UNIVERSIDADE D
COIMBRA

Cláudia Rodrigues Lopes

COMPUTATIONAL INTELLIGENCE MODELS FOR LENGTH OF STAY PREDICTION

Dissertação no âmbito do Mestrado Integrado em Engenharia Biomédica orientada pelo Professor Doutor Jorge Henriques, pelo Professor Doutor Paulo de Carvalho e pelo Doutor Alberto Bonomi apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Outubro de 2020

UNIVERSIDADE D
COIMBRA

Cláudia Rodrigues Lopes

Computational Intelligence Models for Length of Stay Prediction

Thesis submitted to the Faculty of Science and Technology
of the University of Coimbra for the degree of Master in
Biomedical Engineering with specialization in
Clinical Informatics and Bioinformatics

Supervisors:
Prof. Dr. Jorge Henriques
Prof. Dr. Paulo de Carvalho
Dr. Alberto Bonomi

Coimbra, 2020

This work was developed in collaboration with:

Philips Electronics Nederland B.V.

CHUC - Coimbra Hospital and University Center

CISUC - Center of Informatics and Systems of the University of Coimbra

Acknowledgments

This thesis is the result of the shared knowledge and dedication of a group of people to whom I am very grateful.

First, I would like to thank my advisors, