83.56	81.43	79.70	81.34	84.38	79.07	80.27	73.51	76.52	73.02	75.28
1	53.76	58.68	50.71	55.68	86.53	87.33	84.97	84.74	78.60	78.86	78.29	77.74
2	75.69	75.5	72.4	71.74	-	-	-	-	-	-	-	-
STD	13.30	10.36	12.89	9.99	2.60	1.48	2.95	2.24	2.54	1.17	2.64	1.23

The detailed analysis of table 5.10 allows the observation that models B and C always have better results compared to model A regarding general surgery and orthopaedics. However, since the improvement of none of the models is not unanimous compared to another in other specialties, a specific analysis for gynaecology-obstetrics, urology and otorhinolaryngology should be carried out:

- **Gynaecology-Obstetrics**: In this specialty, class 0 of model C always yields better results compared to class 1 of model A, both in orange. Regarding the classes in blue, model B is on average 0.50 percentage points better than model A. In turn, taking into account the classes in green, model C is, on average, 2.19 percentage points better than the 3-class model.

- **Urology**: In urology, models B and C have an improved outcome compared to model A, when the orange and green classes are compared. Regarding the blue class, model B is on average 1.90 percentage points higher than the model of 3 classes.

- **Otorhinolaryngology**: Orange and green classes have better results in model C compared to the same classes as model A. Regarding blue class, model B has its result improved by 4.77 percentage values in relation to A.

After the detailed analysis of each specialty, it is acceptable to infer that the 2-class models B and C yield better results compared to 3-classes model A in all clinical specialties.

5.2.2.3 Best Model Selection

Next it is necessary to check for each specialty whether the specialty should be included in the general model with the remaining specialties or should have a specific model. Following the same criteria as in previous sections, the standard deviation and the F1-score weighted were used here to select the best model. Based on these conditions, table 5.11 presents these values, indicating in yellow which is the best approach in each case. For further information about the univariate description of surgical patients stratified by specialty for Model B and C, table A.2 and A.3 should be consulted.

Table 5.8: Comparative table of the general model between the application (represented as "With") and the absence (represented as "No") of SMOTE. The results are present for each specialty, taking into account 2 different scopes: intervals, and post-surgical (represented as 'Post')/pre-surgical (represented as 'Pre'). W means F1-Weighted. 0, 1 and 2 represent the F1-Score values for class 0, 1 and 2, respectively. STD stands for Standard Deviation and it represents the standard deviation of F1-score between classes.

	Model A				Model B				Model C			
	Post		Pre		Post		Pre		Post		Pre	
	No	With	No	With	No	With	No	With	No	With	No	With
General Surgery												
W	77.25	77.28	72.79	74.29	86.43	87.82	81.55	83.35	80.51	80.85	79.00	79.48
0	86.77	85.62	82.42	82.22	87.37	87.07	83.81	82.17	71.77	72.51	70.20	70.63
1	43.81	47.11	33.50	40.00	85.63	88.45	79.64	84.36	85.42	85.53	83.94	84.45
2	83.47	83.23	82.14	83.09	-	-	-	-	-	-	-	-
STD	19.52	17.62	23.00	20.11	0.87	0.69	2.09	1.10	6.83	6.51	6.87	6.91
Gynaecology-Obstetrics												
W	66.98	66.75	60.22	61.29	86.29	86.84	83.17	83.17	64.71	65.41	66.47	67.45
0	79.41	73.91	72.08	69.89	75.00	76.19	65.32	69.56	82.79	82.98	83.49	83.57
1	77.73	77.19	70,66	73.29	90.83	91.11	90.34	88.64	15.91	17.98	20.53	23.91
2	19.51	27.91	14.39	16.09	-	-	-	-	-	-	-	-
STD	27.85	22.50	26.87	26.20	7.92	7.46	12.51	9.54	33.44	32.50	31.48	29.83
Orthopaedics												
W	69.80	69.46	68.47	67.45	83.75	84.64	80.59	80.64	80.04	81.63	79.61	79.20
0	78.48	74.24	73.83	69.14	75.90	77.34	69.38	71.43	68.77	71.60	66.80	68.20
1	32.74	37.65	36.97	41.11	87.71	88.32	86.24	85.28	85.40	86.41	85.71	84.43
2	80.97	81.03	79.48	78.73	-	-	-	-	-	-	-	-
STD	22.17	19.05	18.85	15.96	5.90	5.49	8.43	6.93	8.31	7.41	9.46	8.12
Urology												
W	62.37	62.89	49.37	50.40	77.71	77.67	71.86	71.82	71.19	73.65	65.88	70.40
0	63.64	59.92	36.85	37.36	56.87	58.87	36.52	41.27	78.97	80.49	75.26	77.92
1	64.49	65.27	59.88	61.07	85.51	84.70	85.09	83.26	59.35	63.26	51.62	58.99
2	57.95	62.07	45.21	46.47	-	-	-	-	-	-	-	-
STD	2.90	2.20	9.52	9.77	14.32	12.92	24.29	21.00	9.81	8.61	11.82	9.47
Otorhinolaryngology												
W	92.59	92.59	91.86	92.11	93.04	93.64	92.69	93.17	60.21	62.02	73.64	73.79
0	97.42	97.31	96.89	97.10	96.88	97.42	96.59	96.98	71.79	73.17	81.08	78.95
1	0.00	0.00	0.00	0.00	38.63	40.00	37.33	39.13	50.00	52.17	67.07	69.23
2	45.00	48.00	38.46	40	-	-	-	-	-	-	-	-
STD	39.81	39.73	39.83	39.85	29.13	28.71	29.63	28.93	10.90	10.50	7.01	4.86

Table 5.9: Comparative table of the specific models between the application (represented as "With") and the absence (represented as "No") of SMOTE. The results are present for each specialty, taking into account 2 different scopes: intervals, and post-surgical (represented as 'Post')/pre-surgical (represented as 'Pre'). W means F1-Weighted. 0, 1 and 2 represent the F1-Score values for class 0, 1 and 2, respectively. STD stands for Standard Deviation and it represents the standard deviation of F1-score between classes.

	General Surgery											
	Model A				Model B				Model C			
	Post		Pre		Post		Pre		Post		Pre	
	No	With	No	With	No	With	No	With	No	With	No	With
W	77.40	79.12	68.13	67.98	87.06	89.00	83.64	84.39	81.06	81.35	80.80	81.22
0	87.00	86.44	77.22	77.15	85.99	88.29	82.2	83.31	72.58	74.24	72.29	73.44
1	45.07	53.23	28.92	31.11	87.97	89.61	84.85	85.31	85.83	85.34	85.59	85.59
2	82.89	84.01	78.15	76.60	-	-	-	-	-	-	-	-
STD	18.87	15.12	22.99	21.57	0.99	0.66	1.33	1.00	6.63	5.55	6.65	6.08
	Gynaecology-Obstetrics											
W	64.87	64.66	64.46	67.31	86.17	87.88	86.12	87.29	70.64	70.70	63.23	62.38
0	75.76	76.70	73.47	79.80	74.61	78.85	74.35	77.67	84.21	80.93	81.13	76.46
1	75.06	72.06	76.40	74.33	90.81	91.51	90.84	91.15	34.00	43.08	14.89	24.39
2	21.18	26.79	18.82	29.82	-	-	-	-	-	-	-	-
STD	25.57	22.51	26.48	22.38	8.10	6.33	8.25	6.74	25.11	18.93	33.12	26.04
	Orthopaedics											
W	75.30	76.14	71.43	73.83	85.15	85.18	81.86	82.89	80.39	81.94	79.16	80.89
0	79.02	77.28	75.17	75.40	78.13	79.26	73.48	76.15	69.77	72.31	67.95	71.79
1	52.34	55.28	44.84	52.85	88.69	88.16	86.08	86.28	85.45	86.52	84.49	85.22
2	83.46	85.22	81.3	82.65	-	-	-	-	-	-	-	-
STD	13.74	12.66	15.94	12.69	5.28	4.45	6.30	5.07	7.84	7.11	8.27	6.72
	Urology											
W	64.57	65.20	60.78	60.86	82.11	83.99	75.81	77.84	72.85	72.78	68.35	68.48
0	65.32	66.14	62.79	60.45	66.06	71.37	52.17	59.92	81.00	77.89	77.84	74.67
1	66.49	65.94	65.78	63.40	88.11	88.71	84.65	84.55	60.45	65.00	53.93	59.07
2	60.93	63.20	51.27	57.38	-	-	-	-	-	-	-	-
STD	2.39	1.34	6.26	2.46	11.02	8.67	16.24	12.32	10.28	6.45	11.96	7.80
	Otorhinolaryngology											
W	85.56	85.33	75.76	75.71	86.13	88.62	85.10	85.14	83.92	84.39	78.15	87.50
0	87.00	86.44	77.22	77.15	90.41	93.02	89.77	89.64	82.87	83.87	77.42	86.67
1	45.07	53.23	28.92	31.11	25.45	26.19	18.82	21.24	84.85	84.85	78.79	88.24
2	82.89	84.01	78.15	76.60	-	-	-	-	-	-	-	-
STD	18.87	15.12	22.99	21.57	32.48	33.42	35.48	34.20	0.99	0.49	0.69	0.78

Table 5.10: Results of the classification models by specialty taking into account 3 different scopes: intervals, specific(S)/general(G) and post-surgical (represented as 'Post')/pre-surgical (represented as 'Pre'). Some rows are coloured to facilitate comparison between the equivalent intervals of the 3-class model (Model A - blue, orange and green) with those of the 2-class model (Model B - blue and Model C - orange and green). W means F1-Weighted. 0, 1 and 2 represent the F1-Score values for class 0, 1 and 2, respectively.

	Model A				Model B				Model C			
	Post		Pre		Post		Pre		Post		Pre	
	G	S	G	S	G	S	G	S	G	S	G	S
General Surgery												
W	77.28	79.12	74.29	67.98	87.82	89.00	83.35	84.39	80.85	81.35	79.48	81.22
0	85.62	86.44	82.22	77.15	87.07	88.29	82.17	83.31	72.51	74.24	70.63	73.44
1	47.11	53.23	40.00	31.11	88.45	89.61	84.36	85.31	85.53	85.34	84.45	85.59
2	83.23	84.01	83.09	76.60	-	-	-	-	-	-	-	-
Gynaecology-Obstetrics												
W	66.75	64.66	61.29	67.31	86.84	87.88	83.17	87.29	65.41	70.70	67.45	62.38
0	73.91	76.70	69.89	79.80	76.19	78.85	69.56	77.67	82.98	80.93	83.57	76.46
1	77.19	72.06	73.29	74.33	91.11	91.51	88.64	91.15	17.98	43.08	23.91	24.39
2	27.91	26.79	16.09	29.82	-	-	-	-	-	-	-	-
Orthopaedics												
W	69.46	76.14	67.45	73.83	84.64	85.18	80.64	82.89	81.63	81.94	79.20	80.89
0	74.24	77.28	69.14	75.40	77.34	79.26	71.43	76.15	71.60	72.31	68.20	71.79
1	37.65	55.28	41.11	52.85	88.32	88.16	85.28	86.28	86.41	86.52	84.43	85.22
2	81.03	85.22	78.73	82.65	-	-	-	-	-	-	-	-
Urology												
W	62.89	65.20	50.40	60.86	77.67	83.99	71.82	77.84	73.65	72.78	70.40	68.48
0	59.92	66.14	37.36	60.45	58.87	71.37	41.27	59.92	80.49	77.89	77.92	74.67
1	65.27	65.94	61.07	63.40	84.70	88.71	83.26	84.55	63.26	65.00	58.99	59.07
2	62.07	63.20	46.47	57.38	-	-	-	-	-	-	-	-
Otorhinolaryngology												
W	92.59	85.33	92.11	75.71	93.64	88.62	93.17	85.14	62.02	84.39	73.79	87.50
0	97.31	86.44	97.10	77.15	97.42	93.02	96.98	89.64	73.17	83.87	78.95	86.67
1	0.00	53.23	0.00	31.11	40.00	26.19	39.13	21.24	52.17	84.85	69.23	88.24
2	48.00	84.01	40.00	76.60	-	-	-	-	-	-	-	-

5.3 Current Model Vs. Classification Model

The main ambition of this dissertation were the models represented in Tables 5.12, 5.13 and 5.14 that present the comparison between the best ML models in post-surgical and pre-surgical scope and the current HBA model using the mean of LOS *per ICD9 Diagnosis*. Nonetheless, a contrasting approach is still presented in relation to HBA model, in which the median and mode of LOS *per ICD9 Diagnosis* is

Table 5.11: Comparative table between all the classification models developed regarding its F1-Score Weighted and Standard Deviation (STD) of F1-score between classes. The best models for each specialty, in each scope - pre-surgical (represented as 'Pre') and post-surgical (represented as 'Post') - are shown in yellow. G represents the general model and S the specific model.

