UNIVERSITAT DE BARCELONA

FUNDAMENTALS OF DATA SCIENCE MASTER'S THESIS

Predicting Intensive Care Unit Length of Stay via Supervised Learning

Author:
Jordi ALBIOL MOSEGUI

Supervisor:
Dra. Laura IGUAL MUÑOZ

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamentals of Data Science

in the

Facultat de Matemàtiques i Informàtica

September 2, 2018

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Predicting Intensive Care Unit Length of Stay via Supervised Learning

by Jordi ALBIOL MOSEGUI

Healthcare is a traditional sector that is demanding, nowadays, a profound change regarding tasks and ways of work. The explotation of data-based analytical techniques together with computational capabilities are potential candidates to lead part of that demanding change. This can cause an innovation to the sector with considerable social impact. In any case, it is necessary to take into account the specific characteristics of the clinical data: quality, volume, access and multimodality.

In this Master Thesis, an analysis of the data from critical patients was carried out in order to study the influence of several observables to determine their Length of Stay in the Intensive Care Unit. Try to solve that problem can help a lot not only the physicians from the mere investigation purposes point of view but also the healthcare sector because Intensive Care Unit logistics counts and it can become very important.

Acknowledgements

I would like to give a message of thanks for the assistance provided over the semester to the research group of the Intensive Care Department of the Hospital Universitari Joan XXIII of Tarragona:

María Amparo Bodi Saera
Alejandro H Rodríguez
Josep Gómez Álvarez

And with the help of the Supervisor:

Laura Igual Muñoz, Associate Professor at the Department of Mathematics and Computer Science at the Universitat de Barcelona

Contents

Chapter 1

Introduction

The present work consist in an analysis over data collected in an Intensive Care Unit (ICU) and the evaluation of predictive models applied to health data area. Instead to do a classical study of the incidence in mortality of the variability of possible observables of the patients such as the variability of glucose over time during an ICU stay, presence or not of comorbidities[1] together with other clinical and demographic variables of the patients, it is investigated the relationship of this observables as being the cause or part of the cause to the ICU Length of Stay (LOS) for each single admission episode.

As part of this work a methodology has been defined, developed a set of extraction, transformation and loading (ETL) processes, which obtain and transform data from the original database and convert it in a more tractable way to perform analysis on that data. Python programming language was used as well as some of their available libraries together with IPython Jupyter Notebook that facilitates a lot the tasks associated with any kind of data analysis and the development and implementation of predictive models. The database used for the study, MIMIC-III **[1]** **[2]**, comes from a database health service with anonymized data to avoid identification of patients.

Some review about the research centered on the MIMIC-III Critical Care Database is presented, establishing a background understanding of the current state of research at the intersection of predictive modeling and hospital inpatient data as well as a review of the methods utilized in this project, key findings and recommendations for further investigation.

1.1 Context and justification

Why LOS? Extended inpatient hospital LOS is associated with increased cost, higher readmission rates, and increased mortality, as well as greater probabilities of contracting a hospital-acquired infection (HAI) **[3]**. Each of these metrics represents significant indicators for hospital performance. Thus, LOS serves as a critical performance indicator for these measures and can provide insights ranging from cost analysis[2] to patient outcomes. This project leverages hospital inpatient data from the MIMIC-III Critical Care Database to explore supervised statistical learning models in order to predict LOS.

A significant portion of the research centered on the MIMIC-III database has focused on developing models for patient mortality, the likelihood of septicemia **[4]** and readmission prediction. Health services and policy research pertaining to hospital costs has fueled the popularity of these questions. For instance, septicemia is the primary driver of HAIs and represents the most expensive inpatient hospital cost at $20.3 billion in 2011 **[5]**. The essential question

[1] Comorbidities refer to the presence of one or more additional diseases, pathologies or disorders in a patient co-occurring with the one diagnosed as principal; in the countable sense of the term, a comorbidity is each additional disease, pathology or disorder.

[2] MIMIC-III data was generated in USA, where the health sector manage costs in a very different way from here Spain. Depending on the bussiness case, the strategy to adopt can change drastically although for sure it is an important problem that needs to be adressed.

of the research documented in this paper is distinguished in its investigation of LOS, a key performance indicator for many of the models in the existing literature.

The LOS predictive models have been developed for comparison with existing, healthcare industry-standard predictive models – namely, the Acute Physiology and Chronic Health Evaluation (APACHE) and the Simplified Acute Physiology Score (SAPS) predictive scoring systems in the ICU. APACHE is a risk assessment and predictive severity of disease classification system that utilizes a point score comprised of physiologic measurements, age, chronic health status and other variables to provide risk estimates for hospital mortality. Similarly, SAPS is a point score for ICU patient severity based on physiological variables, age, type of admission and disease-related variables that calculates mortality probability estimates **[6]**. Additionally, this project employs a similar conceptual framework to the Henry et al. (2015) research **[5]**, and essentially strives to provide a "warning score" in terms of LOS with the intention to be used not as a substitute of the existing ones but as a potential support.

Health is one of the sectors in which the use of new technologies present more possibilities for innovation and greater social impact. In Europe alone, public spending on health is expected to increase from 8% of GDP in 2000 to 14% in 2030 **[7]** and it will have to face major challenges in the coming decades, such as the aging of the population, migratory crises, risks of new epidemics and the increasing cost of clinical technologies. In 2012, the Poneman Institute estimated that 30% of digital data stored worldwide corresponded to health data **[8]**. Extract knowledge of this huge amount of data is considered the fastest, most efficient and least expensive way to put hospital performance to the next level with the indirect benefits that this will carry out: increase the satisfaction of the ICU inpatient admissions and of course their health.

The use of information technologies and data analysis is a fundamental factor in moving towards this goal, although the characteristics of the sector imply that the automated treatment of health data and the carrying out of studies and research present specific difficulties, among them **[9]**:

- **Data quality**: The quality of the data recorded and its structure are critical in order to apply analysis tools, but it is not the priority when it comes to caring for patients. Many of the reports, annotations, among others, are recorded in the form of free text or unstructured annotations. There is a conflict between the rapidity and immediate utility of the clinical data and the ease of treatment and ex post analysis.

- **Data volume**: Because of the automated way of recording all sorts of clinical information (laboratory results, scanners, diagnostic imaging devices, vital sign measurements made at a rate of 1 data point per minute, etc.), comes in action the generation of a considerable and growing amount of data.

- **Multimodal data**: Health sector is characterized by a diversity of data types (images, free text, clinical annotations, multiple coding standards, real-time measurements, temporary graphs, etc.), obtained by systems with very different characteristics and with more or less possibilities of integration with each other.

- **Data access**: The characteristics of health organizations and the requirements of health data protection laws establish a series of limitations that make it practically impossible to carry out studies or data analysis integrating different independent sources or in a public and accessible way.

Due to these particular characteristics that health data sector presents, the two most important fronts that allow progress in the treatment of health data are: the interoperability of different systems **[10]**, **[11]** and the new Big Data technologies, including in them techniques that are not properly Big Data, but that usually enter by extension within the same denomination: ETL (Extraction, Transformation and Load), advanced data analysis, machine learning, statistical models, among others.

1.2 Goals

The study proposal starts from the problematic and context indicated and marked as main lines. General objectives are the evaluation of methodologies and appropriate tools to:

- Automate the selection, extraction and preprocessing of data from a real clinical care database.
- Perform the tasks of parameter search, training, validation and comparison between different models of supervised learning appropriate to the data of origin.
- Application of predictive models and selected tools.
- Obtain knowledge from the constructed models.

The main objectives are:

- Design models capable to predict patients' LOS in the ICU, having as main predictor variables those commonly used in clinical practice for the assessment of the state of the patient. Emphasis has been placed to the first value recorded for each variable selected and time evolution is not contemplated. The selected variable indicators must be obtained during the first 24 hours of stay in ICU.
- Apply the obtained models to analyze the importance of the selected predictors in the prediction of LOS. In addition, a model selection process will be performed in order to select the one with the best performance.

In the realization of the proposal and following clinical criteria, it was decided to include, in addition to the measurements related to the assessment of the state of the patient, other variables obtained in the first day of ICU stay: demographic data, scores, comorbidities, admission diagnosis, mechanical ventilation, daily basic analytics such as excess of bases, hemoglobin, lactose, bicarbonate, among others. Also with medical criteria it was decided that the population included in the study must meet the following criteria:

- Patients older than 16 years of age.
- Patients admitted in the ICU.
- With coded main diagnosis.
- Admissions with either a survivor or a mortal end.[3]

Once established the baselines, this project pretends to be practical, getting rid of lots of mathematics and long technical explanations in order to focus on the practical procedures. In fact, those interested in the results of the present project comes from medicine studies so using a plain vocabulary is essential so they can understand the general idea, purposes and results.

[3] It cannot be included variables that are not known in an hypothetical real admission.

1.3 Approach and methodology followed

For the development of the study, MIMIC-III, an anonymized clinical database of free use, whose characteristics and suitability for this work will be described in the next chapter, has been used.

The main focus that has guided the work has been to use easily replicable technologies in a research environment and that the developed processes can be reused with small modifications for another selection of different variables or even with another database source. This criterion has been used to decide the use of Open Source software for the design of the architecture with Python as the main data analysis programming language and the development of the ETL processes through Python libraries, as will be detailed in the corresponding chapters.

In a first step, the methodology to be followed was defined based on existing standards. The technical architecture needed to carry out the analyzes was designed and constructed, and the MIMIC-III database was analyzed in order to obtain the necessary information to develop the ETL processes that obtain and select the data, carrying out several iterations of preparation and modeling until arriving at the documentation report phase.

The CRISP-DM methodology (Cross Industry Standard Process for Data Mining) **[12]** is still the most used in data mining projects. This methodology divides a data mining project into several phases carried out iteratively until reaching the desired results. Broadly speaking, the approach followed and its different phases are shown in Figure 1.1 where it is shown the life cycle of data mining projects according to this methodology:

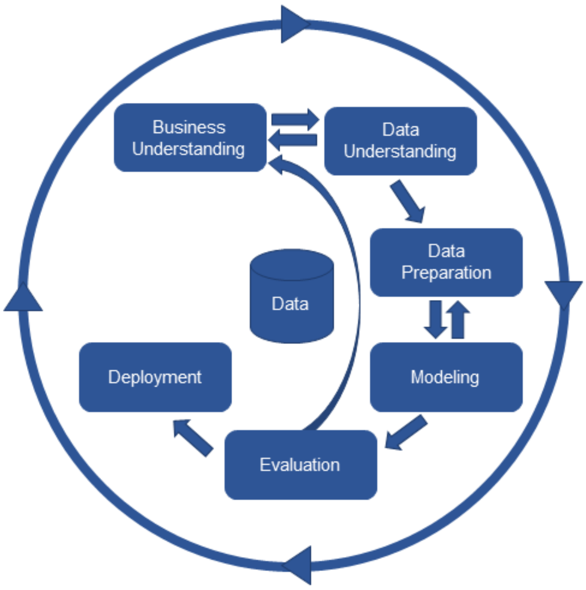


FIGURE 1.1: CRISP-DM Methodology

The different cycles of the model always begin at the knowledge of the sector or Bussiness/Data Understanding and end with the Deployment of the product. During the rest of the phases, forward-back jumps are continuously produced, since the results of one phase influence both the previous and the subsequent ones.

The first Business Understanding phase focuses on understanding the objectives of the project, the context and carrying out the preliminary plan. In the Data Understanding phase, the tasks related to the data collection, the exploratory analysis and the quality review of the data are included. During the Data Preparation phase, the data that will be used for the successive phases are selected, the data is cleaned if necessary, derived data are obtained from the initials, the structure is modified and general all the tasks included in the pre-processing of the data before being used by the different models. Once the selected and preprocessed data is available, the Modeling phase is carried out, in which the different predictive models are used. Within this phase includes the partitioning of the data in training and validation subsets, the search of the appropriate parameters for each model and the comparison of the models. In the Evaluation phase, the results obtained are reviewed and evaluated and it is determined if the phase of implantation is passed or any of the previous processes is refined by starting the cycle at another point with the required modifications.

Finally, once all the previous steps have been developed, the Deployment stage will be carried out, depending on the specific characteristics of the project: the writing of a report, a publication, the production deployment of a system, its modification or maintenance, among others.

1.4 Little description of the contents

This report has been structured in the following way. In Chapter 2 it is presented in detail the database that has served as data source, MIMIC-III. Next in Chapter 3, it is described the ETL processes developed to obtain the data and process them adequately for posterior analysis. Chapter 4 exposes some data preprocessing including data cleaning, the models used together with the evaluation techniques that guide us to the model selection process. Finally, in the last chapter, Chapter 5, it is described the most relevant results of the analyzes carried out, closing with the conclusions, drawbacks and possible lines of future research.

In Appendices A B you can look for additional information regarding feature variables treated while in Appendix C there is a link direction where is located all the code produced during the development of the project. If there is any doubt regarding technical details you can look in the code attached because the present report does not pretend to be super exhaustive.

Chapter 2

MIMIC-III Critical Care Database

MIMIC-III database is described introducing context, structure, content and the interest that it presents in order to do studies like the present.

2.1 About MIMIC-III

The difficulties in carrying out studies and research using real clinical data are firstly those related to the special treatment that privacy demands, and secondly, the existence of organizational and technological barriers, both of which issues makes a difficult access to data regarding not only judicial restrictions but also its treatment and analysis.

To avoid these difficulties, it is available to the researcher and educational community the MIMIC-III (Medical Information Mart for Intensive Care III) database, openly accessible at https://mimic.physionet.org/gettingstarted/dbsetup/. This database is an evolution of the MIMIC-II database created by the Laboratory of Computational Physicology of The Massachusetts Institute for Technology (MIT) with the goal of providing tools for the creation of clinical knowledge through the application of data analysis techniques.

MIMIC-III is a large, freely-available relational database comprising deidentified health-related data associated with over forty thousand patients who stayed in Intensive Care Units at Beth Israel Deaconess Medical Center (Boston, Massachusetts). The data spans June 2001 - October 2012. The current version of the database is v1.4, released on 2 September 2016, which is the version used in this study.

As summarized in Figure 2.1, this database includes information on demographic data of patients, laboratory test results, vital sign measurements, procedures, medications, caregiver notes, imaging reports, mortality (both in and out of the hospital), manual evolution annotations regarding events, discharge reports, prescription, and so on. The database, although deidentified, still contains detailed information regarding the clinical care of patients, so must be treated with appropriate care and respect. Data regarding patients identification, clinical staff and dates are anonymous to comply with the protection laws of data, although by containing real clinical information and by the special beware that this type of data must be treated in order to access and download the database is necessary to complete a formal registration process, overcome an online course about the ethical and legal considerations of the research on human specimens and sign an agreement on use adequate data, as indicated on the MIMIC-III website https://mimic.physionet.org/gettingstarted/access/.

MIMIC-III supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement and electronic tool development. It is notable for three factors:

- It is freely available to researchers worldwide.
- It encompasses a diverse and very large population of ICU patients.
- It contains high temporal resolution data including lab results, electronic documentation and bedside monitor trends and waveforms.

