Universitat Oberta de Catalunya

Comparison of machine learning algorithms in prediction of patients' survival using a health record database

Núria Jolis Orriols

Area 2 – Data analysis
Master in Bioinformatics and Biostatistics

Nuria Pérez Álvarez
(Carles Ventura Royo)

2nd June 2022

FINAL WORK CARD

Title:	Comparison of machine learning algorithms in prediction of patients' survival using a health record database
Author:	Núria Jolis Orriols
Tutor:	Nuria Pérez Álvarez
SRP:	Carles Ventura Royo
Date of delivery:	2nd June 2022
Studies:	Master in Bioinformatics and Biostatistics
Area:	Area 2 – Data analysis
Language:	English
Number of credits:	15
Keywords:	machine learning, heart failure, predictive models, R

Abstract

L'aprenentatge automàtic és una àrea emergent que crea sistemes informàtics que, mitjançant l'ús d'algoritmes i models estadístics, són capaços d'aprendre de dades existents i fer inferències sobre noves dades. El desenvolupament de models d'aprenentatge automàtic ha estat una eina per treballar amb grans bases de dades com els registres sanitaris electrònics per millorar la qualitat de l'assistència sanitària, l'eficiència, la investigació clínica i la reducció de costos.

L'objectiu principal d'aquest TFM ha estat inferir sobre la predicció de la supervivència de pacients d'un registre de salut digital mitjançant la implementació de tres algoritmes de classificació d'aprenentatge automàtic. Per fer-ho, s'ha desenvolupat un protocol bàsic per a principiants en l'aprenentatge automàtic que consta de sis passos: (1) una estudi exploratori de les dades amb anàlisi estadístic univariant i bivariant, (2) neteja i curació de les dades perquè puguin ser analitzades pels models, (3) anàlisi multivariant per conèixer la relació i interacció de les variables predictives amb la variable resposta, (4) aplicació de 3 dels models de classificació d'aprenentatge automàtic més comuns: SVM, ANN i RF, (5) validació mitjançant la tècnica de validació "k-fold cross-validation", (6) finalment una avaluació i comparació del rendiment dels models generats a partir de paràmetres com la precisió balancejada i l'AUC.

Abstract

Machine learning is an emerging area that creates computer systems that by using algorithms and statistical models are capable of learning from existing data and making inferences to new data. The development of machine learning models has been a tool for working with large databases such as electronic health records to improve healthcare quality, efficiency, clinical research and capture billing data.
The main objective of this TFM has been to infer on patients' survival prediction using an electronic health record and through the implementation of three machine learning classification algorithms. To do this, a basic protocol for beginners in machine learning has been developed which consists of six steps: (1) an exploratory analysis of the data with univariate and bivariate statistical analysis, (2) cleaning and curing of the data so that it can be analyzed, (3) multivariate analysis to know the relationship of predictive variables and their interaction with the response variable, (4) application of 3 of the most common machine learning classification models, (5) validation using k-fold cross-validation technique, (6) finally an evaluation and comparison of the generated models by means of some parameters such as balanced accuracy and AUC.

Contents

List of Figures

List of Tables

Chapter 1

Summary

The report is divided into the following chapters:

Chapter 1: Summary. Brief summary of the project.

Chapter 2: Introduction. Includes the context and rationale of the work, the objectives, the strategy carried out, and the detailed planning of the tasks.

Chapter 3: State of the art. Introduction to machine learning and supervised classification algorithms, focusing on the models used (SVM, ANN and RF); background of the methods for handling missing values; and presentation of the database, the "problem" to classify and its importance in healthcare.

Chapter 4: Methodology. Detailed information about the database and the steps followed to conduct the EDA, the data curation, the application of ML models, and the parameters used for the comparison.

Chapter 5: Results. Tables and results of the methodology described in the fourth chapter.

Chapter 6: Discussion. Discussion and reflection of the work: from the fulfillment of the objectives to the results obtained.

Chapter 7: Conclusions. Work's closure. Summary of the previous chapters, analysis of the things learned and assessment of the objectives set.

Chapter 8: Glossary. Definition of the most relevant terms and acronyms used within the report.

Chapter 9: Bibliography. A numbered list of the bibliographical references used within the report.

Chapter 2

Introduction

2.1 Context and rationale

During the last decades, there has been tremendous technology development together with an increase in the amount of available data. Therefore, new areas dedicated to the use and study of this data rose. One example of these emerging areas is Machine Learning (ML).

Machine learning can be defined as an artificial intelligence technology that tries to emulate human intelligence by learning from the surrounding environment [11]. It creates computer systems that, by using algorithms and statistical models, are capable of reviewing data, looking for patterns, inferring future behaviors through a process of training, and also improving automatically by learning from new data. These techniques are being applied successfully in diverse fields such as pattern recognition, object detection, text interpretation, computer vision, finance, entertainment, computational biology, and medical and biomedical applications [11]. Specifically, in the area of medical and biomedical applications, machine learning algorithms are being developed with great interest because they can have a big impact in predicting patients' outcome, patients' diagnosis, improving healthcare response time, etc. Therefore, they can help the health system in being more efficient, more objective, and reduce economic expenses.

Another emerging area that is helping to implement machine learning algorithms is the development of Electronic Health Records (EHR). EHRs are patients' health data systematically collected in a digital format aimed to improve healthcare quality, efficiency, clinical research and capture billing data [7]. They present both new challenges as well as opportunities because although their use is worldwide extended and more information is available, they are still not universally standardized; the databases are usually incomplete and they do not share the same collecting process, amount of data per patient, and they do not take into account a possible bias [12].

This work consists of a comparative study between some of the most commonly used classification algorithms in ML: Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Random Forest (RF). The data used is an EHR about heart failure patients admitted

to the Intensive Care Units (ICU) and the algorithms will be applied to predict the patients' outcome.
During the process, I learned about EHRs; managing, processing and curating a database with R software so it can be used for prediction analysis; programming ML algorithms and their evaluation.

2.2 Objectives

1. Learn machine learning classification techniques and models applicable to predictive analysis.

   1.1 Conduct a search in Google, Scholar Google and Pubmed with the keywords: "predictive models", "machine learning", "classification algorithms", "neural network/support vector machine/random forest".

   1.2 Describe the basic principles of three supervised classification algorithms: SVM, ANN and RF.

   1.3 Search for scientific reports in which classification models of ML have been applied to an EHR to predict a diagnosis or an outcome.

   1.4 Obtain a basic code of each of the algorithms in R.

2. Explore data clean-up and data curation methods using R software for converting an EHR into an analyzable database.

   2.1 Obtain an EHR of interest.

   2.2 Search in Google, Scholar Google, and Pubmed with the keywords: "data clean-up", "handling of missing values in R", "exploratory data analysis", "univariate analysis" and "multivariate analysis.

   2.3 Conduct an Exploratory Data Analysis (EDA) to get to know the database and the possible predictors.

   2.4 Apply data clean-up and data curation methods to obtain four statistically analyzable datasets.

3. Conduct a study to predict the outcome of a classification by comparing the performance of three classification algorithms.

   3.1 Apply the three different algorithms on the four databases.

   3.2 Evaluate the results obtained with each algorithm on each dataset.

   3.3 Validate the best models using k-fold cross-validation technique.

   3.4 Compare the results for the three methods of handling the missing values.

   3.5 Compare and determine the most optimal classification model.

2.3 Applied methodology

Two possible methodologies were identified at the beginning of the project:

1. Start by choosing a proper EHR for classification analysis and then select the three or four more suitable predictive models to apply to it.

2. Start with a learning phase of classification predictive models in ML, choose the more interesting or more common, and then select an EHR to apply them on it.

The second strategy was the selected one because I did not have a specific EHR of interest, and what I wanted to learn with this work was to apply some of the most common classification algorithms to a real database. Therefore, I considered it best to start with a learning phase of the topic instead of spending lots of time looking for a suitable database which in the end, it is not very realistic.
Secondly, I chose the models that I found more interesting or more common for classification: SVM, ANN and RF.
Third, an Exploratory Data Analysis (EDA) was conducted on the database to study the variables and their relationship with the primary response.
Fourth, to complement the work, a search for strategies to deal with missing values was also carried out. Then, four datasets were created using some methods and the classification models were applied to them.
Finally, a comparison between the algorithms' performance on the datasets was carried out to determine the best method for handling missing values and the most optimal algorithm.

2.4 Project planning

The total time to develop the project was about 14 weeks from the 16th of February to the 2nd of June. To organize the project's development, the tasks were divided into seven phases: project contents definition, work plan development, theoretical learning phase, EDA, data clean-up and curation, application and comparison of the ML models, and drafting the report.

The project contents definition was accomplished in one week and consisted in describing the specific area of the project and justifying the rationale.
The work plan development was accomplished in twelve days and consisted in describe and timing the different tasks that have to be performed and defining the limits of the project.
The theoretical learning phase was carried out for four weeks and consisted in learning about the state of the art of the ML prediction models; looking for a base code in R for the selected methods (SVM, ANN, and RF); and searching for methods to deal with the missing values.
The EDA was accomplished in about three weeks and consisted in selecting an EHR, and exploring the variables and their relationship with the primary feature.
The data clean-up and data curation was carried out in one week. Three methods to handle missing values were used to obtain statistical analyzable datasets.
The application of ML algorithms was complete in four weeks and consisted in training the models, performing predictions, evaluating and validating the results, improving the models if possible and finally comparing the results on the different datasets.
The draft of the report has been done since the start of the theoretical learning phase to gather all the information about the project in one document. Only the contents of the results, the discussion, and the conclusions were left for the last two weeks, together with time for unplanned tasks.

A Gantt graphic of the detailed tasks can be seen in Figure 2.1.

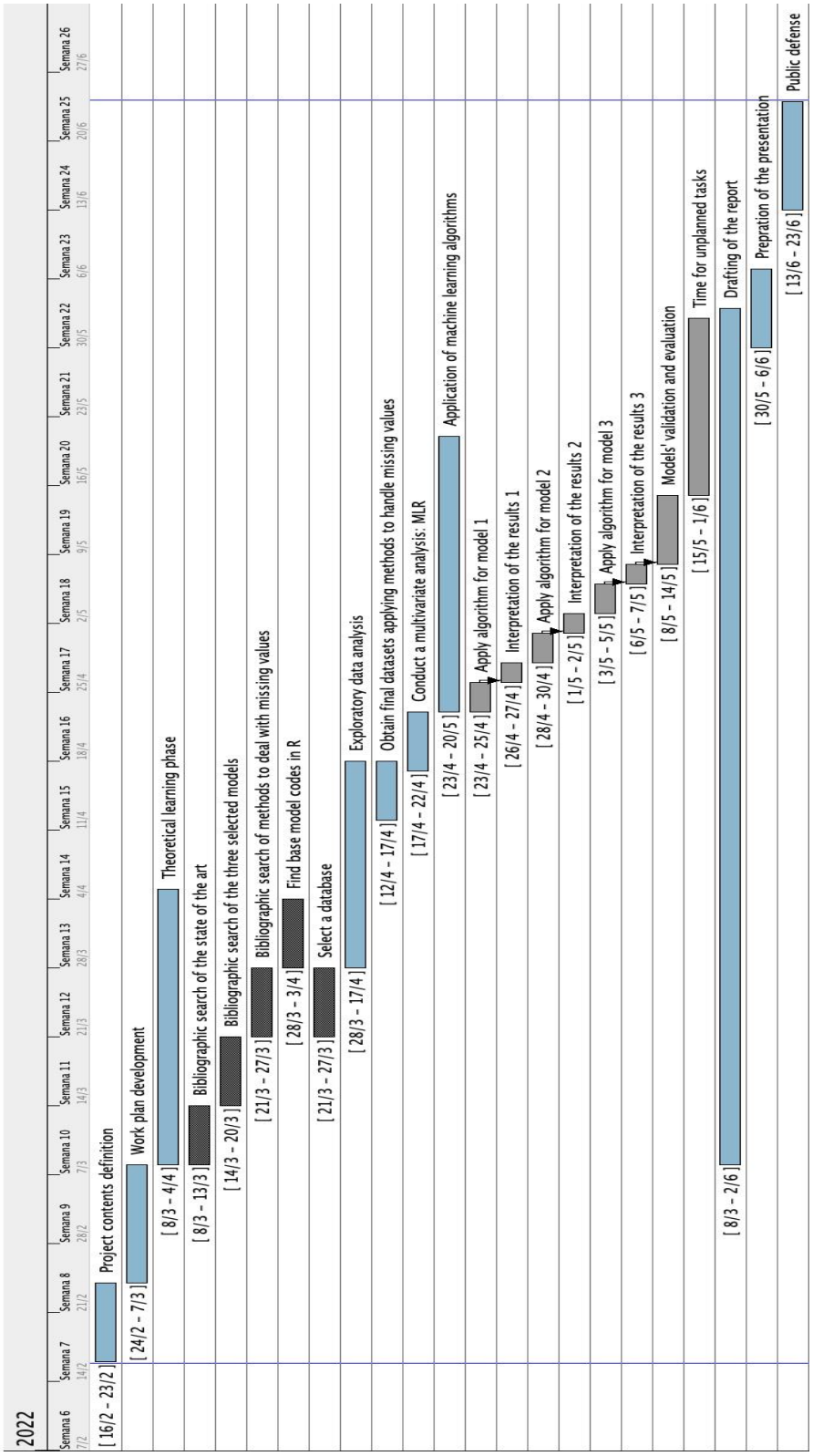

Figure 2.1: Gantt's diagram detailing the planned tasks generated using the free software "GanttProject".

Risk analysis

Initially, an assessment of the risks that could arise during the work was carried out:

- Time could be a limiting factor. Therefore, the initial project planning will consist of applying the three classification models with basic and default parameters. If there is time left during the designated time for unplanned tasks, models selection for choosing better parameters will be intended.

- Because working with EHR can be laborious, data management and cleaning may take longer than expected. If this is the case, the database could be simplified or it could be decided to change it to a simpler one, ready for analysis.

- As the project strategy is to first select the ML models that I want to learn and then applied them to an EHR, it might happen that the models chosen could not be the best for the particular database. If that is the case and there is the time during the designated two weeks for unplanned tasks, a fourth classification model could be added.

2.5 Brief summary of contributions

This work has contributed to the development of a basic protocol for machine learning beginners with specific techniques for the analysis of EHR databases and for the generation, evaluation and comparison of machine learning prediction models according to the scientific question that was wanted to address. In this case, it was the prediction of the survival of patients admitted to the ICU with heart failure.

In addition, it provides an updated state of the art and bibliography on machine learning and its algorithms, and on electronic health records.

Importantly, all code and data can be found in Github: `https://github.com/njolis/TFM_UOC.git`.