	Model B				Model C			
	Post		Pre		Post		Pre	
	G	S	G	S	G	S	G	S
General Surgery								
F1-Weighted	87.82	89.00	83.35	84.39	80.85	81.35	79.48	81.22
Standard Deviation	0.69	0.66	1.10	1	6.51	5.55	6.91	6.08
Gynaecology-Obstetrics								
F1-Weighted	86.84	87.88	83.17	87.29	65.41	70.70	67.45	62.38
Standard Deviation	7.46	6.33	9.54	6.74	32.5	18.93	29.83	26.04
Orthopaedics								
F1-Weighted	84.64	85.18	80.64	82.89	81.63	81.94	79.20	80.89
Standard Deviation	5.49	4.45	6.93	5.07	7.41	7.11	8.12	6.72
Urology								
F1-Weighted	77.67	83.99	71.82	77.84	73.65	72.78	70.40	68.48
Standard Deviation	12.92	8.67	21.00	12.32	8.61	6.45	9.47	7.80
Otorhinolaryngology								
F1-Weighted	93.64	88.62	93.17	85.14	62.02	84.39	73.79	87.50
Standard Deviation	28.71	33.42	28.93	34.20	10.50	0.49	4.86	0.78

also considered. This point aims to effectively compare the adoption of a simpler model, with a single variable, to a ML model.

Regarding the evaluation of the general model, pre-surgical and post-surgical surgical results are shown in table 5.12, whereas the evaluation taking into account each specialty is showed in table 5.13 (post-surgical) and 5.14 (pre-surgical).

5.4 Patient-Related and Procedure-Related Risk Factors

The last ambition of this dissertation is to compare the results of models only containing patient-related features with models only having procedure-related features. As a complement, the results of HBA model and the model that considers all features (patient and procedure-related) are also shown in Table 5.15 and 5.16, in post-surgical and pre-surgical context, respectively.

Table 5.12: Comparative table between the developed general model and the current model in place at HBA, including median and mode. F1 stands for the F1-score.

	General Model - Post Surgical									
	Model B					Model C				
	HBA			RF	# Cases	HBA			RF	# Cases
	Mean	Median	Mode			Mean	Median	Mode		
F1 Weighted	80.62	83.52	81.17	86.01	2766	68.01	66.62	63.06	77.80	1528
F1, Class 0	76.05	81.69	81.10	84.38	1238	60.40	58.76	58.95	76.52	693
F1, Class 1	84.33	85.01	81.23	87.33	1528	74.33	73.15	66.47	78.86	835
	General Model - Pre Surgical									
F1 Weighted	80.62	83.52	81.17	82.74	2766	68.01	66.62	63.06	76.62	1528
F1, Class 0	76.05	81.69	81.10	80.27	1238	60.40	58.76	58.95	75.28	693
F1, Class 1	84.33	85.01	81.23	84.74	1528	74.33	73.15	66.47	77.74	835

Table 5.13: Comparative table of post-surgical scope between the best ML model (by specialty) and the current model in place at HBA, including median and mode. F1 stands for the F1-score.

	General Surgery									
	Model B					Model C				
	HBA			RF	# Cases	HBA			RF	# Cases
	Mean	Median	Mode			Mean	Median	Mode		
F1 Weighted	84.47	86.03	84.99	89.00	672	73.94	72.17	69.09	81.35	364
F1, Class 0	82.55	85.03	84.72	88.29	308	57.69	53.54	50.93	74.24	131
F1, Class 1	86.10	86.87	85.21	89.61	364	83.08	82.64	79.30	85.34	233
	Gynaecology-Obstetrics									
F1 Weighted	77.41	83.03	82.14	87.88	363	68.25	62.14	60.42	70.70	259
F1, Class 0	51.70	67.05	70.27	78.85	104	83.05	77.00	76.85	80.93	189
F1, Class 1	87.74	89.45	86.90	91.51	259	28.28	22.03	16.07	43.08	70
	Orthopaedics									
F1 Weighted	76.81	80.66	81.22	85.18	597	74.27	70.14	59.83	81.94	397
F1, Class 0	61.13	72.47	75.16	79.26	200	57.14	47.39	40.88	72.31	128
F1, Class 1	84.71	84.79	84.27	88.16	397	82.42	80.96	68.85	86.52	269
	Urology									
F1 Weighted	65.29	71.72	73.66	83.99	426	44.66	54.90	54.26	72.78	310
F1, Class 0	15.94	45.87	59.59	71.37	116	37.01	55.13	64.81	77.89	187
F1, Class 1	83.75	81.39	78.93	88.71	310	56.28	54.55	38.22	65.00	123
	Otorhinolaryngology									
F1 Weighted	90.39	92.80	90.83	93.64	486	36.66	34.12	35.42	84.39	32
F1, Class 0	94.69	96.99	96.47	97.42	454	17.39	9.52	0.00	83.87	15
F1, Class 1	29.41	33.33	10.81	40.00	32	53.66	55.81	10.81	84.85	17

Table 5.14: Comparative table of pre-surgical scope between the best ML model (by specialty) and the current model in place at HBA, including median and mode. F1 stands for the F1-score.

	Model B					Model C				
	HBA			RF	# Cases	HBA			RF	# Cases
	Mean	Median	Mode			Mean	Median	Mode		
General Surgery										
F1 Weighted	84.47	86.03	84.99	84.39	672	73.94	72.17	69.09	81.22	364
F1, Class 0	82.55	85.03	84.72	83.31	308	57.69	53.54	50.93	73.44	131
F1, Class 1	86.1	86.87	85.21	85.31	364	83.08	82.64	79.30	85.59	233
Gynaecology-Obstetrics										
F1 Weighted	77.41	83.03	82.14	87.29	363	68.25	62.14	60.42	67.45	259
F1, Class 0	51.70	67.05	70.27	77.67	104	83.05	77.00	76.85	83.57	189
F1, Class 1	87.74	89.45	86.90	91.15	259	28.28	22.03	16.07	23.91	70
Orthopaedics										
F1 Weighted	76.81	80.66	81.22	82.89	597	74.27	70.14	59.83	80.89	397
F1, Class 0	61.13	72.47	75.16	76.15	200	57.14	47.39	40.88	71.79	128
F1, Class 1	84.71	84.79	84.27	86.28	397	82.42	80.96	68.85	85.22	269
Urology										
F1 Weighted	65.29	71.72	73.66	77.84	426	44.66	54.90	54.26	70.40	310
F1, Class 0	15.94	45.87	59.59	59.92	116	37.01	55.13	64.81	77.92	187
F1, Class 1	83.75	81.39	78.93	84.55	310	56.28	54.55	38.22	58.99	123
Otorhinolaryngology										
F1 Weighted	90.39	92.80	90.83	93.17	486	36.66	34.12	35.42	87.50	32
F1, Class 0	94.69	96.99	96.47	96.98	454	17.39	9.52	0.00	86.67	15
F1, Class 1	29.41	33.33	10.81	39.13	32	53.66	55.81	66.67	88.24	17

Table 5.15: Comparative table of post-surgical scope between the current model in force at HBA, the best ML model: patient-related features model and procedure-related features model. In this table, “All” is referring to the ML model combining patient and procedure-related features. F1 stands for the F1-score.

	Model B					Model C				
	HBA	RF			# Cases	HBA	RF			# Cases
	Mean	Pac.	Proc.	All		Mean	Pac.	Proc.	All	
General Surgery										
F1 Weighted	84.47	79.35	88.26	89.00	672	73.94	78.86	77.32	81.35	364
F1, Class 0	82.55	78.25	87.40	88.29	308	57.69	70.72	69.14	74.24	131
F1, Class 1	86.10	80.28	88.98	89.61	364	83.08	83.44	81.92	85.34	233
Gynaecology-Obstetrics										
F1 Weighted	77.41	75.37	87.84	87.88	363	68.25	67.73	70.01	70.70	259
F1, Class 0	51.70	55.56	78.64	78.85	104	83.05	78.22	80.00	80.93	189
F1, Class 1	87.74	83.33	91.54	91.51	259	28.28	39.42	43.03	43.08	70
Orthopaedics										
F1 Weighted	76.81	67.16	84.07	85.18	597	74.27	68.98	78.19	81.94	397
F1, Class 0	61.13	52.63	77.90	79.26	200	57.14	54.21	66.67	72.31	128
F1, Class 1	84.71	74.48	87.18	88.16	397	82.42	76.01	83.68	86.52	269
Urology										
F1 Weighted	65.29	76.12	74.87	83.99	426	44.66	60.97	69.72	72.78	310
F1, Class 0	15.94	55.51	57.03	71.37	116	37.01	68.75	74.73	77.89	187
F1, Class 1	83.75	83.84	81.54	88.71	310	56.28	49.15	62.10	65.00	123
Otorhinolaryngology										
F1 Weighted	90.39	87.68	87.46	93.64	486	36.66	71.79	74.80	84.39	32
F1, Class 0	94.69	91.87	92.06	97.42	454	17.39	68.97	76.47	83.87	15
F1, Class 1	29.41	28.28	22.22	40.00	32	53.66	74.29	73.33	84.85	17

Table 5.16: Comparative table of pre-surgicalscope between the current model in force at HBA, the best ML model: patient-related features model and procedure-related features model. In this table, “All” is referring to the ML model combining patient and procedure-related features. F1 stands for the F1-score.