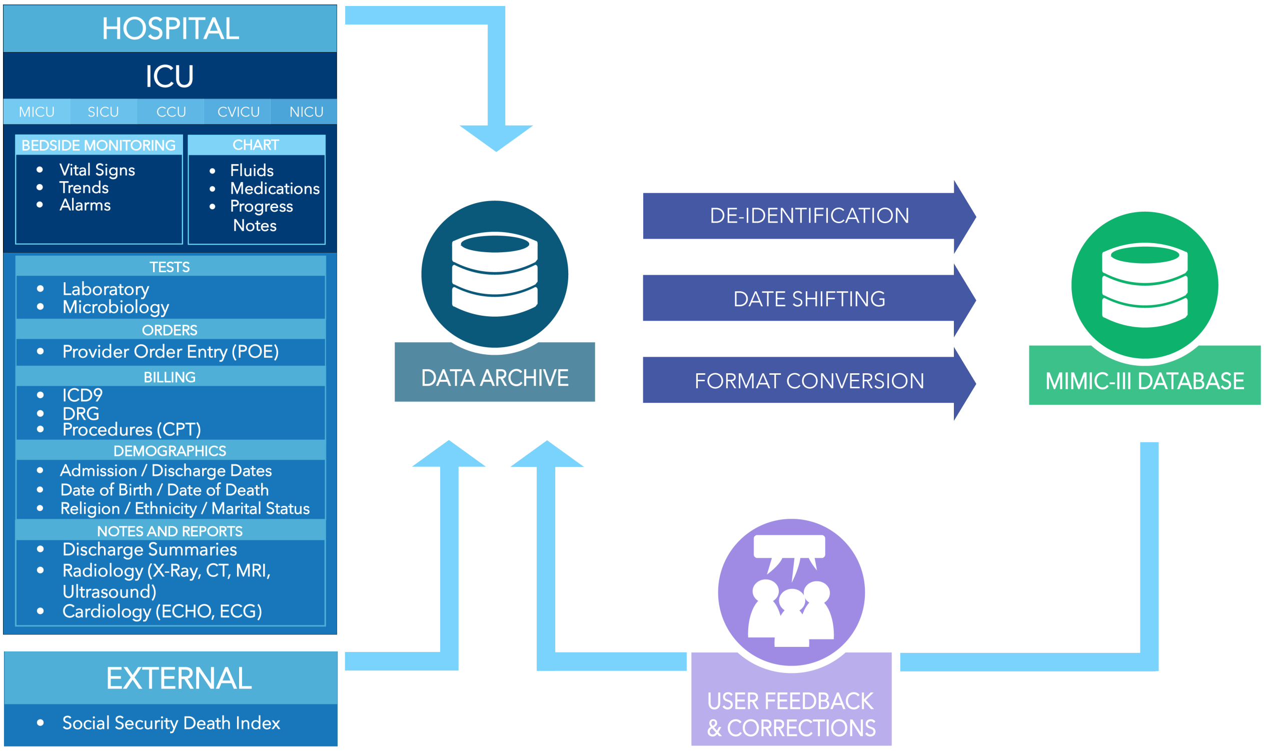


FIGURE 2.1: Model of MIMIC-III construction.

In https://mimic.physionet.org/about/mimic/ the instructions to get the database and load it into the MySQL, Oracle and PostgreSQL are provided, as well as the possibility to obtain a virtual machine that contains the MIMIC-III database and a preconfigured PostgreSQL server that can be useful in some environments because of its ease of use.

MIMIC-III database data have been obtained, preprocessed and anonymized from the different systems of the hospital, so it is representative of the heterogeneity of the data, although there has been a cleaning and elimination of inconsistencies that facilitates the analysis of the information in posterior stages.

MIMIC-III stands out as the only BBDD of its type freely accessible **[1]**, and contains data from more than a decade, which allows studies both in medium and long term and there are no restrictions on the type of study to perform. Thanks to these facilities you can find multiple publications and researches carried out based on the data of MIMIC-II and III.

Among the multiple advantages of MIMIC-III is the existing documentation, such as the guides available on the main website and the book Secondary Analysis of Electronic Health Records **[13]** available freely online, which makes a complete tour of the phases and techniques of data analysis and application of models taking MIMIC-III as a basis.

It is also important to highlight the existence of an active community of researchers **[14]** who have implemented several SQL views to obtain new aggregate data from the original one, such as, comorbidities, scores, and so on.

The types of data that are available in the MIMIC-III database are mainly those detailed in Table 2.1 (adapted from https://www.nature.com/articles/sdata201635/tables/3).

MIMIC-III database has been constructed from the data of different departmental applications of the hospital, management of the ICU and records of the *Social Security Administration Death Master File*. One of the important characteristics to take into account during the exploration and use of MIMIC-III is that the specific data of the ICUs were registered through two different systems: Metavision and Carevue. Both were consolidated in MIMIC-III, although

Class of data	Description
Billing	Coded data recorded primarily for billing and administrative purposes. Includes Current Procedural Terminology (CPT) codes, Diagnosis-Related Group (DRG) codes, and International Classification of Diseases (ICD) codes.
Descriptive	Demographic detail, admission and discharge times, and dates of death.
Dictionary	Look-up tables for cross referencing concept identifiers (for example, International Classification of Diseases (ICD) codes) with associated labels.
Interventions	Procedures such as dialysis, imaging studies, and placement of lines.
Laboratory	Blood chemistry, hematology, urine analysis, and microbiology test results.
Medications	Administration records of intravenous medications and medication orders.
Notes	Free text notes such as provider progress notes and hospital discharge summaries.
Physiologic	Nurse-verified vital signs, approximately hourly (e.g., heart rate, blood pressure, respiratory rate).
Reports	Free text reports of electrocardiogram and imaging studies.

TABLE 2.1: Classes of data available in the MIMIC-III critical care database.

given their different capacities and characteristics, the dictionaries of data used are different and some data is kept in different tables according to the system of origin. This point is important because it implies that almost all queries to the MIMIC-III database must be duplicated and developed with different criteria according to the registration system used.

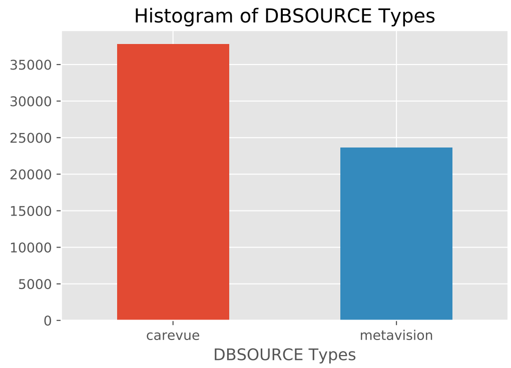


FIGURE 2.2: ICU admission instances for Carevue and Metavision databases

Another important point to have in mind is that Carevue database source of MIMIC-III has a lot of missing data in relevant variables needed to use in this study. For that major reason it is going to work only with the Metavision data. It is newer, more competent, more consistent and most important, with full of data recorded in important variables for this study. To have an idea, Metavision data contains 23620 different ICU admissions whereas the full database has 61532. In Figure 2.2 it is shown the numbers.

The only standardized dictionaries in MIMIC-III are those corresponding to the CIE9 coding, those related to billing (GRDs) and laboratory tests, which have a field mapped with LOINC, the rest of the items used do not have a standardized or documented correspondence, many of them being result of the introduction of free text. This point constitutes one of the greatest difficulties in the use of MIMIC-III since it forces a search effort on thousands of non-standardized items to identify the variables to be extracted. To facilitate this task, it is important to analyze if the variables to be obtained from the MIMIC-III database are already among those used in the views and queries available in the MIMIC-CODE repository **[14]**. Later on, in the annexes, the main variables used in this report will be detailed.

The technique used to anonymize data from the MIMIC-III database is also important when it comes to developing queries, mainly dates and certain free-text annotations. In MIMIC-III the intervals of dates and time for each episode and patient are kept consistent, moving the real dates to one year between 2100 and 2200. For patients over 89 years of age, the possibility of knowing the real age is also eliminated.

2.1.1 MIMIC-III main tables

MIMIC-III is structured in a relational manner and contains 26 linked tables through patient identifier

File Name	Shape	Brief summary
`ADMISSIONS`	(58976, 19)	The `ADMISSIONS` table gives information regarding a patient's admission to the hospital.
`CALLOUT`	(34499, 24)	The `CALLOUT` table provides information about ICU discharge planning.
`CAREGIVERS`	(7567, 4)	This table provides information regarding care givers. For example, it would define if a care giver is a research nurse (RN), medical doctor (MD), and so on.
`CHARTEVENTS`	(330712483, 15)	`CHARTEVENTS` contains all the charted data available for a patient.
`CPTEVENTS`	(573146, 12)	The `CPTEVENTS` table contains a list of which current procedural terminology codes were billed for which patients. This can be useful for determining if certain procedures have been performed (e.g. ventilation).
`D_CPT`	(134, 9)	This table gives some high level information regarding current procedural terminology (CPT) codes. Unfortunately, detailed information for individual codes is unavailable.
`D_ICD_DIAGNOSES`	(14567, 4)	This table defines International Classification of Diseases Version 9 (ICD-9) codes for diagnoses. These codes are assigned at the end of the patient's stay and are used by the hospital to bill for care provided.
`D_ICD_PROCEDURES`	(3882, 4)	This table defines International Classification of Diseases Version 9 (ICD-9) codes for procedures. These codes are assigned at the end of the patient's stay and are used by the hospital to bill for care provided.
`D_ITEMS`	(12487, 10)	The `D_ITEMS` table defines ITEMID, which represents measurements in the database.
`D_LABITEMS`	(753, 6)	`D_LABITEMS` contains definitions for all ITEMID associated with lab measurements in the MIMIC database.

DATETIMEEVENTS	(4485937, 14)	DATETIMEEVENTS contains all date measurements about a patient in the ICU.
DIAGNOSES_ICD	(651047, 5)	This table defines ICD-9 codes for diagnoses. The ICD codes are generated for billing purposes at the end of the hospital stay.
DRGCODES	(125557, 8)	This table defines HCFA-DRG and APR-DRG codes which provide information regarding Diagnosis-Related Group recorded primarily for billing and administrative purposes.
ICUSTAYS	(61532, 12)	This table gives information regarding ICU hospital stays.
INPUTEVENTS_CV	(17527935, 22)	This table contains data of fluid input events (serums, intravenous medication, insulin, etc.) regarding Carevue database source associated to ICU episodes.
INPUTEVENTS_MV	(3618991, 31)	This table contains input data for patients.
LABEVENTS	(27854055, 9)	Contains all laboratory measurements for a given patient, including out patient data.
MICROBIOLOGYEVENTS	(631726, 16)	Contains microbiology information, including tests performed and sensitivities.
NOTEEVENTS	(2083180, 9)	This table contains all notes for patients took in a manual way by their caregivers.
OUTPUTEVENTS	(4349218, 13)	This table contains output data for patients.
PATIENTS	(46520, 8)	This table contains hospitalization-independent data for all patients such as, gender, date of birth, etc.
PRESCRIPTIONS	(4156450, 19)	This table contains medication related order entries, i.e. prescriptions.
PROCEDUREEVENTS_MV	(258066, 25)	This table contains procedures for patients
PROCEDURES_ICD	(17527935, 22)	Contains ICD procedures for patients, most notably ICD-9 procedures. The ICD codes are generated for billing purposes at the end of the hospital stay and are recorded for all patient hospitalizations.
SERVICES	(73343, 6)	The SERVICES table describes the service that a patient was admitted under. This service admission can be elective or caused due to a number of reasons, including bed shortage.
TRANSFERS	(261897, 13)	This table contains physical locations for patients throughout their hospital stay.

TABLE 2.2: Tables of MIMIC-III. Highlighted in ligh grey are the ones used to perform the analysis. More information: https://mimic.physionet.org/mimictables/ (left menu: "Tables in MIMIC")

2.1.2 MIMIC-III derived tables

One strong point of MIMIC-III, as mentioned earlier, is the presence of an active researcher community that develops scripts in order to obtain additional data from the original MIMIC-III database. In this study several of them have been used directly and others to obtain useful information in the selection of variables performed. This views are available at the MIMIC-CODE repo **[14]** https://github.com/MIT-LCP/mimic-code/tree/master/concepts and they are materialized views that must be executed or installed prior to the execution of the ETL.

Class of data	Description
comorbidity	These scripts derive binary flags indicating the presence of various comorbidities using billing codes (ICD-9) assigned to the patient at hospital discharge.
firstday	The first day subfolder contains scripts used to calculate various clinical concepts on the first day of a patient's admission to the ICU, such as the highest blood pressure, lowest temperature, etc. This folder contains many useful scripts which can be adapted to capture data outside the first day.
sepsis	Definitions of sepsis, a common cause of mortality for intensive care unit patients.
severityscores	Severity of illness scores which summarize the acuity of a patient's illness on admission to the intensive care unit (usually in the first 24 hours).
durations	Start and stop times for administration of various treatments or durations of various phenomena, including: medical agents which have a vasoactive effect on a patient's circulatory system, continuous renal replacement therapy (CRRT), and mechanical ventilation.

TABLE 2.3: MIMIC-III used views.

The MIMIC-III database contain a variety of derived tables which simplified the use of the database itself. The database also contain derived parameters commonly required by studies, such as severity scores. The creators of MIMIC-III made a conscious decision to not include any derived tables or calculated parameters as far as is possible. Instead, they encourage the community to produce and share scripts which can be run to create these tables or parameters. This has many advantageous: it keeps the distinction between raw data and calculated data, it encourages users to validate the scripts which derive the data, and allows for as many scripts as is conceivable without cluttering the database for all users. We have provided a set of scripts at the mimic-code repository, which can be found at **[14]**. In this study some of them have been used directly.

2.2 Feature space extraction criteria

Now that we know more about MIMIC-III database is the moment to think about which are the variables of interest. Once the variables are identified we have to define a process to extract them, transform them and load them in a proper and confortable way. This process is detailed in the next Chapter, Chapter 3.

From the original bunch of potential variable candidates that it can be included, physicians from Tarragona Hospital selected a set of variables they know are important from their medicine knowledge, experience or intuition. This are the selected variables:

1. General: Age, BMI, Temperature, Blood Pressure, Heart Rate, Respiratory Rate, Ventilation Status, FiO2, PO2, PCO2, Peripheral Capillary Oxygen Saturation, Arterial pH, Lactate, Sodium, Urine Output, Creatinine, Urea, Glucose, Albumin, Bilirrubin, Hematocrit,

Platelets, WBC, GCS, ICU LOS, Gender, Ward IN, Patient Type, Principal Diagnostic, PrevHospDays.

2. Comorbidities: Chronic Renal Failure, Cirrhosis, Hepatic Failure, Metastatic Carcinoma, Lymphoma, Leukemia/Myeloma, Immunosupression, AIDS.

3. Scores: SAPS, SAPSII, APSIII, OASIS, SOFA.

Once the ETL process is defined and converted to an automated process, it can be thought about modify some of the original variables either dropping a set of them or including additional ones. In posterior stages of the project it is going to be seen the truly relevant variables that guide to a possible deployment process and the variables that are just noise.

Chapter 3

ETL

This section introduces the ETL followed. It is described, in a general way, the procedure and its operation, as well as the structure of the developed processes. Within the methodology used, the use of an ETL is a fundamental piece that allows to obtain the virgin data from the original database, to perform part of the preprocessing, to group and consolidate the data and it is easily adaptable to other databases of origin modifying the input queries.

3.1 Context

As a first step before performing any data analysis task, it is necessary to obtain the raw data and process it to obtain consistent data that can be used in an analysis. Generally, it is considered that 80% of the time devoted to a data analysis project is used in the process of cleaning and preparing the data **[15]** **[16]**. Figure 3.1 shows the steps of a data analysis **[17]**.

