ESCUELA TÉCNICA SUPERIOR DE
INGENIERÍA DE TELECOMUNICACIÓN

DEGREE IN BIOMEDICAL ENGINEERING (ENGLISH)

END OF DEGREE PROJECT

PREDICTION MODEL OF SURGICAL SITE INFECTION IN PATIENTS UNDERGOING HIP ARTHROPLASTY

AUTHOR: JUAN BASILIO DE LA TORRE MOSQUERA
TUTOR: CRISTINA SOGUERO RUÍZ
CO-TUTOR: ÓSCAR ESCUDERO ARNANZ

ACADEMIC YEAR: 2021/2022

"To achieve a great dream, first, it is necessary a great aptitude to dream; then, persistence, which is the belief in one ´s dream."

— Dr. Hans Selye

Acknowledgements

Firstly, I would like to express my gratitude to Hospital Universitario Ramón y Cajal for providing the database needed for the development of this work as well as to Dr. Aranaz, Jefe de Servicio de Medicina Preventiva y Salud Pública, and Dra. C. Díaz-Agero Pérez, médico adjunto de Medicina Preventiva y Salud Pública, for their kindness and help.

Furthermore, I would like to thank all the teachers and mentors from the Rey Juan Carlos University for contributing to my education and professional formation. Specially, to my mentors in this project, Cristina Soguero Ruiz and Óscar Escudero Arnanz, for their incredibly support and attendance during the development of this work and for endless patient with my constant messages and stresses. Thanks to all the things that I have learnt from you, machine learning has become one of my favourites areas of study.

This project marks the end of one of the most wonderful stages of my life but also one of the most challenging and exasperating ones and I want to mention my *Team* because without them it would have been much harder, thank you for the support, for helping me to calm down in the hard moments and for all the fun that we have had and probably will.

My biggest gratitude goes to my parents which have always been there for me, providing everything that I needed, putting up with my temper in the stressful moments and supporting me in everything they could. But mainly for believing in me in every step of the way, for giving me the opportunity to further develop my career after finishing this degree and trusting me undoubtedly going *all-in* together with me. And last but not least, to my siblings which are the best thing I have and will have the rest of my life.

Abstract

Infections are one of the main concerns when it comes to public health. In fact, surgical site infections (SSI) are one of the most usual type of them and the main cause of morbidity among surgical patients. Moreover, hip arthroplasty (HA) is one of the most successful and useful operations nowadays and one of the main reason why it is hampered is precisely SSI.

Therefore, it would be quite positive and advantageous for the health of the patients to be able to predict, based on different parameters of surgery, instances before and after surgery, whether a patient will suffer an infection or not. Furthermore, the fact of suffering an infection needs a total or partial change of the infected prosthesis which strongly increases the costs for the health system. Thus, it has a wide application on the clinical practice where the clinicians, when they discharge a patient, will be able to know with a certain accuracy if the patient will suffer an infection.

To this end, a database provided by Hospital Universitario Ramón y Cajal of patients undergoing HA between the years 2010 and 2020 has been used after applying the corresponding preprocessing and feature engineering techniques, to remove inconsistent data and generate useful one, respectively. Therefore providing data which will naturally allow the application of our models. This preprocessing stage has been followed by the generation of a visual analysis through barplots and boxplots to be able to understand better the role of each variable, its distribution and importance regarding the further prediction.

After the preparation of the data, it can be introduced in our models. The methodology used in this project goes from undersampling and smote in the case of treating with imbalanced classes. One hot encoding for feature engineering and getting binary columns from categorical ones. Regarding feature selection, we have applied the mutual information method. And finally we will end up with supervised machine learning (ML) models to perform the prediction together with a *k-fold cross validation* to find the best hyperparameters. We have used the methods of logistic regression, decision trees, random forest, XGBoost and multilayer perceptron.

Finally, the best general results were provided by linear regression which provides 94,2% of accuracy and 95.5% of specificity. Random forest is the ones that provides better sensitivity of 78%. Overall, they are satisfying results that demonstrate the ability and usefulness of ML algorithms for the prediction of infection. Therefore, with improvement and more studies on this area, it could possibly become an essential asset in the prevention of SSI regarding HA.

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

AACD Association of Anesthesia Clinical Directors

ADASYN Adaptive Synthetic Sampling

ANN Artificial Neural Networks

AUC Area Under The Curve

BMI Body Mass Index

BP BackPropagation

CV Cross-Validation

DT Decision Tree

EHR Electronic Health Record

FE Feature Engineering

FN False Negative

FP False Positive

FS Feature Selection

HAI Healthcare-Associated Infections

HA Hip Arthroplasty

HELICS Hospitals in Europe Link for Infection Control through Surveillance

HURYC Hospital Universitario Ramón y Cajal

ICU Intensive Care Unit

INCLIMECC Indicadores clínicos de mejora contínua de la calidad

IPSE Improving Patient Safety in Europe

LR Logistic Regression

MI Mutual Information

MLP MultiLayer Perceptron

ML Machine Learning

NHSN-CDC National healthcare safety network - Center for disease control and prevention

NHSN National Healthcare Safety Network

NRC National Research Council

NSQIP National Surgical Quality Improvement Program

PHA Partial Hip Arthroplasty

RF Random Forest

ROC Receiver Operating Characteristic

SMOTE Synthetic Minority Oversampling Technique

SSI Surgical Site Infection

SVM Support Vector Machine

THA Total Hip Arthroplasty

TKA Total Knee Arthroplasty

TN True Negative

TP True Positive

WHO World Health Organization

XGBoost Extreme Gradient Boosting

Chapter 1

Introduction and objectives

In this first chapter of the project, a solid preface is carried out exposing the reasons that motivated the project and its context. Furthermore, its objectives, what is intended with it, and an analysis of the procedures implemented for achieving them. Later, a Gantt´s diagram is showed, breaking down each phase of the study and the time during it was performed. And finally, the structure of this report is detailed and briefly explained.

1.1 Context and motivation

Healthcare-associated infections (HAI) are a prime concern in public health problem. The Recommendation of the European Council of June 9, 2009 on patient safety, comprehends the HAI´s prevention and control (2009/C151/01) [11]. This document urged the countries of the European Union to adopt and implement a strategy for the prevention and control of HAIs, highlighting the importance of creating or strengthen active surveillance programs at the regional/national levels that allow to obtain national reference data, as well as evaluate and guide prevention policies and control [12] [13].

In 2000–2002, harmonised methods for the surveillance of two targeted infection types, surgical site infections (SSIs) and HAIs in intensive care units (ICUs), were developed by the network HELICS (Hospitals in Europe Link for Infection Control through Surveillance). These methods belong also to the Improving Patient Safety in Europe (IPSE) project [14].

European Parliament and the October 22th of 2013 Council´s decision No. 1082/2013/EU, on serious cross-border threats to health, in its article 2, states that, among the categories of threats to health, to which the public health measures must be applied, find "microbial resistance

and infections associated with healthcare related to communicable diseases". Moreover, in article 6, it includes them together with the communicable diseases in the Epidemiological Surveillance Network established in the scope of the European Union [15].

The second most typical HAI type is SSIs. and the main cause of morbidity among surgical patients, especially after abdominal and colorectal [16], cardiovascular [17], oncological [18], trauma [19] or orthopedic [20] surgeries. Despite improvements in infection control practices, such as better operating room ventilation, sterilization methods, barriers, surgical technique, and the availability of antimicrobial prophylaxis, SSIs remain a significant cause of morbidity, prolonged hospitalization, unplanned readmissions after surgery, high use of additional resources, and death [21]. Seventy-five percent of SSI-associated deaths are due solely to the SSI, which constitutes approximately 20% of all HAIs and is linked to a 2-to 11-fold increase in the risk of mortality. [22]. SSI is the costliest HAI type. Studies conducted in the USA put the cost of hospitalization at $3.3 billion annually, 9.7 days longer than before, and thus more than $20,000 more expensive per admission. [23] [24] [25].

Effective SSI prevention protocols are the best approach, and the majority of them are adapted to the needs of the nations in which they were produced [26]. There is a need for global guidelines, applicable to all settings. The World Health Organization (WHO) released global recommendations for SSI prevention in 2016 [27], which included carrying out specific tests per patient before a specific medical operation. There are many multiple ways to anticipate the onset of SSIs and lower their frequency, including various risk models and prevention tactics. As an example, High-income nations have recognized the need of data collecting through centralized surveillance technology for the prevention of SSI. [28] [29] [30]. These risk models seek to anticipate SSIs and direct subsequent measures to avert more severe consequences.

Several researchers have already focused on predicting SSIs from electronic health records (EHRs). They have employed patient demographics, medical history, and surgical information [31], with models that could aid in the development of new preoperative preventive guides. and surveillance of SSIs [32]. Increase of complex comorbidities seen in surgical patients has been seen and It is anticipated that employing evidence-based methods, about half of SSIs may be avoided [33].

Among the most effective operations of our day is hip arthroplasty (HA). For a variety of causes, including infection, wear, loosening, fracture, or instability, HA components may need to be replaced (through revision surgery). Preoperative patient characteristics, implant factors, and surgical factors can all impact the requirement for future revision surgery. HA revisions had previously been shown to be ineffective in relieving pain and function as the original operation,

to have a high risk of later revision, to be expensive to the health system, and to expose patients to the added agony and inconvenience of another operation [34].

In fact, suffering an infection requires a revision surgery for solving it [35]. During this procedure, the prosthesis that has produced the infection is changed partially or completely depending on the time that it has taken to appear. Thus, the appearance of infections highly increases the costs for the health system and their prediction can mean a huge save of money. It has been proved that the revision surgery caused by an infection will end up costing 3.6 times more than the first one [35].

For that reason, minimizing the number of complications, in our case, infections, can represent a huge advance in this type of surgery which, nowadays, even more and more people is undergoing [36]. This study performed a predictive mathematical model of SSIs in patients undergoing hip replacement. Surgical quality improvement programs to improve patient safety are a need in the near future. With a solid infection prevention and control programs, many of these infections are preventable.

1.2 Objectives

The major goal of this project is to be able to anticipate, based on a series of variables collected by the hospital staff, if a patient undergoing any kind of hip replacement surgery is likely to suffer from SSI of the prosthesis site or not, through the development of a predictive model. For achieving this, a database of patients operated between 2010 and 2020 provided by Hospital Universitario Ramón y Cajal (HURYC) is used.

The additional objectives that have been posed are to better know which are the most relevant characteristics of patients with SSI in patients undergoing hip replacement. This will allow us to implement preventive measures to avoid them and improve the effectiveness and efficiency of the procedure.

1.3 Methodology

For achieving the objectives mentioned in the previous section, the following steps have been carried out:

- Investigation of previous studies and current advances regarding this kind of studies for

hip replacement surgeries. Development of the state of art.

- Development of the document for the approval of this project by the Ethics Committee for Clinical investigation of HURYC.
- Preprocessing of the final variables with the aim of eliminating inconsistent data and apply several techniques for the increasing of data quality. Together with a Feature engineering (FE) process.
- Exploratory and descriptive analysis of the variables in order to deepen the knowledge about its distribution and role regarding the outcome.
- Application of machine learning (ML) methods for the prediction of infection in those patients undergoing HA. The following techniques will be used: logistic regression (LR), decision trees (DT), random forest (RF), XGBoost (XGB) and multilayer perceptron (MLP).
- Analysis of the results obtained from a clinical point of view taking into account the values of every metric and the conclusions drawn from this project. Moreover, the development of future lines of this project indicating those possible improvements.

Gantt´s Diagram

A Gantt´s Diagram is bar chart that basically shows the stages of a project in the vertical axis and the amount of time each one occupied in the horizontal one. As we can see in Figure 1.1, it is an indicative scheme for the approximate duration of each phase that has been carried out in this project.

The idea of this project came up quite early where the state of art was revised and examined to make sure it would be feasible. However, a long time passed until we started with the procedure of the ethical Committee due to the fact of talking with the hospital clinicians and managers so that they agree with the development of this project.

Furthermore, as we can see, the most time-consuming phases were ethics committee approval and the preprocessing, together obviously with the development of this document.

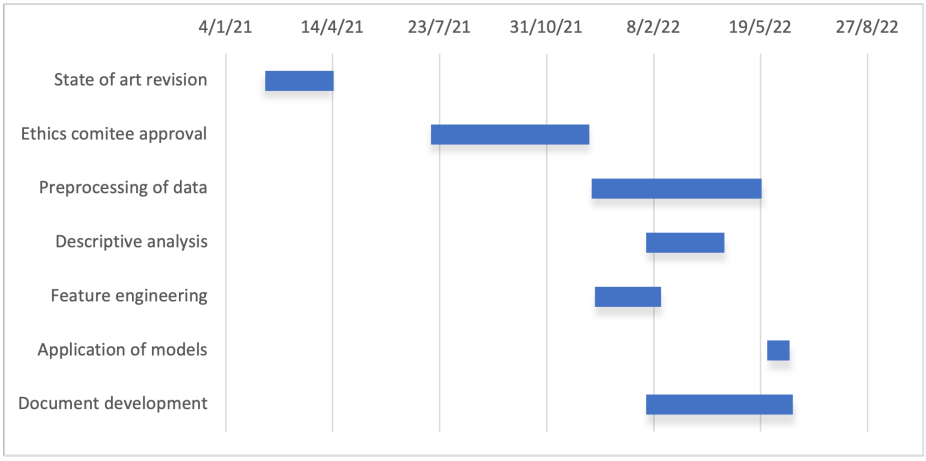

Figure 1.1: Gantt´s Diagram showing the project´s phases.

1.4 Project structure

Once the main objectives and methods of the project have been correctly stated, it is necessary to establish its structure consisting in the following chapters:

- **Chapter 1: Introduction and objectives**. In this first introductory chapter, there can be found four sections. In the first one, the project´s context is described as well as the inspiration of the same. In the second section, the purpose of it are brought up and the different techniques that will be performed to achieve the objectives. In the third one, it is located the Gantt´s diagram which indicates the duration of each part of this work´s process. Finally, in the last section, the project´s structure is found briefly explaining each part.

- **Chapter 2: Previous concepts**. In this chapter, clinical basic concepts are explained with the aim of easing the understanding of the reader of this work. Specifically, the definitions of HA surgery, SSI and some main characteristics to take into account about a surgery of this kind. In addition, a brief introduction of the state of art is performed by developing some studies similar to this one.

- **Chapter 3: Methods**. In this third chapter, the different concepts about ML are introduced followed by the different techniques and procedure used during all the preprocessing and *FE*. Moreover, methods to imbalanced classes solutions are explained as well as the ML models such as LR, MLP, DT and a couple of ensemble methods.

- **Chapter 4: Exploratory data analysis**. In the fourth chapter, a deep explanation of the database and its variables is performed as well as its descriptive visual analysis. Furthermore, the *FE* and preprocessing procedures are developed describing all the changes applied to the dataset.

- **Chapter 5: Experiments and results**. In this part of the project, the several tests carried out and the different steps of them are detailed. Moreover, their results are profoundly studied and analyzed.

- **Chapter 6: Conclusion and future sights**. In the last part of this project, results are summarized and the applicable conclusions obtained are presented. Additionally, some future sights are proposed about this field of study.

Chapter 2

Previous concepts and state of art

In this chapter, a description of the main concepts of the study is performed with the aim of easing the understanding of the project and the range of variables analyzed. Firstly, definition of HA surgery is exposed. Afterwards, SSI explanation is presented. And finally, different features of the surgery are developed with its subsequent influence on the appearance of further complications.

2.1 Hip arthroplasty surgery

HA surgery, hip prosthesis or HA, is a surgical procedure carried out by an orthopaedic surgeon that consists of the removing the injured part of hip joint and replacing it them with new, artificially synthesised parts. It is mainly performed in order to reduce pain and recover the function and mobility by mimicking the function of the normal hip joint. HA surgery may be needed if you have a disease, such as arthritis, osteonecrosis, or if you suffer from an injury or bone disease that will end up fracturing your hip [37].

The hip joint is a spheroid joint and is the body´s most mobile and one of the biggest one. It is composed by a meeting between the head of the femur which acts as a ball and the pelvis of the hip bone called the acetabulum which acts as a socket. Therefore, what is looked for in this type of surgery is to substitute the acetabular head and introduce the prosthesis in the femur for its fixation [38].