	General Surgery									
	Model B					Model C				
	HBA	RF			# Cases	HBA	RF			# Cases
	Mean	Pac.	Proc.	All		Mean	Pac.	Proc.	All	
F1 Weighted	84.47	80.83	80.36	84.39	672	73.94	77.14	73.37	81.22	364
F1, Class 0	82.55	80.25	78.57	83.31	308	57.69	67.95	63.12	73.44	131
F1, Class 1	86.10	81.33	81.87	85.31	364	83.08	82.30	79.14	85.59	233
	Gynaecology-Obstetrics									
F1 Weighted	77.41	77.96	81.21	87.29	363	68.25	64.63	64.01	67.45	259
F1, Class 0	51.7	61.54	66.99	77.67	104	83.05	78.91	75.46	83.57	189
F1, Class 1	87.74	84.56	86.92	91.15	259	28.28	26.09	33.09	23.91	70
	Orthopaedics									
F1 Weighted	76.81	64.54	81.04	82.89	597	74.27	65.53	74.68	80.89	397
F1, Class 0	61.13	49.05	72.33	76.15	200	57.14	46.69	61.30	71.79	128
F1, Class 1	84.71	72.35	85.42	86.28	397	82.42	74.49	81.05	85.22	269
	Urology									
F1 Weighted	65.29	74.62	68.61	77.84	426	44.66	66.41	60.35	70.40	310
F1, Class 0	15.94	50.70	44.72	59.92	116	37.01	73.39	65.17	77.92	187
F1, Class 1	83.75	83.57	77.56	84.55	310	56.28	55.79	53.03	58.99	123
	Otorhinolaryngology									
F1 Weighted	90.39	89.71	84.67	93.17	486	36.66	75.00	87.50	87.50	32
F1, Class 0	94.69	94.10	89.28	96.98	454	17.39	73.33	87.50	86.67	15
F1, Class 1	29.41	27.40	19.30	39.17	32	53.66	76.47	87.50	88.24	17

Chapter 6

Discussion

The goal of this work was to compare results between the developed ML models and the HBA current model. Some decisions were made during the development of the models.

RF was the algorithm chosen because it is considered to have potential due to its efficiency. According to the literature review of this work, RF was the algorithm that obtained the best results for the classification task for some of the articles [41, 42, 44, 45]. Also, the fact that RF is capable of handling binary, categorical and numerical variables and also dealing with outliers and being indifferent to non-linear features contributed to its choice.

Since the dataset of this work was imbalanced, SMOTE was used in the training phase. The choice for this method is justified by its approach of creating new synthetic instances and not just copies, making the samples more general. Also due to this imbalance, F1-score was the metric adopted because it is the harmonic mean of precision and recall. Although other metrics could be used, it was decided to only use F1-score since it offers a more lucid understanding of the results.

Feature selection comprised one part of this work. From the analysis from table 5.1 to 5.6, it is possible to verify that the variable *MeanDiagnosis* is common to all models, effectively showing its importance. This demonstrates, regardless of the final results, that the use of this feature by the current HBA model is probably a good option. Also *difSuggestRealSurgery* is present in all or almost every model. This leads to the conclusion that the time difference between the day of surgery and the suggested day has an effect on the LOS, regardless of the specialty. As there is sometimes a long waiting list, patients do not have the surgery scheduled for the time they really need it. This can lead to an aggravation of the health problems resulting in a longer hospital stay.

Age, *nrDiagnSec*, *neoplasmsDiagSec* and *Maximum Blood Pressure* are also selected by most models. In fact, an older person tends to stay in hospital longer as well as a patient with a higher number of diagnoses that may indicate a greater degree of comorbidities. In turn, surgeries for patients with neoplasms may also result in more complications, increasing the LOS.

Likewise, patients with a higher *Maximum Blood Pressure* may indicate the existence of other pathologies, which may increase the degree of risk of surgery and, consequently, the LOS.

It is possible to note that, with the exception of model C, in post-surgical and pre-surgical models (tables 5.5 and 5.6, respectively), otorhinolaryngology is always considered as an important feature in the general model. This indicates that the majority of these patients belong to a specific class, in this case, most otorhinolaryngology patients have a LOS of 1 day. This can be ascertained from table A.2.

Regarding tables 5.1, 5.3 and 5.5, it is also possible to denote that *timeOt* and *exitOperating* are selected in almost or every model. This can be justified because it is implied that a longer surgery is a more complex and severe surgery, with exceptions, leading to a longer stay. In addition, the time when the patient leaves the operating theatre will also be important because, if a surgery ends late, they will inevitably have to stay longer in hospital.

Regarding the influence on SMOTE, from tables 5.7, 5.8 and 5.9, it is possible to conclude that the application of SMOTE always enhances the outcome. This result is expected due to the imbalance between classes. So the existence of a larger number of samples in previously minority classes during the training phase can lead to an improved forecast in the testing phase.

During this work, two approaches were developed - 2-class and 3-class model - in order to assess the best model between 3-classes model (A) and 2-classes model (B and C). Regarding the general model, from table 5.7, it is possible to infer that a 2-class model enhances the results in relation to a 3-class model. Regarding the models evaluated according to specialty, table 5.10, the distinction for the 2-class model regarding to general surgery and orthopaedics specialties is clear when comparing the results between these models. Notwithstanding, for the remaining specialties a more detailed analysis is required, carried out in section 5.2.2.2. This study highlights the improvement of 2-class models compared to 3-class models, in all specialties, both in pre-surgical and post-surgical environments. The reason behind this decision may be related to the enhanced performance of RF algorithm when dealing with a 2-class problem than a multi-class problem.

Due to the need demanded by hospital managers for an increased importance to the LOS depending on the specialty, 2 types of models were created, specific and general, so that the best possible results could be obtained. In response to this, a thorough analysis by specialty was completed in order to select the best models for pre-surgical and post-surgical environments, as shown in Table 5.11. According to this table, general surgery, gynaecology-obstetrics, orthopaedics and urology benefit the specific model over the general. Through the analysis of Table 5.3, one possible reason for this situation is due to the fact that few features are common to the specific model and general model and also the existence of a subset of exclusive features. Otorhinolaryngology is the only model in these described conditions that presents better results when inserted in the model with other specialties. Table 5.3 can justify this preference since there is no exclusive subset of features for this specialty.

Likewise, the preferred models for specialties in pre-surgical mode remain the same. In the case of gynaecology-obstetrics, orthopaedics and urology there are extensive subsets of exclusive variables for the specific models, as shown in Table 5.4, which may justify the choice for the specific model. In turn, although otorhinolaryngology have a small subset of exclusive features, again benefits more from the interaction with other specialties, favouring the results of the general model.

In turn, in relation to model C, in the post-surgical mode, the specific models obtained better results than the general model for all specialties. Table 5.5 shows that gynaecology-obstetrics, urology and otorhinolaryngology have an extensive subset of exclusive features, which may then justify the choice for the specific model. However, although the remaining specialties analysed do not have a large set of exclusive variables or even have one, in model C, the contact with other specialties in the general model may justify the disadvantageous relationship when included in a general model.

As for the pre-surgical mode of Model C, as shown in table 5.6, orthopaedics and otorhinolaryngology present a large number of exclusive features, which, again, may serve as a justification for choosing the specific model. Conversely, although general surgery also presents better results regarding the specific model, there is no information in table 5.6 to substantiate such a choice. One possible reason for this interval is the detrimental relationship between this specialty and others specialties inserted in the same model. Finally, it is possible to denote that gynaecology-obstetrics and urology obtain better results when included in the general model. As similar as previous cases, these models may benefit from contact with other clinical areas.

It is important to note that sometimes the difference in results between the two types of models is not large (sometimes less that 1 percentage point), so there may not be an unequivocal justification of clinical nature for this, given the very close performance.

In summary, the analysis of Table 5.11 demonstrates that, in the majority of the models, the existence of exclusively post-surgical variables makes the specific model preferable in the case of post-surgical models. However, in the absence of post-surgical variables, that is, in the case of pre-surgical models, there is an increased trend in the choice of the general model, thus benefiting from the interaction between different clinical areas. This event may imply the relevance of the presence of exclusively post-surgical variables in ML models.

As the main objective of this work required, the comparative analyse between the ML models developed and the HBA model, there are presented in Table 5.12, 5.13 and 5.14. In order to compare the results between simple models using only one variable and ML models, not only the hospital approach (mean), but also the median and mode were considered.

Regarding Table 5.12, is it possible to conclude that, taking into account only the mean, the developed ML model is always better. For the median, the result is similar, except for the pre-surgical model B, where the HBA model is better. Regarding mode, also the ML model have better results except for class 0 of the pre-surgical model. Converting to concrete numbers, regarding model B, it is possible to predict correctly more 137 and 48 cases in post-surgical and pre-surgical mode, respectively, adopting the developed model than when using the HBA model. This represents 11% and 4% of the dataset considered, respectively. In addition, for model C, it is possible to correctly predict an additional 136 (20% of the dataset) and 118 (17% of the dataset) cases in post-surgical and pre-surgical environment, respectively, adopting the ML model.

Also, after choosing the best models for each specialty, it can be concluded that the development of ML models shows better results in every specialty, comparing with HBA model, as can be perceived in table 5.13 and 5.14, with an average improvement of 13.87 percentage points for the postoperative model (table 5.13) and 12.32 for the preoperative model (table 5.14).

More in particular, regarding table 5.13, Model B, RF algorithm is, in average, 9.06, 22.72 and 5.01 percentage points better than HBA method, in relation to F1-Weighted, F1 – Class 0 and F1 – Class 1, respectively. In gross numbers, this converges to 169 additional correct cases compared to the HBA model. These numbers represents about 7% of the dataset evaluated considered (2544 cases). In turn, in Model C, the model developed is, on average, 18.68, 27.39 and 12.21 percentage points better than HBA model, regarding F1-Weighted, F1-score for Class 0 and F1-score for Class 1, respectively. This percentage translates into about 134 new correct cases in relation to the HBA model, what it represents about 10% of the total number of cases in question (1362 cases).

Similarly, the improvement in the ML model in relation to the HBA model is also observed in table 5.14. In particular, in Model B, there was an enhancement of, on average, 6.71, 18.12 and 3.37 percentage points in relation to F1-Weighted, F1-score for Class 0 and F1-score for Class 1, respectively, when using the ML Model. This represents an increase in 92 correctly predicted cases or 4% of the dataset. Also in Model C, there was an improvement of, on average, 17.94, 28.22 and 7.65 percentage points regarding F1-Weighted, F1-score for Class 0 and F1-score for Class 1, respectively, when the developed model was selected. This depicts 129 new cases predicted correctly, a number that represents 9% of the dataset.

The number of cases successfully predicted were deducted from Table A.4, in *Appendix* section, corresponding to the confusion matrices of the HBA model and the developed ML models. For further information, this table should be consulted.

Although these values represent a very small percentage, they are a reflection of greater control of bed management, which leads to a decrease in hospital expenditure. This dataset only represents a small number of the current number of surgeries in the hospital in question, so, working with a larger number of samples may lead to a better perception of the real improvement when adopting ML models.

It is also possible to verify an improvement in the post-surgical model compared to the pre-surgical model. This improvement is reflected in 2.82 percentage points for model B and 0.74 percentage points for model C. Likewise, taking into account the general model, there is also an improvement in the results in the post-surgical scope, in this case of 3.27 and 1.17 percentage points in model B and C, respectively. This may be due to the fact that the exclusive variables of the surgical model are important variables for the correct forecast of LOS.

The work also evaluated the impact of patient and procedure related variables in the LOS.

Tables 5.15 and 5.16 represent the growing importance in the insertion of data of different natures. In every specialty, the ML model presents a better performance when its input comprises data related to the patient and also data related to procedures, as can be confirmed by the tables under analysis. Regarding Table 5.15, it is possible to verify that the integration of the two types of features in a ML model adds an average improvement of 9.68 percentage points compared to the exclusive use of patient-related variables and 3.83 to the use of procedure-related variables. In turn, in Table 5.16, it is also possible to observe an average improvement of 7.67 percentage points compared to the model that only uses patient features and 5.72 for the model with only procedure-related variables, when both variables are incorporated into the model.