Chapter 3

State of the art

In cognitive science, learning is typically referred to as the process of gaining information through observations. ML is part of an Artificial Intelligence (AI) study field interested in the development of computer algorithms to transform data into intelligent action [23]. In other words, the machine is trained on some data, and then, the algorithms are applied so that the machine can make predictions and learn, respectively, on the given datasets [10]. It can be considered to learn if it can gather experience by doing a certain task and improve its performance in doing similar tasks in the future [23]. The formal definition presented by Tom Mitchell states that "A computer program learns from experience E for some performance measure P and some task T, if its performance on T, as measured by P, improves with experience ." [27].

The basic learning process consists of four interrelated components that try to emulate the process in which humans learn to a large extent [6]:

- Data storage or input data.
- Abstraction: translation of stored data through deriving a conceptual map or a model into broader representation concepts. Also called training.
- Generalization: creates knowledge from abstracted data and inferences that drive action in a new context so it can be used to take future decisions.
- Evaluation: a mechanism to measure the utility of learned knowledge and inform potential improvements.

According to the nature of the data labeling, machine learning can be divided into supervised (labeled), unsupervised (no labeled) and semi-supervised (partially labeled) [11]. Supervised machine learning, also called predictive machine learning, is one of the most established areas and consists of the use of training data upon which the machine builds a predictive model that can be used in test data to assign a label for each record in the test data (Figure 3.1) [6]. Supervised learning can be divided into classification when the feature is categorical, and regression, when the variable is numerical. Common classification algorithms are k-nearest

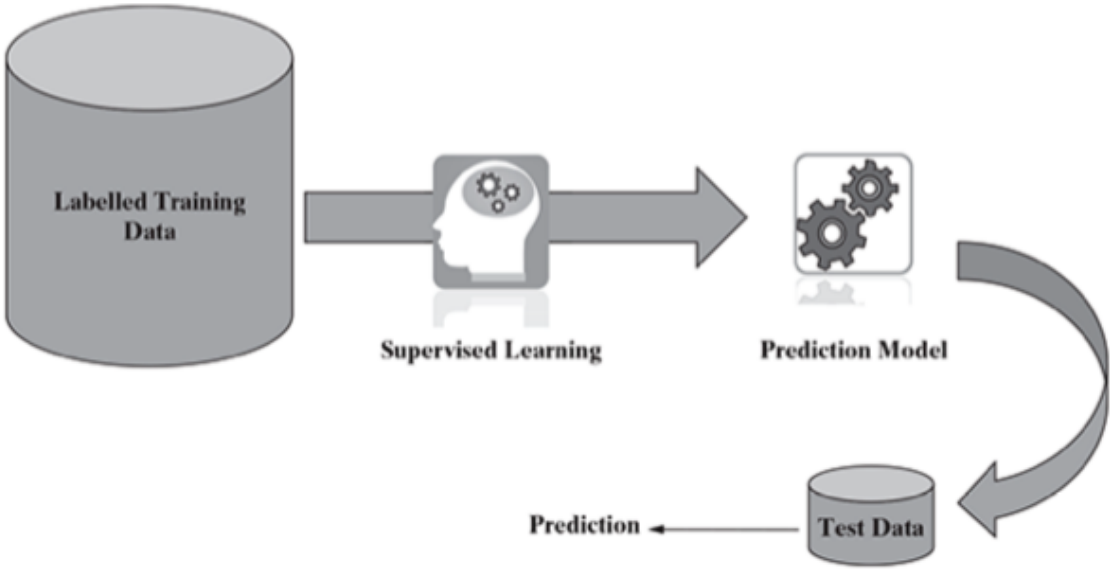

Figure 3.1: Supervised learning [6].

neighbor (KNN), Naive Bayes, decision trees, ANN, SVM, RF, etc. Popular regression algorithms are linear regression, logistical regression and polynomial regression.
Among the classification algorithms, this thesis has focused on SVM, ANN and RF because they have been found to be a good start in ML supervised classification algorithms and that they have been used in many studies related to the purpose of medical prediction or classification [1], [2], [8], [14], [26], [32], [36], [39] [40].
The application of ML in healthcare is an active area of research. Internationally, the adoption of EHRs is increasing due to strategies and agencies that incentivize their use as it provides access to a large number and variety of variables that enable high-quality classifications and predictions while ML offers the methods to handle the large volumes for high-dimensional data that are typical in healthcare settings. ML applied to EHRs can generate actionable insights from improving upon risk score systems to predicting the onset of disease, to streamlining hospital operations. As a result, its application is at the forefront of modern clinical informatics in science and medicine [3].

An example of an EHR is the Medical Information Mart for Intensive Care (MIMIC-III) database, which contains information on ICU-admitted heart failure patients. Predictors of the mortality for these patients remain poorly characterized and this project intends to shed some light on it.

Heart failure is the terminal phase of heart disease and it is the major cause of cardiovascular morbidity and mortality. Therefore is a threat to human health and social development [24].
As a life-threatening disease, heart failure patients may require immediately life-saving care

only available in ICUs and identifying those at a higher risk of poor outcomes can still be improved.

Chapter 4

Methodology

This project has been developed using Rmarkdown, Rstudio software (Version 1.4.1106) and several specific R packages.

The methodology has followed seven steps: (1) obtaining an EHR, (2) exploratory data analysis, (3) data clean-up and data curation, (4) multivariate analysis, (5) application of the supervised classification algorithms, (6) models validation and evaluation, and (7) models performance comparison.

4.1 EHR

The dataset used in the project is a publicly available subset extracted from the EHR MIMIC-III database (version 1.4, 2016) by Li et al., 2021 in a .csv format [24].

The MIMIC-III database (version 1.4, 2016) contains clinical data associated with 46.520 patients and 58.976 admissions to the ICU of the Beth Israel Deaconess Medical Center in Boston, Massachusetts (USA). The data was acquired during routine hospital care between 1 June, 2001 and 31 October, 2012 [18].

From this database, Li et al., 2021 extracted using Structured Query Language queries (SQL) with PostgreSQL (V.9.6) a subset based on demographic characteristics, comorbidities, vital signs and laboratory values of 1177 adult subjects (older than 18 years old) that were admitted to the ICU and suffered from heart failure. Heart failure was identified by manual review of the ICD-9, which is the diagnostic code description based on the ninth revision of the International Classification of Diseases developed by the World Health Organization. Patients without ICU record, data missing for left ventricular ejection fraction (LVEF) or N-terminal pro-brain natriuretic peptide (NT-proBNP) were excluded from the data. Figure 4.1 shows the flowchart of the selection of the patients.

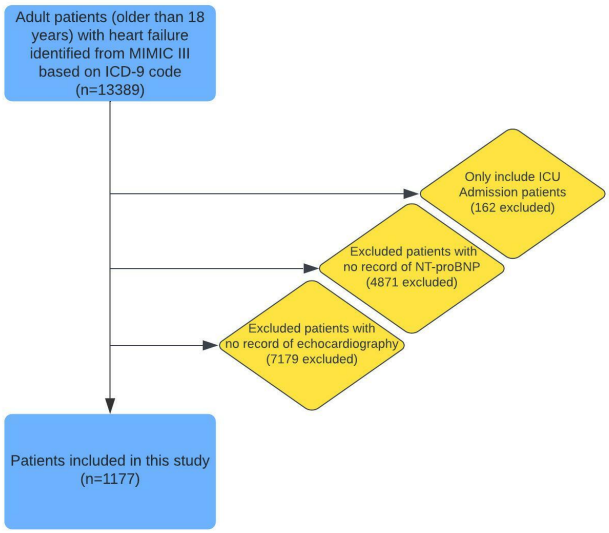

Figure 4.1: Flowchart of patients selection [24].

4.2 EDA

The dataset provided by Li et al., 2021 contains 1177 observations and 51 variables. Before approaching the EDA, a first data clean-up was performed and a final dataset of 1176 observations and 48 variables was obtained (see section 4.3 for more details). The 48 variables have been divided into five groups: primary response, demographic features, vital signs, comorbidities and laboratory works. A table of the variables group, name and a brief description can be found in Appendix A.

The primary response is the variable called outcome, in-hospital mortality defined as the vital status at the time of hospital discharge in survivors and non-survivors. It is the variable whose behavior shall be modeled and the 47 variables left are considered candidate predictors.

4.2.1 Exploratory data and bivariate analysis

For all variables, descriptive analyses of the data have been performed. Numerically, absolute and relative frequencies in addition to counting and percentage of the missing values are described for the 11 qualitative variables. Mean, median, extreme values, and absolute frequency and percentage of the missing values are described for the 37 quantitative variables. The exploratory data visualization of the categorical variables has been approached with bar plots whereas for numerical features, boxplots have been chosen to observe its distribution depending on the outcome.

The bivariate analysis consist in the study of the relationship of two variables, the primary response (outcome) and one predictor.

As far as I know, in order to assess group comparisons, Li et al., 2021 used a Wilcoxon rank-sum test for continuous variables and two-sided Pearson's chi-squared test or Fisher's exact tests for categorical variables. Therefore, they did no apply multivariate testing [24].

In this thesis, to assess differences between the outcome and other categorical variables groups' proportions, a Fisher's exact test has also been performed. For numerical variables, an assessment of the normality distribution of the variables has been done using Q-Qplots and Shapiro-Wilk tests. According to the results, none of the variables present a normal distribution and in consequence, the non-parametric Wilcoxon rank-sum test was performed when comparing two groups. The R package "statix" was used to conduct the tests. The results of Q-Qplots and Shapiro-wilk test can be found on the complementary material in the Github repository provided in section 2.5.

4.3 Data clean-up and data curation

The first part of the data clean-up consisted in removing the variables "group" and "ID" as they did not provide relevant information about patients' health. In addition, an observation that is missing for the outcome variable is also deleted. Finally, a new variable called mean arterial pressure (MAP) was created to combine systolic and diastolic blood pressure.This variable can be defined as the average arterial pressure throughout one cardiac cycle, systole, and diastole, and it can be estimated as follows:

$$MAP = DP + \frac{SP - DP}{3}$$

Where DP is the diastolic blood pressure and SP is the systolic blood pressure [9].

The data curation consists of the study and handling of missing values. Missing data can be defined as the data value that is not stored for a variable in the observation of interest. It is a is relatively common problem in almost all research and can have a significant effect on the conclusions that can be drawn from the data [19]. Moreover, it is one of the biggest challenges in building EHR-based models because many algorithms are very sensitive about it and they can bias the results and reduce their accuracy.

The choice of a statistical method to deal with missing observations depends on the type of missing variables and the assumed missing data mechanism [16], [17]. Formally there are three types of missing values:

1. Missing At Random (MAR): The probability of missing values depends on observed variables but it is not related to the specific missing values. There is a relationship, some

pattern, between the missing data and the observed values which may cause bias in further analysis.

2. Missing Completely At Random (MCAR): The probability of missing values is the same for all the observations and they are completely independent of other data. Even though this is rarely the case, when it is assumed, the statistical analysis remains unbiased.

3. Missing Not At Random (MNAR): The probability of missing values is related to their values. There is a pattern in missing data that the observed data can not explain. It may also result in bias in statistical analysis.

There are several methods for dealing with the missing observations and some of them are described below:

- Deletion: probably the most widely used method. It involves the deletion of observations or features containing missing values. It is not generally recommended, especially if the missing value is of the type MNAR because it can bias the results as it might delete some useful data from the dataset and often results in a substantial decrease in the sample size and loss of power. It may be appropriate for missing data related to the primary outcome of the study [30].

- Imputation: It consist in substitute missing values with meaningful replacements. Used in case of MAR and MNAR.

  - Single imputation: are simple approaches for handling missing data and are popular in practice, however, in most cases they are not guaranteed to provide valid inferences [16].

    * Replacing with the mean: the most common for quantitative features but it is not appropriate if there are outliers.
    * Replacing with median: recommended for quantitative features in case of outliers and skewed data distribution.
    * Replacing with mode: used for categorical features.

  - Multiple imputation: allow for uncertainty in the estimated values and can be thought of in three distinct steps: imputation, analysis and pooling of the results. Two examples of multiple imputation methods are:

    * KNN (K-nearest neighbor) imputation: a machine learning-based method that uses a Euclidean distance to find the nearest neighbor. It can predict the attributes using the most frequent value or the average among neighbors. The main disadvantage is that searches through the complete dataset, so it has limited scope when it comes to larger ones. Works well with both discrete and continuous attributes [17].
    * MICE (Multiple Imputation by chained equations): is a less biased method at the cost of being computationally expensive. It is a semi-parametric Markov

Chain Monte Carlo (MCMC) approach which assumes MAR data is replaced with a set of plausible values which contain the natural variability and uncertainty of the right values. Missing values are estimated by creating a series of regression or other suitable models [16], [21].

In this thesis dataset, there are 1901 missing values (3.4%) found in 18 out of 48 variables. Even though the percentage of missing values is a 3.4% and it could be considered negligible, ML classification models are very sensitive to missing values. Therefore, in order to be able to make the predictions and to study some of the methods mentioned above, three datasets have been generated using three different methods. Altogether, the ML classification algorithms have been applied to four different datasets:

(A) Dataset with no methods applied.

(B) Dataset with listwise deletion (complete cases).

(C) Dataset with imputation by KNN.

(D) Dataset with imputation by MICE-PMM.

The dataset A has been obtained in the first part of the data clean-up and data curation.

The dataset B has been obtained using the complete.cases() function of the R package "stats".

The dataset C has been obtained using the kNN() function of the R package "VIM". As the optimal K value (number of nearest neighbors) is usually found as the square root of N (the total number of samples), a K=34 has been set as the dataset consist in 1176 observations.

The dataset D has been generated using the R package "MICE" with mice() function and the PMM (Predictive Mean Matching) method. The MICE-PMM is one of the MICE variations suggested to impute non-normally distributed data [16]. A specific seed (12345) is fixed to be able to replicate the results. Because of the use of the seed the results are somewhat dependent on this initial choice, to reduce this effect a higher number of multiple imputations is selected by changing the default parameter m=5 to m=30.

4.4 Multivariate analysis: multiple logistic regression

Multiple or multivariable logistic regression (MLR) is a modeling method that can be used to estimate the relationship between a binary dependent variable (Y) and several independent variables (X). Its goal is to find an equation that best predicts the probability of obtaining a particular value of the Y variable as a function of the X variables using maximum likelihood estimation. The independent variables can be continuous, categorical, or ordinal.

The logistic regression method assumes that:

1. The variable outcome (Y) is a binary or dichotomous variable.
2. There is a linear relationship between the logit of the outcome and each predictor variables. The logit function is logit(p) = log(p/(1-p)), where p are the probabilities of the outcome. There are more options for transformations but the ease of calculation and the interpretation (of the logit) as the logarithm of the ODDS in favor of success, make logistics the most used.
3. There is no influential values (extreme values or outliers) in the continuous predictors.
4. There is no high intercorrelations.