It is a very common surgery performed worldwide [37]. It offers really good clinical outcomes in the short term and even in the long one with very high implant success rates in 20-year and 35-year follow-ups corresponding to >80% and 78% respectively. On the other hand,

among the causes of implant failure, infection is one of the most common causes of revision and thus, its prevention would be a great advance.

2.2 Surgical site infection

For explaining what a SSI is, first, it is needed to establish the wound classification as on it depends the later SSI. The classification of surgical wounds is done depending on the degree of contamination at the time of the procedure and it is determined by a person involved in the surgery (surgeon, nurse, etc.) based on the wound class schema adopted within a organization. The four wound classifications available within the National Healthcare Safety Network (NHSN) and the National Research Council (NRC) application are [39] [40]:

- **Clean:** Surgical wounds that are not infected where no inflammation is seen and none entry is made into the respiratory, alimentary, genital, or urinary tracts. Clean wounds are also typically stitched shut and, when needed, drained using closed drainage. If they meet the requirements, surgical incisional wounds that result from blunt trauma should be classified under this heading.
- **Clean-contaminated:** Surgical wounds where the pulmonary, gastrointestinal, genitourinary, or urinary tracts are penetrated under sterile circumstances and without exceptional contamination. This group includes procedures involving the biliary tract, appendix, vagina, and oropharynx, providing that there's no sign of infection or significant technical error.
- **Contaminated:** Exposed, recent, unintentional wounds. This category also includes incisions where acute, nonpurulent inflammation is observed as well as surgeries with significant sterile procedure breakdowns, large gastrointestinal tract spillage, or both.
- **Dirty-infected:** Wounds from previous trauma that still have devitalized tissue, as well as wounds with active clinical infections or visceral perforations. According to this definition, this kind of infection is caused by germs that were present in the operating field before to the procedure.

As it can be inferred, the higher the degree of contamination of a surgical wound, the higher the probability of having a infection [41].

SSI are those infections that happen after a surgery in the part of the body that was intervened. They can be superficial if they involve only the skin or can become more serious by involving tissues under the skin, organs or implanted material such as a hip prosthesis [40]. The following classification of SSIs is established [40]:

- **Superficial incisional SSI:** In this type of infection, the event´s date falls within 30 days after any NHSN surgical procedure establishing day 1 as the procedure date. Only skin and subcutaneous tissue is affected and the patient must suffer one of the following at least [40]:
  - Superficial incision´s purulent drainage.
  - Organism(s) detected using a culture-based or non-culture-based microbiological testing method from a specimen collected aseptically from the superficial incision or subcutaneous tissue and used for clinical diagnosis or therapy.
  - Surgeon´s deliberately opened superficial incision and the superficial incision or subcutaneous tissue is not tested using a culture-based approach or non-culture-based approach. The patient also exhibits at least one of the symptoms or indicators listed below: regional soreness or tenderness; high fever.
  - Physician´s superficial incisional SSI discovery.
- **Deep incisional SSI:** In this type of infection, the event's date falls between 30 and 90 days after the NHSN surgical procedure. Additionally, it includes the incision's deep soft tissues such as fascia, the patient must suffer one of the following at least [40]:
  - Deep incision´s purulent drainage.
  - Deep incision spontaneously dehiscing or when it is purposefully opened or aspirated by a surgeon or doctor, the organism(s) found in the deep soft tissues of the incision are identified using a culture-based or non-culture-based microbiological testing method that is carried out for the purpose of clinical diagnosis or treatment, or it is not carried out. This criterion is not met by a culture or non-culture based test from the deep soft tissues of the incision that yields a negative result. The patient also exhibits at least one of the symptoms or indicators listed below: regional soreness or tenderness; high fever.
  - A deep incision abscess or other indication of infection that is found by a gross anatomical, histological, or imaging examination.

- **Organ/Space SSI:** In this type of infection, the event's date falls 30 or 90 days following the NHSN surgical procedure. Additionally, it includes any area of the body that is opened or moved during surgery that is deeper than the fascial and muscle layers the patient must suffer one of the following at least [40]:
  - Drain´s purulent drainage from one placed into the organ/space such as T-tube drain.
  - Using a culture-based or non-culture-based microbiological testing technique, an organism (or organisms) isolated from fluid or tissue within the organ or space is (are) identified for the purpose of clinical diagnosis or therapy (e.g., not using Active Surveillance Culture/Testing (ASC/AST)).
  - A finding of infection proof on a gross anatomical or histological examination, an imaging test, or other indication of infection involving the organ or space. Furthermore, it satisfies at least one of the requirements listed in the Surveillance Definitions for Particular Types of Infections for a given organ or space infection location.

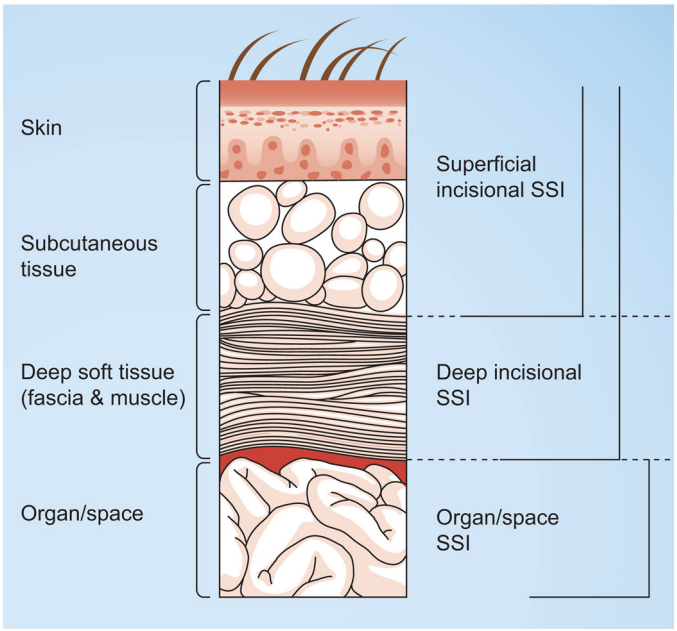

Figure 2.1: SSI types classification depending on the depth. Obtained from [1].

2.3 Surgery characteristics

Regarding the surgical procedure, there are several concepts that have a relevant influence in the later appearance of an infection and therefore, need to be explained for further understanding.

ASA physical status classification

The ASA Physical Status Classification System [42] is used to assess and communicate the patient´s pre-anesthesia medical co-morbidities. It has been used for 60 years now and it does not only predict the perioperative risks but also other factors (type of surgery, frailty or level of deconditioning). It is considered a relevant variable for determining the presence of postoperative issues such as infection.

The definitions and examples shown in the Figure 2.2 are guidelines for the clinician. Primarily to enhance interactions and evaluations at a particular school as anesthesiology departments may decide to create examples unique to their own institution to supplement the ASA-approved examples.

ASA Classification	Definition
ASA I	A normal healthy patient
ASA II	A patient with mild systemic disease without functional limitation
ASA III	A patient with severe systemic disease with functional limitation
ASA IV	A patient with severe systemic disease with functional limitation and constant threat to life
ASA V	Dying patient, unlikely to survive 24 hours

Figure 2.2: ASA physical status classification system.

The clinical decision of assigning a physical status classification level is based on multiple factors such as if the patient has obesity, is a usual smoker or drinker and has diseases such as hypertension or diabetes. At first, the physical status may be determined during several instances of the preoperative assessment of the patient, however, the final assignment is made after evaluation the patient on the day of anesthesia care mainly by the corresponding anesthesiologist [43]. Usually it is not difficult to distinguish between types of ASA.

Duration of the surgery

The duration of a surgery refers to the time span in minutes that the surgery lasts. It is stated, as defined by the Association of Anesthesia Clinical Directors (AACD) [44], as the interval

between: the Procedure/Surgery Start Time (PST), which is the moment at which the procedure begins; and the Procedure/Surgery Finish Time (PF), which corresponds to the moment when it finishes, the procedure is considered as ended once all instruments are completed and verified as correct, all postoperative radiologic studies to be done in the operating room are completed, all dressings and drains are secured, and the physicians/surgeons have completed all procedure-related activities on the patient [44].

According to the Centers for Disease Control and Prevention, the duration of an operation is one of the reliable predictors of SSI risk, in fact, it has been noticed [45] that the revision risk for THA patients with procedures lasting more than 240 minutes is significantly higher than that for the median group. Longer procedures may constitute an indicator of perioperative complications, complex surgery, inexperienced surgical team, poor standardization programs, or patients' preexisting conditions. Moreover, a patient's body mass index (obesity level) has also been shown to directly increase operative time [45].

Preoperative antimicrobial prophylaxis

Preoperative antimicrobial prophylaxis refers to the antimicrobial agent that is given to the patient just before the beginning of surgery with the aim of decreasing the risk of postoperative infections [46]. The use of antimicrobial agents for prophylaxis constitutes the most effective method of minimizing the prevalence of postoperative wound infections after HA surgery [47]. The agent is selected by the surgeons and infection disease experts based on the agent´s efficacy against most common pathogens causing SSI for a specific operation [48]. Specifically, in our case, HA, the ideal prophylaxis administration would be a one that would possess excellent activity against *Staphylococci* and *Streptococci*, whose timing is just before the skin incision, with intravenous administration and the exact duration to avoid antimicrobial toxicity [47].

In the database of use in this project, prophylaxis is classified as suitable or unsuitable. This last term is further classified as unsuitable in choice, timing, duration, indication, beginning and administration according to antibiotic policy guidelines issued by the antibiotics committee of the hospital. It will be deeply explain in Chapter 4.

Antimicrobial prophylaxis is one topic on which the *Indicadores Clínicos de Mejora Continua de la Calidad* (INCLIMECC) system focuses given its known importance in reducing the risk of SSIs [49].

2.4 State of art

Before finishing this chapter, it is essential to look over preceding works belonging to the same field. After a deep search, the following similar studies have been found:

- *Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty?[50]:* In this work, the main goal was to use ML methods and large national databases to develop and validate close risk-prediction models for mortality and complications after total joint arthroplasty (TJA). For achieving it, it was used a National Surgical Quality Improvement Program (NSQIP) database as well as least absolute shrinkage and selection operator (LASSO) regression methods. Relatively accurate predictive models of 30-day mortality and cardiac complications after elective primary TJA were developed that provided the most accurate and rigorously validated TJA-specific prediction models that were available in 2019 [50].

- *Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty [51]:* The main objective of this study was to develop a ML-based ensemble algorithm for being able to predict major complications after THA. Using a database of patients who underwent primary THA at any California-licensed hospital between 2015 and 2017 and an automated ML framework called AutoPrognosis. The result was an algorithm that improved risk prediction when comparing it to logistic regression and standard benchmark models [51].

- *Prediction of Early Periprosthetic Joint Infection After Total Hip Arthroplasty [52]:* In this one, the purpose was to develop a parsimonious risk prediction model for periprosthetic joint infection (PJI) within 90 days after THA. Logistic LASSO regression using ASA class, BMI, age, sex and comorbidities presence as variables obtaining a good calibrated model with an area under the curve (AUC) of 0.68 [52].

These three were the most similar ones found, being the last one [52] quite new as it was published in 4th March 2022. However, there are also few other models specifically for predicting infection in total knee arthroplasty (TKA) and joint arthroplasty in general with quite impressive results [53] [54].

In comparison to the aforementioned works, this project aims to develop a ML model to predict exclusively infection in patients that undergo HA surgeries by using the variables pro-

vided by INCLIMECC, something provided by every Spanish hospital. Therefore, this project has a national scalability.

Chapter 3

Methods

In this third chapter, the methodology employed during this project is detailed. Including an introduction of the main concepts of ML as well as the design and evaluation of the model. Furthermore, FE techniques are explained together with the ones for dealing with imbalanced classes and categorical variables as well as those that have been used for performing FS. And last but not least, predictive ML algorithms of the model and their figures of merit will be also described.

3.1 Fundamental concepts of Machine Learning

The basic goal of many scientific fields is to model the relationship between a collection of observable quantities (inputs) and a set of variables associated to these inputs (outputs) [55]. By measuring the observable events, it is possible to predict the value of the desired variables once a mathematical model has been established. Unfortunately, many real-world occurrences are far too complicated to be represented properly as a closed form input–output relationship. By evaluating the given data and optimizing a problem-dependent performance objective, ML algorithms can automatically develop a computer model of these complex interactions [55].

For being able to build a ML model, data is needed, usually in the form of a dataset. Datasets are composed of N samples, in our case patients, and d features or labels. Each sample is associated with a set of features that define it, hence the i-th sample is expressed as $x^i = [x^i_1, x^i_2, ..., x^i_d]$, a vector of d features for the sample number i. Each feature composes a dimension of the feature space and the exact value of a feature for a precise data point places the point in a defined place in this dimension of the space.

The dimensionality of the resulting feature vector and space is proportional to the amount of feature collected for the dataset [56]. If the dimensionality is too high, it can hinder the correct learning of the model and therefore end up with the curse of dimensionality.

Moreover, before continuing, it is essential to explain the training and test sets. Databases used in ML problems are usually divided in two sets, training set is the one used to train the model so that it can learns from experience and test set is the one used to evaluate the learning of our model and being able to assess its performance. In most cases, training and test sets are usually composed by approximately 70% and 30% of the samples, respectively. In this project, the same division percentages are going to be applied.

ML techniques can be divided into two main groups:

- **Supervised learning**: In this kind of problems, the output values corresponding for each patient are known, called labels. The general goal of these techniques are to map inputs to outputs in order to predict the output from new unseen data whose input values are observed. Here we find two main categories of problems: classification, when the input values are categorical, and regression, when the input values are continuous/numeric [57].
- **Unsupervised learning**: In this kind, there are no labels presented so the input is known but not the output. This type is usually used to discover hidden patterns in the data. Once we introduce new data into the model, in order to identify the label of the data, it employs the previously learned features, hence, it can be used as a preprocessing step for a supervised learning problem. Some of the main methods used are clustering and dimensionality reduction [56].

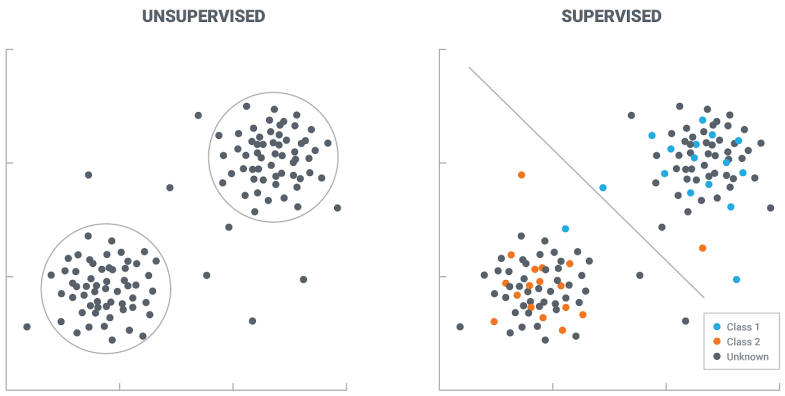

Figure 3.1: Supervised vs unsupervised learning. Obtained from [2].

3.2 Design and evaluation

As already stated, for this project, supervised learning techniques are going to be applied as the clinical database provided contains the labels indicating whether each patient suffered infection or not. In addition, a previously FE and preprocessing stage was applied to the database in order to fix the inconsistencies and prepare the database for the algorithms.

Afterwards, all the instances of the database are divided into train and test sets assigning 70% of the patients to the training set and 30% to the test one. In addition, there has been generated 5 different subsets of the dataset and thus, 5 different training and test sets. This will allow us to increase the reliability of the results as it is computed as the mean between all of them, but also, we can obtain the standard deviation as well.

This train sets are the ones used by the algorithm to learn as it has been previously stated. However, they are also used for hyperparameter tuning of each algorithm that we are implementing by using a technique called *Cross-Validation* (CV) and happens at the same time that the training stage. With the implementation of this, we are able to find the best value of each hyperparameter for our model to subsequently use those values for its training.

For carrying out this CV, we have used the *model_selection.GridSearchCV* from the *scikit learn* python library which has allowed us to vary the scoring depending on our interests.

The CV technique carried out is a *k-fold cross-validation* which means that we are performing it with k subsets which in our case, we have used it with $k = 5$. The working of this technique, as we can see in Figure 3.2, consists of division of the dataset into k different subsets, each one containing *k-1* parts for the training process and 1 for testing one. Also, in every subset, the testing and training part varies.