In Table 5.15, it is also possible to verify that, when only patient variables are used, the performance of the model is generally lower than when using the procedure variables by 5.85 percentage points (F1-Weighted). Also, the analysis of Table 5.15 reflects that it is better to use a ML model with input corresponding only to procedure variables than to use the current model of the hospital, with an improvement of 10.04 percentage points regarding F1-Weighted. Equivalently, the same is true about patient-related variables, showing an improvement of 4.19 percentage points.

Similarly, from Table 5.16 (pre-surgical environment), it is possible to denote an enhancement of 6.37 and 4.42 percentage points, when using the procedure-related model and patient-related model, respectively, compared to the HBA model. However, it is possible to conclude that, as verified in Table 5.15, when only patient-related features are used, in general, the performance of the model is lower than when using the procedure variables by 1.94 percentage points (F1-Weighted). This shows the enormous capabilities of this type of model, even if limited to a certain nature of variables.

It is not possible to make a direct comparison with the study described in the literature review since the metric selected in this case was AIC [54], however, it is possible to draw the same conclusion as the article: in general, the procedure-related variables have more influence on the LOS than patient-related variables.

Since the results *per* specialty are not presented in the literature and the metric adopted is different, a direct comparison between the previously described studies and the models developed is not possible. Also, since the features provided in the studies referenced in the literature review are not the same those used in this work (or are not disclosed), it is not possible to make an exhaustive analysis regarding the comparison of the selected features. However, it is possible, at least, to infer that age was one of the features that proved to be relevant in the models developed, as well as in some of the studies referenced [38, 39].

Despite the promising results, this dissertation presents some limitations. The dataset used was extracted from operational database with other purposes and in many cases with non-mandatory fields. In one of the first data cleaning tasks, over half of the dataset was deleted because there was no indication of discharge dates. This may have happened because the patients were operated on as an outpatient, however, as it cannot be clarified, the decision to remove all cases with this missing feature was made.

The definition of some features may also vary from institution to institution. For example, in one study the time of surgery was defined as the time difference between the last suture and the first cut, however, in other entities, this time may be the difference between leaving and entering the operating theatre. These small differences in some terms may also compromise the results between different hospitals. The classification of the main diagnosis in ICD-9-CM, but the secondary diagnoses in ICD-10-CM, as well as the classification of the main procedure in ICD-10-PCS and the associated procedures in ICD-9-CM, may also differ from the nomenclature adopted in other institutions and, consequently, originate different results.

difAdmissionDaySurgery shows that over 80% of patiens are admitted two or more days before surgery. However, according to HBA, a vast majority of the patients are admitted in the same day or one day before the surgery. The presence of inconsistent values for the variable *difAdmissionDaySurgery* (created and deleted in data preparation phase) demonstrates the possibility of other types of errors in the data.

The existence of social cases, where patients may stay in the hospital longer than the predicted due to non-health reasons, can also weaken the results since the forecast for the LOS is not fulfilled by factors outside the patient's health scope.

In order to fill the possible gaps in this panoply of data, one of the objectives of the meetings with hospital managers was to clarify the existence of possible unreliable variables. Although this task has prevented the insertion of some incorrect data, this objective may not have been fully achieved. In some features, commands were inserted to avoid the existence of unrealistic values, such as, patients weighing 999 Kilograms.

It is, in any case, crucial to be aware of the possibility of the existence of dubious data, such as not updating last minute changes in the patient's surgical profile. This also highlights the challenges of using data with a secondary use in a real world scenario. Consequently, data quality could have been improved if a suitable dataset would have been available to be fulfilled with the LOS goal.

The use of other ML algorithms or even deep learning can considerably improve the results, as alternative to RF. The approach used to select the most important features can also be modified. Also, new techniques or criteria may be added in order to test which is the most suitable.

Most of the cases that did not contain values in some important features were erased because the existence of possible errors not directly detectable in other variables of the subject was considered. Even so, the adoption of another method for substituting missing values, such as using the K-Nearest Neighbors, as mentioned in chapter 2, could have improved the results and increased the final dataset size.

Categories that grouped a smaller number of diagnoses could also have been used, in order to make it more specific. For example, instead of having created a variable that would accommodate all cases with diagnoses related to the circulatory system, two attributes could have been created: one for cardiac diseases and one for other cardiac-related problems (such as diseases in the blood vessels).

The adoption of a classification approach also limits the outcomes of this study since, although the results are promising, these models place, for example, in the same interval patients who remain hospitalised 4 or 5 days. Later studies with a regressive approach can also enrich this work and achieve better outcomes.

Chapter 7

Conclusion

The bed management problem is a current and not easy-to-solve problem. One of the best ways to tackle the excessive cost of the hospital in this regard is the timely control of the patients' LOS. In this context, HBA presents a method that allows, with some precision to control the patients' stay, based on the mean LOS *per* diagnosis. However, in some specialties this precision is extremely low. The ambition of this project was to develop a model that predicted how long patients, who required surgery, remained in the hospital. Although with some limitations that can be overcome with future work, the results of this dissertation discloses a promising view about ML models in the hospital reality. These showed better performances compared to the model currently in place, allowing the correct additional forecast of hundreds of patients.

This work aimed to fill some of the gaps in the literature, such as, when considering patient and procedure-related features. In addition, since there are not many studies regarding the prediction of hospital stay in surgical cases, this dissertation also aimed to enrich this type of approach.

This dissertation also concerned about alerting to the influence of the work of health professionals on the development of these models, since only with their collaboration the data will be viable and reliable. Fortunately, nowadays EHR are becoming a reality in our country. Thus, taking into account the results presented, this work also contributes to an evolution in the development of artificial intelligence models that can improve the results of current models used in Portugal.

The results achieved in this work make possible to conclude that post-surgical models provide better predictor outcomes than pre-surgical ones. Also, according to the tables analysed in the previous section, it is still possible to infer that a combination of patient data with procedure data allows a better estimate of LOS.

This work is not presented as a final model, as it lacks further studies and improvements, nevertheless, it represents an indication of the promising path of artificial intelligence in health.

Finally, it is possible to conclude that the objective for which this dissertation was stated has been fulfilled. The future implementation of artificial intelligence models in hospitals may, hopefully, alleviate the current problem in bed management and mitigate the financial burden that this issue brings to hospitals.

References

[1] C. Kruse, A. Stein, H. Thomas, and H. Kaur, “The use of electronic health records to support population health: A systematic review of the literature,” *Journal of Medical Systems*, vol. 42, 2018.

[2] A. R. R. Andrade, “Adoption of electronic health records in the portuguese healthcare system in the presence of privacy concerns,” Master’s thesis, Universidade Católica Portuguesa, 2017.

[3] F. Sadoughi, T. Khodaveisi, and H. Ahmadi, “The used theories for the adoption of electronic health record: a systematic literature review,” *Health and Technology*, vol. 9, 2018.

[4] N. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, 2012.

[5] V. Raposo, “Electronic health records: Is it a risk worth taking in healthcare delivery?” *GMS Health Technology Assessment*, vol. 11, 2015.

[6] O. Ayaad, A. Alloubani, E. ALhajaa, M. Farhan, S. Abuseif, A. Al Hroub, and L. Akhu-Zaheya, “The role of electronic medical records in improving the quality of health care services: Comparative study,” *International Journal of Medical Informatics*, vol. 127, 2019.

[7] H. P. D. F. Fonseca, “Gerir as camas hospitalares sempre a pensar no utente,” http://hff.min-saude.pt/gerir-as-camas-hospitalares-sempre-a-pensar-no-utente/, 2019, accessed on 19/09/2020.

[8] R. Schmidt, S. Geisler, and C. Spreckelsen, “Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources.” *BMC Medical Informatics and Decision Making*, vol. 13, 2013.

[9] S. Suthummanon and V. Omachonu, “Drg-based cost minimization models: Applications in a hospital environment,” *Health Care Management Science*, vol. 7, 2004.

[10] S. Aghajani and M. Kargari, “Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department,” *Hospital Practices and Research*, vol. 1, 2016.

[11] M. Guzman Castillo, “Modelling patient length of stay in public hospitals in mexico,” Ph.D. dissertation, University of Southampton, 2012.

[12] G. Schuh, G. Reinhart, J.-P. Prote, F. Sauermann, J. Horsthofer, F. Oppolzer, and D. Knoll, "Data mining definitions and applications for the management of production complexity," *Procedia CIRP*, vol. 81, 2019.

[13] N. M. d. C. Cunha, "Metodologia de selecção de segmentações diversificadas: Um caso de aplicação de técnicas de data mining em dados de consumo para avaliação de portfólios de cartões bancários," Master's thesis, Faculdade de Economia da Universidade Do Porto, 2009.

[14] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000.

[15] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "Dmme: Data mining methodology for engineering applications – a holistic extension to the crisp-dm model," *Procedia CIRP*, vol. 79, 2019.

[16] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," *Springer*, vol. 1, 2001.

[17] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.

[18] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. USA: Morgan Kaufmann Publishers Inc., 2011.

[19] G. Batista and M.-C. Monard, "A study of k-nearest neighbour as an imputation method." in *Hybrid Intelligent Systems*, 2002.

[20] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*, 1st ed. USA: O'Reilly Media, Inc., 2018.

[21] A. Suppers, A. van Gool, and H. Wessels, "Integrated chemometrics and statistics to drive successful proteomics biomarker discovery," *Proteomes*, vol. 6, 2018.

[22] M. Kuhn and K. Johnson, *Applied predictive modeling*, 1st ed. USA: Springer, 2013.

[23] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics (Oxford, England)*, vol. 23, 2007.

[24] M. Beraha, A. M. Metelli, M. Papini, A. Tirinzoni, and M. Restelli, "Feature selection via mutual information: New theoretical insights," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019.

[25] C. Molnar, "Interpretable machine learning: A guide for making black box models explainable," https://christophm.github.io/interpretable-ml-book/, 2019, accessed on 17/10/2020.

[26] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, 2010.

[27] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, 2002.

[28] X. Li, W. Chen, Q. Zhang, and L. Wu, “Building auto-encoder intrusion detection system based on random forest feature selection,” *Computers Security*, vol. 95, 2020.

[29] H. Liu and E. Haig, “Induction of classification rules by gini-index based rule generation,” *Information Sciences*, vol. 436-437, 2018.

[30] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, 2001.

[31] M. Blackwell, “Multiple hypothesis testing: The f-test,” https://mattblackwell.org/files/teaching/ftests.pdf, 2008, accessed on 16/10/2020.

[32] F. Doshi-Velez and B. Kim, “A roadmap for a rigorous science of interpretability,” *ArXiv*, vol. abs/1702.08608, 2017.

[33] A. NG, “Machine learning yearning: Technical strategy for ai engineers, in the era of deep learning,” https://d2wvfoqc9gyqzf.cloudfront.net/content/uploads/2018/09/Ng-MLY01-13.pdf, 2018, accessed on 16/10/2020.

[34] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st ed. Wiley-IEEE Press, 2013.

[35] M. Maalouf and T. Trafalis, “Rare events and imbalanced datasets: An overview,” *IJDMMM*, vol. 3, 2015.

[36] S. J. Dattagupta, “A performance comparison of oversampling methods for data generation in imbalanced learning tasks,” Master’s thesis, Nova Information Management School, 2017.

[37] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res. (JAIR)*, vol. 16, 2002.