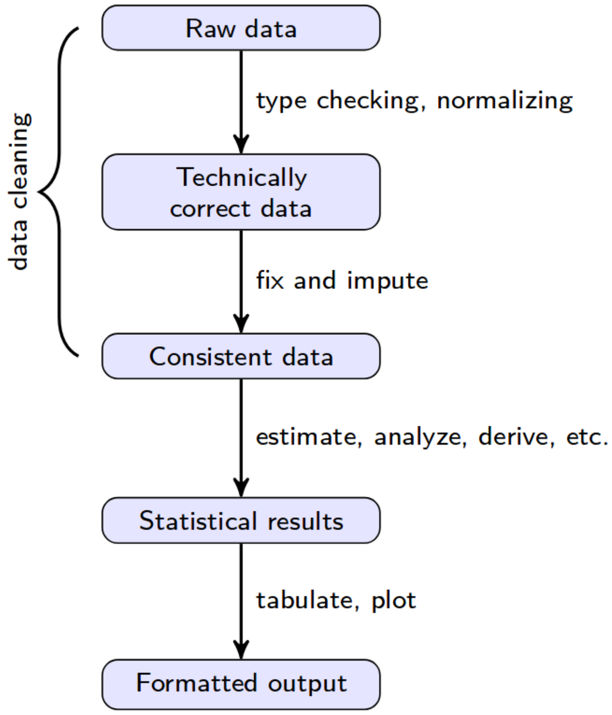


FIGURE 3.1: Statistical analysis value chain

In this study, the data source is the MIMIC-III database, which stores them in a relational structure that it is necessary to convert to an appropriate structure for statistical analysis or tidy, that is, to transform the structure of multiple tables with different number of rows and columns linked by keys[1] to a new structure where each variable is an independent column, each single ICU admission instance forms a row and each different type of problem approach constitutes a table.

There are several technical alternatives to perform this processing. In this study it is used the capabilities of the Python programming language to obtain, transform and filter the data sets and convert them into tidy data. Considering that the original data source in this case is obtained from a SQL database, it was also possible to obtain the data from a Python session directly through both pandas and SQL. The details of the ETL followed are exposed in the next section.

3.2 Technical details

It has to be considered, fixed the business and data understanding, the features to be extracted. That is, to understand the different nature of each table that constitutes the entire MIMIC-III database and where the desired features are located. That points were fixed in the previous Chapter, Chapter 2.

Among all the 26 comma separated files within MIMIC-III is constituted, a first step to classify them in order to proceed with the ETL is to inspect their size. There are several of them that are huge[2] and others that are medium to tiny. On the one side, the small files does not present any kind of problem to be read directly as a pandas DataFrame in order to preprocess for posterior analysis.

Why pandas? Pandas is an open source library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language which has long been great for data munging and preparation, but less so for data analysis and modeling. Pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language. Combined with the excellent IPython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity and the ability to collaborate.

Among all the pandas library highlights[3] it will be of particular interest:

- A fast and efficient DataFrame object for data manipulation.
- Tools for reading and writing data between in-memory data structures and different formats.
- Aggregating or transforming data with high performance merging, joining and grouping by engine allowing split-apply-combine operations on data sets.

On the other side, the big files present a problem when RAM memory limitations are present and the use of pandas library approach is not the adequate one in that cases.

To perform the ETL process among the big files it was used another Python library, sqlite3[4]. SQLite provides a lightweight disk-based database that does not require a separate server process and allows accessing the database using a nonstandard variant of the SQL query language. With the use of sqlite3, it can be transformed any comma separated file to a database file of extension .db, to then define an SQL query on that file. The desired database output, which

[1] To understand better the relationships between tables it is useful to look at https://mit-lcp.github.io/mimic-schema-spy/relationships.html, which is kind of an interactive graph view.

[2] The notion of size is referred to the RAM capabilities of the personal laptop used for the development of the project in relation to the file aiming to be treated.

[3] https://pandas.pydata.org/

[4] https://docs.python.org/3/library/sqlite3.html

will be a subset of the entire original database, hence smaller, can be converted to a pandas DataFrame which will be able to be allocated in RAM memory. The obtained DataFrame can be then combined in a desired way with other pandas DataFrames.

It was used that approach because the use of pandas contributes simplicity and ease treatment of data in order to convert to tidy, which is the main goal of the ETL process before jumping to data analysis and modeling.

Chapter 4

Data analysis and model evaluation

As stated in 2.1, the project was developed only with Metavision data because of the data quality and availability, althought some aspects required to be fixed. In this chapter is presented some of the most relevant ones along with some tips about filling missing data and some exploratory analysis. Problem understanding and model evaluation is also presented.

4.1 Data preparation, data cleaning and exploratory data analysis

Patients who are older than 89 years old at any time in the database have had their date of birth shifted to obscure their age and comply with HIPAA. The median age for the patients whose date of birth was shifted is 91.4. I did a very simple assumption: assign to that group of patients a random triangular probability distribution function comprised between 89 and 100 years of age. It is reasonable given the information about the median of an unknown distribution for that patients's age. This is important because we must work with numerical age values because age is a potential relevant feature candidate.

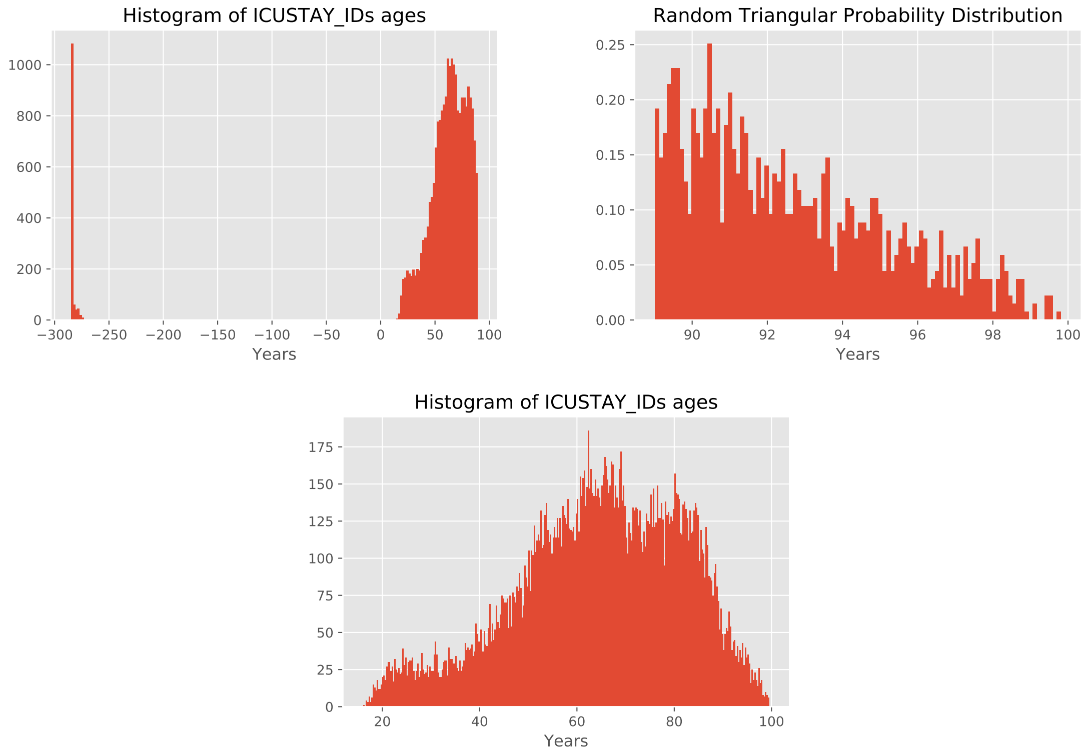


FIGURE 4.1: ICU admission ages preparation.

Regarding categorical admission types, because of the similar nature of the categories URGENT and EMERGENCY and the so few URGENT observations, the URGENT category was assigned to the EMERGENCY one. This treatment is performed for all categorical variables where some category has so little observations. Another example is for categorical insurance types where Self Pay and Private are of the same nature as well as Government and Medicare categories.

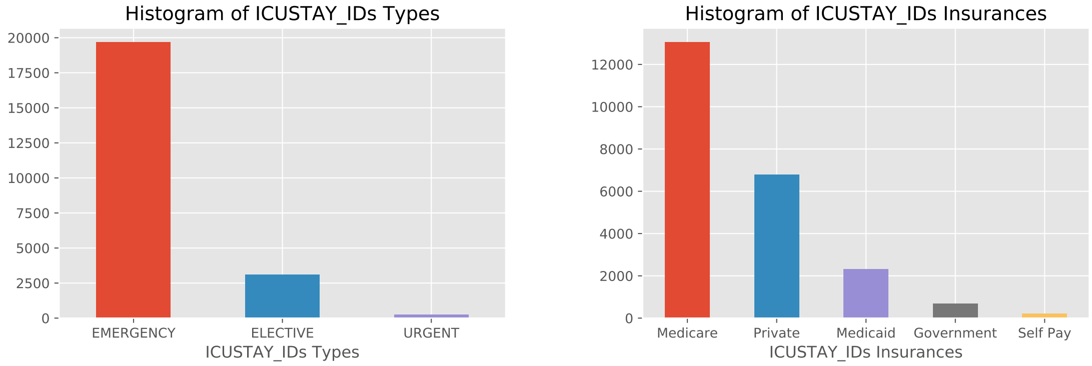


FIGURE 4.2: ICU admission types and insurance types.

Regarding times we have two major ones to be considered. The time that lasts for the entire hospital admission and the time that only lasts at the ICU. The two next histograms will give an idea about it:

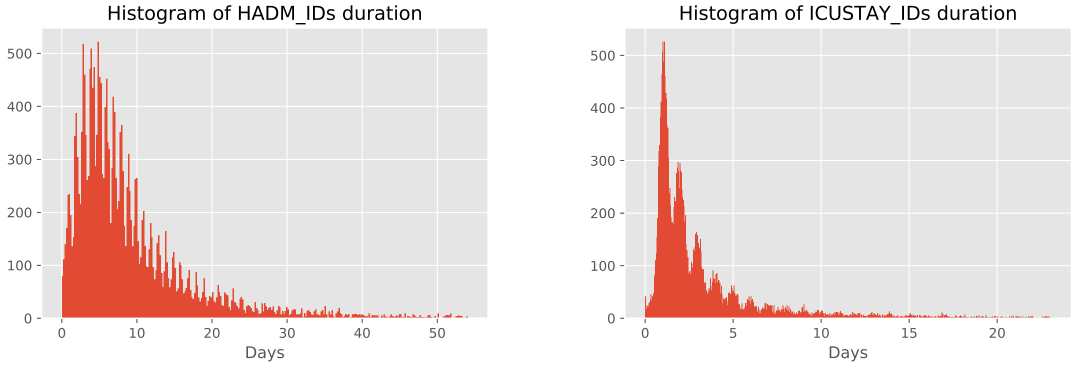


FIGURE 4.3: Hospital admissions durations (left) and ICU admissions LOS (right)

4.1.1 Feature engineering and missing values treatment

Once the Dataframe is completed with all the desired features we have to distinghish between the categorical and the numerical variables.

For the categorical variables One Hot Encoding technique was performed once that the variables were correctly treated as explained before. Regarding feature engineering some new variables were defined from others. This is the case of the body mass index (BMI) which is computed from admission height and weight. In the strict sense of a feature engineering procedure the age needed the treatment exposed before along with the fact that the age must be computed from the admission time and the date of birth of patients, that is, is a derived feature. Another new feature is the time previous to an ICU admission which is another variable defined from others. In that case from both the hospital admission and the ICU admission times.

Regarding the procedure of filling the missing values, two strategies were performed. The first one was to take advantage of the subject identification. Because there are subjects with more than one ICU admission, and there are some missing values for a given admissions I grouped by subject identification and I filled the missing values with a propper operation either taking the most frequent value of the rest of available observations for the considered variable or simply the mean. Depending on the case it was used one or another. The problem of that strategy is that only a certain small group of missing values can be filled. When the first strategy was done, the second one is to do the same but instead of grouping by subject id, group by gender[1]. With this procedures we ensure that all the missing values will be filled.

As an example of the procedure, when the BMI was tried to be computed from the weight and height variables, it was encountered that the admission weights were poor on missing values whereas the admission heights had a considerable amount of missing values. Once the strategies were performed it can be seen that the resulting histograms are pretty reasonable.

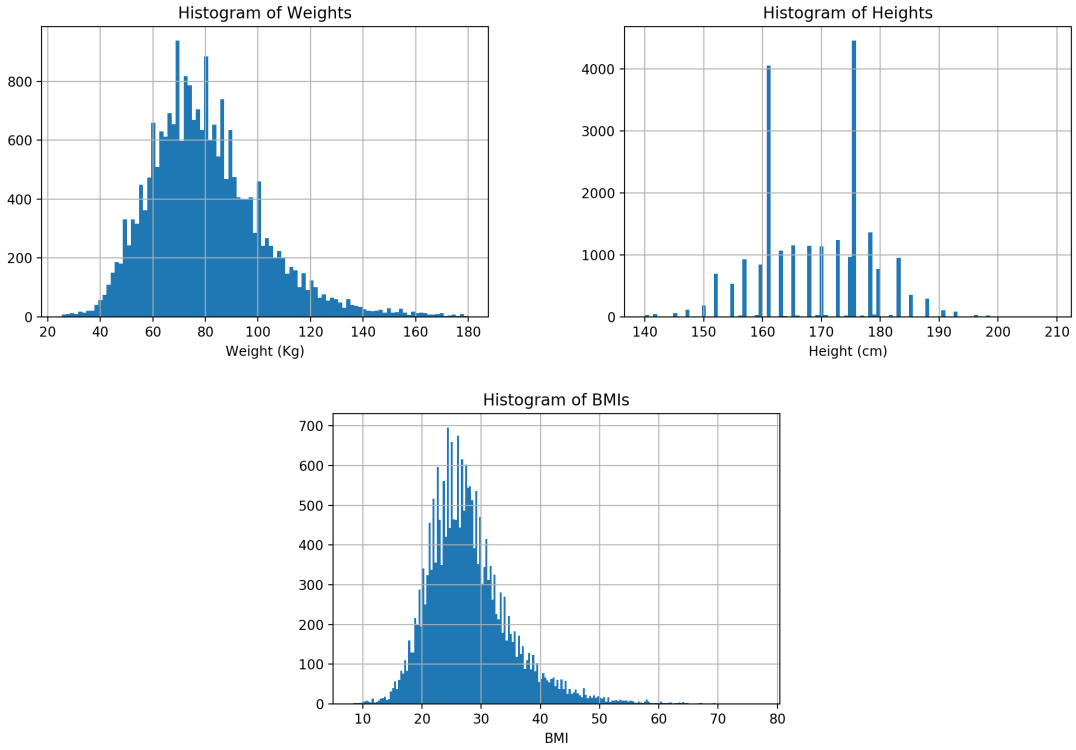


FIGURE 4.4: Weight, Height and BMI histograms.

Regarding outliers, instead of perform an automatic procedure once defined some criteria, for instance given a variable remove all the values far away from some multiple of the standard deviation with respect the mean/median, I did it manually, first plotting each variable versus the LOS to construct some visual intuition (value range and correlation with LOS) and then define a criteria following my opinion. This will give some freedom and the outlier treatment procedure will not be fussy. Actually, the admission instances which contains outliers in one or more variables are preserved. Instead, what it was done is to convert the found outliers to missing values in order to be filled with the strategies described before. With the use of that tricks we ensure that the number of admission instances is maintained.