Therefore, logistic regression estimates the probability of an event occurring, such as survival or not survival, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1, so it is logical to use a Bernoulli distribution. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure [28], [34], [38]. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$Y_i \ Br(p_i) \ \ i.i.d$$

$$logit(p_i) = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i$$

Where p_i represents the probability of success, X_i are the explanatory variables, β_i represents the change in the logit of the probability associated with a change in one unit in X_i, $i = 1, ..., n$ and n is the number of explanatory variables.

Relationship between ODDS and linear predictor:

$$\frac{p_i}{1-p_i} = exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i)$$

This expression defines a multiplicative model for the ODDS. So a change of a unit in X_i would mean that the ODDS would be multiplied by exp(β_i).

In this work, the multiple logistic regression has been implemented using the glm() function of the R package "stats" on the four datasets. The best model for each dataset has been determined by a stepwise procedure using the stepAIC() function of the R package "MASS". It selects models to minimize the AIC (Akaike Information Criterion),which in statistics, is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated as follows:

$$AIC = -2log(likelihood) + 2k$$

Where k is the number of model parameters. It is a weight between the likelihood of the model and a penalty for using too many independent variables. The lower the AIC, the better the model. R allows the AIC to be used as a criterion when selecting the independent variables to include or remove from a model: the goal is to minimize the AIC as much as possible.

The best-fitted model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.

For further information on AIC consult [5] and [41].

4.5 Classification using ML algorithms

In this project, three supervised classification algorithms have been modeled (SVM, ANN and RF) on the four different datasets mentioned in section 4.3.
Before generating the models, data partition has been performed having 70% of the data randomly selected for training the models and the 30% left for testing. In addition, to ensure analysis reproducibility, a seed has been defined.

4.5.1 Support Vector Machine

SVM is a supervised learning algorithm used to solve regression and classification problems. Its underlying idea is based on finding the optimal "hyperplane" that separates observations belonging to one class from another based on patterns of information about those observations called features [31]. The hyperplane drawn corresponds to a n-dimensional space in which the mean-squared error is minimized, and the margin of separation between the two classes is maximised [10].

When input data can not be linearly separated, a kernel (or non-linear) function is used to transform the support vectors to a higher-dimensional feature space. Then, a linear classifier is used for classification [31]. Some well-known kernel functions include the Radial bases function (RBF), the polyomial, the Gaussian, the sigmoid, etc. The Gaussian and Laplace RBF and Bessel kernels are general-purpose kernels used when there is no prior knowledge about the data. The linear kernel is useful when dealing with large sparse data vectors (usually in text categorization). The polynomial kernel is popular in image processing and the sigmoid kernel is mainly used as a proxy for neural networks [20].

In this case, the SVM algorithm has been implemented using the ksvm() function of the "kernlab" R package and several kernels such as linear, Gaussian radial and polynomial have been tested seeking the best performance.

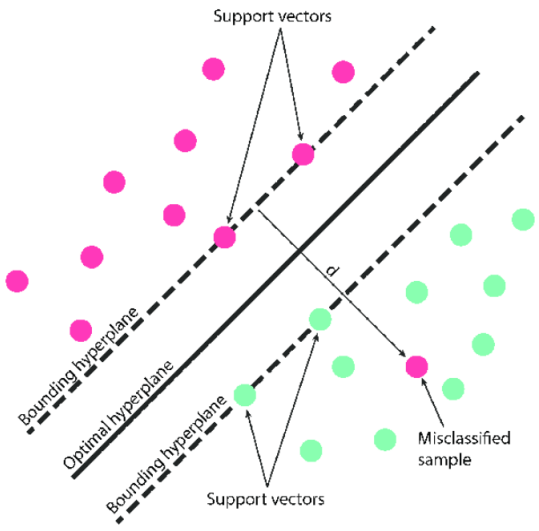

Figure 4.2: General classification hyperplane representation of SVM algorithm[4]

4.5.2 Artificial Neural Network

ANN models the relationship between a set of input signals and an output signal using a model derived from our understanding of how a biological brain responds o stimuli from sensory inputs [23]. It mimics biological learning networks. The basic processing elements are called artificial neurons or nodes. Each node has its own input, from which it receives communications from other nodes and/or from the environment and its own output, from which it communicates with other nodes or with the environment. Finally, each node has a function f through which it transforms its own global input into output [15].

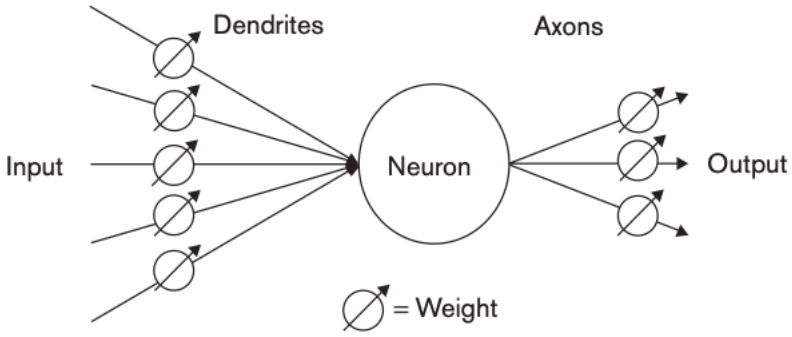

Figure 4.3: Diagram of a single processing element (PE) containing a neuron, weighted dendrites, and axons to process the input data and calculate an output [15].

The basic architecture consists of three types of neuron layers: input, hidden, and output layers. The most popular architecture is the feed-forward network, the signal flow is from input to output units, strictly in a feed-forward direction. The data processing can extend over

multiple (layers of) units, but no feed-back connections are present [15].

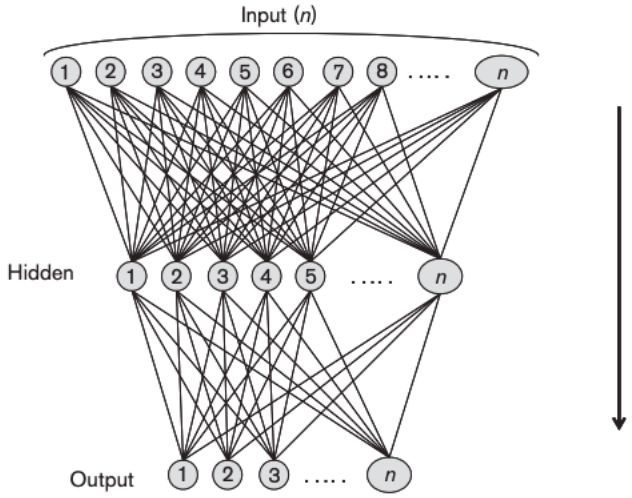

Figure 4.4: Feed-forward neural network architecture [15].

In supervised learning, an input vector is presented together with a set of desired responses, one for each node, at the output layer. The neuron impulse is computed as the weighted sum of the input signals, transformed by the transfer function. A forward pass is done, and the errors or discrepancies between the desired and actual response for each node in the output layer are found. These are then used to determine weight changes in the net according to the prevailing learning rule. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm [35].

In this work, the ANN algorithm has been implemented using the compute() function of the "neuralnet" R package. Models with one and three hidden nodes have been tested seeking the best accuracy.

4.5.3 Random Forest

The decision tree is a technique of a supervised learning algorithm that is used for classification. The algorithm groups attribute depending upon the values in order of their ascending or descending order. The decision tree consist of branches and nodes where the node represents attributes of a group that is to be classified, and the branch displays the value which a node can take [10]. The RF is an ensemble of decision trees combined to get more accurate predictions. It is a non-linear classification algorithm. It is called random because it chooses predictors randomly at a time of training, and forest because it takes the output of multiple trees to make a decision [23]. Figure 4.6 shows the main idea of the algorithm which steps are as follows:

1. Draw a ntree bootstrap sample of size n (randomly choose n samples from training data).

2. Grow a decision tree for each bootstrap sample by choosing the best split based on a random sample of mtry predictors at each node.

3. Predict new data using majority votes for classification and average for regression based on ntree trees.

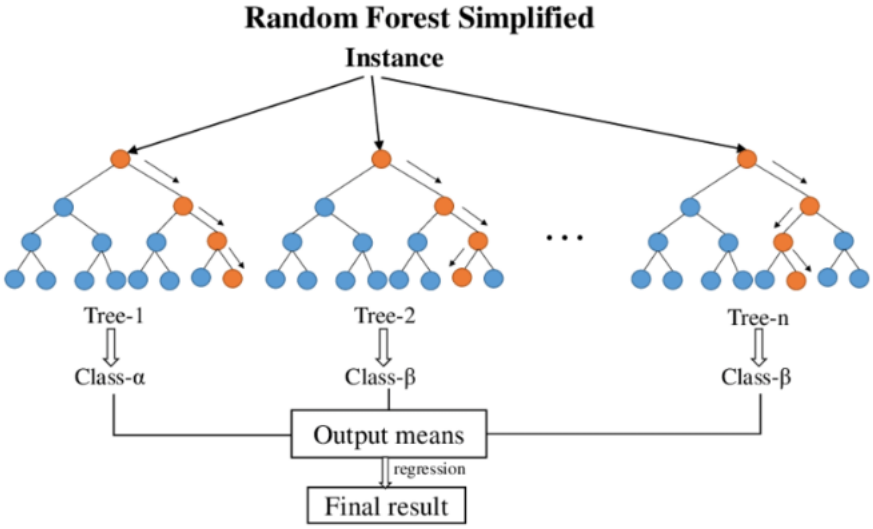

Figure 4.5: Basic idea of random forest [13].

One of its main advantages is that its avoids overfitting. In addition, it can deal with a large number of features and it helps to identify the important variables.

It mainly contains two user-friendly parameters: ntree (number of trees) and mtry (number of variables randomly chosen as candidates at each split).

In this work, RF models have been applied using the randomForest() function of the "randomForest" R package. The randomForest () function has been used with the default parameters of number of trees = 500 and mtry = sqrt(p); where p=number of variables.

4.6 k-fold cross-validation for models' validation

Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to train a model and the other used to validate the model. k-fold cross-validation is a special case of cross-validation and it is one the most common techniques for model evaluation and model selection in machine learning practice. The main idea is that the data is first randomly split into k equal-sized subsets or folds. Then, the train-then-test procedure is repeated k times: each time, one of the k subsets is used as a test set, and the rest of the k – 1 subsets are used for learning or to form the training set [33].

For learning algorithms' evaluation or validation, the model uses k-1 folds of data in each iteration for learning and subsequently models are asked to make predictions about the data in the validation fold. The performance of the algorithm in each fold is tracked using a predetermined performance metric-like accuracy. Eventually, the cross-validation performance is the compute of the arithmetic mean or other methodologies over the k performance estimates from the validation sets.

The idea is to reduce the bias by using more training data in contrast to setting aside a portion of the dataset as test data and test folds are not overlapping. In practice, k-fold cross-validation technique is more used for model selection.

Kohavi's experiments on various real-world datasets suggest that 10-fold cross-validation offers the best trade-off between bias and variance [22] .

After fitting the ML algorithms into training data and predicting the primary response classification with the test sets, the models with the best performance have been selected for being validated with the method of k-fold cross-validation. For conducting this validation technique, the "caret" R package has been used.

4.7 Parameters for models' evaluation

When performing classification predictions, there are four possible outcomes:

- True Positive (TP): predicted positive that are actually positive.
- True Negative (TN): predicted negative that are actually negative.
- False Positive (FP): predicted negative that are actually negative.
- False Negative (FN): predicted negative that are actually positive.

These four outcomes can be obtained by plotting a confusion matrix between the predictions made by the model on the test data and their actual class. The confusion matrix is performed with the function confusionMatrix () of the R package "caret". In addition to the outcomes, the confusion matrix also report several metrics used to evaluate the model [23], [29], [37]:

- Accuracy: indicates the % of the correct classified observations. TP + TN / TP + TN + FP + FN.
- Sensitivity or recall: indicates the true positive rate = TP / (TP+FN).
- Specificity: indicaes the true negative rate = TN / (FP + TN).
- Precision: indicates the predictive positive value = TP / TP + FP.

- F-Measurement: mean of sensitivity and precision = (2 x Precision x Recall) / (Precision + Recall).
- Cohen's Kappa statistic: calculated as the accuracy but normalizing at the baseline the random chance of the dataset. According to the scheme a value provided by Landis and Koch (1977) a kappa <0 indicates no agreement, 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.
- Balanced Accuracy: the average of sensitivity and specificity and its use is recommended when facing imbalanced data = (Sensitivity + Specificity) / 2.

In addition to these estimates, to analyse a model efficiency and accuracy two more parameters can be used:

- ROC (Receiver Operating Characteristics): is a plot confronting the recall (true positive rate) against the false positive rate (1-specificity) at various threshold settings.
- AUC (Area Under the Curve): is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. For this metric, a value of 0,5 indicates the classifier is not better than random guessing.

Chapter 5

Results

5.1 Exploratory data analysis

5.1.1 Exploratory data and bivariate analysis

All Wilcoxon-rank tests results can be found on the complementary material in the Github repository provided in section 2.5.

In addition, for all contrast tests, a significance level of 0,05 will be considered.

Summary tables of the variables and their descriptive statistics for the categorical variables are shown in Table 5.1 and 5.2: MIMIC-III subset of qualitative variables I and II respectively. For numerical variables results are found in Table 5.3 and 5.4: MIMIC-III subset of quantitative variables I and II respectively.