[38] O. Khosravizadeh, S. Vatankhah, P. Bastani, R. Kalhor, S. Alirezaei, and F. Doosty, “Factors affecting length of stay in teaching hospitals of a middle-income country,” *Electronic Physician*, vol. 8, 2016.

[39] A. Mohammad, Z. Asghar, R. Abas, R. Fatemeh, and A. H. F. Elah, “Analysis of factors affecting length of stay in public hospitals in lorestan province, iran,” *Hakim Research Journal*, vol. 12, 2010.

[40] T. Kapadohos, E. Angelopoulos, I. Vasileiadis, S. Nanas, A. Kotanidou, A. Karabinis, K. Marathias, and C. Routsi, “Determinants of prolonged intensive care unit stay in patients after cardiac surgery: A prospective observational study,” *Journal of Thoracic Disease*, vol. 9, 2017.

[41] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, “Analysis of length of hospital stay using electronic health records: A statistical and data mining approach,” *PLOS ONE*, vol. 13, 2018.

[42] T. Daghistani, R. El Shawi, S. Sakr, A. Ahmed, A. Thwayee, and M. Al-Mallah, “Predictors of in-hospital length of stay among cardiac patients: A machine learning approach,” *International Journal of Cardiology*, vol. 288, 2019.

[43] P. Rezaei, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare Informatics Research*, vol. 19, 2013.

[44] M.-T. Chuang, Y.-H. Hu, and C.-l. Lo, "Predicting the prolonged length of stay of general surgery patients: A supervised learning approach," *International Transactions in Operational Research*, vol. 25, 2016.

[45] M.-T. Chuang, Y.-H. Hu, C.-F. Tsai, C.-l. Lo, and W.-C. Lin, "The identification of prolonged length of stay for surgery patients," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015.

[46] H. FadyS, T. Taha, and M. M., "Applying data mining technique for predicting incubator length of stay in egypt and usa," *International Journal of Computer Applications*, vol. 71, 2013.

[47] P. Grünwald, *The Minimum Description Length Principle*. USA: The MIT Press, 2007.

[48] M. Elbattah and O. Molloy, "Using machine learning to predict length of stay and discharge destination for hip-fracture patients," in *Lecture Notes in Networks and Systems*, 2016.

[49] P.-F. Tsai, P.-C. Chen, Y.-Y. Chen, H.-Y. Song, H.-M. Lin, F.-M. Lin, and Q.-P. Huang, "Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network," *Journal of Healthcare Engineering*, vol. 2016, 2016.

[50] C.-S. Yang, C.-P. Wei, Y. Chi chuan, and J.-Y. Schoung, "Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages," *Decision Support Systems*, vol. 50, 2010.

[51] H. Maharlou, S. Rostam Niakan Kalhori, S. Shahbazi, and R. Ravangard, "Predicting length of stay in intensive care units after cardiac surgery: Comparison of artificial neural networks and adaptive neuro-fuzzy system," *Healthcare Informatics Research*, vol. 24, 2018.

[52] R. Lafaro, S. Pothula, K. Kubal, V. Pothula, S. Yuan, D. Maerz, L. Montes, S. Oleszkiewicz, A. Yusupov, R. Perline, and M. Inchiosa, "Neural network prediction of icu length of stay following cardiac surgery based on pre-incision variables," *PLOS ONE*, vol. 10, 2015.

[53] R. Laureano, N. Caetano, and P. Cortez, "Previsão de tempos de internamento num hospital português: aplicação da metodologia crisp-dm," *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. 13, 2014.

[54] N. Piuzzi, G. Strnad, W. Esa, W. Barsoum, M. Bloomfield, P. Brooks, C. Higuera, M. Joyce, M. Kattan, A. Klika, V. Krebs, N. Mesko, M. Mont, T. Murray, G. Muschler, R. Nickodem, P. Patel, J. Schaffer, K. Spindler, and R. Molloy, "The main predictors of length of stay after total knee arthroplasty: Patient-related or procedure-related risk factors," *The Journal of Bone and Joint Surgery*, vol. 101, 2019.

[55] A. C. do Sistema de Saúde, "Volume i – princípios gerais," http://www.acss.min-saude.pt/wp-content/uploads/2016/12/UCGIC-AF-20110511-Vol-I-Principios-Gerais.pdf, 2011, accessed on 16/10/2020.

[56] M. da Saúde, "Perguntas frequentes sobre a gestão da lista cirúrgica," http://www.chmt.min-saude.pt/wp-content/uploads/sites/27/2018/04/UGA-Perguntas-Frequentes-SIGA.pdf, 2018, accessed on 16/10/2020.

[57] N. C. for Health Statistics, "Classification of diseases, functioning, and disability," https://www.cdc.gov/nchs/icd/index.htm, 2020, accessed on 15/10/2020.

[58] ICD9Chrisendres, "Diseases and injuries," http://icd9.chrisendres.com/index.php, 2020, accessed on 16/10/2020.

[59] S. N. de Saúde, "Codificação clínica," http://www.acss.min-saude.pt/2016/07/22/portal-da-codificacao-3/, 2020, accessed on 15/10/2020.

[60] C. for Disease Control and Prevention, "Classification of diseases, functioning, and disability," https://www.cdc.gov/nchs/icd/index.htm, 2019, accessed on 16/10/2020.

Appendix

This appendix presents 3 tables. All the variables that were part of the final models and its respective description are introduced in Table A.1. Also, if a variable is patient-related is represented by a “(P)”. Otherwise, it is procedure or structural-related.

Tables A.2 and A.3 represent the univariate analysis of surgical patients stratified by specialty. Since the 2-classes models obtained better results, the tables below reflect the univariate analysis taking into account models B and C. The values are given as the number of cases, whenever it is a categorical feature. When not explicitly mentioned, the values refer to category 1, not category 0. When the variable is numerical, the values are given as the median, with the interquartile range in parentheses.

Table A.1: Description of all features that can be used in the model, as well as its identification as patient or procedure-related.

Feature	Description
Age (P)	Age of the patient at the time of the act in question
Gender (P)	Identification of the patient’s gender: Masculine (0), Feminine (1)
Origin (P)	Identification of the origin of the act: Internal (0), External (1)
SurgeryMode (P)	Identification of the type of surgery: Urgent (0) or Conventional (1)
Weight (P)	Weight in the preoperative evaluation
Height (P)	Height in the preoperative evaluation
MaximumBloodPressure (P)	Maximum Blood Pressure in the preoperative evaluation
MinimumBloodPressure (P)	Minimum Blood Pressure in the preoperative evaluation
InpatientCare (P)	Case Regime: Outpatient (0) or inpatient Care (1)
Medicaldeviceplacementproposal	Medical device placement proposal: No (0) or Yes(1)
Specificneedsforperioperative supportorspecialtechniques	Specific needs for perioperative support or special techniques: No(0) or Yes (1)
Continued on next page	