[1] It is common in ICU variable observations to distinguish between typical values for males and for females. Inspecting numbers it is true that some variables had strong dependency with the gender. One example known for everybody is the height and the weight.

What I like from that kind of plots is that it can be defined some kind of transparency to each observation which help to construct a visual intuition about the relation of the plotted predictor against the target variable. A couple of examples are shown below:

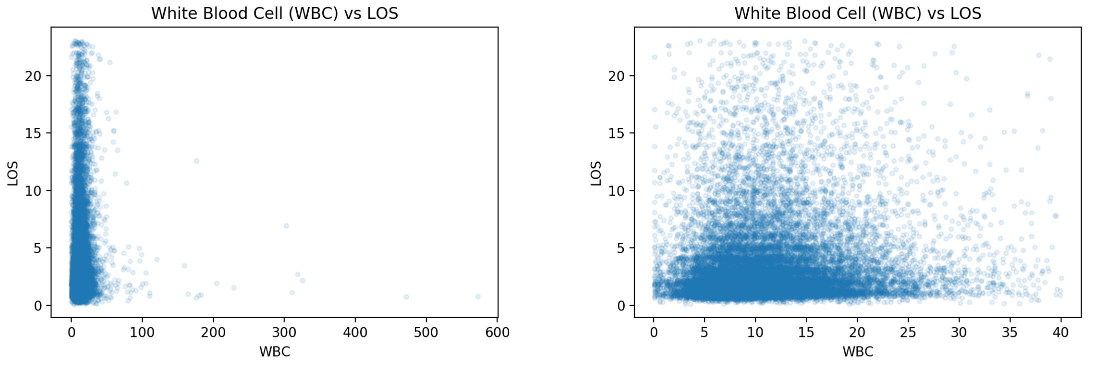


FIGURE 4.5: WBC vs LOS before and after outlier removal.

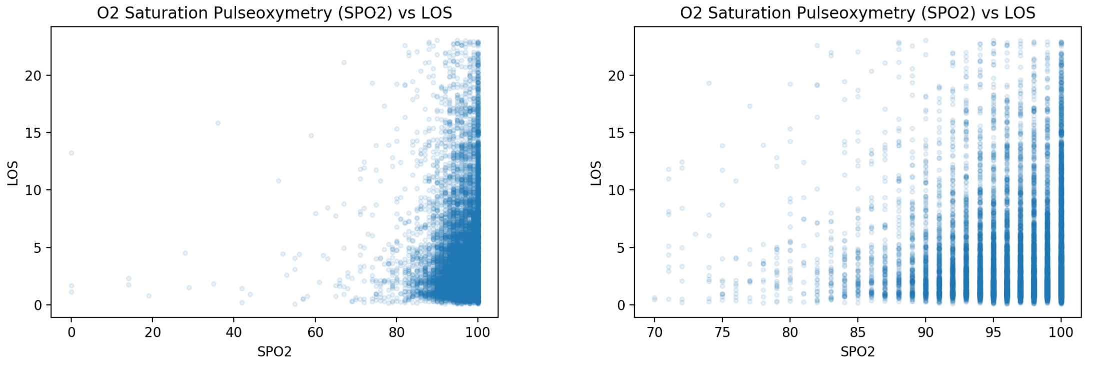


FIGURE 4.6: SPO2 vs LOS before and after outlier removal.

We can see that instances with low and high admission WBC tends to have short LOS whereas LOS tends to take a wider range for instances with normal WBC admission values. On the other hand, for the SPO2 admission variable, lower values tends to have shorter LOS whereas normal to high values of admission SPO2 tends to take a wider range of LOS values.

All this procedures either outlier removal, treatment of missing values and feature engineering were fundamental in order to have the more complete dataframe possible.

4.2 Problem statement

Histograms of Figure 4.3 represents occurrency frequencies of time spent in the Hospital and ICU LOS. We can see that LOS is shorter because a general admission hospital includes LOS together with additional time spent in other hospital departments that are not the ICU. Both histograms are showed after outlier removal with the same criteria of the used before. Also, in order to preserve coherence, the observations that presents negative times are removed. Either for the presence of outliers or for the incoherence among admission times in this cases the admissions with this characteristics are completely removed because of the fact that LOS is the target label.

Arrived at that point comes the following problem: *Is it possible to perform supervised learning with success*[2] *in order to predict, for a given admission, the LOS associated?*

There is no direct reason to think in a pessimistic way because we have the two principal needs to try to give an answer to that question:

1. We have data.
2. We have labels associated to that data.

Given the nature of the label as being a continuous one the natural way to abord the problem is via a supervised, regression machine learning approach. But, in contrast, there is the possibility of artificially transform the labels to be made discrete binary ones by defining a cutoff value, with results being designated as positive or negtive depending on whether the resultant value is lower or higher than the cutoff[3].

However, such conversion causes a loss of information, as the resultant binary classification does not tell how much above or below from the cutoff a value is. As a result, when converting a continuous value that is close to the cutoff to a binary one, the resultant positive or negative predictive value is generally higher than the predictive value given directly from the continuous value. In such cases, the designation of the test of being either positive or negative gives the appearance of an inappropriately high certainty, while the value is in fact in an interval of uncertainty. For example, think about the urine concentration of hCG as a continuous value. A urine pregnancy test that measured 52 mIU/ml of hCG may show as "positive" with 50 mIU/ml as cutoff, but is located in an interval of uncertainty, which may be apparent only by knowing the original continuous value. On the other hand, a test result very far from the cutoff generally has a resultant positive or negative predictive value that is lower than the predictive value given from the continuous value. For example, a urine hCG value of 200000 mIU/ml confers a very high probability of pregnancy, but conversion to binary values results in that it shows just as "positive" as the one of 52 mIU/ml.

At first sight, choose a classification approach to the problem instead of a regression one with the consequent loss of information could seem a problem to face but there are two main reasons in the context of this study to think about that made the problem a classification one could be a clever proposal and not a bad one at all:

1. Physicians works with qualitative LOS[4].
2. The classification approach might also be simpler than the regression one which is a more complex problem that generally performs worst than the same problem converted to a classification one using the same data.

4.3 More about binary classification

Once that binary classification approach is chosen for our problem, we have to think about the metrics we need to assess performance. This is a crucial tradeoof that occurs so frequent with classification problems. In essence it comes because of the unbalanced dataset and the value of miss classify a particular class. Depending on the cutoff value chosen, we are going to have a balanced, unbalanced or drastically unbalanced dataset.

[2] The notion of success for this problem is going to be stated nextly.

[3] When we are talking about categorical data, we have to further distinguish between nominal and ordinal features. Ordinal features can be understood as categorical values that can be sorted or ordered. In contrast, nominal features don't imply any order. Because our new categorical labels are nominal there is absolute freedom to categorize them in a binary form, they just act with the purpose of become distingibles.

[4] Giving them the information that a given admission will spent shorter or longer than a LOS cuttoff value without being worried about an exact value could be of big utility

If we know the probability distribution[5], as is the case (and it will be a normal situation because they are giving us the data), and we fix a cuttoff in some value we are going to know, inmediately, the number of data instances belonging to each class. Depending on the relative number of instances belonging to each class we are going to deal with a balanced or unbalanced labeled dataset. For instance, if we fix a cuttoff exactly in the mean value of the probability distribution we are dealing with a completely balanced classification problem, that is, we are dealing with the same number of admissions tagged as LOS 1 (LOS below the cuttoff) or LOS 0 (LOS above the cuttoff), so selecting at random 1 or 0 as well as selecting everytime the same class, we are going to get an overall fraction of instances that are correctly categorized of a half. That notion of metric performance is what is called accuracy and is defined as:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \tag{4.1}$$

Here $\hat{y}_{\mathrm{i}}$ is the predicted class label for the ith observation using $\hat{f}$, our estimate for f (some fixed but unknown function that defines the problem deterministically), n is the number of ICU admissions, the i index represents each single independent ICU admission, $I(y_{\mathrm{i}} = \hat{y}_{\mathrm{i}})$ is an *indicator variable* that equals one if $y_{\mathrm{i}} = \hat{y}_{\mathrm{i}}$ and zero if $y_{\mathrm{i}} \neq \hat{y}_{\mathrm{i}}$. If $I(y_{\mathrm{i}} = \hat{y}_{\mathrm{i}}) = 1$ then the ith observation was classified correctly by our classification method; otherwise it was missclassified. Hence Equation 4.1 computes the fraction of correct classifications. An alternative way to define accuracy metric is the following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

Where TP and TN are positive and negative instances correctly classified by the method, FP are negative instances incorrectly classified as positives and FN are positive instances incorrectly classified as negatives[6]. Equations 4.1 4.2 tells the same history; as more accuracy the method strives the more overall correctly classified instances we get. In that line, accuracy ranges between 0 and 1.

When we use accuracy, we assign equal cost of miss classify instances independently of the class where it belongs, that is, we assign equal cost to FP and FN. When the dataset is not balanced there is a great way to lower the cost. Predict that every instance belongs to the majority class will get an accuracy of the relative presence of that class with respect the entire dataset.

The problem starts when the actual costs that we assign to every error is not equal. If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests. In general, there is no a measure better than the rest, the best measure is derived from the needs. In a sense, it is not a machine learning question, but a business question. It is common that two people will use the same data set but will choose different metrics due to different goals. Accuracy is a great metric. Actually, all metrics are great and must evaluate each metric if the problem is not clear. However, at some point you will need to decide between using one model or another. There you should use a single metric that best fits your need.

Is accuracy a good metric to choose for our problem? In principle, yes it is. It is a great metric because the cost of missclassify one class or the other is not a major problem because we are trying to guess LOS, that at the end is only time. But just in principle. Imagine that

[5] We work with histograms not with probability distributions. Strictly speaking, a histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a quantitative or continuous variable. Just a language abuse.

[6] The convention used to codify classes was choosen in the following manner. Positive classes are categorized as 1 and plotted in orange, Negative classes are categorized as 0 and plotted in blue. In that line, it is used indifferently Positive or 1, Negative or 0, meaning lower or higher from the cuttoff value definition respectively.

we construct a drastically unbalanced dataset by the definition of a high cuttoff value (as in the bottom of Figure 4.7). Patients belonging to the minority class (the ones with longer LOS or N) are more important to correctly classify (because of the low costs associated to a TP are expected but if the classification process fails, giving a FP, there could be billing surprises[7]) so we must think about dealing not only with accuracy metric when alternative scenarios are considered. Have in mind also the falses associated to the majority class (FN), they correspont to patients that will spent short LOS (P) but the method fails, which turns into savings because the expected costs were higher. We are going to present that kind of scenarios in Section 4.5. For the moment we just try to understand the LOS prediction where a FP has the same cost of a FN meaning that accuracy metric is enough and it will be used to assess performance in order to select a good model.

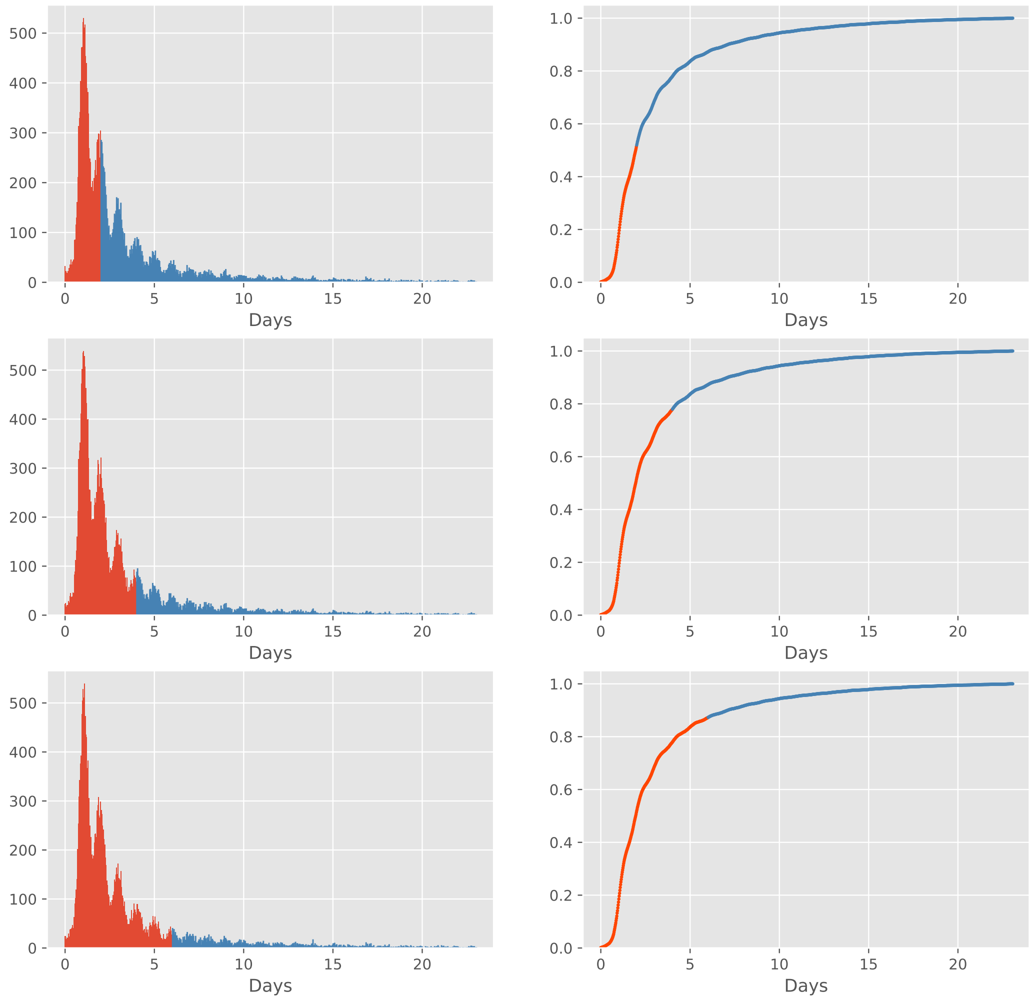


FIGURE 4.7: ICU LOS histograms and CDF with cuttoff values at 2, 4 and 6 days[8]

In Figure 4.7 it is shown the ICU LOS where some cuttoff values are defined (2, 4, 6 days). It is shown also the cumulative distribution function (CDF) associated to each histogram in order to have a visual idea of how much the unbalancing is.

[7] Without mention the more probabilities of contracting an HAI due to the unexpected longer ICU stay.

[8] Proportion of the Majority Class with respect to the entire data set: 0.5099 (top), 0.7769 (middle), 0.8717 (bottom).

Actually it can be defined all sorts of cuttoff values because of the numerical nature of LOS. Given a need, a cuttoff value will be defined. This is the logic. Then the method will perform a guess (representing short stays vs long stays). This will be an indicator that can be used in addition to the others that already exists, such as Apache IV LOS guessing https://intensivecarenetwork.com/Calculators/Files/Apache4.html. Finally, to assess performance could be adequate the accuracy metric which simply measures the fraction of all instances that are correctly categorized and there is no priority among classes.