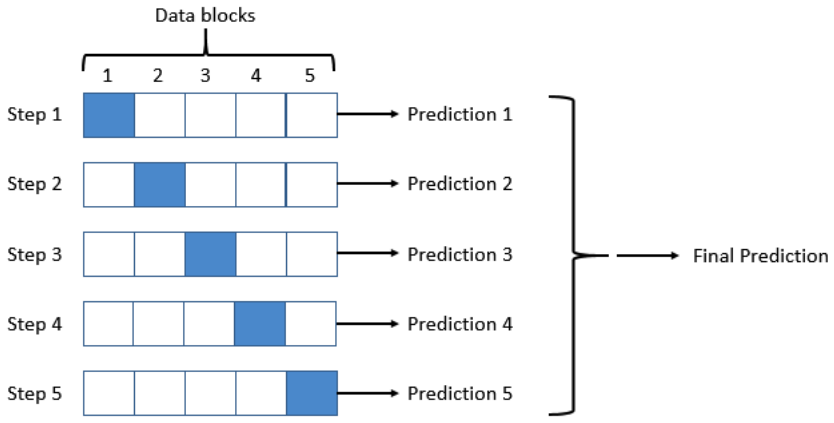

Figure 3.2: *5-fold cross-validation* working. Obtained from [3].

3.3 Feature engineering

The urge to convert non-normally distributed linear regression inputs created FE. For linear regression, such transformations can be useful. George Box and David Cox presented a method [58] for identifying which of numerous power functions would be a helpful change for the outcome of linear regression in their key work in 1964 [58].

ML algorithms need to be introduced a certain number of features to train the model [59]. One of the most difficult tasks is determining the appropriate amount and type of such features from the attributes of a dataset. Each feature in the database has a more relevant or more irrelevant value going from useless to very significant variables. Nonetheless, the key to achieve the best possible result and fitting for ML is to detect the optimal set of the selected feature set with the finest matching of the feature's value [59].

Therefore, FE constitutes a crucial step in the ML data preparation process. The representation of the feature vector has a significant impact on ML performance [60]. As a result, designing preprocessing pipelines and data transformations takes up a lot of time and effort when it comes to deploying ML algorithms [58].

It consists on constructing suitable features from other given features that will lead to an improved predictive performance. FE involves the application of transformation functions such as arithmetic and aggregate operators on existing features with the aim of generating new ones. Transformations make it easier to scale a feature or to turn a non-linear relationship between a feature and a target class into a linear relationship [60].

Regarding this project, several FE methods were applied such as those related to obtaining new features like the age of each patient or the times of the surgical process, but also one hot encoding.

One hot encoding

One hot encoding is a method of FE to convert a categorical variable into several binary variables as we can see in figure [61]. It transforms our data in order to be ready to be introduced into a model and to achieve better results. It works by changing each categorical value of a feature into a new binary column as it can be visualized in 3.3.

This technique becomes really useful for data having no relationship between each other. The order of numbers is understood by ML algorithms as an attribute of significance. In other

words, a larger number will be interpreted as better or more significant than a smaller number.

While this is useful in some ordinal scenarios, other input data lacks a ranking for category values, which can cause problems with predictions and performance. That's when one hot encoding becomes the key to solve them.

In general, thanks to one hot encoding, training data becomes more useful and expressive. By using binary columns, much better results are obtained in most cases [62].

S1 ASA
3
2
1
5
4

One hot encoding →

S1 ASA 1	S1 ASA 2	S1 ASA 3	S1 ASA 4	S1 ASA 5
0	0	1	0	0
0	1	0	0	0
1	0	0	0	0
0	0	0	0	1
0	0	0	1	0

Figure 3.3: Example of one hot encoding working with ASA feature.

3.4 Methods to deal with imbalanced classes

One of the main issues facing the database which it has been worked with in this study, something that happens in most of the previously similar studies on same field topics, was the great difference in number between the infection and non-infection patients. Only a 3% of patients that underwent a HA surgery between 2010 and 2020 suffered later infection which actually, it is a quite good percentage meaning that the immense majority of them had a correct adaptation of the implant.

The imbalanced classes problem [63] is often characteristic of two class classification problems in real world. The rarity is an important issue in the field of ML because, very often the most important cases belong to minority class. When a classification system is trained on such data, the prediction accuracy for the majority class cases is extremely likely to be significantly greater than the prediction accuracy for the minority class instances. This is because the model tries to reduce global quantities such as the overall error rate, which is heavily affected by the prior probabilities of the two classes.

For tackling this problem, several techniques can be applied [63] with the aim of ending up with balanced classes. To avoid excessive bias, a 60%-40% needs to be achieve at least but obviously the best case scenario is having a 50%-50% database. The techniques used to pass from 97%-3% to 50%-50% in this project has been undersampling and oversampling.

3.4.1 Undersampling

It consists on reducing the number of samples in the majority class until it reaches the size of the minority one or any other desired size. As a results, the number of instances is greatly reduced which may not be beneficial to a certain model, however, training time is also reduced. Another drawback is that we are eliminating members from the majority class, so it is possible the loss of valuable information if samples that could be useful to our classifier in building an accurate model are eliminated [63].

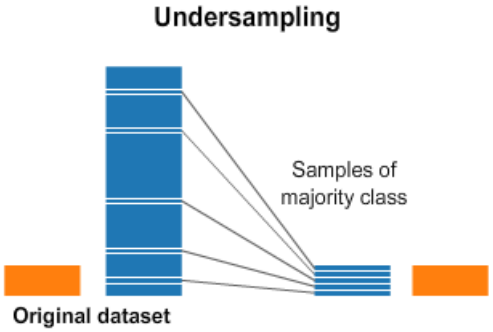

Figure 3.4: Undersampling working. Obtained from [4]

3.4.2 Oversampling

In this case, it is sought to increase the number of samples of the minority class. This process can be performed through two different approaches: heuristic and non-heuristic. In the non-heuristic one the oversampling is obtained by replicating randomly selected samples belonging to the minority class, while in the heuristic approach, the choice of samples to eliminate or to duplicate is informed rather than random [64].

The advantage of oversampling is that we keep all samples from the minority and majority classes, thus no information from the original dataset is lost. The drawback is that the size of it is substantially increased. As a result, we extend the training period and increase the amount of memory necessary to retain the training set [63].

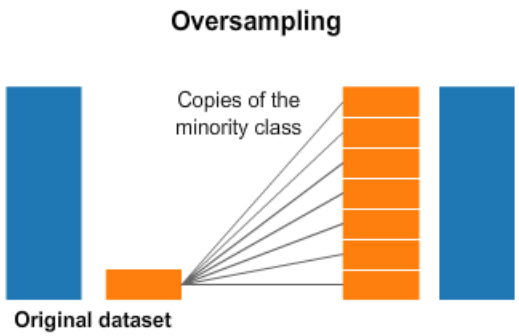

Figure 3.5: Oversampling working. Obtained from [4]

3.4.3 SMOTE

The last technique to solve an imbalanced problem is Synthetic Minority Oversampling Technique (SMOTE) [65]. In contrast to undersampling and oversampling whose generation of new instances is completely random, this method carries out the creating of new instances by operation in "feature space" instead of "data space", corresponding to a less application-specific manner. This means that the minority class is oversampled by taking the minority class instances and inserting synthetic ones along the line segment and joining the k minority class nearest neighbors. The value of k, as it is the amount of samples that we take, depends on the amount of oversampling that we want to perform [65].

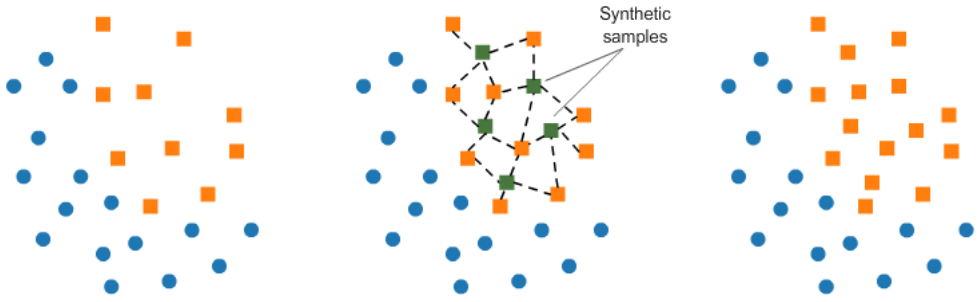

Figure 3.6: SMOTE technique working. Obtained from [5]

Synthetic samples are generated by calculating the difference between a sample and its nearest neighbor and multiplying it by a random number between 0 and 1. This makes the selection of a random point in the segment of the line between two precise features which increases the generalizability of the decision region and subsequently, less specific decision regions are learned. The effect of this a model is that it can generalize better than if we perform

random oversampling [65]. That is why SMOTE method is the one that will be employed in this work.

Regarding this work, there was an extremely huge imbalance and performing a single undersampling or oversampling would not be suitable. On one hand, with undersampling we can potentially remove important examples of the majority class, as it is mentioned above, which will derive in an underfitting of the majority class. On the other hand, oversampling would lead to overfitting since the increasing amount of minority class samples can lead the learner to identify similar but more specific regions in the input space as the decision region for the minority class [64].

Therefore, for solving the imbalanced classes issue in our dataset, primary undersampling of the majority class is performed followed by a later oversampling of the minority one, in order to level the two classes.

3.5 Feature selection

Feature selection (FS) is one of the most used and crucial data preprocessing approaches, and it's become an essential part of the ML process. It consists of detecting those most important features and ruling out the irrelevant, redundant, or noisy ones [66]. This process is usually performed in order to achieve the following:

- Increase interpretability. By reducing the number of features, the model becomes simpler and thus, easier to understand. Moreover, there is a better understanding of the underlying process.
- Speed up algorithms training and output calculations which can be quite useful in real time environments.
- Reduce dimensionality and therefore avoid overfitting. Apart from being more complex, ML algorithms with too many features tend to overfit.

The procedure of FS is composed of four stages, namely subset generation, subset evaluation, stopping criterion, and result validation. Subset generation is a search process using the certain search strategy. By using specific evaluation criterion, a subset feature is created and further evaluated with the prior best feature subset. If the new one is better than the previous, then it is replaced. This process is repeated until a certain stopping criteria is fulfilled and after it, the best final subset is validated using synthetic data or real-world data set.

There are three main approaches for FS depending on the way they work:

- *Filter methods*: This kind of methods provides a way of evaluation feature subsets without implicating the use of a ML algorithm through the use of a independent measure [66]. They basically evaluate the relevance of each variable by individually evaluating if each one is important to discriminate the target by ranking them according to a predefined relevance score, so that low-scored variables are removed. Some of benefits of these kind of techniques are their simplicity and speed, however, they do not take into account possible interactions between the variables [67].

- *Wrapper methods*: The only difference between wrapper and filter methods is the evaluation criteria. These approaches employ a learning algorithm in order to evaluate each subset, a selected ML algorithm acts as a black box in order to score feature subsets according to their ability to perform prediction. Wrapper techniques require to define a classification algorithm, a relevance criterion and a searching procedure in which we can have brute force, randomized and greedy strategies. The drawback of this kind of methods is that they depend on the ML algorithm that was used and that they have a high computational cost[67].

- *Embedded methods*: These ones create a lower computational cost than the previous wrapper. It collects feature dependencies and it takes into account the relations between input and the output feature. Furthermore, it looks for variables that enable a better local discrimination capacity [66]. In this type of methods, the FS process is incorporated into the training process of the model, thus relying on the internal design of the learning algorithm for intrinsically selecting. They can be combined with wrapper´s greedy strategies to get hybrid methods that constitute very efficient schemes for FS [67].

In this project, we have focused in filter methods as we have applied the Mutual Information one.

3.5.1 Mutual information

Mutual information (MI) is filter method for FS which is used for ranking features according to a certain criteria. This criteria used is entropy which is a key measure of information widely employed in various fields due to its capacity of measuring the random variables´ uncertainty and scaling the quantity of information supplied by them effectively [68].

The MI between two random variables X and Y is a measure of the amount of knowledge on Y supplied by X. If both features are independent, X contains no information about Y and thus, the MI is zero [69]. Mathematically, MI is defined as:

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{3.1}$$

where $p(x,y)$ is the joint probability of X and Y, and $p(x)$ and $p(y)$ correspond to the marginal probabilities of X and Y respectively [70].

By writing it depending on entropies, we have:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X;Y) \tag{3.2}$$

being $H(X)$ and $H(Y)$ the marginal entropies of X and Y respectively. The conditional ones measure the uncertainty in X after knowing Y and viceversa, they are $H(X|Y)$ and $H(Y|X)$. Finally, the joint entropy for X and Y is $H(X;Y)$ [70].

3.6 Machine learning methods

ML algorithms aim to perform certain tasks by learning a certain model from a set of training samples and variables that are collected and represented in a dataset. Then, in this chapter, used supervised ML algorithms in this project are deeply explained.

3.6.1 Logistic regression

LR is standard probabilistic statistical classification model. It produces a binomial outcome by giving the probability of an event to happen in terms of 0 (not happening) and 1 (happening) based on several input variables [71]. In our case, predicting if the patient is going to suffer infection after surgery (1) or not (0). The result of a LR model is a probability but we need to convert it to a binary outcome, to achieve this, the following logistic function is used so if the

probability is higher than 0.5, the outcome would a 1 and otherwise, it would be a 0:

$$p(t) = \frac{1}{1+e^{-t}} \tag{3.3}$$

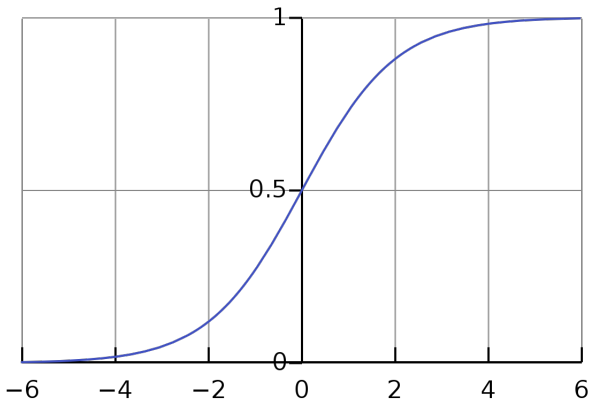

Figure 3.7: Logistic function curve. Obtained from [6]

Regarding the training of a LR, its objective is to set the parameter vector θ in order for the model to calculate high probabilities for positive samples (y = 1) and low odds for negative ones (y = 0). This is represented in the following cost function:

$$c(\theta) = \begin{cases} -log(\hat{p}) & if \quad y = 1 \\ -log(1-\hat{p} & if \quad y = 0 \end{cases} \tag{3.4}$$

Being y the outcome of the function. If the model estimates a probability near 0 for a positive case, the cost will be high, and if the model guesses a probability near 1 for a negative instance, cost will be even higher. If the estimated probability is near 0 when negative samples or near 1 in the case of a positive ones, the cost will be close to 0, which is exactly what we want [72].

When using LR models, there is a problem of overfitting of the training data, especially when the data are high dimensional and the training data are sparse. This issue can be reduced by using regularization [73]. It is performed by adding to the function to be optimized during learning a penalty term dependent on the model parameters and which allows to find a balance between a solution that does not overfit the design cases and provides the lowest possible error.

The impact of this regularization term [67] is controlled by weighting the penalty term by an adjustable multiplicative parameter λ which is called the regularization parameter. It imposes certain constraints on the solution, for solving overfitting (complex boundaries), we need a high

λ to smooth them.

The two simplest and most used techniques for this are Ridge and Lasso which differ on that Ridge has the penalty term squared [67]:

$$\begin{aligned} Ridge : l(w) &= logL(w) + \lambda ||w||^2 \\ Lasso : l(w) &= logL(w) + \lambda ||w||^1 \end{aligned} \tag{3.5}$$

3.6.2 Decision trees

DTs are a non-parametric supervised learning method used both for classification and regression. The predictor space is stratified or segmented into a number of simple regions. To construct a prediction for a given observation, we usually utilize the mean or mode of the training observations in the region where it is located. These methods [67] are known as DT methods because the set of splitting criteria used to segment the prediction space may be described in a tree. One of the main advantages of tree-based methods is that, regarding its interpretation, they are quite simple and useful. Nevertheless, they usually fall behind when competing with some other supervised learning techniques [67].

The building of a DT given a training set X, y, with $\mathbb{X} \in \mathbb{R}^{NxD}$ and $y \in \mathbb{R}^{Nx1}$ carries out in the following manner:

1. Begin with the root node that contains all training samples.

2. Apply a threshold test on the root node, so two branches are created. By selecting the predictor x_d and the threshold s that best splits the predictor space into two regions.