Table A.1 – continued from previous page

Feature	Description
MainSurgeonCategory	Category of the Main Surgeon: Resident (0) or Specialist (1)
difSuggestRealSurgery	Difference, in days, between the suggested for surgery and the day of surgery
timeOt	Difference, in hours, between the exit and the entry in the block
timeAn	Difference, in hours, between the end of anaesthesia and anaesthetic induction
timeSurgery	Difference, in hours, between the last suture and the first cut
exitOperating	Time (Hour) when patients left the operating theatre
nrDiagnsec (P)	Number of secondary diagnoses
nrProcedassoc	Number of associated procedures
Readmission (P)	Number of readmissions in the last 6 months
Readmission_ever (P)	Number of readmissions in total
BMI (P)	Body Mass Index in the preoperative evaluation
MeanDiagnosis	LOS according to the mean *per* diagnosis
infectiousDiagSec (P)	Secondary Diagnosis concern certain infectious and parasitic diseases ICD-10-CM code: No(0) or Yes(1)
neoplasmsDiagSec (P)	Secondary Diagnosis concern neoplasms ICD-10-CM code: No(0) or Yes(1)
bloodDiagSec (P)	Secondary Diagnosis concern diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism ICD-10-CM code: No(0) or Yes(1)
endocrineDiagSec (P)	Secondary Diagnosis concern endocrine, nutritional and metabolic diseases ICD-10-CM code: No(0) or Yes(1)
mentalDiagSec (P)	Secondary Diagnosis concern mental, behavioral and neurodevelopmental disorders diseases ICD-10-CM code: No(0) or Yes(1)
nervousDiagSec (P)	Secondary Diagnosis concern diseases of the nervous system ICD-10-CM code: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
eyeDiagSec (P)	Secondary Diagnosis concern diseases of the eye and adnexa ICD-10-CM code: No(0) or Yes(1)
earDiagSec (P)	Secondary Diagnosis concern diseases of the ear and mastoid process ICD-10-CM code: No(0) or Yes(1)
circulatoryDiagSec (P)	Secondary Diagnosis concern diseases of the circulatory system ICD-10-CM code: No(0) or Yes(1)
respiratoryDiagSec (P)	Secondary Diagnosis concern diseases of the respiratory system ICD-10-CM code: No(0) or Yes(1)
digestiveDiagSec (P)	Secondary Diagnosis concern diseases of the digestive system ICD-10-CM code: No(0) or Yes(1)
skinDiagSec (P)	Secondary Diagnosis concern diseases of the skin and subcutaneous tissue ICD-10-CM code: No(0) or Yes(1)
musculoDiagSec (P)	Secondary Diagnosis concern diseases of the musculoskeletal system and connective tissue ICD-10-CM code: No(0) or Yes(1)
genitourinaryDiagSec (P)	Secondary Diagnosis concern diseases of the genitourinary system ICD-10-CM code: No(0) or Yes(1)
pregnancyDiagSec (P)	Secondary Diagnosis concern pregnancy, childbirth and the puerperium ICD-10-CM code: No(0) or Yes(1)
perinatalperiodDiagSec (P)	Secondary Diagnosis concern certain conditions originating in the perinatal period ICD-10-CM code: No(0) or Yes(1)
congenitalDiagSec (P)	Secondary Diagnosis concern congenital malformations, deformations and chromosomal abnormalities ICD-10-CM code: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
abnormalfindingsDiagSec (P)	Secondary Diagnosis concern symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified ICD-10-CM code: No(0) or Yes(1)
injuryDiagSec (P)	Secondary Diagnosis concern injury, poisoning and certain other consequences of external causes ICD-10-CM code: No(0) or Yes(1)
morbidity (P)	Secondary Diagnosis concern external causes of morbidity ICD-10-CM code: No(0) or Yes(1)
specialcodes (P)	Secondary Diagnosis concern special purposes ICD-10-CM code: No(0) or Yes(1)
healthstatus (P)	Secondary Diagnosis concern factors influencing health status and contact with health services ICD-10-CM code: No(0) or Yes(1)
hypertension (P)	Secondary Diagnosis concern hypertension: No(0) or Yes(1)
hyperlipidaemia (P)	Secondary Diagnosis concern hyperlipidaemia: No(0) or Yes(1)
overweight (P)	Secondary Diagnosis concern overweight: No(0) or Yes(1)
unspecifiedObesity (P)	Secondary Diagnosis concern unspecified obesity: No(0) or Yes(1)
type2Diabetes (P)	Secondary Diagnosis concern type 2 diabetes: No(0) or Yes(1)
nicotineAddiction (P)	Secondary Diagnosis concern nicotine addiction: No(0) or Yes(1)
historyChemoterapy (P)	Secondary Diagnosis concern history chemoterapy: No(0) or Yes(1)
procAndInterProcSec	Associated Procedure concern procedures and interventions, not elsewhere classified ICD-9-CM code: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
nervousProcSec	Associated Procedure concern operations to the nervous system ICD-9-CM code: No(0) or Yes(1)
endocrineProcSec	Associated Procedure concerns operations to the endocrine system ICD-9-CM code: No(0) or Yes(1)
eyeProcSec	Associated Procedure concerns operations to the eye ICD-9-CM code: No(0) or Yes(1)
otherMiscProcSec	Associated Procedures concerns other miscellaneous diagnostic and therapeutic procedures ICD-9-CM code: No(0) or Yes(1)
earProcSec	Associated Procedure concerns operations to the ear ICD-9-CM code: No(0) or Yes(1)
nose&mouthProcSec	Associated Procedure concerns operations to the nose, mouth, and pharynx ICD-9-CM code: No(0) or Yes(1)
respiratoryProcSec	Associated Procedure concerns operations to the respiratory system ICD-9-CM code: No(0) or Yes(1)
cardiovascularProcSec	Associated Procedure concerns operations to the cardiovascular system ICD-9-CM code: No(0) or Yes(1)
hemic&lymphaticProcSec	Associated Procedure concerns operations to thehemic and lymphatic system ICD-9-CM code: No(0) or Yes(1)
digestiveProcSec	Associated Procedures concerns operations to the digestive system ICD-9-CM code: No(0) or Yes(1)
urinaryProcSec	Associated Procedure concerns operations to the urinary system ICD-9-CM code: No(0) or Yes(1)
maleGenitalProcSec	Associated Procedure concerns operations to the male genital organs ICD-9-CM code: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
femaleGenitalProcSec	Associated Procedure concerns operations to the female genital organs ICD-9-CM code: No(0) or Yes(1)
obstetricalProcSec	Associated Procedure concerns obstetrical procedures ICD-9-CM code: No(0) or Yes(1)
musculoskeletalProcSec	Associated Procedures concerns operations to the musculoskeletal system ICD-9-CM code: No(0) or Yes(1)
intergumentaryProcSec	Associated Procedure concerns operations to the intergumentary system ICD-9-CM code: No(0) or Yes(1)
miscProcSec	Associated Procedure concerns miscellaneous diagnostic and therapeutic procedures ICD-9-CM code: No(0) or Yes(1)
Aveiro (P)	Aveiro district is the patient's district: No (0) or Yes (1)
Azores (P)	Azores district is the patient's district: No (0) or Yes (1)
Beja (P)	Beja district is the patient's district: No (0) or Yes (1)
Braga (P)	Braga district is the patient's district: No (0) or Yes (1)
Bragança (P)	Bragança district is the patient's district: No (0) or Yes (1)
CasteloBranco (P)	Castelo Branco district is the patient's district: No (0) or Yes (1)
Coimbra (P)	Coimbra district is the patient's district: No (0) or Yes (1)
Unknown (P)	patient's district is not known: No (0) or Yes (1)
foreignCountry (P)	The pacients lives in other country: No (0) or Yes (1)
Faro (P)	Faro district is the patient's district: No (0) or Yes (1)
Guarda (P)	Guarda district is the patient's district: No (0) or Yes (1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
Madeira (P)	Madeira district is the patient's district: No (0) or Yes (1)
Leiria (P)	Leiria district is the patient's district: No (0) or Yes (1)
Lisboa (P)	Lisboa district is the patient's district: No (0) or Yes (1)
Portalegre (P)	Portalegre district is the patient's district: No (0) or Yes (1)
Porto (P)	Porto district is the patient's district: No (0) or Yes (1)
Santarém (P)	Santarém district is the patient's district: No (0) or Yes (1)
Setúbal (P)	Setúbal district is the patient's district: No (0) or Yes (1)
VianaDoCastelo (P)	Viana do Castelo district is the patient's district: No (0) or Yes (1)
VilaReal (P)	Vila Real district is the patient's district: No (0) or Yes (1)
Viseu (P)	Viseu district is the patient's district: No (0) or Yes (1)
Évora (P)	Évora district is the patient's district: No (0) or Yes (1)
SDC (P)	SDC is the department (inside the hospital) where the act takes place: No (0) or Yes (1)
hospitalisation (P)	Hospitalisation is the department (inside the hospital) where the act takes place: No (0) or Yes (1)
emergency (P)	Emergency is the department (inside the hospital) where the act takes place: No (0) or Yes (1)
SIGLIC_1 (P)	SIGLIC level is 1: No (0) or Yes (1)
SIGLIC_2 (P)	SIGLIC level is 2: No (0) or Yes (1)
SIGLIC_3 (P)	SIGLIC level is 3: No (0) or Yes (1)
SIGLIC_4 (P)	SIGLIC level is 4: No (0) or Yes (1)
AngiologyAndVascularSurg (P)	Angiology and Vascular Surgery is the specialty of the surgery: No (0) or Yes (1)
CardioThoracicSurg (P)	Cardio-Thoracic Surgery is the specialty of the surgery: No (0) or Yes (1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
GeneralSurgery (P)	General Surgery is the specialty of the surgery: No (0) or Yes (1)
ReconstructiveAndAestheticPlasticSurg (P)	Reconstrutive and Aesthetic Surgery is the specialty of the surgery: No (0) or Yes (1)
Gynaecology-Obstetrics (P)	Gynaecology-Obstetrics is the specialty of the surgery: No (0) or Yes (1)
Ophthalmology (P)	Ophthalmology is the specialty of the surgery: No (0) or Yes (1)
Orthopaedics (P)	Orthopaedics is the specialty of the surgery: No (0) or Yes (1)
Otorhinolaryngology (P)	Otorhinolaryngology is the specialty of the surgery: No (0) or Yes (1)
Urology (P)	Urology is the specialty of the surgery: No (0) or Yes (1)
Room1	Surgery took place in room 1: No(0) or Yes (1)
Room2	Surgery took place in room 2: No(0) or Yes (1)
Room3	Surgery took place in room 3: No(0) or Yes (1)
Room4	Surgery took place in room 4: No(0) or Yes (1)
Room5	Surgery took place in room 5: No(0) or Yes (1)
Room6	Surgery took place in room 6: No(0) or Yes (1)
Room7	Surgery took place in room 7: No(0) or Yes (1)
Room8	Surgery took place in room 8: No(0) or Yes (1)
Room9	Surgery took place in room 9: No(0) or Yes (1)
caesareanRoom10	Surgery took place in caesarean room 10: No(0) or Yes (1)
caesareanRoom11	Surgery took place in caesarean room 11: No(0) or Yes (1)
deliveryRoom1	Surgery took place in delivery room 1: No(0) or Yes (1)
deliveryRoom2	Surgery took place in delivery room 2: No(0) or Yes (1)
BilateralLatDiagMain (P)	Bilaterality was the laterality of the main diagnosis: No(0) or Yes (1)
rightLatDiagMain (P)	Right was the laterality of the main diagnosis: No(0) or Yes (1)
leftLatDiagMain (P)	Left was the laterality of the main diagnosis: No(0) or Yes (1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
noLatDiagMain (P)	No laterality associated to the main diagnosis: No(0) or Yes (1)
BilateralLatProcMain	Bilaterality was the laterality of the main procedure: No(0) or Yes (1)
rightLatProcMain	Right was the laterality of the main procedure: No(0) or Yes (1)
leftLatProcMain	Left was the laterality of the main procedure: No(0) or Yes (1)
noLatProcMain	No laterality in the main procedure: No(0) or Yes (1)
Diagnosis	Diagnosis was the objective of the surgery: No(0) or Yes(1)
ImprovementOrmitigation	Improvement or Mitigation was the objective of the surgery: No(0) or Yes(1)
ResolutionOrHealing	Resolution or Healing was the objective of the surgery: No(0) or Yes(1)
WatchOrFollow	Watch or Follow was the objective of the surgery: No(0) or Yes(1)
General	General was the anaesthesia type: No(0) or Yes(1)
Local	Local was the anaesthesia type: No(0) or Yes(1)
RegionalCentral	Regional Central was the anaesthesia type: No(0) or Yes(1)
RegionalPeripheral	Regional Peripheral was the anaesthesia type: No(0) or Yes(1)
Sedation	Sedation was the anaesthesia type: No(0) or Yes(1)
AssistantSurgeon1Spec	Assistant Surgeron 1 was specialist: No(0) or Yes(1)
AssistantSurgeon1Res	Assistant Surgeron 1 was resident: No(0) or Yes(1)
AssistantSurgeon2Spec	Assistant Surgeron 2 was specialist: No(0) or Yes(1)
AssistantSurgeon2Res	Assistant Surgeron 2 was resident: No(0) or Yes(1)
sunday	Surgery was on a Sunday: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
wednesday	Surgery was on a Wednesday: No(0) or Yes(1)
thursday	Surgery was on a Thursday: No(0) or Yes(1)
saturday	Surgery was on a Saturday: No(0) or Yes(1)
monday	Surgery was on a Monday: No(0) or Yes(1)
friday	Surgery was on a Friday: No(0) or Yes(1)
tuesday	Surgery was on a Tuesday: No(0) or Yes(1)
april	Surgery was in April: No(0) or Yes(1)
august	Surgery was in August: No(0) or Yes(1)
december	Surgery was in December: No(0) or Yes(1)
february	Surgery was in February: No(0) or Yes(1)
january	Surgery was in January: No(0) or Yes(1)
july	Surgery was in July: No(0) or Yes(1)
june	Surgery was in June: No(0) or Yes(1)
may	Surgery was in May: No(0) or Yes(1)
march	Surgery was in March: No(0) or Yes(1)
november	Surgery was in November: No(0) or Yes(1)
october	Surgery was in October: No(0) or Yes(1)
september	Surgery was in September: No(0) or Yes(1)
DeviatedNasalSeptum (P)	Main Diagnosis concerns deviated nasal septum: No(0) or Yes(1)
Cholelithiasis (P)	Main Diagnosis concerns cholelithiasis: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
E_V_codes (P)	Main Diagnosis concerns supplementary classification of factors influencing health status and contact with health services and supplementary classification of external causes of injury and poisoning ICD-9-CM code: No(0) or Yes(1)
blood (P)	Main Diagnosis concerns diseases of the blood and blood-forming organs ICD-9-CM code: No(0) or Yes(1)
circulatory (P)	Main Diagnosis concerns diseases of the circulatory system ICD-9-CM code: No(0) or Yes(1)
congenital (P)	Main Diagnosis concerns congenital anomalies ICD-9-CM code: No(0) or Yes(1)
digestive (P)	Main Diagnosis concerns diseases of the digestive system ICD-9-CM code: No(0) or Yes(1)
endocrine (P)	Main Diagnosis concerns diseases of the endocrine system ICD-9-CM code: No(0) or Yes(1)
genitourinary (P)	Main Diagnosis concerns diseases of the genitourinary system ICD-9-CM code: No(0) or Yes(1)
infectious (P)	Main Diagnosis concerns infectious and parasitic diseases ICD-9-CM code: No(0) or Yes(1)
injury (P)	Main Diagnosis concerns injury and poisoning ICD-9-CM code: No(0) or Yes(1)
muscular (P)	Main Diagnosis concerns diseases of the musculoskeletal system and connective tissue ICD-9-CM code: No(0) or Yes(1)

Continued on next page

Table A.1 – continued from previous page

Feature	Description
neoplasms (P)	Main Diagnosis concerns neoplasms ICD-9-CM code: No(0) or Yes(1)
nervous (P)	Main Diagnosis concerns the nervous system and sense organs ICD-9-CM code: No(0) or Yes(1)
pregnancy (P)	Main Diagnosis concerns complications of pregnancy, childbirth, and the puerperium ICD-9-CM code: No(0) or Yes(1)
respiratory (P)	Main Diagnosis concerns diseases of the respiratory system ICD-9-CM code: No(0) or Yes(1)
skin (P)	Main Diagnosis concerns diseases of the skin and subcutaneous tissue ICD-9-CM code: No(0) or Yes(1)
sympt_signs (P)	Main Diagnosis concerns symptoms, signs, and ill-defined conditions ICD-9-CM code: No(0) or Yes(1)
p_Administration	Main Procedure concerns administration ICD-10-PCS code: No(0) or Yes(1)
p_Extracorporeal	Main Procedure concerns extracorporeal assistance and performance ICD-10-PCS code: No(0) or Yes(1)
p_Imaging	Main Procedure concerns imaging ICD-10-PCS code: No(0) or Yes(1)
p_Measurement/Monitoring	Main Procedure concerns measurement and monitoring ICD-10-PCS code: No(0) or Yes(1)
p_medical/surgical	Main Procedure concerns medical and surgical ICD-10-PCS code: No(0) or Yes(1)
LOS	Target Feature: LOS

Table A.2: Univariate Descriptions of Surgical Patients stratified by specialty for Model B. * or special techniques. The values are given as the number of cases, whenever it is a categorical feature. When the variable is numerical, the values are given as the median, with the interquartile range in parentheses.