Table 5.1: MIMIC-III subset of qualitative variables I

Variable group	Variable	Categories	Descriptive statistic		NAs	
			n	%	Count	%
Primary response	outcome	Survivor	1017	86.48	0	0
		Non-survivor	159	13.52		
Demographic features	gender	Female	618	52.55	0	0
		Male	558	47.45		

Table 5.2: MIMIC-III subset of qualitative variables II

Variable group	Variable	Categories	Descriptive statistic n	Descriptive statistic %	NAs Count	NAs %
Comorbidities	atrialfibrillation	Yes	531	45.15	0	0
		No	645	54.85		
	CHD.with.no.MI	Yes	101	8.59	0	0
		No	1075	91.41		
	COPD	Yes	89	7.57	0	0
		No	1087	92.43		
	deficiencyanemias	Yes	399	33.93	0	0
		No	777	66.07		
	depression	Yes	140	11.90	0	0
		No	1036	88.10		
	diabetes	Yes	495	42.09	0	0
		No	681	57.91		
	hyperlipemia	Yes	447	38.01	0	0
		No	729	61.99		
	hypertensive	Yes	844	71.77	0	0
		No	332	28.23		
	renal.failure	Yes	429	36.42	0	0
		No	747	63.52		

Table 5.3: MIMIC-III subset of quantitative variables I

Variable	N	Mean	Min	Median	Max	NAs	%NAs
Demographic features							
BMI	962	30.19	13.35	28.31	104.97	214	18.2
age	1176	74.05	19.00	77.00	99.00	0	0.0
Vital signs							
heart.rate	1164	84.58	36.00	83.61	135.71	12	1.0
respiratory.rate	1164	20.80	11.14	20.37	40.90	12	1.0
temperature	1158	36.68	33.25	36.65	39.13	18	1.5
SP.O2	1164	96.27	75.92	96.45	100.00	12	1.0
urine.output	1141	1899.28	0.00	1675.00	8820.00	35	3.0
m.a.p	1161	79.02	51.16	77.30	129.01	15	1.3

Table 5.4: MIMIC-III subset of quantitative variables II

Variable	N	Mean	Min	Median	Max	NAs	%NAs
Laboratory							
Blood count							
hematocrit	1176	31.91	20.31	30.80	55.42	0	0.0
RBC	1176	3.57	2.03	3.49	6.58	0	0.0
MCH	1176	29.54	18.12	29.75	40.31	0	0.0
MCHC	1176	32.86	27.82	32.99	37.01	0	0.0
MCV	1176	89.90	62.60	90.00	116.71	0	0.0
RDW	1176	15.95	12.09	15.51	29.05	0	0.0
leucocyte	1176	10.72	0.10	9.68	64.75	0	0.0
platelets	1176	241.52	9.57	222.67	1028.20	0	0.0
neutrophils	1032	80.12	5.00	82.47	98.00	144	12.2
basophils	917	6.23	0.10	0.30	675.00	259	22.0
lymphocyte	1031	12.23	0.97	10.47	83.50	145	12.3
Coagulations							
PT	1156	17.49	10.10	14.64	71.27	20	1.7
INR	1156	4.07	0.87	1.30	975.00	20	1.7
Chemistry							
creatine.kinase	1011	246.94	8.00	89.50	42987.50	165	14.0
creatinine	1176	16.00	0.27	1.33	975.00	0	0.0
urea.nitrogen	1176	36.29	5.36	30.61	161.75	0	0.0
glucose	1159	148.80	66.67	136.40	414.10	17	1.4
blood.potassium	1176	4.18	3.00	4.11	6.57	0	0.0
blood.sodium	1176	138.90	114.67	139.25	154.74	0	0.0
blood.calcium	1175	8.50	6.70	8.50	10.95	1	0.1
chloride	1176	102.29	80.27	102.52	122.53	0	0.0
anion.gap	1176	13.92	6.64	13.67	25.50	0	0.0
magnesium.ion	1176	2.12	1.40	2.09	4.07	0	0.0
Venous blood							
pH	885	7.38	7.09	7.38	7.58	291	24.7
bicarbonate	1176	26.91	12.86	26.50	47.67	0	0.0
lactic.acid	948	8.36	0.50	1.62	975.00	228	19.4
PCO2	883	45.54	18.75	43.00	98.60	293	24.9
Heart specific							
EF	1176	48.71	15.00	55.00	75.00	0	0.0
NT.proBNP	1176	11011.04	50.00	5837.75	118928.00	0	0.0

Primary response

As mentioned, the primary outcome of the study is in-hospital mortality, defined as the vital status at the time of hospital discharge in survivors and non-survivors. The barplot in Figure

5.1 shows that there is an 86,41% (1017) of survivors and a 13,51% (159) of non-survivors. Therefore, the data is not balanced.

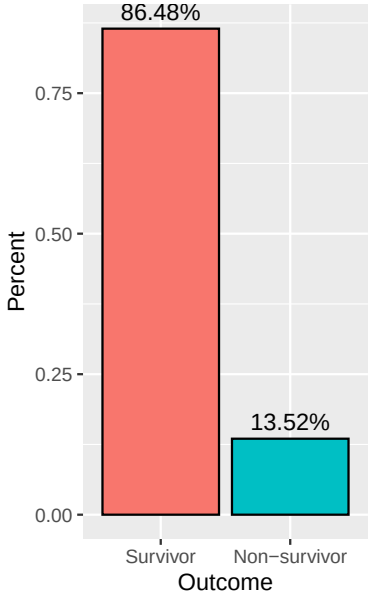

Figure 5.1: Percentage of Patients' outcome.

Demographic features

- Age and gender: The age of the patients in this study ranges from 19 to 99 (Table 5.3) and there are 52,55% of females and 47,45% of males (Table 5.1). Figure 5.2 a) shows that among survivors, the incidence is higher between 60 and 90 years old, and in non-survivors, between 80 and 90 years old; independently of the gender. Wilcoxon test for the age results in a p-value of 0,017 indicating that there are significant differences between age and outcome. Fisher's exact test for gender presents an odds ratio of 0,88, close to 1, and a p-value of 0,44 meaning that there is no significant evidence that gender and outcome are different (Table 5.5).

- BMI: Patients' BMI ranges from 13,35 to 104,95 and presents an 18,3% of missing values (Table 5.3). Although the median of the two groups looks similar in Figure 5.2 b), the p-value of the Wilcoxon test is 0.017 meaning that there are significant differences between BMI and outcome.

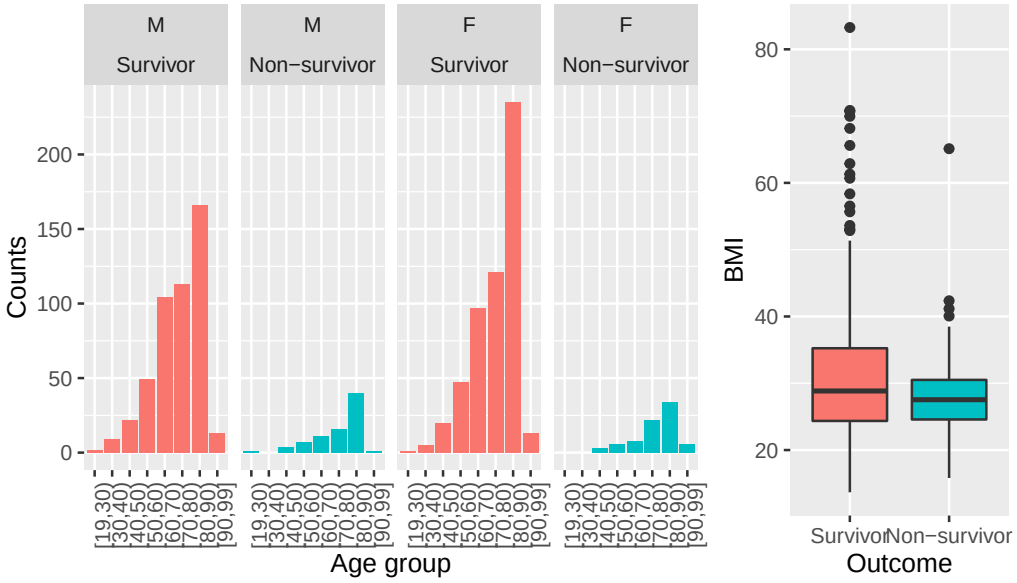

Figure 5.2: From left to right: a) Barplot of age groups by gender and outcome. b) BMI's distribution by outcome.

Table 5.5: Gender's Fisher's exact test results

Categorical Class	outcome	Categorical Class	n	TOTAL	Freq	OR	pvalor
gender	Survivor	M	478	1017	47.00	0.88	0.44
		F	539		53.00		
	Non-survivor	M	80	159	50.31		
		F	79		49.69		

Vital signs

The group of vital singles comprises six variables: MAP, heart rate, respiratory rate, oxygen saturation, temperature and urine output.

The boxplots in Figure 5.3 indicate that the group of non-survivors presents elevated heart and respiratory rates, lower MAP and urine output, and similar temperature and oxygen saturation compared to survivors.
All p-values resulting from the Wilcoxon-rank test are <0,05; therefore, there is statistical evidence to affirm that there are differences between the outcome and all vital signs.

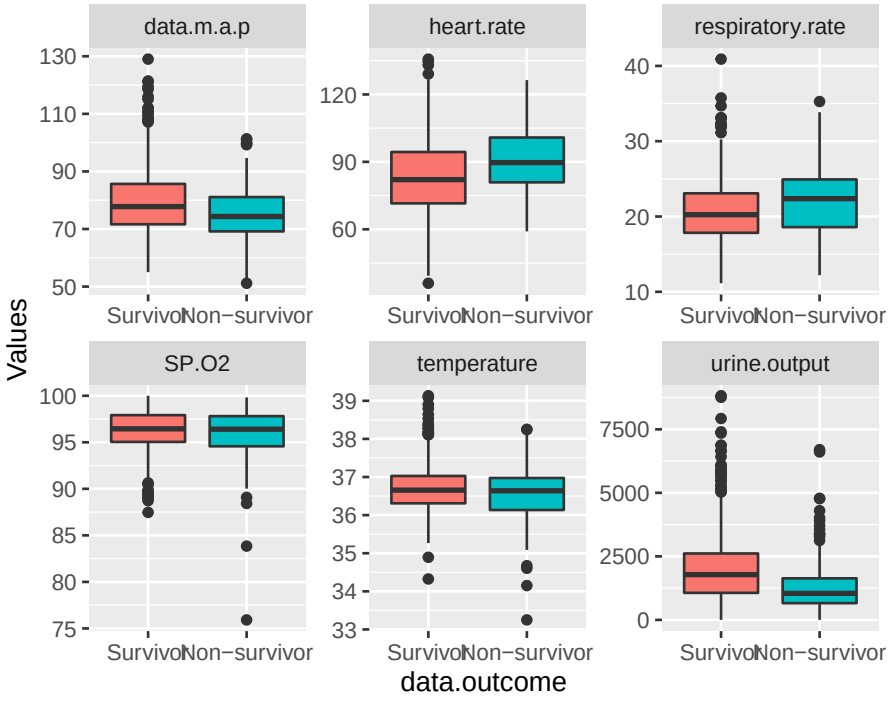

Figure 5.3: Vital signs' distribution by outcome.

Comorbidities

The group of comorbidities comprises nine attributes: atrial fibrillation, coronary heart disease with no myocardial infarction, chronic obstructive pulmonary disease, depression, diabetes mellitus, hyperlipidemia, hypertension, hypoferric anemia and renal failure.

Barplot in Figure 5.4 shows that the most prevalent comorbidity is hypertension which is found in almost 75% of the patients followed by atrial fibrillation, diabetes, hyperlipemia, renal failure and deficiency anemia, which are present between 30% and 50% of the patients. Finally, coronary heart disease with no myocardial infarction, chronic obstructive pulmonary disease and depression are suffered by less than 10% of the patients.

Bar plots in Figure 5.5 display that in general, the presence of comorbidities is smaller among non-survivors than survivors except for atrial fibrillation, which is more prevalent in non-survivors than survivors. In the case of coronary heart disease with no myocardial infarction, chronic obstructive pulmonary, diabetes and depression, there seems to be very little difference between the incidence among survivors and non-survivors. These results are mainly confirmed by Fisher's test results (Table 5.6), in which the p-value of the variables coronary heart disease with no myocardial infarction, chronic obstructive pulmonary and diabetes resulted in >0,05; therefore, there is significant evidence that these features are no different compared to the primary response. The p-value of diabetes is 0,04, and although it is statistically significant, it is very close to the limit. The p-values of the rest of the comorbidities are <0,05.

Patients with atrial fibrillation have 1,81 times more likely to die whereas patients with hypertension, deficiency anemia, depression, hyperlipemia and renal failure have 1,56, 1,96, 1,39 and 2,08 times more likely to survive (Table 5.6).

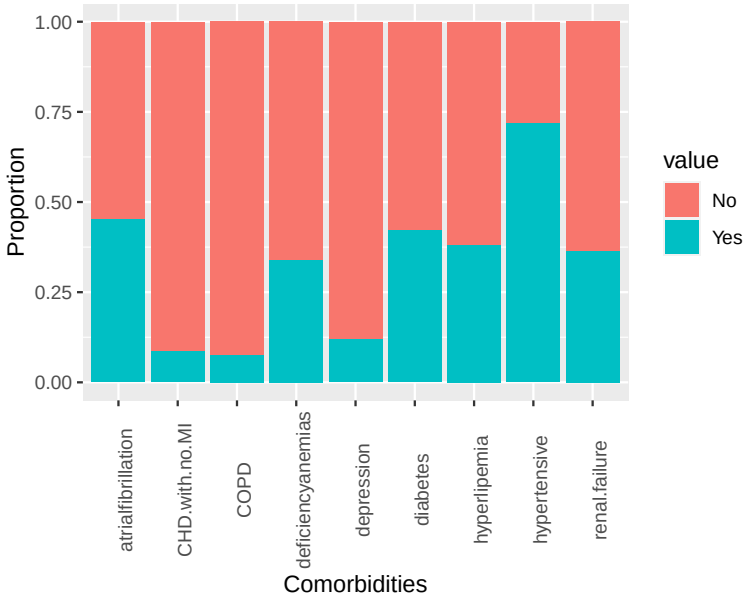

Figure 5.4: Percentage of the presence of the comorbidities.

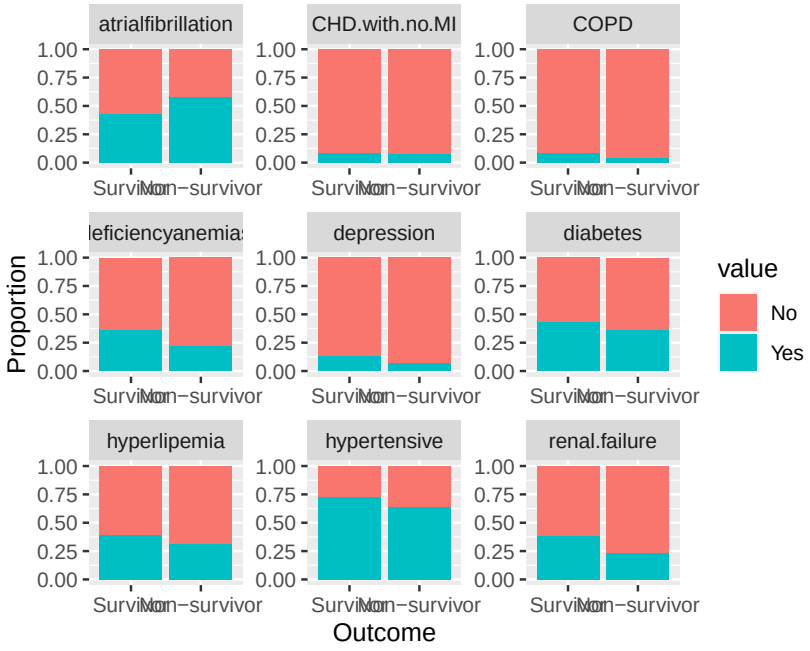

Figure 5.5: Comorbidities' proportion by outcome.

Table 5.6: Comorbidities' Fisher's exact test results.