3. Recursive splitting: on each branch, apply a threshold test. This can performed until every sample is correctly predicted but to avoid this, which would increase the complexity and decrease the performance maybe leading to overfitting too, stopping criteria is used.

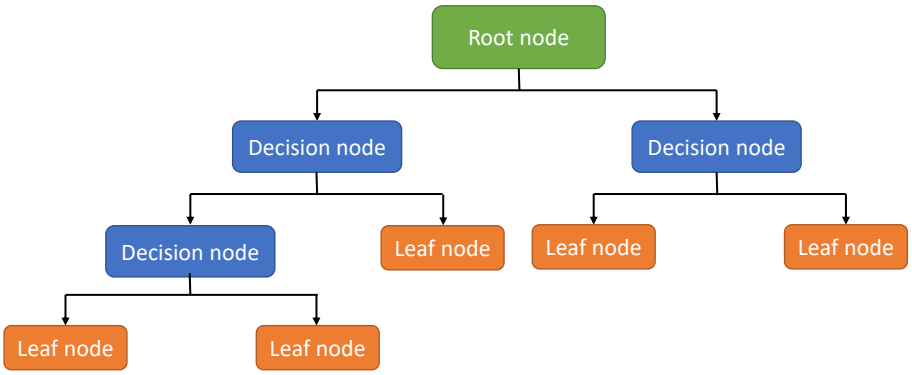

Figure 3.8: Example of a DT´s structure.

The threshold is usually selected by setting all the possible values for it, evaluating the figure of merit of each one, which in the case of classification we have the classification error rate, Gini index and cross-entropy (deeply explained afterwards), and finally choosing the one that provides the best figure of merit.

The most straightforward way of tackling overfitting is using a certain stopping criteria and thus, avoiding the tree to grow to its full size, it is called pruning [74]. Some of this pruning techniques are establishing a maximum depth for the tree and when it is achieved, the tree stops growing; or setting the minimum number of samples required to form a leaf node, which is basically a terminal node that has no additional nodes coming from it.

In the case of this project, as we are dealing with a classification problem, classification DTs are going to be implemented. Regarding this type, they are almost the same as regression ones but with the difference of predicting a qualitative outcome rather than a quantitative one. The prediction is carried out by assigning an observation to the majority class in the region of that instance. In the results interpretation, it is usually of interest not only the class prediction of a certain region, but also the proportions of that precise class among the training set that fall into that region [67].

To grow a DT, the *classification error rate* is used as classification metric. We intend to classify an observation of a certain region into the majority class in that precise region, thus the *classification error rate* is basically the proportion of training observations in a given location that do not correspond to the most prevalent class of a problem of K categories [67]:

$$E = 1 - max_k(p_k). \tag{3.6}$$

Being p_k the proportion of training observations at a node n for the kth class.

Apart from that, we have also other classification metrics [67]:

- Gini index: it is a measure of total variance across the K classes. It is a measure of homogeneity or purity, a small value indicates that a node is containing the majority of the observations from the same class and thus, it is more pure.

$$G = \sum_{k=1}^{K} p_k(1 - p_k) \tag{3.7}$$

- Cross-entropy: It will take a value close to zero if the $p'_k s$ are all near zero or near one. Therefore, like the Gini index, the smaller the value, the purer is the *n-th* node.

$$D = -\sum_{k=1}^{K} p_k \log(p_k) \tag{3.8}$$

Regarding the hyperparameters that DTs have, we have a lot of them but the most relevant and the ones that will be employed in this project are the following:

- *max_depth*: Integer number representing the maximum depth of a tree. If None, then the nodes are expanded until all leaves are pure which means that every sample is correctly predicted and this leads to increase the overfitting and complexity of the tree.
- *criterion*: Referring to the two previously explained classification metrics, gini or entropy. In our case, we have chosen gini to perform all the analysis which provided way better results.
- *min_samples_leaf*: It is an integer number that indicates the minimum number of samples needed to be reach to form at a leaf node.

3.6.3 Random forest

Apart from the previously described methods, we have another group of them included in what are called ensemble methods [67]. These techniques basically combine the predictions of several single estimators built with the same training data to increase the robustness and generalizability of a model by introducing diversity. This combination can be done by taking the mean of the predictions of the single estimators with or without weights, by taking the mode of

the predictions or performing majority vote of them, among other ways. In the end, what we are trying with this kind of methods is to decrease the estimator´s variance by averaging the results [75].

We have two major families of ensemble methods:

- Bagging: several estimators are built independently and their predictions are averaged.
- Boosting: several estimators are built sequentially and one tries to reduce the bias of the combined model.

In the case of RF, we are talking about a bagging method that builds a series of uncorrelated trees and then averages them. For a forest composed of B DTs, for $b = 1, ..., B$ [75]:

1. Create a subsample with replacement of the training set.
2. Train a DT on this subsample and for each split, a random sample of $m < D$ predictors is chosen as split candidates (each node only has access to a subset of the input predictors).

In our case, as we are working with a classification problem, the predictions are made by using the majority voting while in regression, averaging is performed.

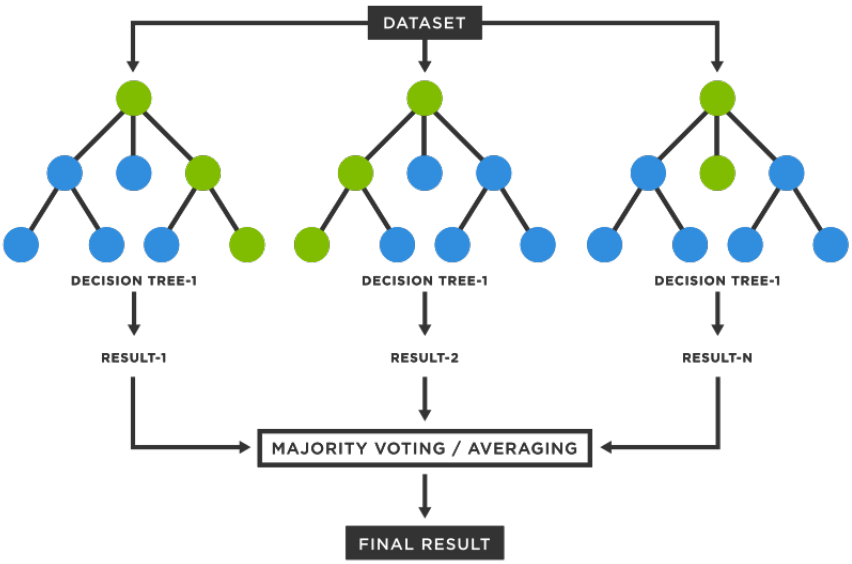

Figure 3.9: RF building example with N DTs. Obtained from [7].

Regarding the hyperparameters of RF, it shares them with the DTs which have been previously explained above. But the only one that we have added is the number of estimators (*n_estimators*) which refers to the number of trees of the forest. It is an extremely relevant parameter used in some cases to avoid overfitting and ease the computational process.

3.6.4 XGBoost

In this case, instead of being a bagging method such as RF, Extreme Gradient Boosting (XGBoost) is a tree boosting ensemble method [76] which is widely used in the data science world to achieve different ML challenges. Gradient tree boosting is a technique that gives brilliant results in many applications, it has been shown to achieve results close to the state-of-the-art on various standard classification benchmarks. The most crucial component in XGBoost's success is its scalability across all scenarios. On a single machine, the system is more than ten times faster than existing popular methods, and it scales to billions of samples in distributed or memory-limited environments. XGBoost's scalability is due to a number of major systems and algorithmic optimizations [76].

As it has previously explained, boosting ensemble methods such as XGBoost work by using sequential estimators to try to reduce the error of the previous one in an additive manner as we can see in the figure 3.10 below.

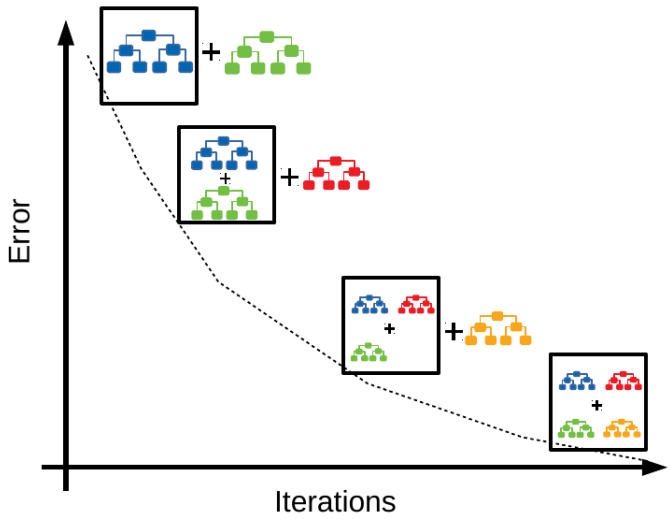

Figure 3.10: Boosting tree technique structure. Obtained from [8].

In a given set of data with n instances and m features $\mathbb{D} = (x_i, y_i)(|\mathbb{D}| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$. In this way, XGBoost predicts the output as:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{3.9}$$

where K is the number of additive functions and F is the space containing the set of classification and regression trees also known as CART [76].

Furthermore, to learn the functions that are used in the model, an objective function with regularization is used:

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.10}$$

where Ω is the penalization parameter that determines the complexity of the model and $\hat{y}_i^{(t-1)}$ will be the prediction of the *i-th* instance in the *t-th* iteration.

Regarding the hyperparameters introduced in this method, we have taken *max_depth* and *n_estimators*, limiting the depth and number of trees as it has been explained in the previous methods but we have added *gamma* which sets the minimum loss reduction needed to carry out a further partition on a leaf node of the tree and the higher it is, the more conservative the algorithm will be.

3.6.5 Multilayer perceptron

Artificial neural networks (ANNs) are neural computation systems that were proposed by McCulloch and Pitts (1943) and Metropolis et al. (1953). They are inspired by the network structure of biological neurons without reaching to the brain´s huge complexity, but we can find two main similarities among biological neural networks and ANNs [77]. First, both networks' building blocks are simple computational devices that are closely interconnected. And second, the network's function is determined by the connections between neurons. They are ML algorithms that, in some applications, achieve a higher performance than humans.

Their use was boosted in the 1980s due to very important advances regarding computational methods about self-organizing properties and information systems. Specifically, a learning procedure was proposed called backpropagation (BP) [77]. This method continuously adapts the weights of the network connections in order to minimize a measure of the difference between

the network's calculated output value and the real output value.

One type of ANN are the MLP, they are fully-connected neural networks very popular and used more than any other type of neural network used in huge range of problems. A MLP, as well as any other ANN, consists on three layers with non-linear computational elements called neurons or processing units. These three layers are: input layer, which collects information to be processed; output layer, where the calculation of the results of the processing are found; and units in between known as hidden layers where the processing occurs [74].

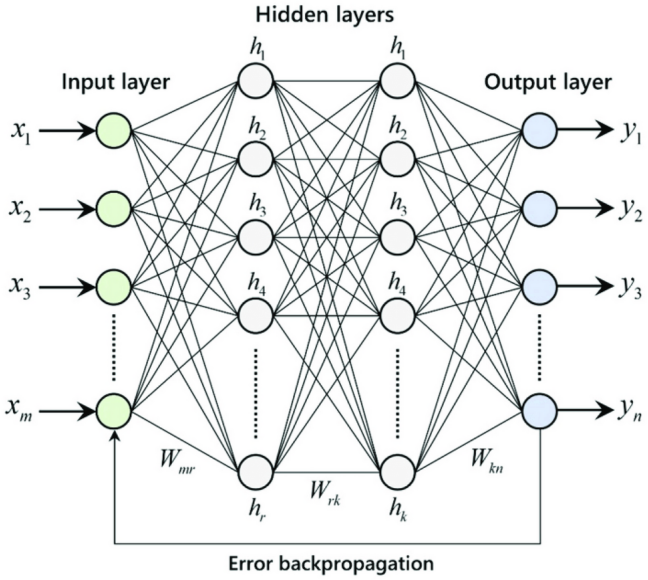

Figure 3.11: ANN architecture with 2 hidden layers. Obtained from [9].

All neurons from one layer are fully interconnected to neurons in the bordering layers. These connections are represented as weights [77]. They are coefficients that adapt with the aim of determining the intensity of the input signal, they establish the influence of the input on the output as some inputs may be more important than others regarding the final result. The amount of neurons in the input layer is established depending on the number of independent features that we have in the model, thus the number of neurons in the output layer is the same than the number of dependent variables. Both the number of hidden layers and their neurons are crucial parameters in the construction of the MLP model because the model´s complexity rely on them .

The training of a MLP [77] is performed in order to optimize a cost function, this is, to minimize the difference between the real target values and the values calculated by our model. If the network returns a bad output, or if the difference is bigger than a previously established threshold, the weights are adapted in order to minimize them. Therefore, training consists

on finding the weights that produce the best outcome. The process of learning for a MLP is composed by a forward propagation step followed by a backward one.

Forward propagation begins with the reception of data by the MLP by way of the input and output neurons and over-viewing it. The net input to a certain neuron of the hidden layer is calculated as the summation of each output of the input layer multiplied by a certain weight. An activation function is applied to calculate the output of that neuron of the hidden layer and the output of a neuron of the output layer. This is performed by adapting the sum of the activation to a value between 0 and 1 and/or by establishing, except if a threshold level of said sum is reached, the activation value to zero [74]. This activation functions are employed to bring out some non-linearity into the model and they can be sigmoids: $sigm(x) = \frac{1}{(1+e^{-x})}$, hiperbolic tangents: $tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$ or Rectified Linear Unit (RELU): $relu(x) = max(0,x)$, among others.

One possible training method is to use a technique called gradient descent. Backpropagation training uses this method to try to locate the absolute minimum of the error surface. It basically propagates the error from the output neurons to the input ones for the calculation of the gradient in each point and thus, for the appropriate weight adjustment. To do so, the chain rule is used to determine each interposed variable's and parameter's gradient. The order of calculations is reversed compared to forward propagation because It is necessary to begin by working with the computational graph's results the way towards the parameters [78].

Regarding the hyperparameters of this method, we have used the following ones:

- *alpha*: Strength of the L2 regularization term. It is used to avoid overfitting by limiting the size of the weights.
- *hidden_layer_sizes*: Number of neurons and hidden layers in the network. In our case as we are working with a small database, we have used one hidden layer and inside several possible numbers of neurons.
- *max_iter*: Referring to the maximum number of iterations. The solver iterates until this number is reached.
- *learning_rate*: It marks the way that the network learns. We have worked with 'adaptive' with consists on keeping the learning rate constant as long as training loss keeps decreasing. It decreases the learning rate each time there is no decrease in the training loss or no increase in the validation score.

3.7 Figures of merit

In this section, the metrics to assess the performance of the classification models is explained in order to allow the accurate interpretation of the results obtained.

3.7.1 Confusion matrix

The first way of assessing our model is by generating a confusion matrix. It consists on a 2x2 table layout relating the number of predicted and actual values in order to see the amount of correctly predicted samples and otherwise. Used generally in classification supervised problems.

As we can see in Figure 3.12, we have:

- **True Positive (TP)**: Referring to the patients that were predicted positive and it is actually true. In our case, those ones that had infection and were predicted as that.
- **True Negative (TN)**: Referring to those patients that were predicted negative and it is actually true. In our case, the ones that did not have infection and were predicted as that.
- **False Positive (FP)**: Referring to those patients that were predicted positive and the actual value is negative. The ones that did not have infection but were predicted as infection.
- **False Negative (FN)**: Referring to those patients that were predicted negative and the actual value is positive. The ones that did have infection but were predicted as non-infection.

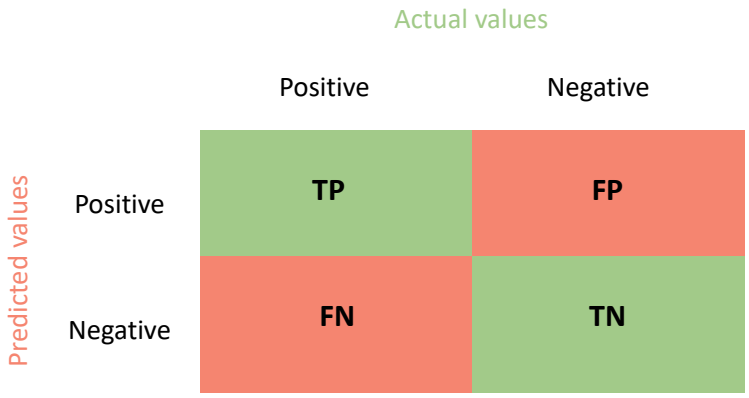

Figure 3.12: Confusion matrix layout.