4.4 Model selection

Once the data is prepared, the problem stated, the metric to assess performance is chosen and there are model candidates is the moment to run them. The candidate models are k-Nearest-Neighbor Classifier, Logistic Regression Classifier, Neural Networks Classifier and Random Forest Classifier. In the case of the Neural Networks Classifier it was used a Multilayer Perceptron for binary classification that uses the sigmoid activation function in order to produce a probability output in the range of 0 to 1 that can easily and automatically be converted to crisp class values[9]. But first let's standarize the data for both the Basic Dataframe and the Rich Dataframe[10].

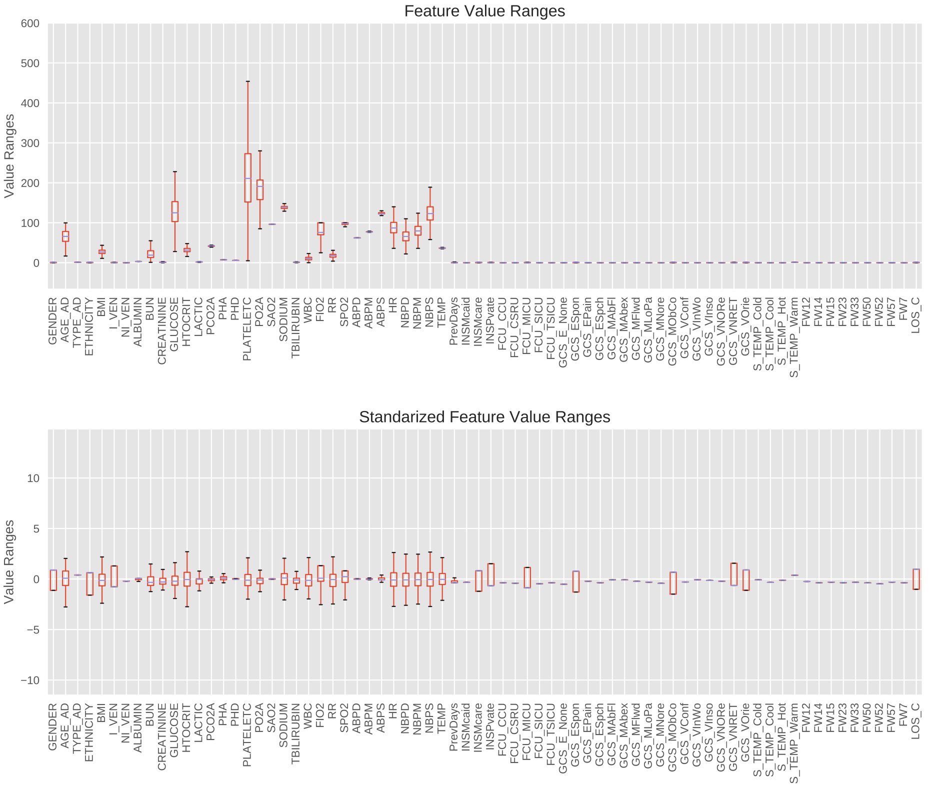


FIGURE 4.8: Feature value ranges with and without standarization for Basic Dataframe[11]

[9] The architecture for each model can be found on Appendix C.

[10] The Rich Dataset consists in the Basic Dataset features but including comorbidities and scores obtained as a derived information from MIMIC-III as detailed in Chapter 2.

[11] In Appendix B you can find the complete feature names corresponding to each single feature abbreviation.

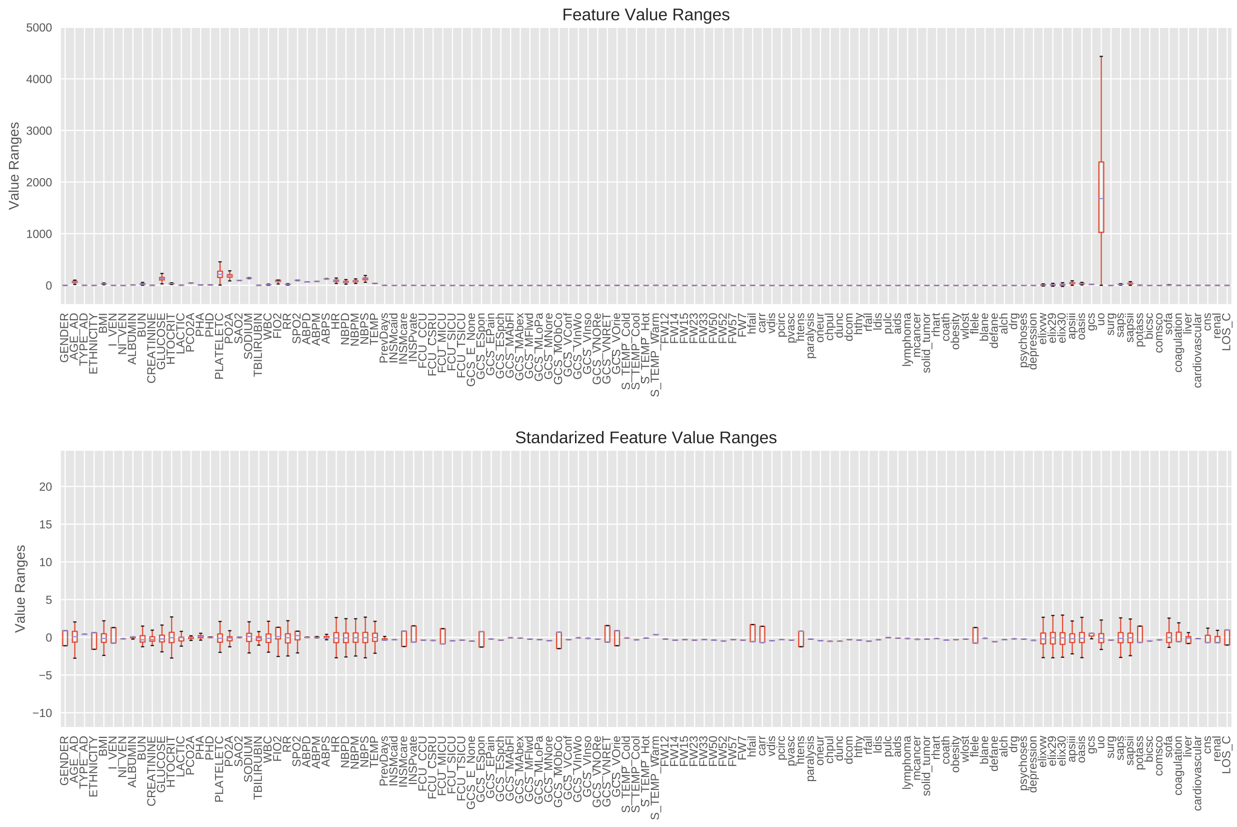


FIGURE 4.9: Feature value ranges with and without standarization for Rich Dataframe

K Fold Cross Validation technique was used. It is a very useful technique for assessing the effectiveness of a model, particularly in cases where there is a need to mitigate overfitting. It is also of use in determining the hyper parameters of a model, in the sense that which parameters will result in lowest test error.

As there is never enough data to train a model, removing a part of it for validation poses a problem of underfitting. By reducing the training data, we risk losing important patterns in data set, which in turn increases error induced by bias. So, what we require is a method that provides ample data for training the model and also leaves ample data for validation. K Fold Cross Validation does exactly that.

In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time one of the k subsets is used as the test/validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness of the model. As can be seen, every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times. This significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set. Interchanging the training and test sets also adds to the effectiveness of this method. K = 10 was used, but nothing is fixed and it can take any value.

The accuracy outcomes for each model are the followings:

TABLE 4.1: Accuracy results for the basic DataFrame[12]

Cuttoff	MC	KN	LR	NN	RF
2 Days	0.5099	0.6254	0.6802	0.7080	0.7293
4 Days	0.7769	0.7898	0.8130	0.8288	0.8458
6 Days	0.8717	0.8733	0.8791	0.8804	0.8916

TABLE 4.2: Accuracy results for the rich DataFrame

Cuttoff	MC	KN	LR	NN	RF
2 Days	0.5099	0.6304	0.6958	0.7218	0.7341
4 Days	0.7769	0.7918	0.8251	0.8347	0.8461
6 Days	0.8717	0.8740	0.8858	0.8891	0.8898

TABLE 4.3: Accuracy results for the rich DataFrame with feature reduction

Cuttoff	MC	KN	LR	NN	RF
2 Days	0.5099	0.6665	0.6842	0.7155	0.7319
4 Days	0.7769	0.8078	0.8214	0.8365	0.8492
6 Days	0.8717	0.8781	0.8838	0.8886	0.8952

Once obtained the results it was selected the Random Forest Classifier due to the accuracy performane achieved. As it can be seen in the tables, the results improve if we use the Rich Dataset which contains more potential data. The results are statistically stable until the third digit after the coma so it can be stated that the models, once fitted to the data, are all robust. In the case of the results obtained using the Rich DataFrame with feature reduction, the criteria to obtain the subset of features was to drop all the features with a feature importance score lower than 0.004. With this criteria the original Rich Dataset consisting in 121 features was reduced to 47 with a cuttoff at 2 days, with a cuttoff at 4 days was reduced to 48 and with a cuttoff at 6 days was reduced to 49. In all cases the feature space is reduced more than the half of the original shape. In the tables it can be seen that the models applied to the Rich Dataframe with feature reduction tends to improve, meaning that some sound is dropped.

PR and ROC curves are included nextly for each situation. In each page we have at the top the results for a cuttoff at 2 days, at the middle the results for a cuttoff at 4 days, at the bottom the results for a cuttoff at 6 days. It is usefull to include the PR curve for the reasons explained in the next section. For now just a kick comment; in that problem, Random Forest Classifier is a great model that not only performs good in terms of accuracy metric but also in terms of Precission and Recall as it can be seen in the PR curve. In all the PR curves, for a given Precision we get more Recall with that method than taking any of the rest. The same for a given Recall the more Precision is achieved with that method than with any of the rest.

[12] MC: Majority Class Classifier
KN: k-Nearest-Neighbor Classifier
LR: Logistic Regression Classifier
NN: Neural Networks Classifier
RF: Random Forest Classifier