Comorbidity	outcome	Class	n	TOTAL	Freq	OR	pvalor
hypertensive	Survivor	No	274	1017	26.94	0.64	0.02
		Yes	743		73.06		
	Non-survivor	No	58	159	36.48		
		Yes	101		63.52		
atrialfibrillation	Survivor	No	578	1017	56.83	1.81	0
		Yes	439		43.17		
	Non-survivor	No	67	159	42.14		
		Yes	92		57.86		
CHD.with.no.MI	Survivor	No	928	1017	91.25	0.85	0.76
		Yes	89		8.75		
	Non-survivor	No	147	159	92.45		
		Yes	12		7.55		
diabetes	Survivor	No	579	1017	56.93	0.74	0.1
		Yes	438		43.07		
	Non-survivor	No	102	159	64.15		
		Yes	57		35.85		
deficiencyanemias	Survivor	No	653	1017	64.21	0.51	0
		Yes	364		35.79		
	Non-survivor	No	124	159	77.99		
		Yes	35		22.01		
depression	Survivor	No	888	1017	87.32	0.51	0.04
		Yes	129		12.68		
	Non-survivor	No	148	159	93.08		
		Yes	11		6.92		
hyperlipemia	Survivor	No	620	1017	60.96	0.72	0.08
		Yes	397		39.04		
	Non-survivor	No	109	159	68.55		
		Yes	50		31.45		
renal.failure	Survivor	No	625	1017	61.46	0.48	0
		Yes	392		38.54		
	Non-survivor	No	122	159	76.73		
		Yes	37		23.27		
COPD	Survivor	No	935	1017	91.94	0.53	0.14
		Yes	82		8.06		
	Non-survivor	No	152	159	95.60		
		Yes	7		4.40		

Laboratory works

The group of laboratory works comprises 29 of the 48 variables on the dataset. As shown in Table A.3 (Annex A), they have been distributed in five groups: blood cell count, coagulation factors, chemistry (substances in blood related to its chemical balance), venous blood measurements and heart specific indicators.

Although looking at the boxplots in Figures 5.6, 5.7, 5.8, 5.9 and 5.10 it seems that the variables MHC, MCHC, MCV, RBC, sodium, chloride, creatine kinase, creatinine, glucose, magnesium and EF have similar medians, all p-values of the Wilcoxon-rank test resulted in <0,05. There are many outliers in many variables, and because of that, results are better concluded with the Wilcoxon test. Therefore, there is statistical evidence to affirm that there are differences between the outcome and all laboratory work variables.

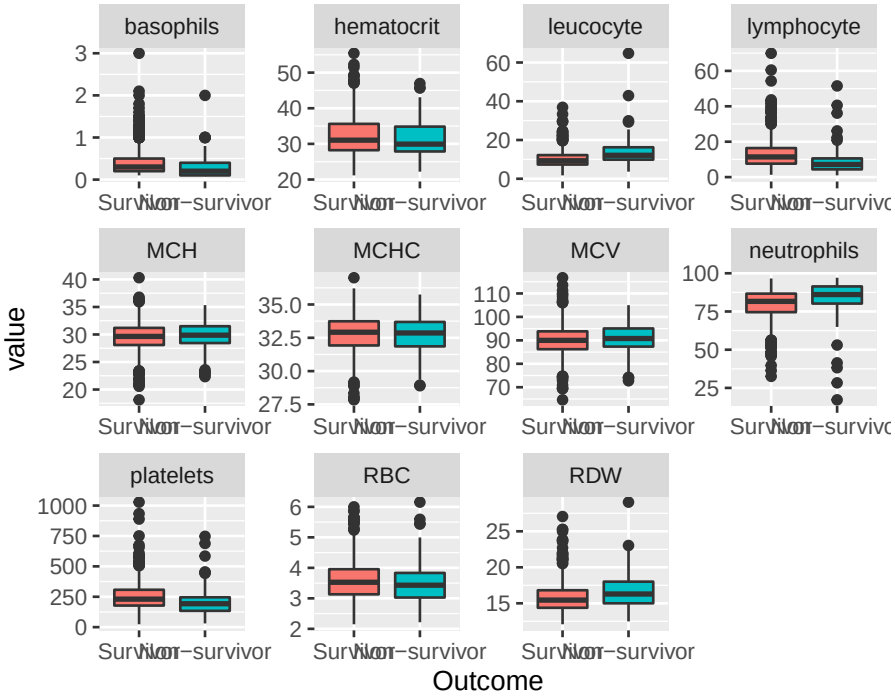

Figure 5.6: Cell count factors by outcome.

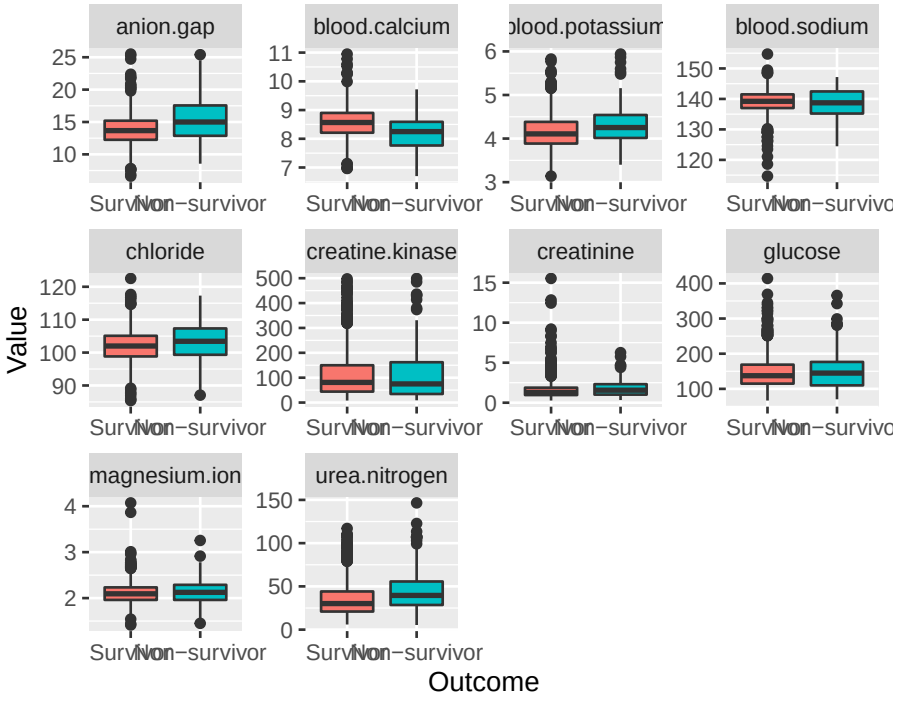

Figure 5.7: Blood chemical substances by outcome.

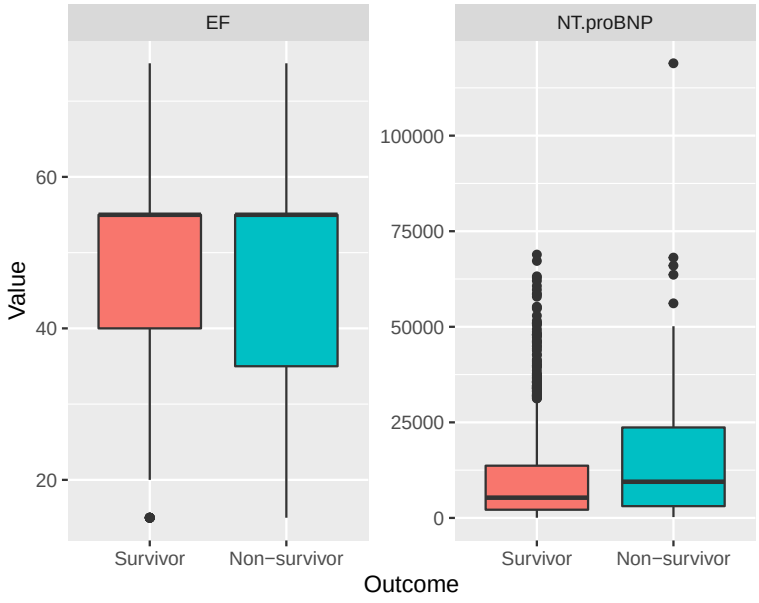

Figure 5.8: Heart specific factors by outcome.

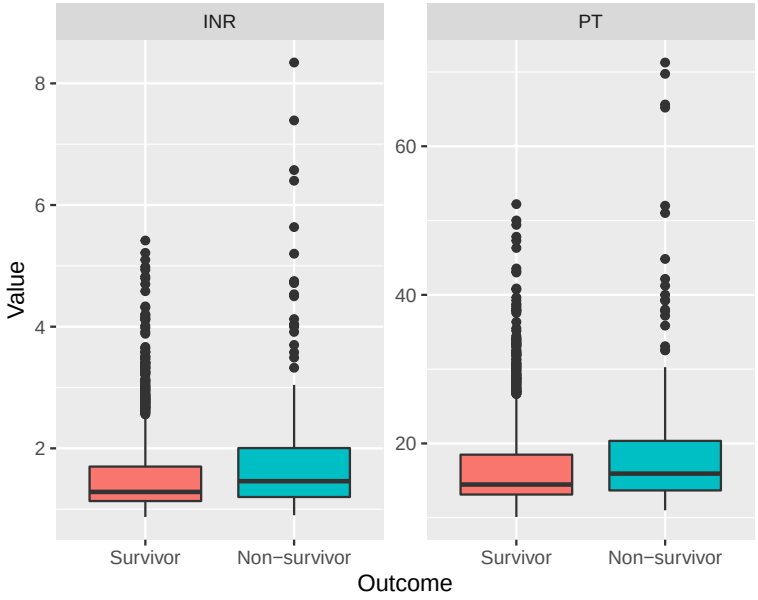

Figure 5.9: Coagulation factors by outcome.

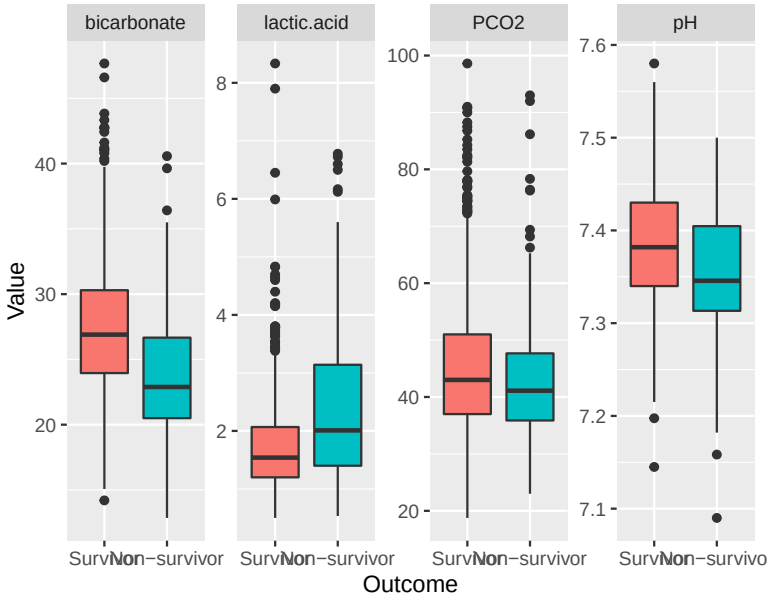

Figure 5.10: Venous blood factors by outcome.

5.2 Data clean-up and data accuracy

As mentioned in the section of methodology, the first part of the data clean-up and data accuracy was to make some modifications in the dataset provided by Li et al., 2021 to finally obtain

a dataset of 1176 observations and 48 variables.

After that, an analysis of the missing values have been conducted and the results show that in this thesi's dataset, there are 1901 missing values (a 3.4%) found in 18 out of 48 variables: BMI, heart.rate, respiratory.rate, SP.O2, temperature, urine.output, basophils, lymphocytes, neutrophils, INR, PT, blood.calcium, creatine.kinase, glucose, lactic.acid, PCO2 and pH.

Of those 18 features, the variables PCO2, pH, basophils, lactic.acid, BMI, creatine.kinase, lymphocytes and neutrophils, have between 10 and 25% of missing values whereas the other 10 features presents less than the 10%.

According to Figure 5.11, some of the variables seem to have a pattern of missing data. Even though the percentage of missing values is very low and it could be considered negligible, it is assumed that we are dealing with MAR or MNAR.

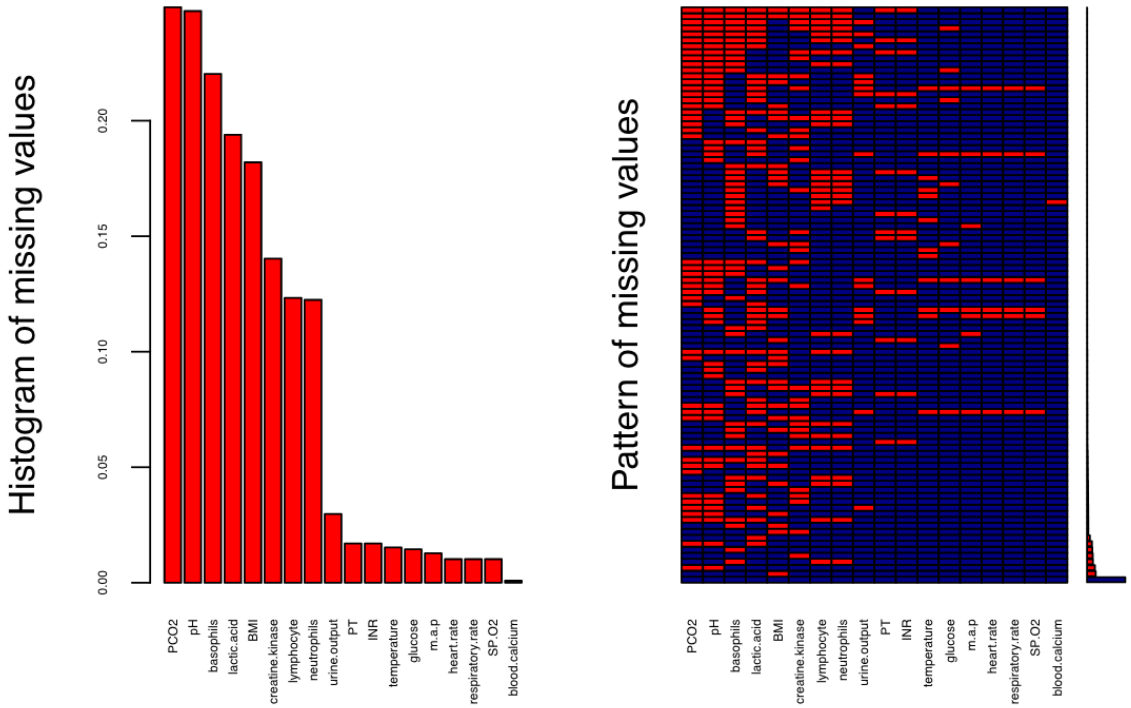

Figure 5.11: Pattern of missing values

According to section 4.3, in orderd to be able to make predictions and to study some of the methods described to deal with missing values, four different datasets have been created. Table 5.7 contains a summary of their properties. The multivariate analysis and the prediction classification with ML models have been applied on all of them.