By means of this matrix, we can extract several evaluation metrics, all of them have values in the 0 to 1 range while 1 corresponds to the perfect score and 0 the lowest possible value:

- Sensitivity/Recall: It represents the ability of the model to detect positive instances. It is expressed as:

$$Sensitivity = \frac{TP}{TP+FN} \tag{3.11}$$

- Specificity: It represents the ability of the model to detect negative instances. It is expressed as:

$$Specificity = \frac{TN}{TN+FP} \tag{3.12}$$

- Accuracy: It represents the ability of the model to predict correctly. It is expressed as:

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \tag{3.13}$$

- Precision: It represents the model's ability in classifying a sample as positive.

$$Precision = \frac{TP}{TP+FP} \tag{3.14}$$

- F1-score: It is used to combine sensitivity and precision into a single value. It indicates the test´s accuracy regarding these two metrics.

$$F1\text{-}score = 2\frac{precision \cdot sensitivity}{precision + sensitivity} \tag{3.15}$$

3.7.2 ROC-AUC

Receiver operating characteristic (ROC) curve is further performance measurement used mainly in classification models. It consists on a graphical plot that represents the differentiation capacity of a given model. The ROC curve is obtained by the plot of the true positive rate (TPR)(Same that Sensitivity) against the false positive rate (FPR)(1 - Specificity).

ROC curve is used to help us select possible optimal models and rule out not optimal ones. There is a link in a direct and natural way of ROC analysis, to a cost-related assessment of decision making when diagnosing. [10].

A single point in the ROC space is determined by the combination of one TPR and one FPR, and the position of a point in the ROC space reflects a tradeoff between sensitivity and specificity. In an ideal situation, a point identified by both TPR and FPF yields a coordinates (0, 1), or the upper left corner of the ROC space. This concept point denotes that the diagnostic test has a 100% sensitivity and 100% specificity. It is also called perfect classification [79].

In theory, any random guess would give a point along the red discontinuous diagonal. A point predicted by a diagnostic test falling into the area above the diagonal means a good diagnostic prediction, otherwise a bad one [79]. It can be seen clearly in Figure 3.13.

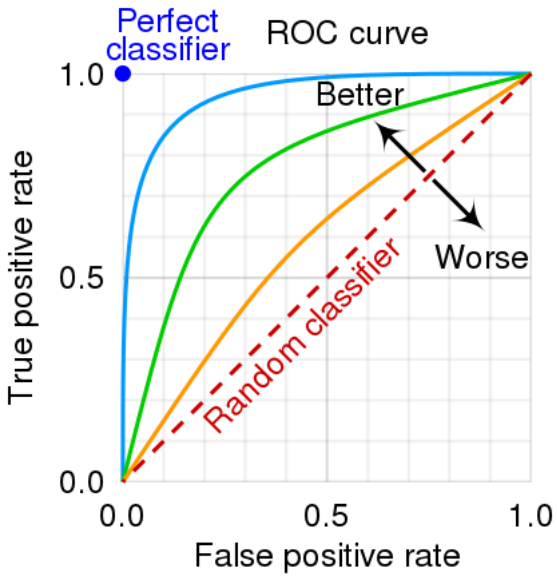

Figure 3.13: ROC curve showing better and worse results. Obtained from [10].

From the ROC curve, we can obtain the AUC. It is basically the area under the ROC curve and the more area, the better. A perfect classification model would have an AUC of 1. It provides a way to measure the accuracy of a prediction model. According to [79], an AUC between 0.9 and 1 is excellent, between 0.8 and 0.9 is good, between 0.7 and 0.8 is worthless and between 0.6 and 0.7 is not good.

Chapter 4

Exploratory data analysis

In this chapter, it will be carried out a description of the dataset of use as well as explanation of the preprocessing stages that were applied to it and finally a visual analysis of the variables with the aim of deepening the knowledge about its variables, distribution and organization but also, for the further understanding of the methods performed.

4.1 Database description

The databases used for the development of this study has been provided by HURYC. Naturally, it is completely anonymized, having been approved by the Ethics Committee for Clinical investigation of Hospital HURYC. The only identifier for each patient is the EHR number, therefore it is impossible to know the identity of the patients under study. The databases come from INCLIMECC from the first year of record until 31st December 2020, however, only patients undergoing surgery between 2010 and 2020 were studied mainly for quality reasons as data between 2000 and 2010 did not represent a reliable source and was very inconsistent.

INCLIMECC is a HAI´s prospective surveillance system that collects data on incidence in surgical and intensive-care unit patients. The procedure is grounded on the National Healthcare Safety Network (NHSN) surveillance system and employs US Centers for Disease Control and Prevention´s standard definitions of infection [48].

Regarding the study population, operated adult patients ($\geq$18 years) are included mainly because no young patients are intervened due to the possible complications of the surgery. The patients are operated both in a programmed or urgent way, during the aforementioned study period. The surgical procedure following the NHSN-CDC model is the hip arthroplasty (HA).

Patients were included in a consecutive way, and a tracking was performed since admission until the date of discharge, recording any revision for infection or any other complication derived from the previous surgery.

The data have been collected by the nursing staff from the Preventive Medicine department, with specific training for that, and supervised by the doctors of mentioned service. The information sources were the EHRs, nursing annotation, vital and clinical signs records, diagnosis techniques and microbiological test results as well as the direct contact with the healthcare team.

The dataset is composed of 2172 patients that have undergone a hip replacement surgery, partial hip arthroplasty, total hip arthroplasty or hip arthroplasty revision. Out of those 2172, only 66 infections were registered corresponding to a 3% of the patients, creating a clear imbalance between the two classes. Furthermore, it contains information about patients discharges and admissions, the surgery and its characteristics and details about the infection, if produced. We can identify the following groups of variables:

- *Variables related to the patient*: contain information concerning the patient and its characteristics such as the date of birth, sex and EHR number.
- *Variables related to the admission*: contain information about the patient´s stay in the hospital. They are related to its admission, discharge, transfer (if existing) and origin.
- *Variables related to the surgery*: contain all the characteristics of each surgery carried out in the hospital. This includes the preparation, prophylaxis, duration and cause among others.
- *Variables related to the infection*: contain information about the infection if existing such as its date, location and microorganisms.

Having state the four classes of variables that have been worked with, they are further explained below. Moreover, mention that these variables correspond to the original ones meaning that they have not been subjected to the preprocessing stage yet.

- **Birthdate**: date of birth of each patient found in the format Month/Day/Year.
- **EHR number**: continuous and numerical variable corresponding to the EHR number of each patient.
- **Sex**: categorical variable indicating the sex of the patient assigning a value of 2 to women and 1 to men.

- **Admission date**: date of admission of each patient found in the format Month/Day/Year.
- **Discharge date**: date of discharge of each patient found in the format Month/Day/Year.
- **Admission**: categorical variable indicating if the admission was programmed (1) or urgent (2).
- **Discharge type**: categorical variable indicating the type of discharge assigning 1 to home, 2 to transfer to another hospital, 3 to voluntary discharge, 4 to death and 7 to elopement.
- **Admission type**: categorical variable indicating the type of admission, it has several cases: 1 (New admission, if patient does not have previous episodes), 2 (Readmission due to complication/surgery, patient who has suffered a complication that needs intervention), 3 (Readmission for chronic disease) and 4 (Major ambulatory surgery).
- **ASA Surgery 1**: categorical variable that indicates the degree of ASA which is a physical status classification system for patients before undergoing surgery. It has five possible degrees: 1 (Healthy patient, no systemic infection), 2 (Mild systemic disease without functional limitation), 3 (Acute systemic disease with functional limitation), 4 (Acute systemic disease with functional limitation and life threat) and 5 (Dying, not likely to survive 24 hours).
- **Service Surgery 1**: categorical variable that indicates the service that performed the surgery 1. There is a code for each service of the hospital but the ones that appear in the dataset are: from 500 to 505 (Traumatology) and from 510 to 515 (Orthopedic traumatology).
- **Revision cause Surgery 1**: categorical variable indicating the cause of revision regarding surgery 1. It can take the values of: 1 (Bleeding), 2 (Infection), 3 (Osteosynthesis material removal), 5 (Post-surgery functional disorders), 9 (Two-times treatment) and 0 (Other).
- **Duration Surgery 1**: numerical and continuous variable that indicates the duration (in minutes) of the surgery 1.
- **Date Surgery 1**: date of surgery 1 found in format Month/Day/Year.
- **Procedure Surgery 1**: categorical variable indicating the procedure that was carried out during surgery 1. It can take a huge amount of values collected in Clasificación Internacional de Enfermedades, 9ª Ed. Modificación Clínica (IDC-9) having a code for each surgical technique but in the database we can only find 81.51 (Total hip arthroplasty), 81.52 (Partial hip arthroplasty) and 81.53 (Hip arthroplasty revision).

- **Prophylaxis Surgery 1**: categorical variable indicating the level of prophylaxis that was administered in surgery 1. It has the following options: 0 (None, not applicable), 1 (Yes, administered), 2 (Treatment, in case of patient subjected to dirty surgery in antibiotic treatment) and 3 (None but indicated).

- **Type of Surgery 1**: categorical variable indicating if the surgery was performed on a programmed (1) or urgent (2) way.

- **Contamination Surgery 1**: Categorical variable indicating the degree of contamination regarding surgery 1. It can take: 1 (Clean surgery), 2 (Clean-contaminated surgery), 3 (Contaminated surgery) or 4 (Unclean/Infected surgery).

- **Assessment of prophylaxis Surgery 1**: categorical variable that indicates the evaluation of the prophylaxis administered in surgery 1. It takes: 1 (Adequate), 2 (Inadequate in duration), 3 (Inadequate in election), 4 (Inadequate in indication), 5 (Inadequate in beginning) and 6 (Inadequate in administration way).

- **Date of Infection 1 of Surgery 1**: date of the infection 1 produced by surgery 1 in the format of Month/Day/Year.

- **Microorganism of Infection 1 of Surgery 1**: categorical variable that indicates the microorganism that caused the infection 1 of surgery 1. It can take a huge amount of values, there is a code for each bacterium, fungus or virus.

- **Location of Infection 1 of Surgery 1**: categorical variable indicating the location of the infection 1 of surgery 1. There is a number designating each possible surgical infection site but in the dataset we have just 6, 7 and 48 being SSI, deep SSI and organ or space SSI, respectively.

- **Type of Infection 1 of Surgery 1**: categorical variable that indicates the type of infection 1 of surgery 1. It can be hospital acquired (1), hospital acquired but in another service (2) and extra hospital acquired (3).

As it can be inferred from the variables, they are repeated for each surgery that the patient would overcome existing a second and even a third surgery. However, the amount of patients that underwent more than one surgery was insufficient to have certain statistical relevance. For that reason, the variables referring to those extra surgeries were removed.

4.2 Feature engineering

As it has been explained in Chapter 3.3, FE constitutes an essential part of the process as the quality of the data has a huge weight in the final result, the success of the project strongly relies on it.

As we can notice from the list of variables developed in Section 4.1, there are four variables that are dates and working with them for python algorithms is harder and may be confusing for them. Therefore, to ease its performance, it is preferred to calculated time differences between certain dates which will allow to get useful information from them. The following features have been extracted:

- **Age**: by subtracting the admission date and the birth date of each patient, the age that the patient had when he/she was admitted is calculated.
- **Preoperative time (Days)**: from surgery and admission dates, the time that a patient is hospitalized before surgery is calculated.
- **Postoperative time (Days)**: from surgery and discharge dates, the time that a patient is hospitalized after undergoing surgery is found.
- **Outcome**: the binary classification feature has been created so that the work can be performed in a supervised way. It has been created by assigning a one if there is a date of infection and a zero otherwise, as the existence of a date indicates that a infection occurred.

These new variables can be extremely determinant when it comes to performing prediction tasks. Timing is crucial in these types of surgeries, for example, if the patient has a longer postoperative time, it is probably due to surgery complications or a poor recovery and may have an influence on the later appearance of infections. Same happens with preoperative one, if the patient needs more days usually means that he/she has not been able to go through surgery before due to its health condition.

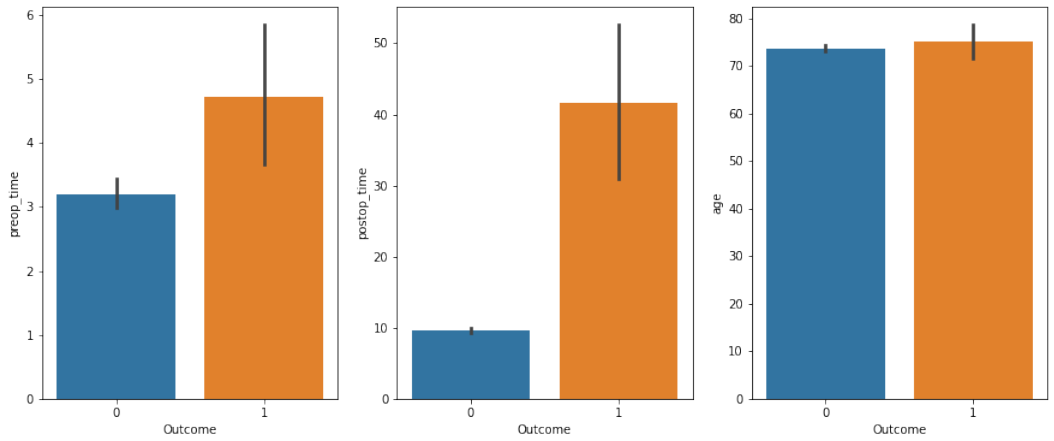

Figure 4.1: Distribution of the engineered new variables, preoperative time, postoperative time and age, in both classes.

As we can see in Figure 4.1, the postoperative time is highly related with the outcome as people that suffer infection 4 times more time than the ones that did not, meaning that the higher the time after the surgery in the hospital, the higher the probability of suffering from infection. Also, we can notice in Figure 4.2 how infection patients tend to have longer postoperative times. However, we can see that age is not a decisive variable and preoperative time do not have as high influence as the postoperative time.

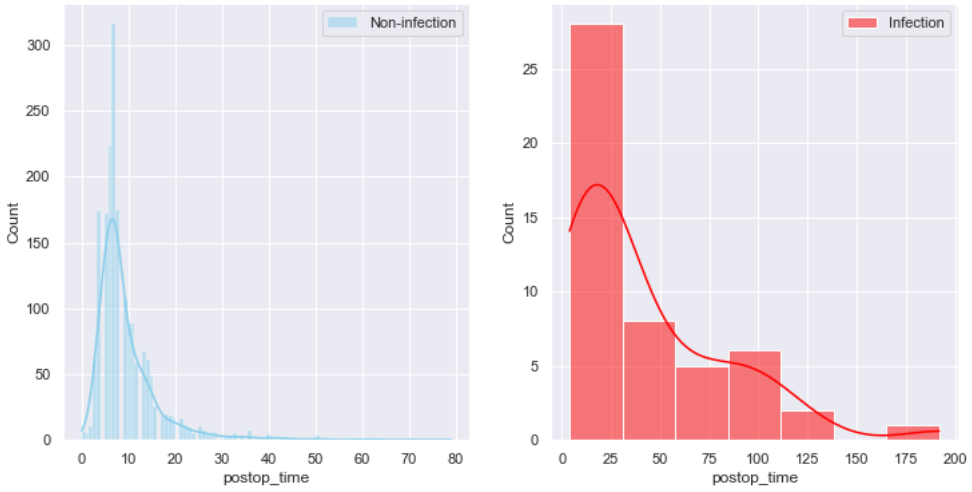

Figure 4.2: Histogram showing the distribution of infection and non-infection patients along the postop_time values.

Apart from creating features, we carried out an extensive procedure of removing the ones that were not useful or did not gave information to the model:

- The features (Admission, discharge, surgery 1, infection 1 and birth dates) used to extract the ones previously explained were evidently removed as, once the time differences have been obtained, they are not useful anymore.
- EHR number and revision cause of surgery 1.
- The data related to the infection such as its location, microorganism and type was also ruled out because they are a posteriori information that will not be used for prediction evidently.