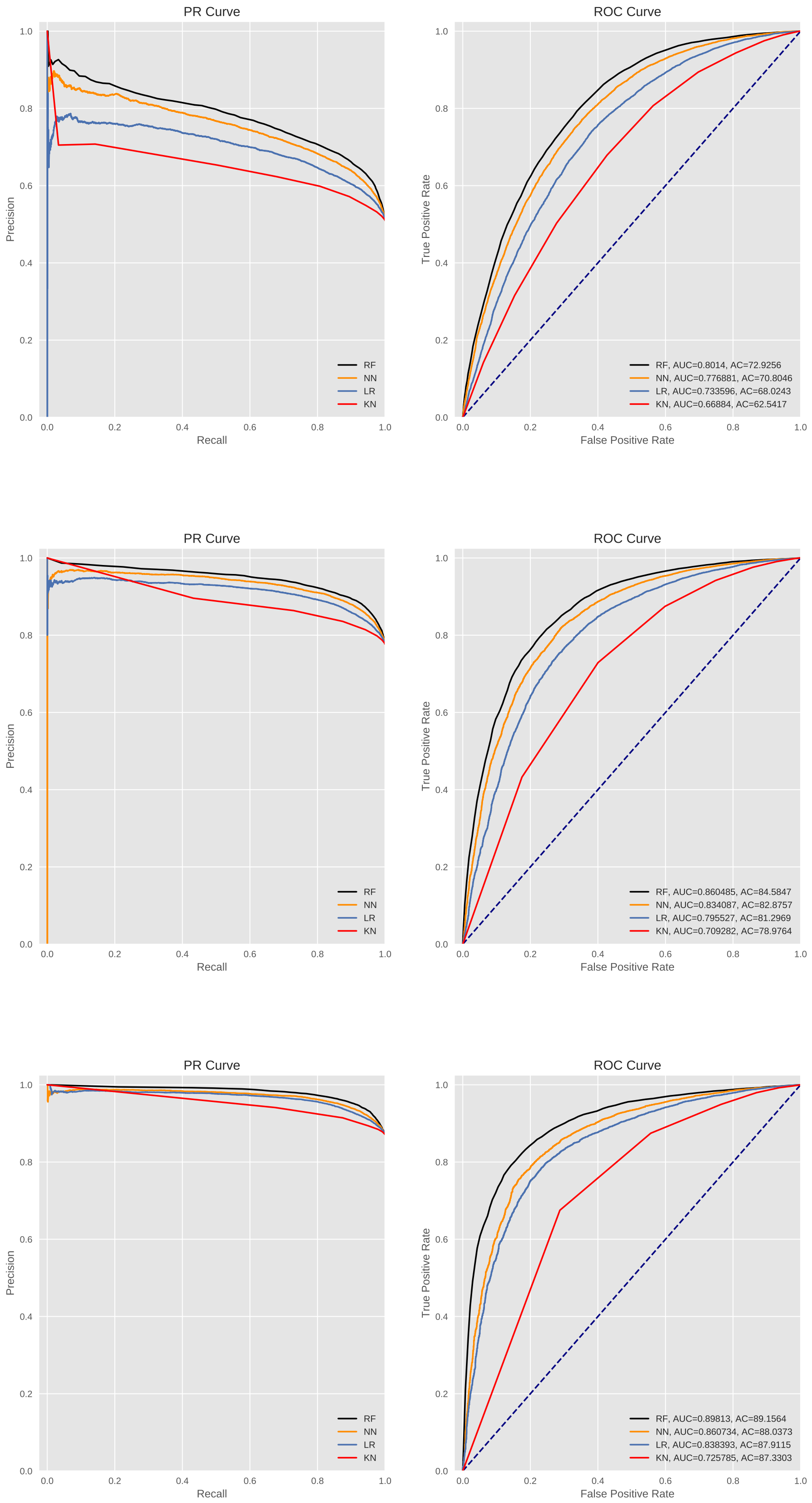


FIGURE 4.10: PR and ROC curves for the Basic Dataframe

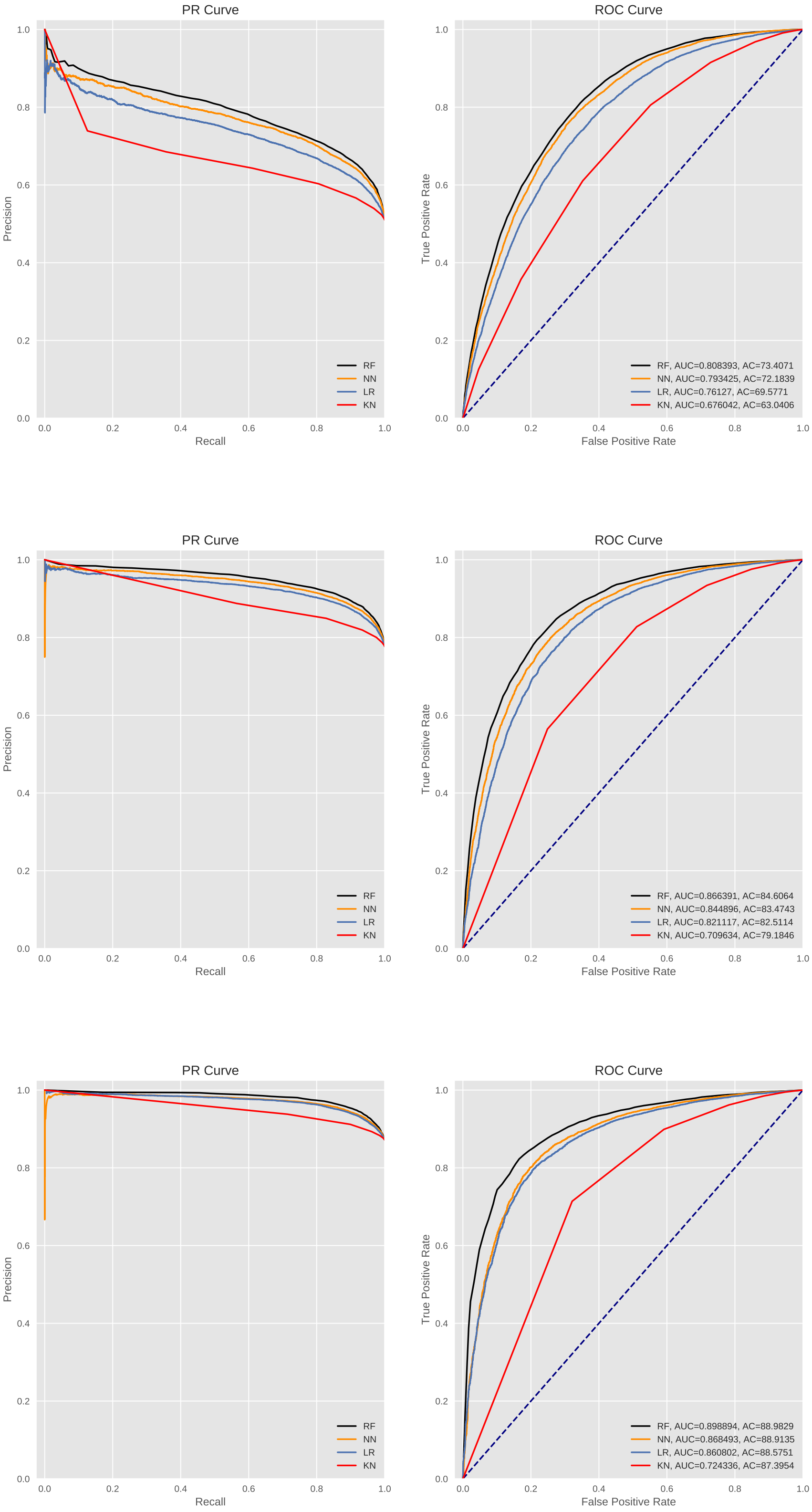


FIGURE 4.11: PR and ROC curves for the Rich Dataframe

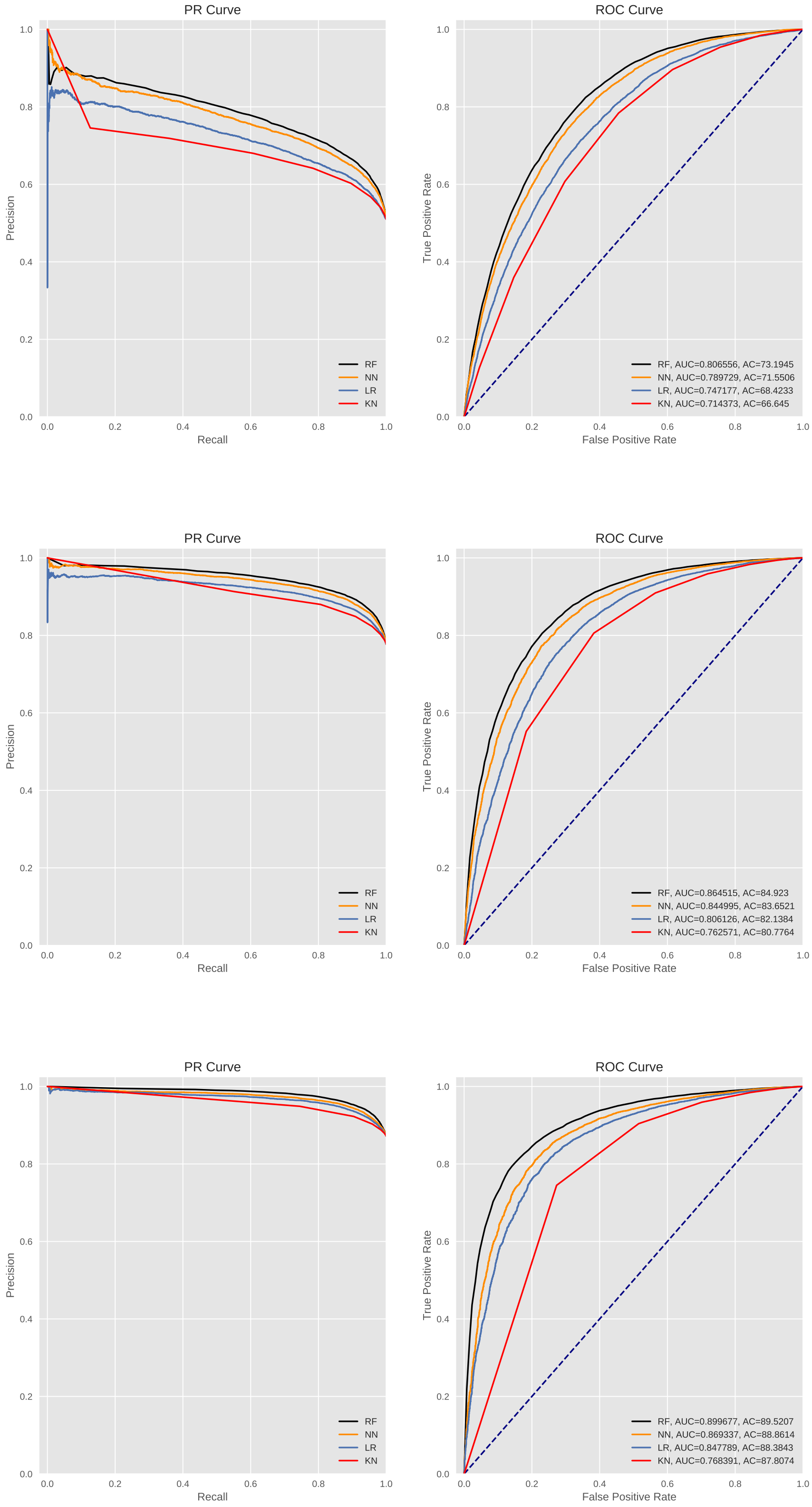


FIGURE 4.12: PR and ROC curves for the Rich Dataframe with feature reduction

4.5 Precision-Recall curve

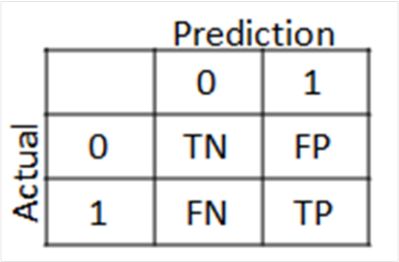
To assess performance, it was used accuracy. The argumentation was related to the equal cost of missclassify one class or the other. The only important point was to correctly classify the maximum number possible of instances independently of the class at which each admission belongs. What if missclassify one class costs more than missclassify the other? To understand the nature of the problem first we need to fix ideas. Given a binary classifier with positive and negative classes, precision and recall metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

That performance metrics can be plotted in order to obtain what is called the Precision-Recall curve or PR curve as it was done in the previous section. Because in the ideal case of having a perfect binary classifier both precision and recall will be 1 (because both FP and FN will be 0), the better a PR curve the closer to the right-top corner.

Given a classification method it can be defined the confusion matrix where are located the basic elements that serves to compute derived performance metrics such as the defined Precision, Recall and Accuracy.



		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

FIGURE 4.13: Confusion matrix convention for that project

To show a practical case, in this section it is going to work with the Rich Dataset with feature reduction using the Random Forest Classifier and setting the cuttoff value at a LOS of 4 days. The results for that example are located in the middle plot of Figure 4.12. If we inspect the obtained results for that case, we got 84.92% in accuracy where the majority class dominates the dataset by 77.69%. So the classification method is able to learn more than 7% in accuracy above from the Majority Class Classifier. The resulting confusion matrix in this case is:

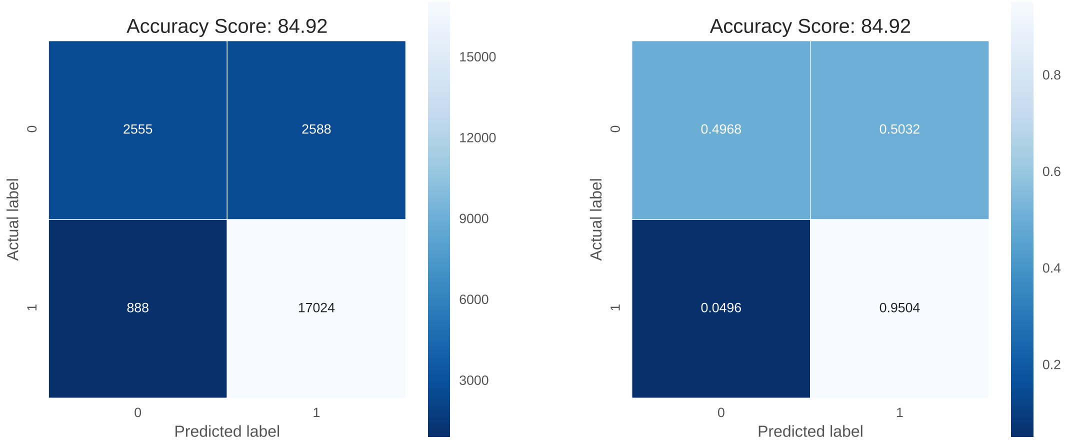


FIGURE 4.14: Confusion matrix and normalized confusion matrix by rows with the default Threshold

This result is the optimal one if we optimize the accuracy metric. What if missclassify instances from the minority class (FP) has more cost than missclassify instances from the majority class (FN)? One approach is the visual that consist in plotting precision and recall as dependent variables of a Threshold which is a value that the classification method use to convert a probability class prediction of an instance to a class itself; if the probability class prediction is above or below with respect the Threshold, it is assigned to one class or the other.

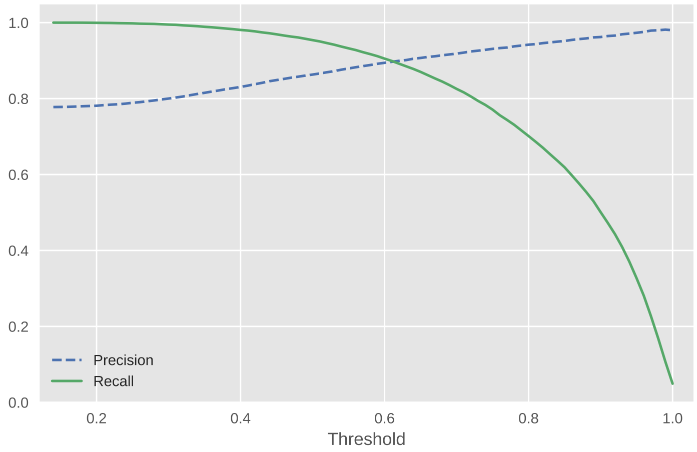


FIGURE 4.15: Precission/Recall plot versus Threshold

It can be computed manually Precision and Recall metrics from the confusion matrix in order to know from the previous plot the actual value of the Threshold that the classifier uses by default in order to optimize accuracy metric. From the confusion matrix:

$$Recall = \frac{TP}{TP + FN} = \frac{17024}{17024 + 888} \simeq 0.9504 \tag{4.5}$$

$$Precision = \frac{TP}{TP + FP} = \frac{17024}{17024 + 2588} \simeq 0.8680 \tag{4.6}$$

From that results we see that the Threshold is around 0.5.

Now you can simply select the Threshold value that gives you the best Precision/Recall tradeoff for your task. Another way to select a good Precision/Recall tradeoff is to plot precision directly against recall as we did before. As you can see, it is fairly easy to create a classifier with virtually any Precision you want: just set a high enough Threshold, and you are done. Hmm, not so fast. A high-Precision classifier is not very useful if its Recall is too low. If someone says “let’s reach 99% Precision,” it should be asked, “at what Recall?”

If we set the Threshold to 0.8 the confusion matrix results to be:

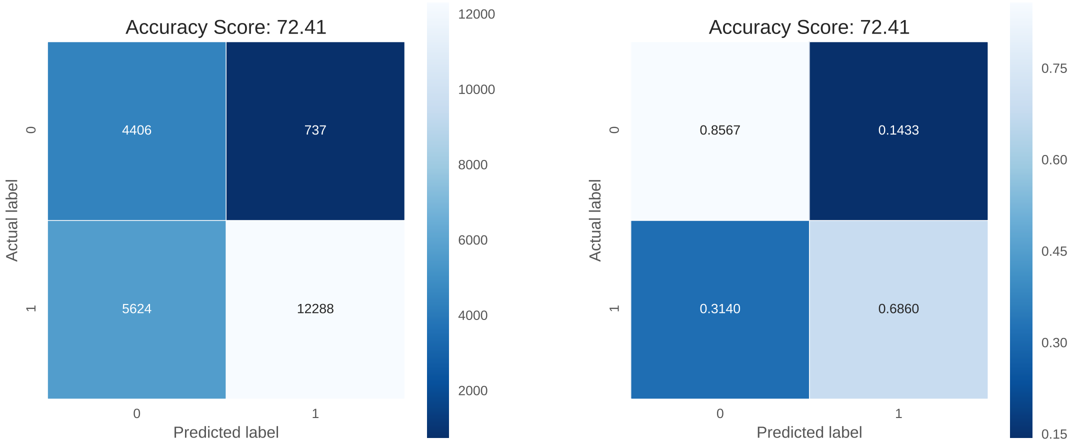


FIGURE 4.16: Confusion matrix and normalized confusion matrix by rows with Threshold of 0.8

And the Precision and Recall metrics are:

$$Recall = \frac{TP}{TP + FN} = \frac{12288}{12288 + 5624} \simeq 0.6860 \tag{4.7}$$

$$Precision = \frac{TP}{TP + FP} = \frac{12288}{12288 + 737} \simeq 0.9434 \tag{4.8}$$

Setting the Threshold to 0.9 the confusion matrix, the Precision and Recall metrics are:

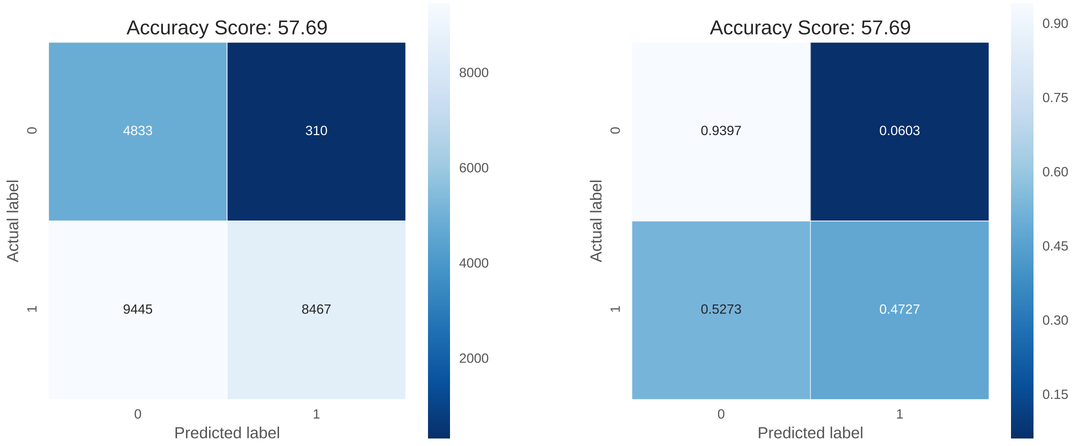


FIGURE 4.17: Confusion matrix and normalized confusion matrix by rows with Threshold of 0.9

$$Recall = \frac{TP}{TP + FN} = \frac{8467}{8467 + 9445} \simeq 0.4727 \tag{4.9}$$

$$Precision = \frac{TP}{TP + FP} = \frac{8467}{8467 + 310} \simeq 0.9647 \tag{4.10}$$

Is important to have a great accuracy when the important instances are missclassified? It can be seen how the accuracy metric is dropping when the Threshold is augmented at the same time that the classification of the minority class is improved because of the higher Precission achieved due to the less presence of FP. If the user's objectives are not clear, the Precision and Recall cannot be optimized whereas Accuracy takes the lead as being the metric capable to be optimized. Elsewhere, if the missclassification costs are defined[13], the problem is converted to an optimization one where Accuracy metric leave the leadership to Precission and Recall metrics.