Table 5.7: Properties of input datasets.

Dataset	Methodology applied	No. of observations	No. of features	% NA
A	-	1176	48	3.37
B	Listwise deletion	428	48	0
C	KNN imputation	1176	48	0
D	MICE-PMM	1176	48	0

5.3 Multivariate analysis: multiple logistic regression

As mentioned, multiple logistic regression has been first performed on the four different datasets with the full model (with all the covariates) and then, a stepwise procedure based on the AIC has been applied to obtain the best model with the variables that contribute the most to explain the primary response. Only the best models determined by the "stepAIC()" R function with the lowest AIC are shown in this section. However, all models can be found on the Github repository mentioned in section 2.5.

The full model is fitted as follows:

outcome $\sim age + gender + BMI + hypertensive + atrial fibrillation + CHD.with.no.MI + diabetes + deficiency anemias + depression + hyperlipemia + renal.failure + COPD + heart.rate + respiratory.rate + temperature + SP.O2 + urine.output + hematocrit + RBC + MCH + MCHC + MCV + RDW + leucocyte + platelets + neutrophils + basophils + lymphocyte + PT + INR + NT.proBNP + creatine.kinase + creatinine + urea.nitrogen + glucose + blood.potassium + blood.sodium + blood.calcium + chloride + anion.gap + magnesium.ion + pH + bicarbonate + lactic.acid + PCO2 + EF + m.a.p.$

The model fitting with the dataset A has only been accomplished with the full model. It is not possible to reduce the model with the "stepAIC()" R function because this dataset contains missing values. The model fitting must apply the models to the same dataset and dimensions and with the missing values the dimensions of the dataset change.
The best MLR model obtained based on the AIC criterion by the R function mentioned above on the dataset B is a model fitted with 20 variables. Table 5.8 shows that the variables age, heart.rate, leucocyte, PT, urea.nitrogen, glucose, chloride, anion.gap, magnesium.ion and PCO2 present a negative coefficient so the probability of surviving is reduced when these variables increase when taking into account all the features included in the model.
The variables BMI, artrialfibrillation (Yes), deficiencyanemias(Yes), renal.failure(Yes), COPD (Yes), temperature, platelets, neutrophils, lymphocyte and blood.calcium present a positive coefficient. Taking into account all variables included in the model, the probability of surviving increases when these variables also increase.

Therefore, this model denotes that the risk factors for non-surviving are being old, presenting

an elevated heart rate, high count of leucocytes, elevated PT, urea nitrogen, glucose, chloride, anion gap, magnesium and PCO2.

Taking into account all variables included in the model, the variables with significant p-value (<0,05) are age, deficiencyanemias, renal.failure, platelets, PT, blood.calcium, anion.gap and PCO2; which means that there is statistical evidence to affirm that they are different from the primary response.

Table 5.8: MLR of dataset B after applying stepwise.

Reduced MLR model of dataset B						
Formula:	outcome ∼age + BMI + atrialfibrillation + deficiencyanemias + renal.failure + COPD + heart.rate + temperature + leucocyte + platelets + neutrophils + lymphocyte + PT + urea.nitrogen + glucose + blood.calcium + chloride + anion.gap + magnesium.ion + PCO2					
AIC:	239.6					
	Estimate	**Odds**	**Std. Error**	**z value**	**Pr(> |z|)**	
(Intercept)	-5.831	0.003	15.437	-0.378	0.706	
age	-0.053	0.948	0.02	-2.667	0.008	**
BMI	0.039	1.04	0.027	1.471	0.141	
atrialfibrillationYes	0.699	2.012	0.461	1.515	0.13	
deficiencyanemiasYes	1.405	4.075	0.518	2.713	0.007	**
renal.failureYes	2.419	11.238	0.545	4.438	0.000	***
COPDYes	1.385	3.996	0.762	1.818	0.069	
heart.rate	-0.024	0.976	0.013	-1.919	0.055	
temperature	0.483	1.621	0.314	1.539	0.124	
leucocyte	-0.072	0.931	0.046	-1.563	0.118	
platelets	0.007	1.007	0.002	3.004	0.003	**
neutrophils	0.069	1.072	0.048	1.428	0.153	
lymphocyte	0.123	1.131	0.068	1.808	0.071	
PT	-0.067	0.936	0.029	-2.331	0.020	*
urea.nitrogen	-0.019	0.981	0.011	-1.67	0.095	
glucose	-0.005	0.995	0.003	-1.554	0.120	
blood.calcium	0.841	2.318	0.366	2.297	0.022	*
chloride	-0.078	0.925	0.043	-1.801	0.072	
anion.gap	-0.346	0.707	0.107	-3.238	0.001	**
magnesium.ion	-1.205	0.3	0.758	-1.589	0.112	
PCO2	-0.064	0.938	0.022	-2.892	0.004	**

The final model conducting MLR with the stepAIC() R function on dataset C presents 18 variables. Table 5.9 shows the results of this model. In this case, the variables heart.rate, respiratory.rate, RDW, leucocyte, PT, urea.nitrogen and blood.potassium present a negative

coefficient so when taking into account all the features included in the model, the probability of surviving is reduced with an increase of these variables. The variables BMI, deficiencyanemias(Yes), renal.failure(Yes), COPD (Yes), temperature, SP.O2, urine.outpu, platelets, lymphocyte, creatinine and blood.calcium present a positive coefficient. Taking into account all variables included in the model, the probability of surviving increases when these variables also increase.

Regarding all variables included in this model, the risk factors for non-surviving are presenting an elevated heart and respiratory rate, elevated blood count of leucocytes and RDW, elevated PT, urea nitrogen and potassium.

The variables with significant p-value (<0,05) are BMI, deficiencyanemias, renal.failure, COPD, heart.rate, SP.O2, urine.output, RDW, leucocyte, platelets, lymphocyte, urea.nitrogen, blood.potassium and blood.calcium; which means that there is statistical evidence to affirm that they are different from the primary response considering all variables included in the model.

Table 5.9: MLR of dataset C after applying stepwise.

Reduced MLR model of dataset C						
Formula:	outcome ~BMI + deficiencyanemias + renal.failure + COPD + heart.rate + respiratory.rate + temperature + SP.O2 + urine.output + RDW + leucocyte + platelets + lymphocyte + PT + creatinine + urea.nitrogen + blood.potassium + blood.calcium					
AIC:	719.43					
	Estimate	**Odds**	**Std. Error**	**z value**	**Pr(> |z|)**	
(Intercept)	-19.498	0.000	7.976	-2.445	0.015	*
BMI	0.031	1.032	0.015	2.059	0.039	*
deficiencyanemiasYes	0.764	2.147	0.238	3.214	0.001	**
renal.failureYes	1.192	3.295	0.257	4.635	0.000	***
COPDYes	1.060	2.887	0.446	2.378	0.017	*
heart.rate	-0.019	0.981	0.007	-2.777	0.005	**
respiratory.rate	-0.041	0.959	0.027	-1.547	0.122	
temperature	0.288	1.333	0.169	1.700	0.089	
SP.O2	0.116	1.124	0.046	2.556	0.011	*
urine.output	0.000	1.000	0.000	2.564	0.010	*
RDW	-0.113	0.893	0.046	-2.433	0.015	*
leucocyte	-0.057	0.945	0.018	-3.080	0.002	**
platelets	0.003	1.003	0.001	3.490	0.000	***
lymphocyte	0.031	1.032	0.016	1.990	0.047	*
PT	-0.023	0.978	0.012	-1.877	0.060	
creatinine	0.002	1.002	0.002	1.273	0.203	
urea.nitrogen	-0.024	0.976	0.005	-4.717	0.000	***
blood.potassium	-0.720	0.487	0.243	-2.967	0.003	**
blood.calcium	0.709	2.032	0.187	3.797	0.000	***

The final MLR model on dataset D has been accomplished fitting 25 variables. Table 5.10 shows that the variables the variables age, MCH, RDW, leucocyte, blood.potassium, blood.sodium, heart.rate, respiratory.rate, lactic.acid, PCO2 and m.a.p present negative coefficients which means that the probability of surviving is reduced when they increase considering all features included in the model. The variables deficiencyanemias(Yes), renal.failure(Yes), COPD (Yes), MCHC, platelets, chloride, bicarbonate, BMI, temperature, SP.O2, urine.output, blood.calcium and m.a.p present a positive coefficient.

With this model, the risk factors for non-surviving are being old, presenting an elevated heart and respiratory rate, elevated blood count of MCH, RDW and leucocyte, and having elevated urea.nitrogen, blood.potassium, blood.sodium, magnesium.ion, lactic.acid and PCO2.

The variables with significant p-value (<0,05) are deficiencyanemias, renal.failure, COPD,

heart.rate, SP.O2, RDW, leucocyte, platelets, lymphocyte, urea.nitrogen, chloride, bicarbonate, blood.calcium, acid.lactic and PCO2; which means that there is statistical evidence to affirm that they are different from the primary response considering the formula of the model.

Table 5.10: MLR of dataset D after applying stepwise.

Reduced MLR model of dataset D						
Formula:	outcome ∼age + deficiencyanemias + renal.failure + COPD + MCH + MCHC + RDW + leucocyte + platelets + urea.nitrogen + blood.potassium + blood.sodium + chloride + magnesium.ion + bicarbonate + BMI + heart.rate + respiratory.rate + temperature + SP.O2 + urine.output + blood.calcium + lactic.acid + PCO2 +m.a.p					
AIC:	712.58					
	Estimate	**Odds**	**Std. Error**	**z value**	**Pr(> |z|)**	
(Intercept)	-21.281	0.000	9.919	-2.146	0.032	*
age	-0.017	0.983	0.009	-1.850	0.064	
deficiencyanemiasYes	0.753	2.123	0.245	3.076	0.002	**
renal.failureYes	1.325	3.762	0.265	4.994	0.000	***
COPDYes	1.144	3.138	0.475	2.407	0.016	*
MCH	-0.081	0.923	0.049	-1.649	0.099	
MCHC	0.198	1.219	0.106	1.878	0.060	
RDW	-0.108	0.898	0.052	-2.093	0.036	*
leucocyte	-0.072	0.931	0.019	-3.767	0.000	***
platelets	0.003	1.003	0.001	3.320	0.001	***
urea.nitrogen	-0.016	0.984	0.006	-2.617	0.009	**
blood.potassium	-0.512	0.599	0.278	-1.838	0.066	
blood.sodium	-0.118	0.889	0.062	-1.920	0.055	
chloride	0.114	1.121	0.055	2.071	0.038	*
magnesium.ion	-0.745	0.475	0.430	-1.732	0.083	
bicarbonate	0.173	1.189	0.050	3.475	0.001	***
BMI	0.021	1.021	0.015	1.418	0.156	
heart.rate	-0.019	0.981	0.007	-2.717	0.007	**
respiratory.rate	-0.040	0.961	0.027	-1.450	0.147	
temperature	0.265	1.303	0.174	1.517	0.129	
SP.O2	0.112	1.119	0.048	2.350	0.019	*
urine.output	0.000	1.000	0.000	1.430	0.153	
blood.calcium	0.782	2.186	0.206	3.800	0.000	***
lactic.acid	-0.003	0.997	0.001	-2.780	0.005	**
PCO2	-0.022	0.978	0.010	-2.288	0.022	*
m.a.p	0.021	1.022	0.011	1.878	0.060	

The overall risk factors common among the three reduced final models are presenting an elevated

heart rate, high count of leucocytes, and elevated urea nitrogen. Elevated respiratory rate, RDW count, magnesium and potassium were common in two out of three models.

5.4 Classification by predictive machine learning algorithms

5.4.1 Support Vector Machine

The classification using SVM has been tested with three different approaches: linear classification, Gaussian radial kernel classification and polynomial kernel classification. Except for kernel specification, the ksvm() function has been conducted with the other default arguments. The parameters obtained with model performances are summarized in Figure 5.12.

Models generated with the dataset A did not converge because it presents missing values.
With linear classification, the model on dataset B present better balanced accuracy, sensitivity, kappa and AUC compared to models on datasets C and D which are very similar and present better overall accuracy, specificity, PPV and NPV.

Parameters' results of linear classification are the same as those obtained by polynomial kernel classification. The performance of the models using radial kernel classification is worst compared to linear or polynomial. Between linear and polynomial classification, the simplest model (linear) is selected as the best one.

After selecting the linear SVM models, a 10-folds cross-validation has been conducted to validate the results on those models. Figure 5.12 shows that the validation results model on dataset C is better with all metrics compared to models on dataset B and D.

Algorithms	Acc	B.acc	Sens	Sepc	PPV	NPV	Kappa	AUC
Linear SVM								
Dataset A	NA	NA	NA	NA	NA	NA	NA	NA
Dataset B	0,852	0,629	0,294	0,964	0,625	0,871	0,328	0,629
Dataset C	0,881	0,597	0,204	0,990	0,769	0,885	0,281	0,597
Dataset D	0,884	0,599	0,208	0,990	0,769	0,888	0,287	0,599
Data B k-fold CV	0,791	0,601	0,304	0,896	0,389	0,856	0,219	0,600
Data C k-fold CV	0,887	0,618	0,245	0,990	0,800	0,891	0,332	0,618
Data D k-fold CV	0,884	0,599	0,208	0,990	0,769	0,888	0,287	0,599
RBD Guassian SVM								
Dataset A	NA	NA	NA	NA	NA	NA	NA	NA
Dataset B	0,852	0,559	0,118	1,000	1,000	0,849	0,182	0,559
Dataset C	0,867	0,520	0,041	1,000	1,000	0,866	0,068	0,520
Dataset D	0,870	0,530	0,063	0,997	0,750	0,871	0,097	0,530
Polynomial SVM								
Dataset A	NA	NA	NA	NA	NA	NA	NA	NA
Dataset B	0,852	0,629	0,294	0,964	0,625	0,871	0,328	0,629
Dataset C	0,887	0,597	0,204	0,990	0,769	0,885	0,281	0,597
Dataset D	0,884	0,599	0,208	0,990	0,769	0,888	0,287	0,599

Figure 5.12: SVM models results. Acc= accuracy; B.acc= balanced accuracy; Sens= sensitivity, Spec= specificity; PPV= positive predicted value; NPV = negative predicted value.

5.4.2 Artificial Neural Network

First, to generate ANN models, variables have been escalated using a normalize function.

The ANN models have been performed on the four datasets with the default parameters of the neuralnet() R function with one and three nodes or neurons in the hidden layer. The algorithm used by default is the backpropagation. The number of hidden neurons affects how well the network can separate the data. A large number of hidden neurons will ensure correct learning but its performance on new data and its ability to generalize is compromised. With too few hidden neurons, the network may be unable to learn the relationships amongst the data and the error will fail to fall below an acceptable level.

The metric parameters obtained with model performances are summarized in Figure 5.13.
As for SVM, the ANN model with dataset A did not converge because of the missing values. Regarding datasets B, C and D the best models are obtained with one node in the hidden layer as all metrics are a little bit higher.