4.3 Preprocessing

Preprocessing stage refers to all that actions carried out on the database for the purpose of transforming it for the later use of ML methods. As mentioned in section 4.2, qualitative data is needed and that is why in this project, there has been paid a special attention on this part. The database needed a well-defined analysis, cleaning and structuring to start being useful and ready to be used for prediction algorithms, thereby, a series of modifications have been applied.

This procedure basically consisted on cleaning the data of missing values. Apart from removing those variables with an excessive percentage of null values (*NaN*) such as the ones of following surgeries as we have mentioned above, there were still some features presenting a small amount of them and that needed to be solved. Therefore, as they were categorical variables, the median of each feature was found and replaced in those blank spaces.

As an special case, *S1 procedure* needed a huge remodelling as there were lots of mistaken instances with surgery codes different from the ones we are working with. Therefore, those patients were eliminated.

4.4 Descriptive analysis of the data

In this section, an extensive visual analysis of the variables is performed with the aim of increasing the knowledge on which ones has a stronger effect in the process of determining if a patient will have infection or not.

Descriptive analysis of continuous variables

Regarding the final continuous variables that will be employed in the study, we have four of them: S1 Duration, postop time, preop time and age.

As it is shown in Figure 4.3, the time (in days) is usually much higher in the case of postoperative one than in the preoperative even doubling or tripling it in some instances. When it comes to the age we can see that the common range of age of patients that undergo HA is around 67 to 83 years old and the normal duration of a surgery of this kind is between 120 and 150 minutes.

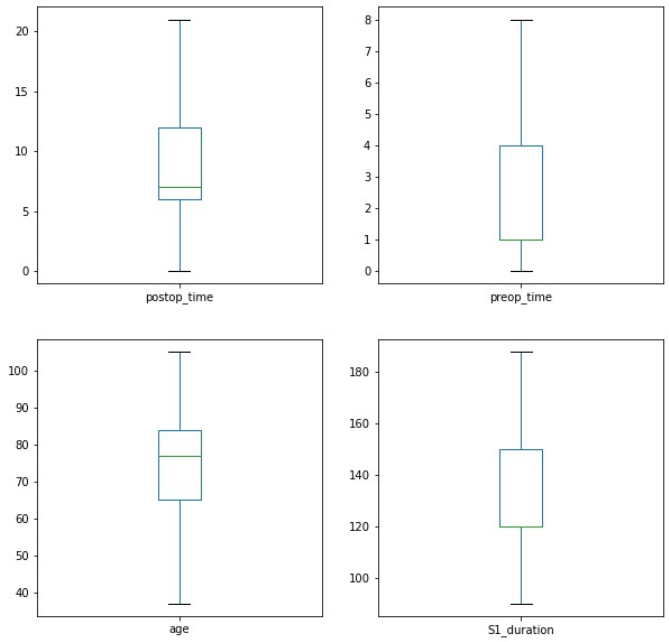

Figure 4.3: Boxplots of the continuous variables, age, postoperative time, preoperative time and surgery duration.

Descriptive analysis of categorical variables

Regarding the final categorical variables, we have the rest ones where we can highlight some of them.

First, the sex seems to be something to take into account as we can see in Figure 4.4. Nearly

80 % of patients with infection are women while slightly more than 20 % are men which is not completely reasoned by the fact that there are more women intervened than men. Therefore, we can state that women are more likely to suffer from infection after a HA surgery.

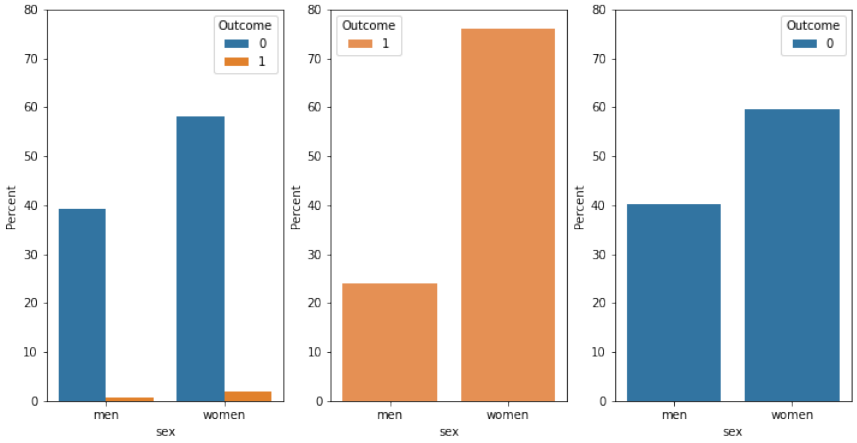

Figure 4.4: Barplot of the sex variable with respect to the outcome.

Second, as we can see in Figure 4.5, contrary to what anyone could expect, almost every patient that has suffered infection after surgery has been intervened in a programmed way which means it was not immediate and not performed with urgency. The fact of being an urgent procedure means it is carried out in a faster manner and the patient has not the possibility of having a ideal preoperative preparation which is thought to increase the risk of infection but seems not.

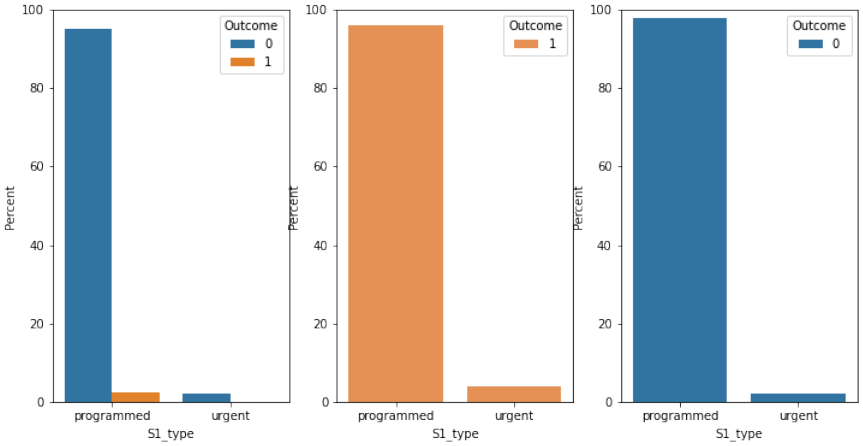

Figure 4.5: Barplot of the S1 type variable with respect to the outcome.

Moreover, one very relevant variable in prediction, as we can see in Figure 4.6, should be mentioned. The prophylactic assessment determines whether the prophylaxis has been performed correctly or not and we can see that the inadequate prophylaxis in duration is duplicated in patients suffering infection compared to the patients that do not.

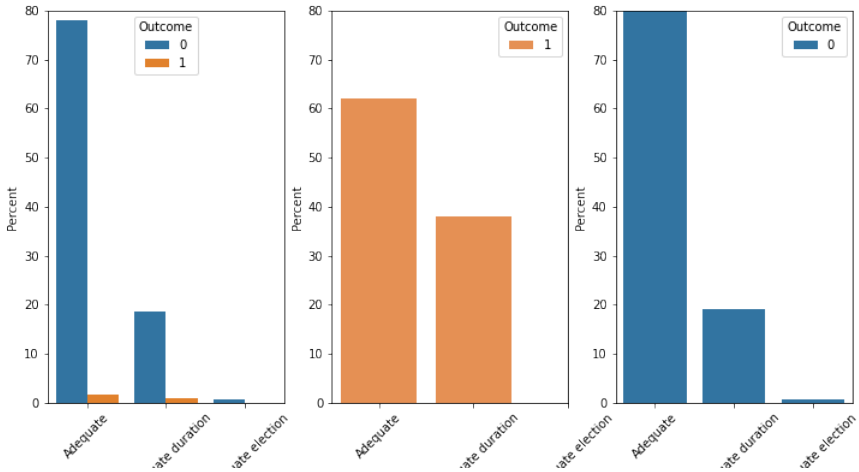

Figure 4.6: Barplot of the S1 prophylaxis assessment variable with respect to the outcome.

Chapter 5

Experiments and results

In this chapter, the experimental stage of the project is presented. Firstly, the classification problem definition is stated. Followed by the experimental configuration that has been used. And finally, the results of the proposed experiments are showed and examined by selecting those metrics providing the best results.

5.1 Problem definition

As it has been mentioned several times in this document, the objective of this project is to be able to predict by basing on several clinical and surgical parameters, once a patient is intervened of HA, if he/she is going to suffer an infection. Therefore, we are dealing with a problem of binary classification where we need to know if a patient´s SSI is going be infected (Outcome = 1) or if it is not (Outcome = 0).

For the achievement of this, a database provided by HURYC was used with information from patients undergoing HA between 2010 and 2020. This dataset is composed of 2172 instances, however, after an extensive preprocessing process previously explained in Chapter 4, we finally ended up with 1923 patients and 14 variables after extensive preprocessing, FE techniques to improve the quality of the data. Also, as we are working with a quite imbalanced dataset with 2 % of infected patients and 98 % of non-infected ones, undersampling and oversampling procedures have been carried out explained in Section 3.4.

5.2 Experimental set-up

For the experimental setup, as it has previously been explained, a modified database is used with the corresponding techniques of FE and preprocessing as it has been detailed in Section 4.3 and 4.2. Once we have our data with certain consistency and quality, we can know perform our experiments.

In the first place, we have divided the dataset into 5 different partitions of train and test. This will allow us to avoid depending just on one set and increase the reliability of the models by computing the mean and standard deviation between the partitions. A 70% of the database is going to be assigned to train and a 30% to test, in every partition.

Afterwards, we need to have a balanced train sets. Hence, an random undersampling of the majority class is performed followed by a SMOTE, both using the python module *imblearn*. Furthermore, we have performed the experiments with three different undersampling strategies: 0.8, 0.6 and 0.25. This numbers refer to the sampling ratio of the first undersampling where we basically reduce the number of samples of the majority class and the higher the number we use, the stronger the reduction will be and thus, the lower the number of instances. After undersampling, we apply a SMOTE which, as it is detailed in Section 3.4.3, balances both classes so that they have the same number of instances.

Therefore, we will end up with the following sets:

Sampling Strategy	Partition	Train (Label 0 - 1)	Test (Label 0 - 1)
0.25	1	136-136	561-16
	2	140-140	562-15
	3	124-124	558-19
	4	156-156	566-11
	5	144-144	563-14
0.6	1	56-56	561-16
	2	58-58	562-15
	3	51-51	558-19
	4	65-65	566-11
	5	60-60	563-14
0.8	1	42-42	561-16
	2	43-43	562-15
	3	38-38	558-19
	4	48-48	566-11
	5	45-45	563-14

Table 5.1: Distributions and partitions used in the experiments.

The experiments that will be carried out are the following ones:

- **Experiment 1 (E1)**: In the first one, the full final dataset is going to be used with all its features.
- **Experiment 2 (E2)**: In the second one, the same will be done with the difference of performing a previous FS process that will enable us to select those characteristics more relevant of each model.

In both experiments, same methods will be used with same hyperparameters. We have used the *GridSearch k-fold cross validation* with k=5 which will allow us to determine the best hyperparameters for each partition of the dataset. The hyperparameters used has been detailed in Section 3.6. For the tuning of each hyperparameter, it has been established a set of possible values that it could take with the aim of avoiding overfitting and being able to achieve the best possible results.

Moreover, we can use cross-validation to manually check for the best hyperparameter and see if our model is overfitting or not. It can be found using validation curves, as we can see in the two example figures 5.1 and 5.2. The first one referring to the hyperparameter max_iter of LR and the second one to the n_estimators of RF method. Also, they are interpreted by checking the difference between the training curve (in blue) and testing curve (in green). If the difference is quite high, the model tends to overfit and by analyzing the curves below, we can make sure that the model is selecting the correct values without overfitting.

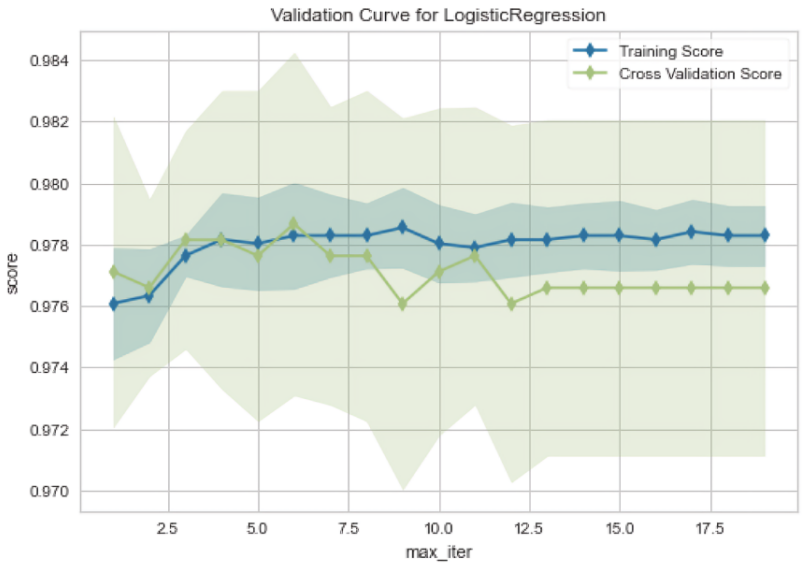

Figure 5.1: Validation curve of the hyperparameter max_iter in the LR method.

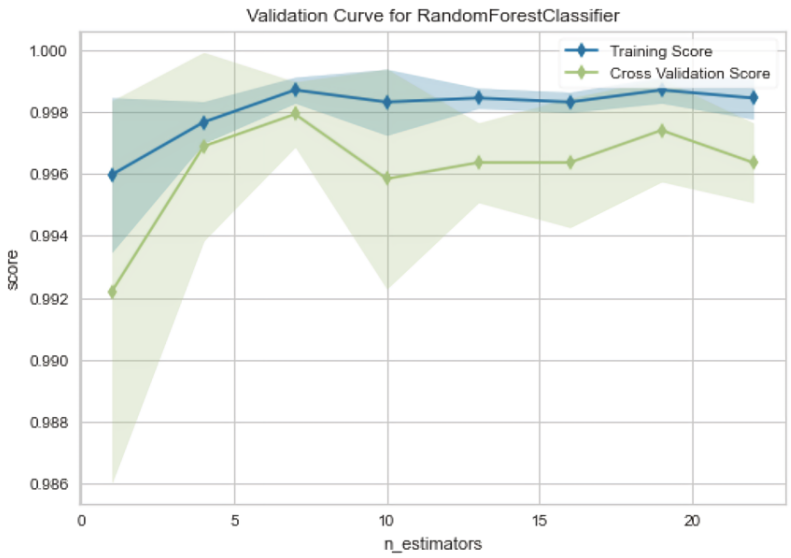

Figure 5.2: Validation curve of the hyperparameter n_estimators in the RF method.

In addition, apart from varying the sampling rate, it has also been varied the scoring criteria of each model. By using the *GridSearch* *k*-fold cross validation, we could also vary the strategy to assess the performance of the cross-validation model on the testing set. In our case, we have evaluated two different scorings: accuracy and ROC-AUC.

5.3 Feature selection results

As it has been explained in the previous section 5.2, FS is implemented in order to carry out the second experiment (E2) where the same balancing and preprocessing techniques have been applied, the only difference with E1 is the set of features we are working with. Regarding the FS methods applied, it has been performed just the MI filter method.

5.3.1 Mutual information

As it has been deeply explained in Section 3.5.1, MI ranks the variables depending on the amount of dependency between two variables which in our case has been done between the outcome and the rest of features. Thus, those ones with the highest values of dependency with the outcome are the ones that are going to remain while the ones with almost none relationship are going to be dropped. For it, the *feature_selection.mutual_info_classif* has been used, obtained from the *scikit learn* library of python.

We can see it Figure 5.3 that there is a huge prevalence of the postoperative time achieving a score of almost 1. It is subsequently followed by admission type and the contamination of the surgery. Then by the duration, preoperative time, age and type of surgery which as it has been mentioned in Section 4.4, this last has a surprising distribution regarding the outcome. Therefore, the features that are gonna be discarded are sex, admission, discharge type, ASA, surgical procedure, prophylaxis and the assessment of the prophylaxis, which do not give almost any useful information to the model. In the end, E2 will be performed with 7 variables instead of 14.