[13] Maybe just merely economical costs associated to longer LOS or logistic costs like patients transference due to the lack of beds even mortality costs...

Chapter 5

Conclusion

There are three general approaches to improve an existing machine learning model:

1. Use more (high-quality) data and feature engineering.
2. Try different algorithms.
3. Tune the hyperparameters of the algorithm.

These are presented in the order in which I tried them. Often, the immediate solution proposed to improve a poor model is to use a more complex model. However, approach inevitably leads to frustration. A complex model is built over many hours, which then also fails to deliver, leading to another model and so on. Instead, we have to question ourselves: "Can we get more data relevant to the problem?" As pointed out in an article titled "The Unreasonable Effectiveness of Data" **[18]**, the amount of useful data is more important to the problem than the complexity of the model. Others have echoed the idea that a simple model and plenty of data will beat a complex model with limited data. If there is more information that can help with our problem that we are not using, the best payback in terms of time invested versus performance gained is to get that data.

In that section I comment that items in relation to the problems I had to face during the development of the project along with the results obtained and compare it with the expected goals the project attemted to achieve. In addition, some comments regarding future investigation on LOS topic that this project did not cover are presented.

5.1 Conclusions

The conclusions can be broken down into three sections. On the one hand, those related to the use of the MIMIC-III database, on the other those associated with the methodology, technical details and tools used and finally those related to the results of the analyzed data and the data itself.

In the first point, the use of the MIMIC-III database to analyze health data is considered to be probably the best alternative to develop studies that can be replicated and shared with the research and educational community, avoiding problems related to privacy and special sensitivity of health data, difficulties of integration between various applications, and so on. In any case, it is necessary to take into account certain specific characteristics of MIMIC-III due to the differences between the US health system and public health systems existing in Spain, mainly, the private nature of US health organizations and their orientation of the use of the ICD9 coding system not so much towards health objectives as towards the management of billing of clinical procedures. Another point to consider in MIMIC-III is the existence of many semi-structured data, manually entered in free text and in two different systems, which forces to make a search effort to obtain certain variables without access to the real systems for validation. Even with these considerations, the MIMIC-III database fully covers its objectives and this work would have been impossible without its existence.

A considerable part of the time spent was dedicated to the tasks of obtaining and processing data and at that point, the ETL processes developed have covered without problems the objectives. The use of Python to perform this type of analysis is justified by the availability of libraries and a large number of models. Regarding the evaluation of predictive models, multiple tests have been carried out until a satisfactory combination has been achieved. In Python there are a huge number of models and not all have stability and scalability characteristics suitable for working with large data sets in reasonable times. In summary, through the evaluation of different tools, technologies and models, a reliable, rapid and scalable combination has been reached for the type of clinical data studied.

An interesting point about the selection of models is that good results have been obtained with a known and classic model and not with another more modern as neural networks. On the one hand, the cost of optimizing parameters and training a neural network is very high and no outstanding results have been obtained compared to the rest of the models. On the other hand, the selection of variables previously carried out, with prevalence of variables that medical experience is related to LOS, may have favored models that use more direct relationships to obtain predictions. The associated advantage is that more interpretable models are obtained although one must always be cautious when quantifying the importance of a variable in the outcome outside linear models.

Finally, in the application of the predictive models it has been obtained that the model with the best results have been the Random Forest Classifier which obtain the best predictions in the complete data set and in the different groups in which this has been divided. The gain of accuracy as indicator was also obtained by including the comorbidities and scores variables in the set of predictors. From the clinical point of view, an interesting predictive power gain is considered, given the ease of obtaining the comorbidities of a patient with the initial diagnosis.

Regarding the analysis of the importance of the variables related to LOS, it is observed that they are almost all of medium importance, except for the variables with greater importance in all models. In general, the variables related to analytical results are more important, both for the basic dataset and in the different groups.

In conclusion, we have achieved models with great predictive capacity, selecting variables related to scores, comorbidities, basic analytical data and demographical data. As indicated, it is necessary to be cautious in the interpretation of the importance of the variables in the different models. *As a final comment, the application of the obtained method is very intuitive: just input the admission variables of an ICU admission instance and the model will give an estimate of a categorical LOS given a LOS cuttoff value*[1]*. Using this approach, the physician would has to define a LOS cuttoff (or a set of them) and the model will work for him or her in an hypotetical deployment. In that line, is kind of amazing that all the used models are capable to learn so much when the problem is so simplified and the data so complex.*

5.2 Drawbacks

The development of this work has met the objectives and initial expectations, especially in the section of technical evaluation and in the development of a methodology of analysis and application of predictive models to health data. The methodology followed has been sufficiently robust and effective for this type of project. The different phases of the work have not been exempt from both technical problems and associated with the data structure of the MIMIC-III database, which have forced to develop to a greater extent than initially estimated ETL

[1] It is assumed that the missclassification costs are fixed, that is, the optimization problem is defined which will define a Threshold value over the PR curve as explained in 4.5.

processes and dedicate time was thought. Among the problems encountered it is worth mentioning, for example, the difficulty to identify some variables in the MIMIC-III database and to understand the structure of the tables of some files.

It were identified three major drawbacks which, interestingly, are not related to problems encountered during the development of the project either technical ones as just described or more theoretical ones. The drawbacks were more related to clinical area and data quality assessment:

- Roughly speaking, LOS is hospital-dependent. That turns into that scaling the obtained models to new databases is a problem; the data recording process depends on procedures and clinical protocols may differ between regions/countries which will produce an hospital dependency modeling.
- The way in which the data was recorded presents a particular difficulty: there are some variables that were registered in different moments from one another's. In a sense, this will produce a correlation breaking process between the time-dependent variables, meaning that this recording procedures has the risk of losing important patterns in the dataset capable to define in a better way the relation between the predictors and the label.
- There are some important variables that presents a considerable large amount of missing values meaning that the inputation process followed can bias the model. Of course that in the ideal case where all the admission variables would be available, the performance metrics obtained would be higher.

5.3 Further investigation

At first sight, the future investigation line regarding LOS topic related to the present project is evident; instead to do a binary classification try a multiclass one or even change to a try of a regression approach[2]. But it is more adequate to start first for the simplest approach because the problem can be converted to a very challenging one and fixing ideas is easier with the use of simple models.

Another line of research would be to combine the obtained models with a powerful mortality classifier. With the combination of a LOS classifier and a mortality one and with the help of some probability theory answers to challenging questions could be addressed.

[2] As it was pointed out in 4.2 that line is not relevant for our purposes.

Bibliography

[1] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. *MIMIC-III, a freely accessible critical care database*. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635

[2] Pollard, T. J. & Johnson, A. E. W. *The MIMIC-III Clinical Database*. http://dx.doi.org/10.13026/C2XW26 (2016).

[3] J. R. Butler. *Hospital cost analysis*. Springer, 1995.

[4] Hassan M., Tuckman H., Patrick R., Kountz D., & Kohn J. *Hospital length of stay and probability of acquiring infection*. International Journal of Pharmaceutical and Healthcare Marketing (2010), 324-338.

[5] Katharine E. Henry, David N. Hager, Peter J. Pronovost and Suchi Saria. *A targeted real-time early warning score (TREWScore) for septic shock*. Science Translational Medicine (2015). DOI: 10.1126/scitranslmed.aab3719. Available at: http://stm.sciencemag.org/content/7/299/299ra122

[6] A. G. Rapsang, D. C. Shyam. *Scoring systems in the intensive care unit: A compendium*. Indian Journal of Critical Care Medicine (2014). DOI: 10.4103/0972-5229.130573. Available at: http://www.ijccm.org/article.asp?issn=0972-5229;year=2014;volume=18;issue=4;spage=220;epage=228;aulast=Rapsang

[7] European Health Parliament. *7 recommendations on Healthcare in Europe*. Available at: https://issuu.com/europeanhealthparliament/docs/ehp_papers_boek_schermversie

[8] *Comment: Health networks - delivering the future of healthcare*. Building better healthcare (2014). Available at: https://www.buildingbetterhealthcare.co.uk/technical/article_page/Comment_Health_networks__delivering_the_future_of_healthcare/94931

[9] *Informe Big Data Technologies in Healthcare: Necesidades, oportunidades y retos en el sector salud*. TicSalut (2017). Available at: Ticsalut: Big data in healthcare

[10] *Interoperabilidad: La torre de babel de los sistemas de salud*. Hackathon Nacional de Salud (2016).

[11] *What is Interoperability?*. HIMSS (2016). Available at: http://www.himss.org/library/interoperability-standards/what-is-interoperability

[12] *CRISP-DM*. CRISP-DM by Smart Vision Europe (2001). Available at: http://crisp-dm.eu/

[13] *Secondary Analysis of Electronic Health Records*. MIT Critical Data. SpringerOpen.

[14] *GitHub - MIT-LCP/mimic-code: MIMIC Code Repository: Code shared by the research community for the MIMIC-III database*. Available at: https://github.com/MIT-LCP/mimic-code

[15] T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Inc., 2003, pp. 99–137.

[16] H. Wickham. *The Journal of Statistical Software*. vol. 59, n^{o} 10, 2014.

[17] E. de Jonge and M. van der Loo. *An introduction to data cleaning with R*. Statistics Netherlands, Discussion Paper, 2013, pp. 7.

[18] Alon Halevy, Peter Norvig, and Fernando Pereira. *The Unreasonable Effectiveness of Data*. IEEE Computer Society (2009). Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf

Appendix A

Feature importances

It is exposed in this Appendix the feature importance scores for the Random Forest Classifier for both the Basic Dataset and the Rich Dataset[1]. Additional information regarding feature importance is added in Appendix B. This information contains complete feature names along with types and units of measurements for each single feature abbreviation used in the present project.

In that series of histograms, the red bar corresponds to the feature importance score value itself whereas the thin black line is the standard deviation associated. There are two sections, the first one is the resulting of apply the method to the Basic Dataset, the second one of apply it to the Rich Dataset. The cuttoffs values are defined for the LOS.

A.1 For the Basic Dataset

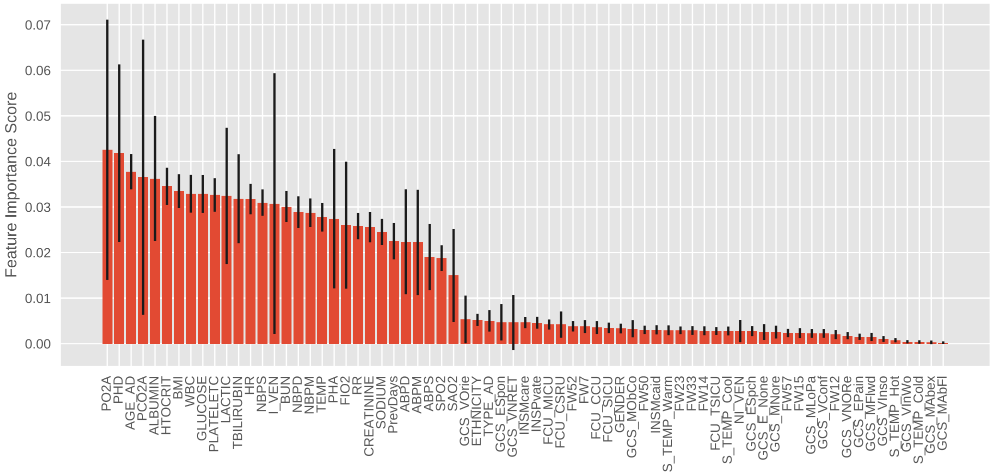


FIGURE A.1: Cuttoff at 2 days

[1] The Rich Dataset consists in the Basic Dataset features but including comorbidities and scores obtained as a derived information from MIMIC-III as detailed in Chapter 2