In this case, a 10-folds cross-validation has been conducted to validate the results of all models with one hidden neuron on datasets B, C and D. After validation, performance in dataset B model decreases and the best model obtained is the one performed on dataset C.

Algorithms	Acc	B.acc	Sens	Sepc	PPV	NPV	Kappa	AUC
One hidden neuron								
Dataset A	NA	NA	NA	NA	NA	NA	NA	NA
Dataset B	0,860	0,765	0,615	0,915	0,615	0,915	0,530	0,765
Dataset C	0,870	0,662	0,368	0,955	0,583	0,899	0,382	0,662
Dataset D	0,862	0,659	0,354	0,963	0,657	0,882	0,389	0,659
Dataset B k-fold CV	0,846	0,592	0,192	0,991	0,833	0,847	0,227	0,592
Dataset C k-fold CV	0,880	0,610	0,228	0,991	0,813	0,883	0,312	0,610
Dataset D k-fold CV	0,855	0,599	0,215	0,982	0,700	0,863	0,273	0,599
Three hidden neurons								
Dataset A	NA	NA	NA	NA	NA	NA	NA	NA
Dataset B	0,839	0,707	0,600	0,915	0,565	0,892	0,434	0,707
Dataset C	0,857	0,647	0,351	0,943	0,513	0,895	0,339	0,647
Dataset D	0,827	0,625	0,323	0,927	0,467	0,873	0,285	0,652

Figure 5.13: ANN models results. Acc= accuracy; B.acc= balanced accuracy; Sens= sensitivity, Spec= specificity; PPV= positive predicted value; NPV = negative predicted value.

5.4.3 Random forest

The classification using a basic RF model has been performed selecting the parameters ntrees=500 (by default) and mtyr= sqrt(p); being p the number of the variables. As with SVM and ANN, the model with dataset A did not converge. The parameters obtained with models performances on datasets B, C and D are summarized in Figure 5.15.

In addition to model's performance parameters, random forest has the advantage of showing the most important variables for the model. Graphics of the variables importance are found in Figure 5.14 in which it can be observed that with the model applied on dataset B the five more important covariates are blood.calcium, lymphocyte, anion.gap, INR and bicarbonate. For model applied on dataset C are lactic.acid, anion.gap, lymphocyte, bicarbonate and urine.output. Finally, for model applied on dataset D are anion.gap, lactic.acid, bicarbonate, blood.calcium and lymphocyte.

Overall, although they are not in the same order, the covariates lymphocyte, anion.gap and bicarbonate are common within the five more important variables between the three models; meaning that they have a strong relationship with the primary response.

The results in Figure 5.15 indicate that the best model obtained with random forest algorithm is with dataset D.

A 10-fold cross-validation has been performed on models with datasets B, C and D. The results of the validation technique confirm that the best model is obtained with dataset D.

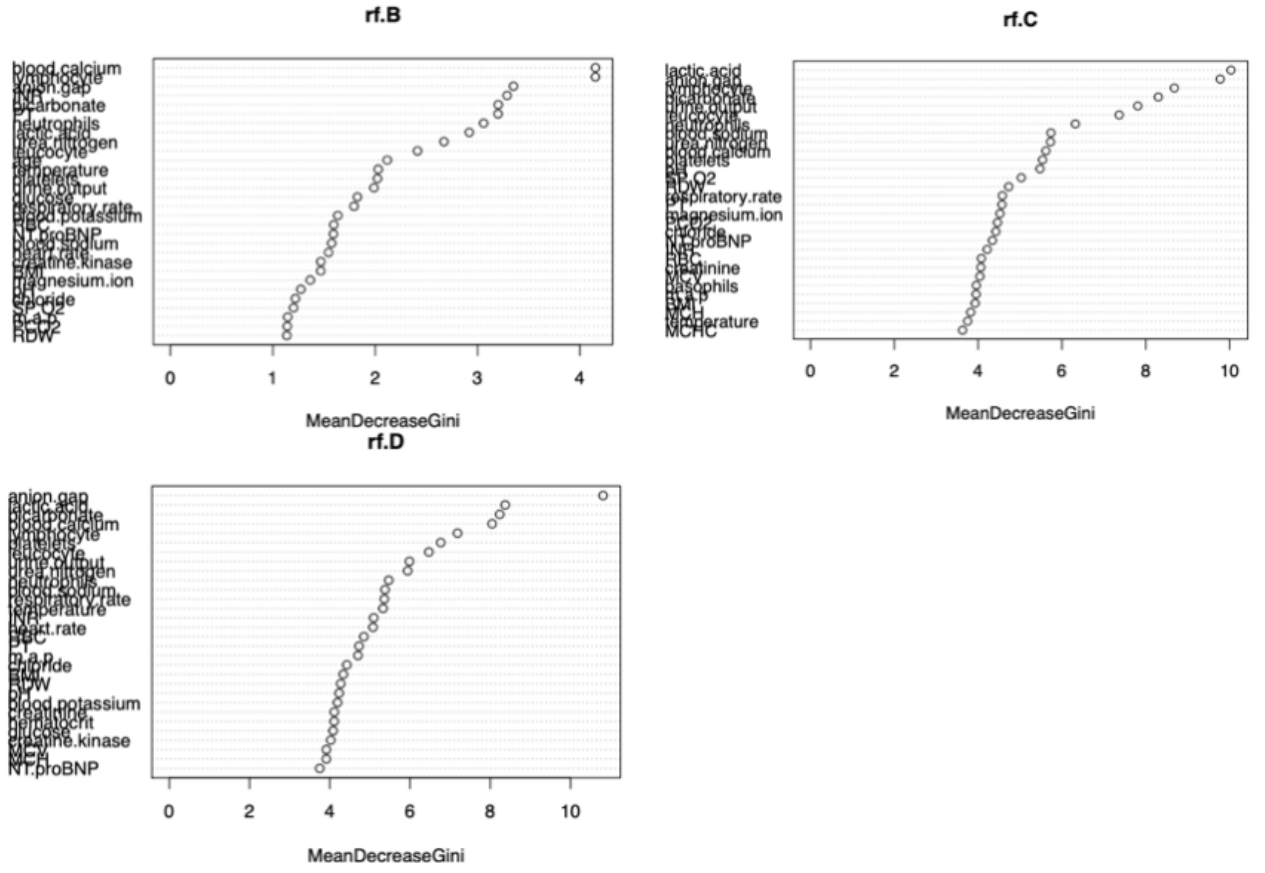

Figure 5.14: Variables impact in RF models.

Algorithms	Acc	B.acc	Sens	Sepc	PPV	NPV	Kappa	AUC
RF								
Dataset A	NA	NA	NA	NA	NA	NA	NA	NA
Dataset B	0,845	0,823	0,800	0,847	0,174	0,991	0,237	0,582
Dataset C	0,875	0,875	0,875	0,875	0,140	0,997	0,211	0,568
Dataset D	0,884	0,886	0,889	0,884	0,167	0,997	0,248	0,582
RF K-fold CV								
Dataset B	0,837	0,578	0,174	0,981	0,667	0,876	0,218	0,578
Dataset C	0,875	0,577	0,160	0,993	0,800	0,878	0,202	0,577
Dataset D	0,887	0,592	0,188	0,997	0,900	0,886	0,276	0,592

Figure 5.15: RF models results. Acc= accuracy; B.acc= balanced accuracy; Sens= sensitivity, Spec= specificity; PPV= positive predicted value; NPV = negative predicted value.

Chapter 6

Discussion

The objectives of this thesis consisted in learning about ML classification algorithms and EHR databases, learning data clean-up and data curation techniques to be able to obtain analyzable datasets and performing a study of binary classification.

6.1 Comments on the database

The selected database was a subset of the MIMIC-III open database created by Li et al., 2021 [24]. Working with this subset was selected in the first place because it is freely accessible and is easy to use for learning objectives as the data extraction of the different MIMIC-III tables was already done.
It was modified during the first phase of data clean-up and data curation and it finally consisted of 1176 observations and 48 variables. The primary response to predict was the outcome, and of those 1176 observations, 86.48% survived and 13.52% died. Therefore, the data was very imbalanced.

Because of the size of the dataset, no variable transformations have been applied and the outliers have not been dealt with as it was considered that it was possible to perform a non-parametric statistical test. This decision might have had a bad impact on the results and these two problems would have been addressed if time had allowed it.

The results of the bivariate analysis found enough statistical evidences to affirm that there were differences between the outcome and all the covariates except for gender, Coronary Heart Disease with no Myocardial Infarction (CHD.with.no.MI), Chronic Obstructive Pulmonary Disease (COPD) and depression.

The study of the missing values showed only 3.4% of the values were missing but because of the sensitivity of machine learning models in front of incomplete datasets, three methods to deal with missing values were explored: listwise deletion, KNN imputation and MICE. A total of four datasets were prepared for the multivariate analysis and the prediction of the classification

algorithms.

6.2 Comments on the multivariate analysis

The multivariate analysis with multiple logistic regression was done to get to know the variables that contribute the most to explaining the outcome of the patients.

The stepAIC() R function did not work with dataset A because of the missing values. A manual stepwise procedure could be done by removing the variables with a higher p-value (one at a time) until the AIC did not decrease.

The best models obtained on datasets B, C and D were fitted with 20, 18 and 25 variables and present an AIC of 239,6, 719,43 and 712,58 respectively. Although datasets C and D have the same dimensions, the models obtained are different because the imputed values are not the same. As both have similar AIC values, the best model between those two would be the MLR on dataset C because it includes fewer covariates and it makes the model more simple.

Regarding all variables included in each model, the automatic "stepwise" logistic regression model determined the following features (common between the three generated models) to be key risk predictors for patients' outcome: elevated heart rate, high count of leucocytes, and elevated urea nitrogen.

However, the results of the multivariate analysis are questioned because many variables with positive estimates do not have clinical or biological sense. A possible explanation would be that the models do not fit correctly because two of the logistic regression assumptions are violated: the outliers of the continuous predictors have not been dealt with, and the intercorrelation between covariates has not been checked either. Therefore, there are too many variables that depend on each other to explain the models and that can guide to misleading results. In addition, a validation of the models could be conducted.

6.3 Comparison of the supervised classification algorithms

Table 6.1 summarizes the best model obtained with each algorithm and its parameters after the 10-folds cross-validation.

Table 6.1: Comparison with the metrics of the best performed predictions.

Algorithms	Dataset	Acc	B.acc	Sens	Sepc	PPV	NPV	Kappa	AUC
Linear SVM	C	0,887	0,618	0,245	0,990	0,800	0,891	0,332	0,618
ANN one HN	C	0,880	0,610	0,228	0,991	0,813	0,883	0,312	0,610
RF	D	0,887	0,592	0,188	0,997	0,900	0,886	0,276	0,592

HN= hidden node; Acc= accuracy; B.acc= balanced accuracy; Sens= sensitivity, Spec= specificity; PPV= positive predicted value; NPV = negative predicted value.

Considering that our particular dataset presents a very imbalanced data, AUC, balanced accuracy and kappa values will be more informative as the general accuracy is misled by the majority class. Taking into account those metrics, we can conclude that the linear SVM model is the algorithm that classifies better the testing set closely followed by ANN with one hidden node. Although they are the best models obtained, the AUC values are relatively >0,5, which means that they are slightly better than random guessing. The values of the Cohen's kappa statistic indicates that the classifiers just provide a fair classification on new data as they are between 0,21–0,40 of Landis and Koch (1977) scheme value. Finally the balanced accuracy is the 62% indicating a regular precision of the classification.

Overall, the models obtained are not very good and considerably poorer than the XGBoost model and LASSO regression model described by Li et al., 2021 in the original article [24]. Considering that in this thesis only basic models have been conducted, probably a further study with best model selection parameters would be able to improve our models.

Reviewing the algorithms individually, the random forest had several advantages although it has been the most computational consuming. It offered a prioritization of the variables which has great value for medical and scientific personnel. The five more used variables were anion.gap, lactic.acid, bicarbonate, lymphocyte and blood.calcium. Strangely non of them are heart specific factors.

Chapter 7

Conclusions

7.1 Conclusions

Here we highlight the main conclusions of the present thesis:

- The bivariate statistical analysis found almost al covariates different from the outcome.
- The multivariate analysis results are inconsistent due to the amount of variables included in each model and probably because outliers and intercorrelation have not been dealt with.
- Non of the models worked with the dataset containing missing values.
- Models performed on the dataset with listwise deletion obtained slightly worst results than models on datasets with imputation methods.
- There is no big difference between imputating with KNN or MICE methods.
- In general, the models obtained in this thesis have a poor performance.
- Linear support vector machine algorithm on dataset C is the best model obtained.

7.2 Future perspectives

Several questions and issues have been left open for future work:

First I would address the analysis and treatment of the extreme values or outliers within the covariates together with a study of the intercorrelations to improve the EDA and the multiple logistic regression models.

An important theme of the thesis and related to one of our objectives was to introduce myself into the new world of machine learning and, as I was not an expert, I performed basic prediction algorithms with its default parameters. A further model selection and new training could easily

achieve models with better results.

In addition, an analysis of PCA (Principal Component Analysis) could be carried out to reduce the covariates to include in the models and maybe this would lead to better prediction outcomes.

Finally, ensemble-based machine learning techniques which are a combination of several base models to produce one optimal predictive model could be explored as it has been reported to be successful for imbalanced datasets [25].

7.3 Planning follow-up

To assess the compliance to the original plan we'll use the milestones described in section 2.4.:

- PAC0 - (23 of February 2022) Definition of the contents of the work
- PAC1 - (7 of March 2022) Workplan
- PAC2 - (21 of April 2022) Workplan - phase 1
- PAC3 - (16 of May 2022) Workplan - phase 2
- PAC4 - (2 of June 2022) memory delivery
- PAC5 - (6 of June 2022) we must have prepared the thesis presentation

Excepting the fifth milestone, which is due for completion after the submission of this thesis, all the other assignments have been completed on time and according to plan.

Regarding the methodology, the commitment to the initial goals has been mainly accomplished and in addition, multivariate analysis has been included. Although in the end, it did not contribute to significant results, it completed the statistical analysis.

However, the exploratory and statistical analysis of such a big dataset has been a challenge and several improvements could be further performed out of the planned milestones.

Overall, with this thesis I have been able to learn and deepen in machine learning supervised classification algorithms and statistical techniques that I had seen more superficially during the master's degree. It provided an introductory guide for anyone who wishes to discover prediction models.