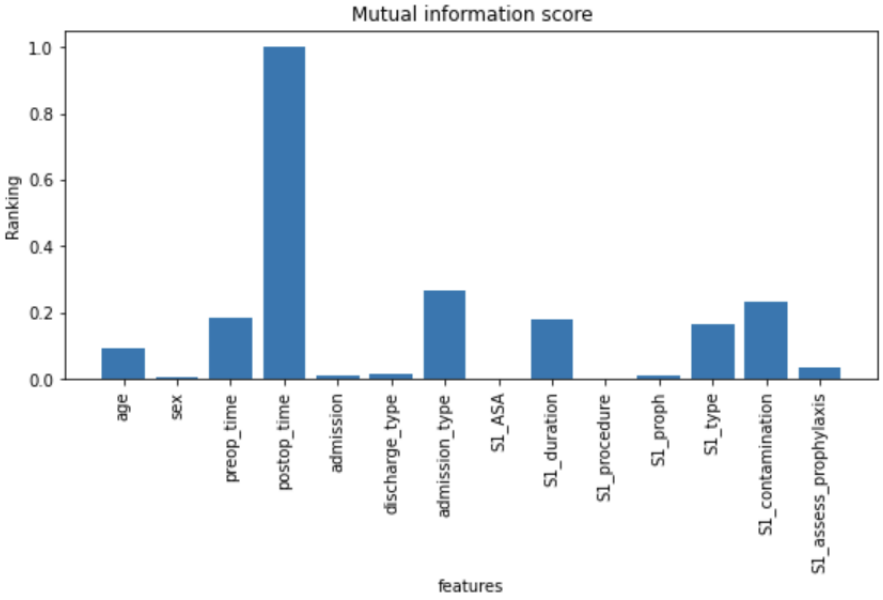

Figure 5.3: MI feature importance showing the ranking of each variable.

5.4 Results

In this section, the results of both experiments are posed and analysed according to the metrics explained in section 3.7. From each experiment, we obtain the values of accuracy, sensitivity, specificity and F1-score and ROC-AUC.

5.4.1 Results of Experiment 1

The results of E1 can be seen in Table 5.2. As it has been stated in Section 5.2, it has been carried out with 15 features and the experimental set-up aforementioned. As we can see, there a division for each sampling strategy and scoring criteria for each type of model.

Best scores for accuracy and specificity have been provided by a LR with a sampling strategy of 25%, meaning that the minority class becomes a 25% of the majority one, (working with approximately 280 instances) and a scoring of ROC-AUC. F1-score´s best was also achieved by the same model but with accuracy scoring. Regarding sensitivity, the best one was obtained by a random forest with a sampling of 80% (84 instances) and a scoring of ROC-AUC and when it comes to AUC, the best result was provided by a MLP with a sampling of 60% (116 instances) and accuracy scoring.

Model	Sampling strategy	Scoring	Accuracy	Sensitivity	Specificity	*F1-score*	AUC
LR	0.25	Accuracy	0.937 ± 0.023	0.538 ± 0.11	0.948 ± 0.02	0.321 ± 0.088	0.743 ± 0.063
		ROC-AUC	0.942 ± 0.019	0.484 ± 0.133	0.955 ± 0.017	0.307 ± 0.061	0.719 ± 0.067
	0.6	Accuracy	0.876 ± 0.028	0.601 ± 0.168	0.884 ± 0.031	0.198 ± 0.022	0.742 ± 0.072
		ROC-AUC	0.882 ± 0.031	0.615 ± 0.155	0.889 ± 0.034	0.215 ± 0.037	0.752 ± 0.069
	0.8	Accuracy	0.829 ± 0.049	0.679 ± 0.125	0.833 ± 0.048	0.176 ± 0.041	0.756 ± 0.077
		ROC-AUC	0.828 ± 0.062	0.629 ± 0.119	0.834 ± 0.06	0.176 ± 0.067	0.731 ± 0.089
DT	0.25	Accuracy	0.856 ± 0.057	0.488 ± 0.108	0.866 ± 0.057	0.171 ± 0.083	0.677 ± 0.069
		ROC-AUC	0.844 ± 0.042	0.518 ± 0.118	0.852 ± 0.04	0.155 ± 0.062	0.685 ± 0.074
	0.6	Accuracy	0.791 ± 0.087	0.645 ± 0.116	0.795 ± 0.092	0.158 ± 0.06	0.72 ± 0.033
		ROC-AUC	0.824 ± 0.092	0.633 ± 0.063	0.829 ± 0.093	0.198 ± 0.101	0.731 ± 0.066
	0.8	Accuracy	0.821 ± 0.118	0.621 ± 0.065	0.827 ± 0.12	0.197 ± 0.088	0.724 ± 0.078
		ROC-AUC	0.798 ± 0.134	0.668 ± 0.113	0.804 ± 0.068	0.197 ± 0.1	0.735 ± 0.091
RF	0.25	Accuracy	0.89 ± 0.012	0.552 ± 0.097	0.899 ± 0.014	0.205 ± 0.032	0.725 ± 0.044
		ROC-AUC	0.901 ± 0.023	0.499 ± 0.095	0.912 ± 0.027	0.21 ± 0.045	0.706 ± 0.038
	0.6	Accuracy	0.815 ± 0.055	0.594 ± 0.219	0.821 ± 0.06	0.144 ± 0.048	0.707 ± 0.094
		ROC-AUC	0.81 ± 0.086	0.646 ± 0.153	0.816 ± 0.091	0.164 ± 0.052	0.731 ± 0.053
	0.8	Accuracy	0.732 ± 0.049	0.746 ± 0.09	0.732 ± 0.049	0.132 ± 0.046	0.739 ± 0.063
		ROC-AUC	0.752 ± 0.049	0.763 ± 0.122	0.753 ± 0.052	0.139 ± 0.024	0.758 ± 0.036
XGB	0.25	Accuracy	0.933 ± 0.012	0.486 ± 0.151	0.945 ± 0.011	0.267 ± 0.049	0.715 ± 0.075
		ROC-AUC	0.924 ± 0.017	0.471 ± 0.134	0.936 ± 0.015	0.242 ± 0.05	0.704 ± 0.069
	0.6	Accuracy	0,816 ± 0,029	0,581 ± 0,106	0,822 ± 0,031	0,139 ± 0,018	0,701 ± 0,044
		ROC-AUC	0,747 ± 0,034	0,605 ± 0,101	0,751 ± 0,034	0,112 ± 0,026	0,678 ± 0,051
	0.8	Accuracy	0.718 ± 0.018	0.64 ± 0.144	0.72 ± 0.021	0.103 ± 0.014	0.68 ± 0.066
		ROC-AUC	0,738 ± 0,06	0,633 ± 0,127	0,741 ± 0,062	0,112 ± 0,023	0,678 ± 0,059
MLP	0.25	Accuracy	0.899 ± 0.008	0.535 ± 0.081	0.909 ± 0.007	0.216 ± 0.047	0.722 ± 0.043
		ROC-AUC	0.88 ± 0.006	0.551 ± 0.122	0.889 ± 0.008	0.19 ± 0.04	0.72 ± 0.058
	0.6	Accuracy	0.835 ± 0.024	0.691 ± 0.122	0.84 ± 0.028	0.177 ± 0.02	0.765 ± 0.0486
		ROC-AUC	0.836 ± 0.043	0.625 ± 0.172	0.842 ± 0.048	0.164 ± 0.024	0.733 ± 0.062
	0.8	Accuracy	0.756 ± 0.07	0.706 ± 0.136	0.757 ± 0.072	0.135 ± 0.028	0.732 ± 0.078
		ROC-AUC	0.792 ± 0.087	0.686 ± 0.097	0.795 ± 0.088	0.177 ± 0.096	0.74 ± 0.077

Table 5.2: Results of E2 with the mean and standard deviation of the metrics of study for different sampling strategies and scorings.

5.4.2 Results of Experiment 2

The results of E2 can be seen in Table 5.3. In this case, it has been carried out with 7 features after a FS process with MI in which the lowest ranking variables have been eliminated.

As we can see, generally better results have been got comparing to E1. LR keep being the model with the more top scores with the highest values in accuracy, F1-score, specificity and

AUC, in fact, all of them with 25% of sampling and accuracy scoring. Furthermore, best AUC is found in DT with 25% and ROC-AUC and best sensitivity in RF with 80% and ROC-AUC. However, this same RF model has lowest scores for accuracy and specificity and the DT 60% and ROC-AUC provides also the worst ones for F1-score and AUC.

Model	Sampling strategy	Scoring	Accuracy	Sensitivity	Specificity	*F1-score*	AUC
LR	0.25	Accuracy	0.921 ± 0.029	0.637 ± 0.109	0.929 ± 0.029	0.306 ± 0.05	0.783 ± 0.056
		ROC-AUC	0.915 ± 0.021	0.626 ± 0.116	0.923 ± 0.02	0.278 ± 0.031	0.775 ± 0.059
	0.6	Accuracy	0.875 ± 0.021	0.678 ± 0.14	0.881 ± 0.022	0.217 ± 0.025	0.779 ± 0.065
		ROC-AUC	0.877 ± 0.036	0.665 ± 0.15	0.883 ± 0.039	0.221 ± 0.028	0.774 ± 0.063
	0.8	Accuracy	0.853 ± 0.061	0.705 ± 0.132	0.858 ± 0.065	0.228 ± 0.101	0.781 ± 0.051
		ROC-AUC	0.858 ± 0.069	0.671 ± 0.141	0.863 ± 0.072	0.234 ± 0.109	0.767 ± 0.062
DT	0.25	Accuracy	0.843 ± 0.045	0.657 ± 0.173	0.848 ± 0.05	0.18 ± 0.033	0.753 ± 0.069
		ROC-AUC	0.826 ± 0.075	0.737 ± 0.182	0.829 ± 0.082	0.197 ± 0.053	0.783 ± 0.053
	0.6	Accuracy	0.837 ± 0.073	0.593 ± 0.079	0.844 ± 0.075	0.179 ± 0.061	0.718 ± 0.04
		ROC-AUC	0.775 ± 0.077	0.651 ± 0.078	0.778 ± 0.081	0.137 ± 0.03	0.714 ± 0.026
	0.8	Accuracy	0.769 ± 0.08	0.731 ± 0.158	0.771 ± 0.085	0.15 ± 0.039	0.751 ± 0.044
		ROC-AUC	0.766 ± 0.117	0.747 ± 0.16	0.767 ± 0.123	0.172 ± 0.077	0.757 ± 0.041
RF	0.25	Accuracy	0.876 ± 0.036	0.59 ± 0.173	0.884 ± 0.037	0.199 ± 0.047	0.737 ± 0.082
		ROC-AUC	0.873 ± 0.033	0.554 ± 0.173	0.883 ± 0.035	0.184 ± 0.034	0.718 ± 0.08
	0.6	Accuracy	0.817 ± 0.025	0.7 ± 0.061	0.82 ± 0.026	0.164 ± 0.01	0.76 ± 0.032
		ROC-AUC	0.828 ± 0.036	0.685 ± 0.092	0.832 ± 0.039	0.174 ± 0.034	0.758 ± 0.037
	0.8	Accuracy	0.811 ± 0.049	0.729 ± 0.133	0.814 ± 0.053	0.17 ± 0.028	0.771 ± 0.052
		ROC-AUC	0.761 ± 0.017	0.789 ± 0.11	0.761 ± 0.018	0.144 ± 0.014	0.775 ± 0.054
XGB	0.25	Accuracy	0.846 ± 0.039	0.619 ± 0.19	0.853 ± 0.04	0.17 ± 0.026	0.736 ± 0.092
		ROC-AUC	0.859 ± 0.03	0.658 ± 0.146	0.865 ± 0.031	0.194 ± 0.022	0.761 ± 0.07
	0.6	Accuracy	0.821 ± 0.062	0.673 ± 0.166	0.825 ± 0.065	0.168 ± 0.043	0.749 ± 0.078
		ROC-AUC	0.825 ± 0.058	0.689 ± 0.138	0.829 ± 0.061	0.175 ± 0.043	0.759 ± 0.062
	0.8	Accuracy	0.766 ± 0.099	0.731 ± 0.158	0.767 ± 0.105	0.155 ± 0.052	0.749 ± 0.049
		ROC-AUC	0.77 ± 0.097	0.745 ± 0.152	0.771 ± 0.103	0.159 ± 0.049	0.758 ± 0.041
MLP	0.25	Accuracy	0.912 ± 0.019	0.608 ± 0.088	0.92 ± 0.018	0.266 ± 0.042	0.764 ± 0.047
		ROC-AUC	0.91 ± 0.018	0.613 ± 0.122	0.918 ± 0.016	0.263 ± 0.05	0.766 ± 0.066
	0.6	Accuracy	0.829 ± 0.052	0.615 ± 0.193	0.835 ± 0.055	0.159 ± 0.024	0.725 ± 0.085
		ROC-AUC	0.836 ± 0.053	0.594 ± 0.207	0.843 ± 0.055	0.162 ± 0.035	0.719 ± 0.098
	0.8	Accuracy	0.812 ± 0.063	0.701 ± 0.157	0.816 ± 0.068	0.171 ± 0.045	0.758 ± 0.055
		ROC-AUC	0,823 ± 0,075	0,66 ± 0,17	0,828 ± 0,076	0,183 ± 0,075	0,744 ± 0,102

Table 5.3: Results of E2 with the mean and standard deviation of the metrics of study for different sampling strategies and scorings.

5.5 Discussion of the results

Once both experiments have been performed in an independent way, we can now compare the best results obtained in both of them as we can see in Tables 5.2 and 5.3.

Firstly, we should start by ruling out F1-score which as we can see is extremely low and has not much relevance as well as precision which has not been included in the tables. Then, we need to focus on sensitivity and specificity which are the most relevant regarding clinical practice, as it has been stated in Section 3.7.1, thus, we can say that the model is highly specific and able to differentiate the patients that are not getting infection with a 94,2% but it is not quite sensible as it is able to differentiate the positive infections with a maximum of 78,9%.

Regarding both experiments, we can see that the FS, despite of being extremely simple with just MI examination, really helped us to get better results as there has been a very considerable improvement in sensitivity values in general as well as in AUC ones. What FS has done is to remove possible useless variables that may introduce noise and thus making the algorithm more able to detect infections because the values of accuracy and specificity barely changed.

To sum up, the best model, generally speaking, we can say that is LR as it has the best values for accuracy, specificity and AUC. Therefore, I believe it would be the most suitable in the clinical practice even though its low sensitivity. Another possible option would be RF as it provide the highest sensitivity, however it also gives a strongly lower accuracy and specificity compared to the one of LR.

Chapter 6

Conclusions and future sights

In this last chapter, some conclusions about this project are drawn together with some of its possible lines that could be developed in the future.

6.1 Conclusions

The objective of this project was to achieve the classification of patients undergoing HA into whether they are suffering infection or not. For this, a dataset provided from HURYC was used containing dates, parameters and information about each patient relating its hospital timings, discharge and admission data. The database was composed of 2173 patients operated between 2010 and 2020.

Firstly, the data was feature engineered with the aim of creating some news variables and getting rid of the ones that were not useful or provided no valid information for the correct prediction. Furthermore, an extensive preprocessing was applied in order to be able to introduce data into the models and remove inconsistent one, this phase has been a time-consuming but really essential part, probably the main reason of our results. This phase includes all the procedures related to the balancing of classes, one hot encoding and the data is further visualized to make it easier to comprehend.

Afterwards, a brief FS process was carried out in order to find those variables more determinant in predicting the outcome. For this, the filter method, Mutual information, was used which allowed us to select a feature vector only with the most important variables.

In the third place, the final dataset was used for the implementation of the algorithms in the two experiments that have been carried out. In the first one (E1) all the features have been

used while in the second one (E2), a prior FS has been applied eliminating those ones that do not give information for classification. In both experiments, same methods and conditions have been applied.

Regarding the results obtained from the algorithms, we have obtained good results regarding specificity and accuracy achieving up to 94,2% of accuracy and 95.5% of specificity which allow us to predict those patients that are not going to suffer infection with a very high reliability. However, the results of sensitivity and AUC have been a bit weaker achieving a maximum of approximately 78% in both cases which indicates some problems for the model in the prediction of infections. In fact, considerably better results have been obtained by doing the FS process, even though it was quite simple and short. Furthermore, we can finally state that those methods who are more complicated and sophisticated do not deliver better results than those simpler, thus less computational cost is required for this classification problem.