Variable	**General Surgery**	**Gynaecology-Obstetrics**	**Orthopaedics**	**Urology**	**Otorhinolaryngology**
Total	(N = 1529)	(N = 775)	(N = 1560)	(N = 979)	(N = 1076)
Age	62 (48, 72)	54 (44,67)	61 (40, 71)	69 (62,75)	41 (26,56)
Weight	70 (64,81)	68 (60,78)	72 (63, 81)	75 (67,83)	70 (60,82)
Height	164 (157,170)	160 (155,165)	163 (156,170)	167 (160,172)	165 (158,173)
Maximum Blood Pressure	140 (125,154)	144 (130, 158)	144 (130,159)	150 (137, 162.5)	136 (122, 148)
Minimum Blood Pressure	80 (70,89)	83 (75, 90)	82 (74,91)	85 (76,94)	80 (71, 89)
difSuggestRealSurgery	48 (21, 108)	43 (20, 77)	85.5 (24,178.25)	58 (32, 130.5)	191 (160.5, 204)
timeOt	2.28 (1.53, 3.72)	2.57 (1.98, 3.22)	2.76 (2.17, 3.44)	1.57 (1.12, 2.23)	1.25 (0.93, 1.77)
timeAn	2.08 (1.32, 3.45)	2.27 (1.75, 2.92)	2.55 (2.00, 3.25)	1.33 (0.92, 2.02)	1.03 (0.75, 1.53)
timeSurgery	1.45 (0.83, 2.42)	1.68 (1.18, 2.25)	1.84 (1.35, 2.43)	0.88 (0.51, 1.54)	0.58 (0.38,1.08)
exitOperating	13.97 (11.83, 16.30)	14.08 (11.99, 17.03)	13.68 (11.76, 16.37)	16.42 (13.68, 18.45)	13.6 (11.5, 17.4)
nrDiagnsec	7 (4,9)	4 (3,6)	4 (2,5)	5 (3,7)	5 (3,7)
nrProcedassoc	0 (0,0)	1 (0,1)	0 (0,0)	0 (0,0)	1 (0,1)
Readmission	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Continued on next page

Table A.2 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
Readmission_ever	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
BMI	26.64 (23.88, 29.78)	26.73 (23.81, 30.12)	27.10 (23.95, 30.43)	26.67 (24.24, 29.43)	25.27 (21.97, 28.71)
MeanDiagnosis	2 (1,5)	3 (2,3)	3 (2,5)	4 (2,5)	1 (1,1)
infectiousDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
neoplasmsDiagSec	0 (0,1)	1 (0,2)	0 (0,0)	0 (0,1)	0 (0,0)
bloodDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
endocrineDiagSec	0 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
mentalDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
nervousDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
eyeDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
earDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
circulatoryDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
respiratoryDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	1 (1,2)
digestiveDiagSec	1 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
skinDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
musculoDiagSec	0 (0,0)	0 (0,0)	1 (1,1)	0 (0,0)	0 (0,0)
genitourinaryDiagSec	0 (0,0)	1 (0,2)	0 (0,0)	1 (0,1)	0 (0,0)
pregnancyDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
perinatalperiod DiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
congenitalDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
abnormalfindings DiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
injuryDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
morbidity	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Continued on next page

Table A.2 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
specialcodes	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
healthstatus	2 (1,4)	0 (0,1)	1 (0,1)	1 (0,5)	1 (0,2)
hypertension	0 (0,1)	0 (0,1)	0 (0,1)	0 (0,1)	0 (0,0)
hyperlipidaemia	0 (0,1)	0 (0,0)	0 (0,1)	0 (0,1)	0 (0,0)
overweight	0 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,1)
unspecifiedObesity	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
type2Diabetes	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
nicotineAddiction	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
historyChemoterapy	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
procAndInterProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
nervousProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
endocrineProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
eyeProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
otherMiscProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
earProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,1)
nose&mouthProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
respiratoryProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
cardiovascularProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
hemic&lymphatic ProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
digestiveProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
urinaryProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
maleGenitalProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
femaleGenitalProcSec	0 (0,0)	0 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)
obstetricalProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Continued on next page

Table A.2 – continued from previous page

Variable		General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
musculoskeletal ProcSec		0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
intergumentary ProcSec		0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
miscProcSec		0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
Gender	0	682	0	625	770	591
	1	847	775	935	209	485
Origin	0	1528	775	1559	975	1073
	1	1	0	1	4	3
Surgery Mode	0	67	11	12	47	8
	1	1462	764	1548	932	1068
InpatientCare	0	137	47	104	124	56
	1	1392	728	1456	855	1020
Medical device	0	1457	644	538	824	1076
placement proposal	1	72	131	1022	155	0
Specific needs for	0	916	503	61	383	1068
perioperative support*	1	613	272	1499	596	8
Main Surgeon	0	363	210	148	258	14
Category	1	1166	565	1412	721	1062
Aveiro		1	0	0	1	2
Azores		3	1	2	0	1
Beja		2	1	3	0	2
Braga		0	0	1	0	0
Bragança		2	0	0	1	1
Castelo Branco		0	3	7	0	0
Coimbra		0	0	3	0	0

Continued on next page

Table A.2 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
Unknown	1	0	1	0	0
foreignCountry	1	0	4	0	0
Faro	2	7	10	1	3
Guarda	2	1	2	0	0
Madeira	0	0	0	0	0
Leiria	11	8	7	25	11
Lisboa	1429	681	1444	923	1028
Portalegre	1	2	3	4	1
Porto	0	0	3	0	1
Santarém	17	22	14	3	14
Setúbal	52	34	47	18	9
Viana Do Castelo	0	1	0	0	1
Vila Real	2	0	0	1	0
Viseu	1	0	4	2	2
Évora	2	14	5	0	0
SDC	862	650	1081	806	882
hospitalisation	667	125	479	171	192
emergency	0	0	0	2	2
SIGLIC_1	1432	597	1371	913	942
SIGLIC_2	94	167	151	51	119
SIGLIC_3	3	8	23	10	9
SIGLIC_4	0	3	15	5	6
Room1	80	12	92	333	2
Room2	1	4	5	2	2
Room3	3	4	0	12	1001
Room4	8	613	0	577	8

Continued on next page

Table A.2 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
Room5	2	0	757	3	0
Room6	0	4	581	0	4
Room7	809	29	4	15	4
Room8	567	18	4	9	7
Room9	35	4	3	9	0
caesareanRoom10	0	0	1	0	0
caesareanRoom11	24	87	111	19	48
deliveryRoom1	0	0	1	0	0
deliveryRoom2	0	0	1	0	0
BilateralLatDiagMain	55	14	61	29	263
rightLatDiagMain	148	73	625	146	167
leftLatDiagMain	123	79	643	129	157
noLatDiagMain	1203	609	231	675	489
BilateralLatProcMain	51	24	58	29	269
rightLatProcMain	150	73	619	147	173
leftLatProcMain	121	70	646	134	160
noLatProcMain	1207	608	237	669	474
Diagnosis	19	2	1	70	50
Improvement Ormitigation	4	47	334	150	327
Resolution OrHealing	1506	722	1225	756	699
WatchOrFollow	0	4	0	3	0
General	1496	771	921	638	1076
Local	17	0	0	1	0
RegionalCentral	1	4	523	313	0

Continued on next page

Table A.2 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
RegionalPeripheral	0	0	115	22	0
Sedation	15	0	1	5	0
AssistantSurgeon 1Spec	1069	660	739	435	841
AssistantSurgeon 1Res	401	90	790	253	82
AssistantSurgeon 2Spec	136	220	174	45	11
AssistantSurgeon 2Res	563	293	263	58	33
sunday	3	0	3	0	0
wednesday	350	2	277	424	327
thursday	360	338	315	1	420
saturday	4	0	24	0	0
monday	296	0	445	478	0
friday	131	229	110	76	121
tuesday	385	206	386	0	208
april	124	63	125	70	94
august	110	50	115	89	95
december	59	29	63	35	39
february	114	72	156	78	89
january	154	79	155	99	99
july	139	58	122	89	74
june	149	74	149	71	95
may	124	65	147	86	99
march	150	84	151	85	112

Continued on next page

Table A.2 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
november	136	61	116	83	84
october	160	69	157	106	114
september	110	71	104	88	82
DeviatedNasalSeptum	0	0	0	0	297
Cholelithiasis	336	0	0	0	0
E_V_codes	51	10	31	1	0
blood	1	0	0	0	0
circulatory	13	0	0	1	0
congenital	41	0	38	4	4
digestive	376	3	1	1	18
endocrine	163	0	1	0	2
genitourinary	8	416	0	571	0
infectious	1	0	0	0	1
injury	2	2	319	0	3
muscular	0	0	1159	0	36
neoplasms	501	336	6	382	119
nervous	0	0	3	0	250
pregnancy	0	2	0	0	0
respiratory	13	0	0	0	339
skin	7	0	1	0	1
sympt_signs	16	6	1	19	6
p_Administration	11	1	3	7	1
p_Extracorporeal	1	0	0	1	1
p_Imaging	4	0	2	3	2
p_Measurement /Monitoring	1	1	2	4	0

Continued on next page

Table A.2 – continued from previous page

Variable		General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
p_medical/surgical		1512	773	1553	964	1072
LOS	0	710	222	552	279	981
	1	819	553	1008	700	95

Table A.3: Univariate Descriptions of Surgical Patients stratified by specialty for Model C. * or special techniques. The values are given as the number of cases, whenever it is a categorical feature. When the variable is numerical, the values are given as the median, with the interquartile range in parentheses.