Chapter 8

Glossary

8.1 List of abbreviations

AIC. Akaike Information Criterion.
AI. Artificial Intelligence.
ANN. Artificial Neural Networks.
AUC. Area Under the Curve.
BMI. Body Mass Index.
EDA. Exploratory Data Analysis.
EHR. Electronic Health Records.
FN. False Negative.
FP. False Positive.
HN. Hideen Node.
ICU. Intensive Care Unit.
KNN. K-Nearest Neighbour.
LVEF. Left Ventrifuclar Ejection Fraction.
MAP. Mean Arterial Pressure.
MAR. Missing At Random.
MCAR. Missing Completely At Random.
MCMC. Markov Chain Monte Carlo.
MICE-PMM. Multiple Imputation by Chained Equations - Predictive Mean Matching.
MIMIC-III. Medical Information Mart for Intensive Care.
ML. Machine Learning.
MLR. Multiple Logistic Regression.
MNAR. Missing Not At Random.
NT-proBNP. N-Terminal Pro-Brain Natriuretic Peptide.
PPV. Positive Predicted Value.
ROC. Receiver Operating Characteristics.
RF. Random Forest.
SVM. Support Vector Machines.
TN. True Negative.

TP. True Positive.

8.2 Brief definitions

Akaike Information Criterion. Is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data.

Artificial Neutral Networks. A computing system inspired by the biological neutral networks that can be applied to both regression and classification problems. Neurons in the network receive, integrate, process and transfer input data to generate a final outcome.

Atrialfibrillation. Atrial fibrillation is a heart condition that causes an irregular and often abnormally fast heart rate. A normal heart rate should be regular and between 60 and 100 beats a minute when you're resting. You can measure your heart rate by checking your pulse in your wrist or neck.

Electronic Health Records. An electronic health record (EHR) is a digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users.

False Negative Incorrectly predicted the number of instances as not required.

False Positive. Number of instance which incorrectly predicted.

Multiple Logistic Regression. Is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.

Machine Learning. Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Random Forest. Is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Support Vector Machines. Supervised learning algorithm that estimates the value or class of an observation by constructing hyperplanes that split the data into fairly homogeneous subsets.

True Negative Correctly predicted the number of instances as not required.

True Positive. Number of instance which correctly predicted.

9 Bibliography

[1] Hiba Asri et al. "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis". In: *Procedia Computer Science* 83 (Jan. 2016), pp. 1064–1069. ISSN: 1877-0509. DOI: 10.1016/J.PROCS.2016.04.224. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877050916302575.

[2] Juan Jose Beunza et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)". In: *Journal of Biomedical Informatics* 97 (Sept. 2019), p. 103257. ISSN: 1532-0464. DOI: 10.1016/J.JBI.2019.103257.

[3] Alison Callahan and Nigam H. Shah. "Machine Learning in Healthcare". In: *Key Advances in Clinical Informatics: Transforming Health Care through Health Information Technology*. Elsevier Inc., July 2017, pp. 279–291. ISBN: 9780128095256. DOI: 10.1016/B978-0-12-809523-2.00019-4.

[4] Joana Cardoso-Fernandes et al. "Semi-automatization of support vector machines to map lithium (Li) bearing pegmatites". In: *Remote Sensing* 12.14 (July 2020). ISSN: 20724292. DOI: 10.3390/rs12142319.

[5] A Chakrabarti and J. K Ghosh. "AIC, BIC and recent advances in model selection." In: *Philosophy of statistics* (2011), pp. 583–605.

[6] Subramanian Chandramouli, Saikat Dutt, and Amit Kumar Das. *Machine learning*. Pearson Education India, 2018, p. 456.

[7] Martin R Cowie et al. "Electronic health records to facilitate clinical research". In: *Clinical Research in Cardiology* 106 (2017), pp. 1–9. DOI: 10.1007/s00392-016-1025-6.

[8] Dhiraj Dahiwade, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach". In: *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019* (Mar. 2019), pp. 1211–1215. DOI: 10.1109/ICCMC.2019.8819782.

[9] Daniel DeMers and Daliah Wachs. "Physiology, Mean Arterial Pressure". In: *StatPearls* (Apr. 2022). URL: https://www.ncbi.nlm.nih.gov/books/NBK538226/.

[10] Devanshi Dhall, Ravinder Kaur, and Mamta Juneja. "Machine learning: A review of the algorithms and its applications". In: *Lecture Notes in Electrical Engineering*. Vol. 597. 2020. DOI: 10.1007/978-3-030-29407-6{_}5.

[11] Issam El Naqa and Martin J. Murphy. "What Is Machine Learning?" In: *Machine Learning in Radiation Oncology*. Springer International Publishing, 2015, pp. 3–11. DOI: 10.1007/978-3-319-18305-3{_}1.

[12] R. S. Evans. "Electronic Health Records: Then, Now, and in the Future". In: *Yearbook of medical informatics* (May 2016), S48–S61. ISSN: 23640502. DOI: 10.15265/IYS-2016-s006.

[13] Guangda Gao et al. *Agricultural Irrigation Area Prediction based on Improved Random Forest Model [U+F020]*. Tech. rep.

[14] Eleni I. Georga et al. "Artificial Intelligence and Data Mining Methods for Cardiovascular Risk Prediction". In: (2019), pp. 279–301. DOI: 10.1007/978-981-10-5092-3{_}14. URL: https://link.springer.com/chapter/10.1007/978-981-10-5092-3_14.

[15] Enzo Grossi and Massimo Buscema. "Introduction to artificial neural networks". In: *European Journal of Gastroenterology and Hepatology* 19.12 (Dec. 2007), pp. 1046–1054. ISSN: 0954691X. DOI: 10.1097/MEG.0B013E3282F198A0.

[16] Alejandra E Guevara Morel et al. "Dealing with Missing Data in Real-World Data: A Scoping Review of Simulation Studies". In: (2022). DOI: 10.21203/rs.3.rs-1619388/v1. URL: https://doi.org/10.21203/rs.3.rs-1619388/v1.

[17] Amelia Ritahani Ismail, Nadzurah Zainal Abidin, and Mhd Khaled Maen. "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare". In: *Journal of Robotics and Control (JRC)* 3.2 (Feb. 2022), pp. 143–152. ISSN: 2715-5072. DOI: 10.18196/JRC.V3I2.13133. URL: https://journal.umy.ac.id/index.php/jrc/article/view/13133.

[18] Alistair E W Johnson et al. "Data Descriptor: MIMIC-III, a freely accessible critical care database". In: (2016). DOI: 10.1038/sdata.2016.35. URL: www.nature.com/sdata/.

[19] Hyun Kang. "The prevention and handling of the missing data". In: *Korean Journal of Anesthesiology* 64.5 (May 2013), p. 402. ISSN: 20056419. DOI: 10.4097/KJAE.2013.64.5.402. URL: /pmc/articles/PMC3668100/%20/pmc/articles/PMC3668100/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/.

[20] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. "Support vector machines in R". In: *Journal of Statistical Software* 15.9 (2006). ISSN: 15487660. DOI: 10.18637/jss.v015.i09.

[21] Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. "SICE: an improved missing data imputation technique". In: *Journal of Big Data* 7.1 (Dec. 2020), pp. 1–21. ISSN: 21961115. DOI: 10.1186/S40537-020-00313-W/FIGURES/9. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00313-w.

[22] Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: (1995). URL: http://robotics.stanford.edu/~ronnyk.

[23] Brett Lantz. *Machine Learning with R Second Edition all your data analysis problems*. 2021. ISBN: 9781784393908.

[24] Fuhai Li et al. "Prediction model of in-hospital mortality in intensive care unit patients with heart failure: Machine learning-based, retrospective analysis of the MIMIC-III database". In: *BMJ Open* 11.7 (2021). ISSN: 20446055. DOI: 10.1136/bmjopen-2020-044779.

[25] Victoria López et al. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". In: *Information Sciences* 250 (Nov. 2013), pp. 113–141. ISSN: 0020-0255. DOI: 10.1016/J.INS.2013.07.007.

[26] Nagaraj M. Lutimath, C. Chethan, and Basavaraj S. Pol. "Prediction of heart disease using machine learning". In: *International Journal of Recent Technology and Engineering* 8.2 Special Issue 10 (2019). ISSN: 22773878. DOI: 10.35940/ijrte.B1081.0982S1019.

[27] Tom M. (Tom Michael) Mitchell. *Machine Learning*, p. 414. ISBN: 0070428077.

[28] Todd G. Nick and Kathleen M. Campbell. "Logistic regression." In: *Methods in molecular biology (Clifton, N.J.)* 404 (2007), pp. 273–301. ISSN: 10643745. DOI: 10.1007/978-1-59745-530-5{_}14.

[29] Jasmina Dj. Novaković et al. "Evaluation of Classification Models in Machine Learning". In: *Theory and Applications of Mathematics & Computer Science* 7.1 (Apr. 2017), Pages: 39 –46. ISSN: 2067-2764. URL: https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158.

[30] Grigorios Papageorgiou et al. "Statistical primer: how to deal with missing data in scientific research?" In: *Interactive CardioVascular and Thoracic Surgery* 27.2 (Aug. 2018), pp. 153–158. ISSN: 15699285. DOI: 10.1093/ICVTS/IVY102. URL: https://academic.oup.com/icvts/article/27/2/153/4995008.

[31] Derek A. Pisner and David M. Schnyer. "Support vector machine". In: *Machine Learning: Methods and Applications to Brain Disorders*. 2019. DOI: 10.1016/B978-0-12-815739-8.00006-7.

[32] G. Purusothaman and P. Krishnakumari. "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease". In: *Indian Journal of Science and Technology* 8.12 (Jan. 2015), pp. -. ISSN: 0974-5645. DOI: 10.17485/IJST/2015/V8I12/58385. URL: https://indjst.org/articles/a-survey-of-data-mining-techniques-on-risk-prediction-heart-disease%20https://indjst.org/.

[33] Payam Refaeilzadeh, Lei Tang, and Huan Liu. "Cross-Validation". In: *Encyclopedia of Database Systems* (2016), pp. 1–7. DOI: 10.1007/978-1-4899-7993-3{_}565-2. URL: https://link.springer.com/referenceworkentry/10.1007/978-1-4899-7993-3_565-2.

[34] Sandro Sperandei. "Understanding logistic regression analysis". In: *Biochemia Medica* 24.1 (2014), pp. 12–20. DOI: 10.11613/BM.2014.003. URL: http://dx.doi.org/10.11613/BM.2014.003.

[35] P. H. Sydenham and Richard Thorn. *Handbook of measuring system design*. Wiley, 2005. ISBN: 0470021438.

[36] J E T Akinsola, Akinsola Jet, and Hinmikaiye J O. "Supervised Machine Learning Algorithms: Classification and Comparison Netiquette of Cyberbullying and Privacy Issues View project The Use Of BIG DATA in Mobile Analytics View project Supervised Machine Learning Algorithms: Classification and Comparison". In: *International Journal of Computer Trends and Technology* 48 (2017). ISSN: 2231-2803. DOI: 10.14445/22312803/IJCTT-V48P126. URL: http://www.ijcttjournal.org.

[37] Jussi Tohka and Mark van Gils. "Evaluation of machine learning algorithms for health and wellness applications: A tutorial". In: *Computers in Biology and Medicine* 132 (May 2021), p. 104324. ISSN: 0010-4825. DOI: 10.1016/J.COMPBIOMED.2021.104324.

[38] Fernando Tusell. "Análisis de Regresión. Introducción Teórica y Práctica basada en R." In: *Universidad del País Vasco* (2011).

[39] Ch Anwar Ul Hassan, Muhammad Sufyan Khan, and Munam Ali Shah. "Comparison of machine learning algorithms in data classification". In: *ICAC 2018 - 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing* (Sept. 2018). DOI: 10.23919/ICONAC.2018.8748995.

[40] K. M. Veena, K. Manjula Shenoy, and K. B. Ajitha Shenoy. "Performance comparison of machine learning classification algorithms". In: *Communications in Computer and Information Science* 906 (2018), pp. 489–497. ISSN: 18650929. DOI: 10.1007/978-981-13-1813-9{_}49/COVER/. URL: https://link.springer.com/chapter/10.1007/978-981-13-1813-9_49.

[41] Scott I. Vrieze. "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)". In: *Psychological Methods* 17.2 (June 2012), pp. 228–243. ISSN: 1082989X. DOI: 10.1037/A0027127.

Appendix A

Variables description

Table A.1: Variables description I

Variables description		
Group	**Name**	**Description**
Primary response	**outcome**	Vital status at hospital discharge
Demographic features	**age**	Are
	gender	Sex
	BMI	Body Mass Index
Comorbidities	**atrialfibrillation:**	Atrial fibrillation
	CHD.with.no.MI	Coronary Heart Disease with no Myocardial Infarction
	COPD	Chronic Obstructive Pulmonary Disease
	deficiencyanemias	Hypoferric anaemia
	depression	Depression
	diabetes	Diabetes mellitus
	hyperlipemia	Hyperlipidaemia
	hypertensive	Hypertension
	renal.failure	Chronic renal insufficiency

Table A.2: Variables description II

Group	Name	Description
Vital signs	**Diastolic blood pressure**	Diastolic blood pressure (mmHg)
	heart.rate	Rythm of the heart (bpm)
	reapiratory.rate	Rythm of breath (bpm)
	SP.02	Oxygen saturation (%)
	Systolic blood pressure	Systolic blood pressure (mmHg)
	temperature	Subject body temperature at admission time (C)
	urine.output	Urine 24-hour volume (ml)
	m.a.p	Mean arterial pressure (Systolic + 2*Diastolic)/3

Table A.3: Variables description III

Group	Name	Description
Laboratory		
Blood count	**basophils**	Basophils (%)
	hematocrit	Red blood cells in blood (%)
	leucocyte	White cells (x10ˆ9/L)
	lymphocyte	Lymphocytes (%)
	MCH	Mean corpuscular haemoglobin (pg)
	MCHC	Mean corpuscular haemoglobin concentration (%)
	MCV	Mean corpuscular volume (fL)
	neutrophils	Neutrophils (%)
	platelets	Platelet count (x10ˆ9/L)
	RBC	Red Blood Cells (x10ˆ12/L)
	RDW	Red blood cell distribution width (%)
Coagulations	**INR**	International Normalised Ratio
	PT	Prothrombin time (s)
Chemistry	**anion.gap**	Blood anion gap (mEq/ml)
	blood.calciun	Blood calcium (mg(dL)
	blood.potassium	Blood potassium (mEq/ml)
	blood.sodium	Blood sodium (mEq/ml)
	chloride	Blood chloride (mEq/ml)
	creatine.kinase	Creatine kinase (IU/L)
	creatinine	Creatinine (mg/dL)
	glucose	Glucose (mEq/ml)
	magnesium.ion	Blood Magnesium ion (mg/dL)
	urea.nitrogen	Blood urea nitrogen (mg/dL)
Venous blood	**bicarbonate**	Blood bicarbonate (mEq/ml)
	lactic.acid	Lactate (mmol/L)
	PCO2	Partial pressure of carbon dioxide in artery (mmHg)
	pH	pH
Heart specific	**EF**	Left Ventricular Ejection fraction (%)
	NT. proBNP	N-terminal pro-brain natriuretic peptide (pg/ml)