Lastly, we can conclude that ML algorithms can become a really useful and promising option to be able to carry out infection prediction by using simple parameters that are accessible and available in any hospital. They can become a tool, apart from highly decreasing the costs for the health system, for increasing the safety and future welfare of all the patients undergoing HA which is, indeed, a kind of surgery that completely changes the life of a person. Thus, with more studies in this field and the acquisition of new works and ideas, it can really transform the HA surgery prospects forever.

6.2 Future lines

This project, as it is said above, supposes some contribution to the possible techniques to improve HA by preventing one of the complications coming from HA, infection. It offers a way of predicting it by using very simple variables that could be taken from almost any hospital in the world performing this kind of surgeries.

In addition, this project is intended to be adapted and developed for its publication in a research journal which may encourage other researchers in continue developing the methods and utilities in this field.

However, despite getting some decent results, there is still a huge room for improvement and for being able to incorporate this into the clinical practice. Some of the things that could be improved are:

- Increase the variables that could be taken from the patient during the preoperative time.

- Make use of other sophisticated models and predicting methods such as Support Vector Machine (SVM).
- Utilize better and refined oversampling techniques such as Borderline Oversampling with SVM and Adaptive Synthetic Sampling (ADASYN).
- Rescale the problem to try to mark, before the surgery is carried out, those patients that are more likely to suffer from infection.
- Increase the FS process with more techniques to increase its reliability and achieve the best possible feature vector. Some of examples of them would be Boostraps confidence intervals or chi-square.

Bibliography

[1] T. Chopra, J. Zhao, G. Alangaden, M. Wood, and K. Kaye, "Preventing surgical site infections after bariatric surgery: value of perioperative antibiotic regimens.," *Expert Rev Pharmacoecon Outcomes Res.*, 2010.

[2] "Machine learning." https://www.informatec.com/en/machine-learning, Accessed: 03-04-2022.

[3] V. García, J. Sánchez, and A. Marqués, "Synergetic application of multi-criteria decision-making models to credit granting decision problems," *Applied Sciences*, vol. 9, pp. 1–15, 11 2019.

[4] "Undersampling and oversampling: An old and a new approach." https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392. Accessed: 03-04-2022.

[5] A. Das, "Oversampling to remove class imbalance using smote," 2019. https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35, Accessed: 28-05-2022.

[6] "Logistic function." https://en.wikipedia.org/wiki/Logistic_function, Accessed: 23-04-2022.

[7] "¿qué es un bosque aleatorio?." https://www.tibco.com/es/reference-center/what-is-a-random-forest, Accessed: 20-05-2022.

[8] W. Huel, "What are ensemble techniques?," 2021. lhttps://morioh.com/p/e108a4521555, Accessed: 4-06-22.

[9] M. H. Esfe, S. A. Eftekhari, M. Hekmatifar, and D. Toghraie, "A well-trained artificial neural network for predicting the rheological behavior of mwcnt–al2o3 (30–70%)/oil sae40 hybrid nanofluid," *Scientific Reports*, vol. 11, 2021.

[10] "Receiver operating characteristic." https://en.wikipedia.org/wiki/Receiver_operating_characteristic, Accessed: 24-04-2022.

[11] "Official Journal of the European Union. Council recommendation of 9 June 2009 on patient safety, including the prevention and control of healthcare-associated infections (HAI).," 2009/C 151/01.

[12] "Official Journal of the European Communities Decision No. 2119/98/EC of the European Parliament and of the Council of 24 September 1998 setting up a network for the epidemiological surveillance and control of communicable diseases in the Community.," 1998:L268/1-6.

[13] "European Centre for Disease Prevention and Control. European surveillance of healthcare-associated infections in intensive care units – HAI-Net ICU protocol, version 1.02.," Stockholm: ECDC; 2015.

[14] "European Centre for Disease Prevention and Control. Surveillance of healthcare-associated infections and prevention indicators in European intensive care units.," Stockholm: ECDC; 2017.

[15] Ministerio de Sanidad, Servicios Sociales e Igualdad., *Grupo de trabajo de la Ponencia de Vigilancia Epidemiológica. Documento marco del sistema nacional de vigilancia de infecciones relacionadas con la asistencia sanitaria. Comisión de Salud Pública del Consejo Interterritorial del Sistema Nacional de Salud.*, 2015.

[16] R. M. e. a. F.J.Medina-Fernández, D.J.Garcilazo-Arismendi, "Validation in colorectal procedures of a useful novel approach for the use of c-reactive protein in postoperative infectious complications," *Colorectal Disease*, vol. 18, no. 3, pp. 0111–0118, 2016.

[17] T. e. a. FK. Maeda, Y. Kanaoka, "Better clinical practice could overcome patient-related risk factors of vascular surgical site infections," *Journal of Endovascular Therapy*, vol. 22, no. 4, pp. 640–646, 2015.

[18] a. K.Rolston, C.Mihu, "Current microbiology of surgical site infections associated with breast cáncer surgery," *Wounds*, vol. 22, no. 5, pp. 132–135, 2010.

[19] R. A. Cooper, "Surgical site infections: epidemiology and microbiological aspects in trauma and orthopaedic surgery," *International Wound Journal*, vol. 10, no. 1, pp. 3–8, 2013.

[20] I. Uçkay, P. Hoffmeyer, D. Lew, , and D. Pittet, "Prevention of surgical site infections in orthopaedic surgery and bone trauma: state-of-the-art update," *Journal of Hospital Infection*, vol. 84, no. 1, pp. 5–12, 2013.

[21] "Surgical site infection event (ssi). national healthcare safety network (nhsn). patient safety component manual.," January 2022.

[22] S. Awad, "Adherence to surgical care improvement project measures and postoperative surgical site infections," *Surgical Infection (Larchmt)*, vol. 13, no. 4, pp. 234–7, 2012.

[23] K. Ban, "American college of surgeons and surgical infection society: Surgical site infection guidelines, 2016 update," *Journal of the American College of Surgeons*, vol. 224, no. 1, pp. 59–74, 2017.

[24] E. Z. et al., "Health care-associated infections. a meta-analysis of costs and financial impact on the us health care system," *JAMA Intern Med*, vol. 173, no. 22, pp. 2039–46, 2013.

[25] C. Suetens, S. Hopkins, and J. K. et al, "Point prevalence survey of healthcare-associated infections and antimicrobial use in european acute care hospitals 2011–2012,," *European Centre for Disease Prevention and Control*, 2013. Stockholm, Sweden.

[26] M. Abbas and D. Pittet, "Surgical site infection prevention: a global priority," *Journal of Hospital Infection*, vol. 93, no. 4, pp. 319–322, 2016.

[27] B. Allegranzi, B. Zayed, and P. B. et al., "New who recommendations on intra operative and postoperative measures for surgical site infection prevention: an evidence-based global perspective," *The Lancet Infectious Diseases*, vol. 16, no. 12, pp. e267–e287, 2016.

[28] C. Brandt, D. Sohr, M. Behnke, F. Daschner, H. Ru¨den, and P. Gastmeier, "Reduction of surgical site infection rates associated with active surveillance," *Infection Control and Hospital Epidemiology*, vol. 27, no. 12, pp. 1347–1351, 2016.

[29] E. L. P. E. Geubbels, N. J. D. Nagelkerke, A. J. M.-D. Groot, C. Vandenbroucke-Grauls, D. Grobbee, and A. D. Boer, "Reduced risk of surgical site infections through surveillance in a network," *International Journal for Quality in Health Care*, vol. 18, no. 2, pp. 127–133, 2006.

[30] W. Staszewicz, M.-C. Eisenring, V. Bettschart, S. Harbarth, and N. Troillet, "Thirteen years of surgical site infection surveillance in swiss hospitals," *Journal of Hospital Infection*, vol. 88, no. 1, pp. 40–47, 2014.

[31] C. van Walraven and R. Musselman, "The surgical site infection risk score (ssirs): a model to predict the risk of surgical site infections," *PLOS ONE*, vol. 8, no. 6, 2013.

[32] P. Kocbek, N. Fijacko, C. Soguero-Ruiz, K. Øyvind Mikalsen, U. Maver, P. P. Brzan, A. Stozer, R. Jenssen, S. O. Skrøvseth, and G. Stiglic, "Maximizing interpretability and cost-effectiveness of surgical site infection (ssi) predictive models using feature-specific regularized logistic regression on preoperative temporal data.," *Computational and Mathematical Methods in Medicine Volume*, 2019.

[33] S. Berríos-Torres, C. Umscheid, and D. B. et al, "Centers for disease control and prevention guideline for the prevention of surgical site infection," *JAMA Surg*, vol. 152, no. 8, pp. 784–791, 2017.

[34] J. Evans, A. Blom, A. Timperley, P. Dieppe, M. Wilson, A. Sayers, and M. Whitehouse, "Factors associated with implant survival following total hip replacement surgery: A registry study of data from the national joint registry of england, wales, northern ireland and the isle of man.," *PLoS Med*, vol. 17, no. 8, 2020.

[35] S. Klouche, E. Sariali, and P. Mamoudy, "Total hip arthroplasty revision due to infection: A cost analysis approach," *Orthopaedics & Traumatology: Surgery & Research*, vol. 96, no. 2, pp. 124–132, 2010.

[36] J. M. Meessen, W. F. Peter, R. Wolterbeek, S. C. Cannegieter, C. Tilbury, M. R. Bénard, H. M. van der Linden, R. Onstenk, R. Tordoir, S. B. Vehmeijer, S. H. Verdegaal, H. M. Vermeulen, R. G. Nelissen, and T. P. Vliet Vlieland, "Patients who underwent total hip or knee arthroplasty are more physically active than the general dutch population," *Rheumatology international*, vol. 37, 2017.

[37] R. Pivec, A. J. Johnson, S. C. Mears, and M. A. Mont, "Hip arthroplasty," *The Lancet*, vol. 380, no. 9855, pp. 1768–1777, 2012.

[38] J. S. Siopack and H. E. Jergesen, "Total hip arthroplasty.," *Western journal of medicine*, vol. 162, no. 3, p. 243, 1995.

[39] E. T. Horan TC, "Definitions of key terms used in the nnis system," *Am J Infect Control*, vol. 112, no. 6, 1997.

[40] National Healthcare Safety Network (NHSN), *Surgical Site Infection Event (SSI). Patient Safety Component Manual*, January 2022.

[41] H. RW, C. DH, M. WM, W. JW, E. TG, and H. TM, "Identifying patients at high risk of surgical wound infection. a simple multivariate index of patient susceptibility and wound contamination," *American Journal Epidemiol*, vol. 121, 1985.

[42] American Society of Anesthesiologists, *ASA physical status classification system*, 2020.

[43] D. DJ, G. A, and G. EH, "American society of anesthesiologists classification," *Treasure Island (FL): StatPearls*, 2022.

[44] R. Donham, W. Mazzei, and R. Jones, "Association of anesthesia clinical directors' procedure times glossary," *American Journal of Anesthesiology*, vol. 23, no. 5S, 1996.

[45] K. Ong, E. Lau, M. Manley, and S. Kurtz, "Effect of procedure duration on total hip arthroplasty and total knee arthroplasty survivorship in the united states medicare population," *J Arthroplasty*, vol. 23, no. 6, 2008.

[46] M. Crader and M. Varacallo, "Preoperative antibiotic prophylaxis.," *StatPearls*, 2022.

[47] A. H. D. MD* and D. O. R. MD**, "The use of prophylactic antimicrobial agents during and after hip arthroplasty, clinical orthopaedics and related research," *American Journal of Infection Control*, vol. 369, pp. 124–138, 1999.

[48] C. D.-A. Pérez, A. R. Rodela, M. P. López, N. L. Fresneña, and V. M. Jodrá, "Quality control indicator working group. surgical wound infection rates in spain: data summary, january 1997 through june 2012," *American Journal of Infect Control*, vol. 42, no. 5, 2014.

[49] E. P. Dellinger, P. A. Gross, T. L. Barrett, P. J. Krause, W. J. Martone, J. E. M. Jr., R. L. Sweet, and R. P. Wenzel, "Quality standard for antimicrobial prophylaxis in surgical procedures," *Clinical Infectious Diseases*, vol. 18, no. 3, 1994.

[50] A. H. H. S. et al., "Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty?," *Clinical orthopaedics and related research*, vol. 477, no. 2, pp. 452–460, 2019.

[51] A. A. Shah, S. K. Devana, C. Lee, R. Kianian, M. van der Schaar, and N. F. SooHoo, "Development of a novel, potentially universal machine learning algorithm for prediction of complications after total hip arthroplasty," *The Journal of Arthroplasty*, vol. 36, no. 5, pp. 1655–1662.e1, 2021.

[52] E. Bülow, U. Hahn, I. Andersen, O. Rolfson, A. Pedersen, and N. Hailer, "Prediction of early periprosthetic joint infection after total hip arthroplasty," *Clin Epidemiol*, vol. 14, pp. 239–253, 2022.

[53] C. Klemt, S. Laurencin, A. C. Uzosike, J. C. Burns, T. G. Costales, I. Yeo, Y. Habibi, and Y.-M. Kwon, "Machine learning models accurately predict recurrent infection following revision total knee arthroplasty for periprosthetic joint infection," *Knee Surgery, Sports Traumatology, Arthroscopy*.

[54] T. L. Tan, M. G. Maltenfort, A. F. Chen, A. Shahi, C. A. Higuera, M. Siqueira, and J. Parvizi, "Development and evaluation of a preoperative risk calculator for periprosthetic joint infection following total joint arthroplasty," *The Journal of Bone and Joint Surgery*, vol. 100, no. 9, pp. 777–785, 2018.

[55] Y. Bastanlar and M. Ozuysal, *Introduction to Machine Learning*. Totowa, NJ: Humana Press, 2014.

[56] S. Badillo, B. Banfai, F. Birzele, I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and Zhang, "An introduction to machine learning.," *Clin. Pharmacol. Ther.*, vol. 107, pp. 871–885, 2020.

[57] P. Cunningham, M. Cord, and S. J. Delany, *Supervised Learning. Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[58] J. Heaton, "An empirical analysis of feature engineering for predictive modeling. southeastcon 2016," pp. 1–6, 2016.

[59] Uddin, M. Fahim, Lee, Jeongkyu, Rizvi, Syed, and S. Hamada, "Proposing enhanced feature engineering and a selection model for machine learning processes," *Applied Sciences*, vol. 8, no. 4, 2018.

[60] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. Turaga, "Learning feature engineering for classification. proceedings of the twenty-sixth international joint conference on artificial intelligence, ijcai-17," pp. 2529–2535, 2017.

[61] "Using categorical data with one hot encoding." https://programmerclick.com/article/7976471743/, Accessed: 05-04-2022.

[62] "Data science in 5 minutes: What is one hot encoding?." https://www.educative.io/blog/one-hot-encoding, Accessed: 05-04-2022.

[63] B. Alexander Yun-chung Liu, *The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets*. 2004.

[64] M. Molinara, M. Ricamato, and F. Tortorella, "Facing imbalanced classes through aggregation of classifiers. 14th international conference on image analysis and processing (iciap 2007)," pp. 43–48, 2007.

[65] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[66] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[67] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning with applications in R*. New York, USA: Springer, 2013.

[68] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 7, pp. 1330–1339, 2009.

[69] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1184–1199, 2011. Advanced Topics in Cloud Computing.

[70] N. Hoque, D. Bhattacharyya, and J. Kalita, "Mifs-nd: A mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, 2014.

[71] S. Ray, "A quick review of machine learning algorithms. 2019 international conference on machine learning, big data, cloud and parallel computing (comitcon)," pp. 35–39, 2019.

[72] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Inc, 2019.

[73] J. Feng, H. Xu, S. Mannor, and S. Yan, "Advances in neural information processing systems," vol. 27, Curran Associates, Inc., 2014.

[74] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques. proceedings of the 2007 conference on emerging artificial intelligence applications in computer engineering: Real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies," (NLD), p. 3–24, IOS Press, 2007.

[75] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer Series in Statistics Springer New York Inc., 2001.

[76] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system. proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining," pp. 785–794, 2016.

[77] "Chapter 7 - artificial neural networks: Multilayer perceptron for ecological modeling," vol. 28 of *Developments in Environmental Modelling*, pp. 123–140, Elsevier, 2016.

[78] A. Z. Rachel Hu, "Dive into deep learning - 4.7. forward propagation, backward propagation, and computational graphs," 2020.

[79] W. Zhu, N. F. Zeng, and N. Wang, "1 sensitivity , specificity , accuracy , associated confidence interval and roc analysis with practical sas," 2010.