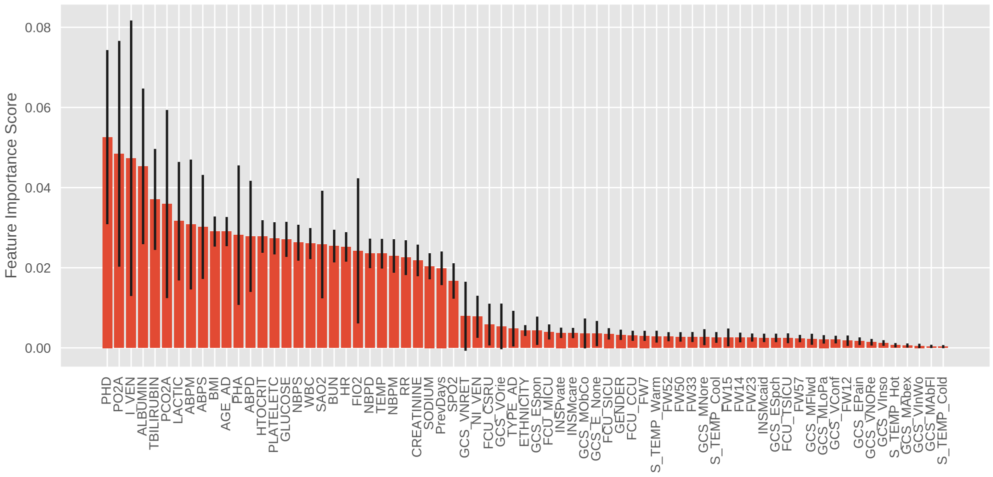


FIGURE A.2: Cuttoff at 4 days

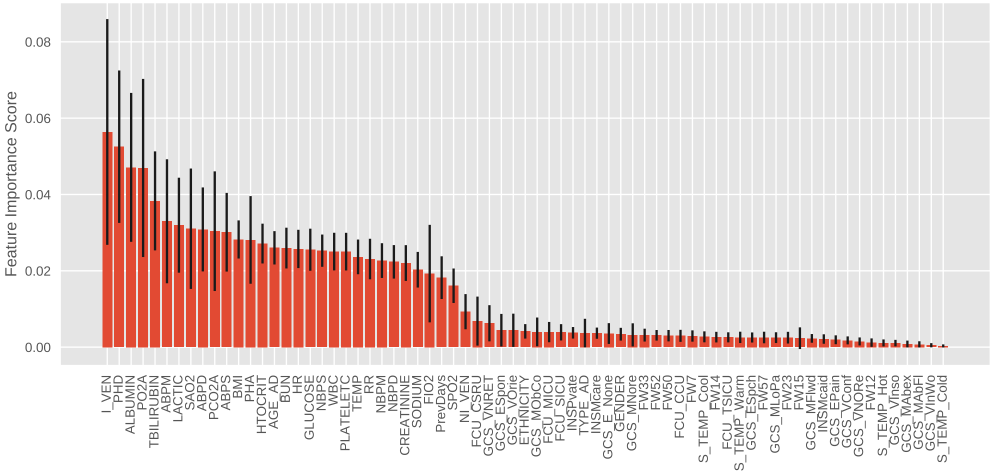


FIGURE A.3: Cuttoff at 6 days

A.2 For the Rich Dataset

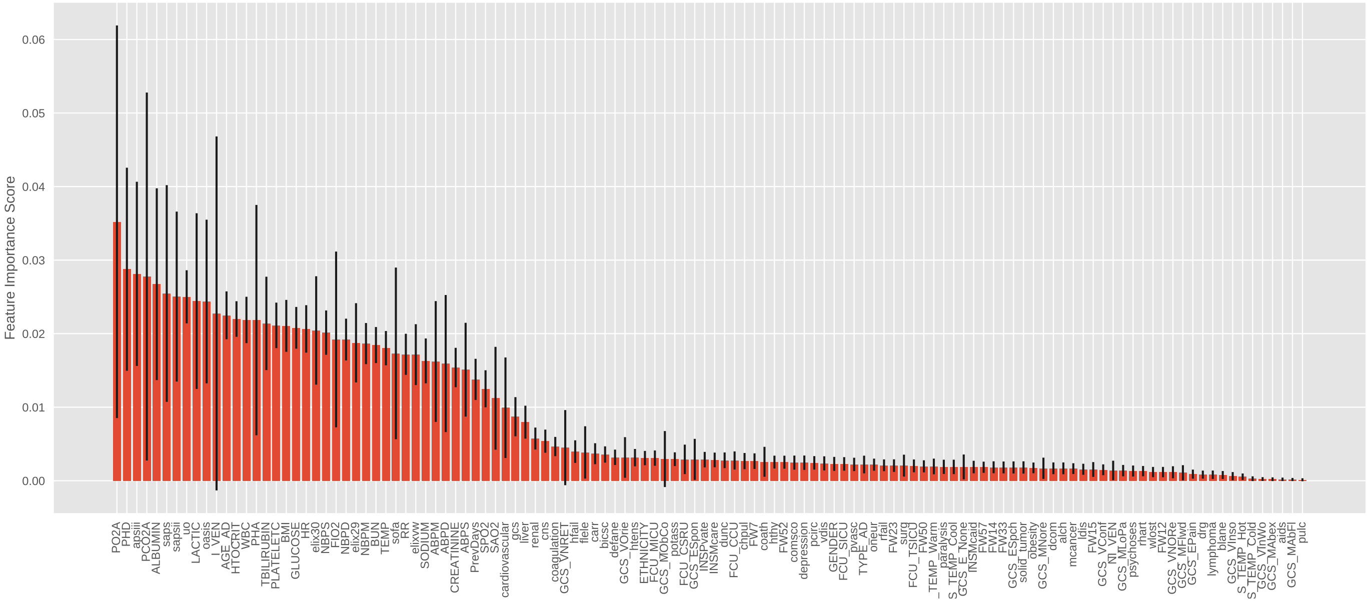


FIGURE A.4: Cuttoff at 2 days

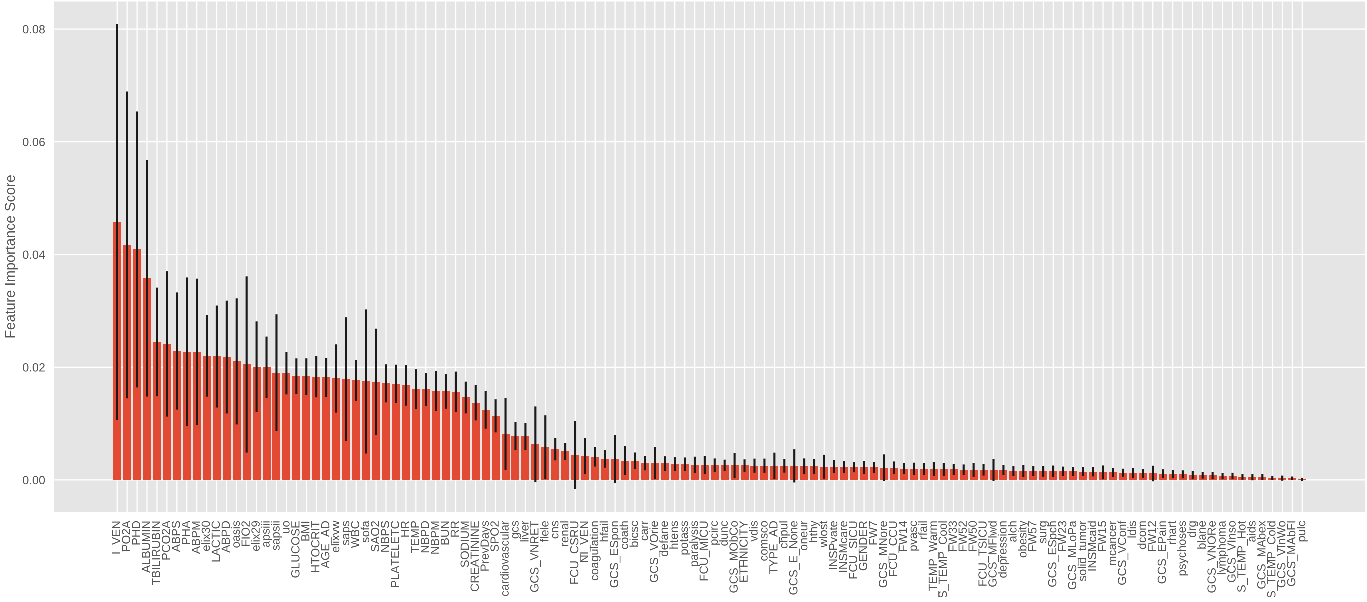


FIGURE A.5: Cuttoff at 4 days

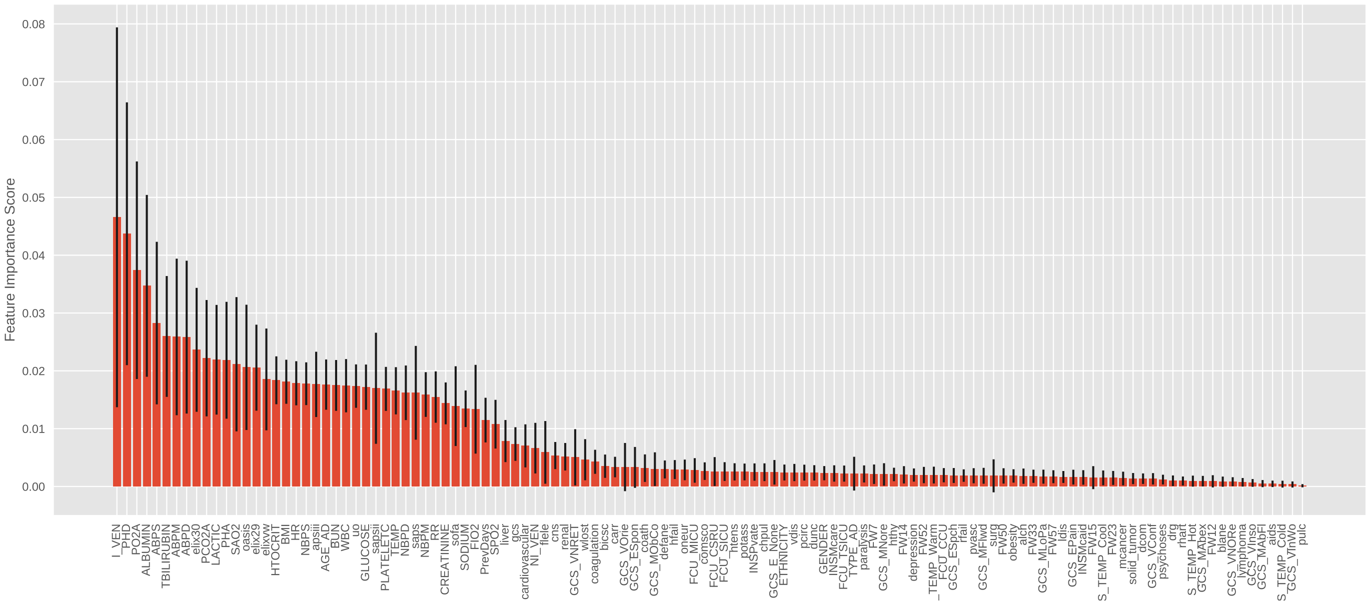


FIGURE A.6: Cuttoff at 6 days

Appendix B

Glossary

It is exposed in this Appendix the abbreviations corresponding to every feature utilized in this project. Here it can be found the abbreviation together with their complete name, feature type and units of measurement. To differentiate between the basic feature variables and the derived ones, the basic ones are presented in capital letters whereas the derived ones are presented in lower case.

B.1 Feature Abbreviations

ABPD Arterial Blood Pressure Diastolic
Numeric - mmHg

ABPM Arterial Blood Pressure Mean
Numeric - mmHg

ABPS Arterial Blood Pressure Systolic
Numeric - mmHg

AGE_AD Age
Numeric - years

ALBUMIN Albumin
Numeric - g/L

BMI Body Mass Index
Numeric - Kg/m^2

BUN Blood Urea Nitrogen
Numeric - mg/dL

CREATININE Creatinine
Numeric - mg/dL

ETHNICITY Ethnicity
Binary - None

FCU_CCU First Careunit CCU
Binary - None

FCU_CSRU First Careunit CSRU
Binary - None

FCU_MICU First Careunit MICU
Binary - None

FCU_SICU First Careunit SICU
Binary - None

FCU_TSICU First Careunit TSICU
Binary - None

FIO2 Inspired O2 Fraction
Numeric - %

FW12 First Wardid 12
Binary - None

FW14 First Wardid 14
Binary - None

FW15 First Wardid 15
Binary - None

FW23 First Wardid 23
Binary - None

FW33 First Wardid 33
Binary - None

FW50 First Wardid 50
Binary - None

FW52 First Wardid 52
Binary - None

FW57 First Wardid 57
Binary - None

FW7 First Wardid 7
Binary - None

GCS_EPain GCS Eyes To Pain
Binary - None

GCS_ESpch GCS Eyes To Speech
Binary - None

GCS_ESpon GCS Eyes Spontaneously
Binary - None

GCS_E_None GCS Eyes None
Binary - None

GCS_MAbFl GCS Motor Abnormal Flexion
Binary - None

GCS_MAbex GCS Motor Abnormal Extension
Binary - None

GCS_MFlwd GCS Motor Flex Withdraws
Binary - None

GCS_MLoPa GCS Motor Localizes Pain
Binary - None

GCS_MNore GCS Motor No Response
Binary - None

GCS_MObCo GCS Motor Obeys Commands
Binary - None

GCS_VConf GCS Verbal Confused
Binary - None

GCS_VInWo GCS Verbal Inappropriate Words
Binary - None

GCS_VInso GCS Verbal Incomprehensible Sounds
Binary - None

GCS_VNORe GCS Verbal No Response
Binary - None

GCS_VNRET GCS Verbal No Response ETT
Binary - None

GCS_VOrie GCS Verbal Oriented
Binary - None

GENDER Gender
Binary - None

GLUCOSE Glucose
Numeric - mg/dL

HR Heart Rate
Numeric - bpm

HTOCRIT Hematocrit (serum)
Numeric - %

INSMcaid Insurance Medicaid
Binary - None

INSMcare Insurance Medicare
Binary - None

INSPvate Insurance Private
Binary - None

I_VEN Invasive Ventilation
Binary - None

LACTIC Lactic Acid
Numeric - mmol/L

LOS_C Length Of Stay Categorical
Binary - None

NBPD Non Invasive Blood Pressure Diastolic
Numeric - mmHg

NBPM Non Invasive Blood Pressure Mean
Numeric - mmHg

NBPS Non Invasive Blood Pressure Systolic
Numeric - mmHg

NI_VEN Non Invasive Ventilation
Binary - None

PCO2A Arterial CO2 Pressure
Numeric - mmHg

PHA Arterial PH
Numeric - None

PHD Dipstick PH
Numeric - None

PLATELETC Platelet Count
Numeric - None

PO2A Arterial O2 Pressure
Numeric - mmHg

PREVDAYS Days Previous to ICU
Numeric - days

RR Respiratory Rate
Numeric - insp/min

SAO2 Arterial O2 Saturation
Numeric - %

SODIUM Sodium Serum
Numeric - mmol/L

SPO2 O2 Saturation Pulseoxymetry
Numeric - %

S_TEMP_Cold Skin Temperature Cold
Binary - None

S_TEMP_Cool Skin Temperature Cool
Binary - None

S_TEMP_Hot Skin Temperature Hot
Binary - None

S_TEMP_Warm Skin Temperature Warm
Binary - None

TBILIRUBIN Total Bilirubin
Numeric - mg/dL

TEMP Temperature
Numeric - °C

TYPE_AD Admission Type
Binary - None

WBC White Blood Cell
Numeric - $x1000/mm^3$

aids Acquired Immune Deficiency Syndrome
Binary - None

alch Alcohol Abuse
Binary - None

apsiii APSIII Score
Numeric - None

bicsc Bicarbonate Score
Numeric - None

blane Blood Loss Anemia
Binary - None

cardiovascular Cardiovascular
Numeric - None

carr Cardiac Arrhythmias
Binary - None

chpul Chronic Pulmonary
Binary - None

cns Central Nervous System
Numeric - None

coagulation Coagulation
Numeric - None

coath Coagulopathy
Binary - None

comsco Comorbidity Score
Numeric - None

dcom Diabetes Complicated
Binary - None

defane Deficiency Anemias
Binary - None

depression Depression
Binary - None

drg Drug Abuse
Binary - None

dunc Diabetes Uncomplicated
Binary - None

elix29 Elixhauser Sid29 Score
Numeric - None

elix30 Elixhauser Sid30 Score
Numeric - None

elixvw Elixhauser Vanwalraven Score
Numeric - None

flele Fluid Electrolyte
Binary - None

gcs GCS Score
Numeric - None

hfail Congestive Heart Failure
Binary - None

htens Hypertension
Binary - None

hthy Hypothyroidism
Binary - None

ldis Liver Disease
Binary - None

liver Liver
Numeric - None

lymphoma Lymphoma
Binary - None

mcancer Metastatic Cancer
Binary - None

oasis Oasis Score
Numeric - None

obesity Obesity
Binary - None

oneur Other Neurological
Binary - None

paralysis Paralysis
Binary - None

pcirc Pulmonary Circulation
Binary - None

potass Potassium Score
Numeric - None

psychoses Psychoses
Binary - None

pulc Peptic Ulcer
Binary - None

pvasc Peripheral Vascular
Binary - None

renal Renal Score
Numeric - None

rfail Renal Failure
Binary - None

rhart Rheumatoid Arthritis
Binary - None

saps Saps Score
Numeric - None

sapsii SapsII Score
Numeric - None

sofa Sofa Score
Numeric - None

solid_tumor Solid Tumor
Binary - None

surg Elective Surgery
Binary - None

uo Urine Output
Numeric - mL/24h

vdis Valvular Disease
Binary - None

wlost Weight Loss
Binary - None

Appendix C

Resources

All the code produced in this project is in the following Github Repository where it can be found both Data Preparation and Model Selection IPython Jupyter Notebooks:

- https://github.com/Jordi588/Final-Project-Master