Variable	**General Surgery**	**Gynaecology-Obstetrics**	**Orthopaedics**	**Urology**	**Otorhinolaryngology**
Total	(N = 819)	(N = 553)	(N =1008)	(N = 700)	(N = 95)
Age	65 (54,74)	52 (44,66)	66 (48, 73)	69.5 (63.75,75)	49 (35,64)
Weight	70 (64.85,81)	68 (60,79)	73.5 (65,81)	74 (67,82)	70 (60,80)
Height	165 (158,170)	160 (155,165)	162 (156,170)	167 (162,172)	168 (160,176.5)
Maximum Blood Pressure	136 (121, 151)	144 (130,158)	147 (133,162)	150 (138,163.25)	139 (119,150.5)
Minimum Blood Pressure	76 (66,85)	84 (75,90)	83 (74,91)	85 (76,93)	78 (69.5,89)
difSuggestRealSurgery	29 (15, 69.5)	49 (21,84)	108 (38,180.25)	62.5 (33,135)	184 (23.5,198)
timeOt	3.55 (2.5, 4.98)	2.77 (2.30, 3.42)	3 (2.47,3.68)	1.73 (1.3, 2.85)	1.92 (1.44,3.15)
timeAn	3.25 (2.30, 4.67)	2.5 (2.03, 3.12)	2.8 (2.28,3.44)	1.5 (1.08, 2.62)	1.58 (1.22,2.82)
timeSurgery	2.23 (1.54, 3.48)	1.83 (1.42, 2.47)	2 (1.58,2.62)	1.08 (0.7,1.92)	1.17 (0.68,2.28)
exitOperating	14.17 (12.33, 16.75)	14.25 (12.13, 17.33)	14.01 (11.85,17)	16.63 (13.67,18.67)	13.88 (11.71,17.02)
nrDiagnsec	8 (5,11)	5 (3,6)	4 (2,6)	5 (3,7)	5 (3,7)
nrProcedassoc	0 (0,0)	1 (0,1)	0 (0,0)	0 (0,0)	1 (0,1)
Readmission	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Continued on next page

Table A.3 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
Readmission_ever	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
BMI	26.42 (23.67, 29.40)	26.67 (23.74, 30.04)	27.68 (24.61,30.8)	26.44 (24.22,29.39)	24.98 (22.49,27.12)
MeanDiagnosis	5 (2,8)	3 (2,3)	5 (2,6)	4 (4,4)	1 (1,4)
infectiousDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
neoplasmsDiagSec	1 (0,1)	1 (0,2)	0 (0,0)	0 (0,1)	0 (0,1)
bloodDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
endocrineDiagSec	0 (0,2)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
mentalDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,1)
nervousDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
eyeDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
earDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
circulatoryDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
respiratoryDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,2)
digestiveDiagSec	0 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
skinDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
musculoDiagSec	0 (0,0)	0 (0,0)	1 (1,1)	0 (0,0)	0 (0,0)
genitourinaryDiagSec	0 (0,0)	1 (1,2)	0 (0,0)	1 (0,1)	0 (0,0)
pregnancyDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
perinatalperiod DiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
congenitalDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
abnormalfindings DiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
injuryDiagSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
morbidity	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Continued on next page

Table A.3 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
specialcodes	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
healthstatus	3 (2,4)	0 (0,1)	1 (0,1)	1 (0,1)	1 (1,2)
hypertension	0 (0,1)	0 (0,1)	0 (0,1)	0 (0,1)	0 (0,1)
hyperlipidaemia	0 (0,1)	0 (0,0)	0 (0,1)	0 (0,1)	0 (0,0)
overweight	0 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,1)
unspecifiedObesity	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
type2Diabetes	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
nicotineAddiction	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
historyChemoterapy	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
procAndInterProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
nervousProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
endocrineProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
eyeProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
otherMiscProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
earProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
nose&mouthProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
respiratoryProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
cardiovascularProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
hemic&lymphatic ProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
digestiveProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
urinaryProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
maleGenitalProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
femaleGenitalProcSec	0 (0,0)	0 (0,1)	0 (0,0)	0 (0,0)	0 (0,0)
obstetricalProcSec	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Continued on next page

Table A.3 – continued from previous page

Variable		General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
musculoskeletal ProcSec		0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
intergumentary ProcSec		0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
miscProcSec		0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
Gender	0	421	0	406	569	63
	1	398	553	602	131	32
Origin	0	818	553	1008	696	95
	1	1	0	0	4	0
Surgery Mode	0	52	11	11	44	4
	1	767	542	997	656	91
InpatientCare	0	52	22	55	43	2
	1	767	531	953	657	93
Medical device	0	783	478	275	620	95
placement proposal	1	36	75	733	80	0
Specific needs for	0	474	370	38	294	87
perioperative support*	1	345	183	970	406	8
Main Surgeon	0	109	152	65	177	1
Category	1	710	401	943	523	94
Aveiro		1	0	0	0	0
Azores		3	1	2	0	0
Beja		2	1	1	0	0
Braga		0	0	1	0	0
Bragança		2	0	0	1	0
Castelo Branco		0	3	6	0	0
Coimbra		0	0	3	0	0

Continued on next page

Table A.3 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
Unknown	0	0	1	0	0
foreignCountry	0	0	3	0	0
Faro	1	7	7	1	1
Guarda	0	1	2	0	0
Madeira	0	0	0	0	0
Leiria	8	5	4	19	2
Lisboa	748	479	923	657	91
Portalegre	1	1	3	3	0
Porto	0	0	2	0	0
Santarém	12	18	8	3	1
Setúbal	36	28	33	13	0
Viana Do Castelo	0	1	0	0	0
Vila Real	2	0	0	1	0
Viseu	1	0	4	2	0
Évora	2	8	5	0	0
SDC	288	459	713	566	68
hospitalisation	531	94	295	132	27
emergency	0	0	0	2	0
SIGLIC_1	728	418	919	679	68
SIGLIC_2	90	126	71	38	21
SIGLIC_3	1	6	13	9	4
SIGLIC_4	0	3	5	4	2
AngiologyAnd VascularSurg	0	0	0	0	0
CardioThoracicSurg	0	0	0	0	0
General Surgery	819	0	0	0	0

Continued on next page

Table A.3 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
ReconstructiveAnd AestheticPlasticSurg	0	0	0	0	0
Gynaecology-Obstetrics	0	553	0	0	0
Ophthalmology	0	0	0	0	0
Orthopaedics	0	0	1008	0	0
Otorhinolaryngology	0	0	0	0	95
Urology	0	0	0	700	0
Room1	39	7	56	236	0
Room2	0	4	1	1	0
Room3	0	1	0	7	89
Room4	1	485	0	424	1
Room5	1	0	521	3	0
Room6	0	0	369	0	1
Room7	463	18	4	10	1
Room8	290	17	2	4	0
Room9	16	2	0	8	0
caesareanRoom10	0	0	1	0	0
caesareanRoom11	9	19	53	7	3
deliveryRoom1	0	0	1	0	0
deliveryRoom2	0	0	0	0	0
BilateralLatDiagMain	23	9	16	19	12
rightLatDiagMain	60	23	428	96	15
leftLatDiagMain	57	19	409	64	29
noLatDiagMain	679	502	155	521	39
BilateralLatProcMain	20	14	13	19	13
rightLatProcMain	62	25	422	98	16

Continued on next page

Table A.3 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
leftLatProcMain	58	12	412	67	31
noLatProcMain	679	502	161	516	35
Diagnosis	11	2	1	40	0
Improvement Ormitigation	2	34	259	112	33
Resolution OrHealing	806	513	748	545	62
WatchOrFollow	0	4	0	3	0
General	813	550	657	443	95
Local	6	0	0	1	0
RegionalCentral	0	3	318	240	0
RegionalPeripheral	0	0	33	15	0
Sedation	0	0	0	1	0
AssistantSurgeon 1Spec	627	470	528	326	81
AssistantSurgeon 1Res	168	63	467	179	4
AssistantSurgeon 2Spec	95	176	148	37	7
AssistantSurgeon 2Res	390	232	205	53	3
sunday	3	0	2	0	0
wednesday	198	2	188	318	30
thursday	173	193	149	1	39
saturday	2	0	24	0	0
monday	156	0	357	339	0

Continued on next page

Table A.3 – continued from previous page

Variable	General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
friday	68	168	20	42	21
tuesday	219	190	268	0	5
april	68	39	87	49	9
august	60	33	74	66	6
december	28	20	35	19	1
february	64	59	94	53	7
january	66	59	103	80	5
july	71	40	83	68	4
june	75	53	100	47	8
may	72	44	91	62	13
march	72	67	98	58	10
november	81	37	67	60	11
october	96	52	110	77	17
september	66	50	66	61	4
DeviatedNasalSeptum	0	0	0	0	20
Cholelithiasis	38	0	0	0	0
E_V_codes	51	5	14	0	0
blood	0	0	0	0	0
circulatory	3	0	0	0	0
congenital	19	0	32	1	1
digestive	146	3	1	1	2
endocrine	94	0	0	0	1
genitourinary	4	303	0	413	0
infectious	1	0	0	0	0
injury	1	2	131	0	0
muscular	0	0	823	0	4

Continued on next page

Table A.3 – continued from previous page

Variable		General Surgery	Gynaecology-Obstetrics	Orthopaedics	Urology	Otorhinolaryngology
neoplasms		411	234	4	279	41
nervous		0	0	1	0	13
pregnancy		0	2	0	0	0
respiratory		12	0	0	0	12
skin		2	0	1	0	1
sympt_signs		7	4	1	6	0
p_Administration		9	0	2	6	0
p_Extracorporeal		1	0	0	1	1
p_Imaging		3	0	2	2	2
p_Measurement /Monitoring		1	1	2	3	0
p_medical/surgical		805	552	1002	688	92
LOS	0	302	410	344	387	40
	1	517	143	664	313	55

Table A.4: Confusion matrices for the HBA model and ML models. T stands for True and P for Predicted. represents the quantity of samples.

	Model B							Model C						
	Post		Pre		HBA		#	Post		Pre		HBA		#
	Predicted							Predicted						
T \ P	0	1	0	1	0	1		0	1	0	1	0	1	
General Model														
0	1045	193	968	270	832	406	1238	554	139	545	148	363	330	693
1	194	1334	206	1322	118	1410	1528	201	634	210	625	146	689	835
Total	1239	1527	1174	1592	950	1816	2766	755	773	755	773	509	1019	1528
General Surgery														
0	279	29	262	46	246	62	308	98	33	94	37	60	71	131
1	45	319	59	305	42	322	364	35	198	31	202	17	216	233
Total	324	348	321	351	288	384	672	133	231	125	239	77	287	364
Gynaecology-Obstetrics														
0	82	22	80	24	38	66	104	157	32	178	11	174	15	189
1	22	237	22	237	5	254	259	42	28	59	11	56	14	70
Total	104	259	102	261	43	320	363	199	60	237	22	230	29	259
Orthopaedics														
0	172	28	166	34	103	97	200	94	34	98	30	66	62	128
1	62	335	70	327	34	363	397	38	231	47	222	37	232	269
Total	234	363	236	361	137	460	597	132	265	145	252	103	294	397
Urology														
0	86	30	71	45	11	105	116	148	39	157	30	47	140	187
1	39	271	50	260	11	299	310	45	78	59	64	20	103	123
Total	125	301	121	305	22	404	426	193	117	216	94	67	243	310
Otorhinolaryngology														
0	454	0	449	5	428	26	454	13	2	13	2	2	13	15
1	24	8	23	9	22	10	32	3	14	2	15	6	11	17
Total	478	8	472	14	450	36	486	16	16	15	17	8	24	32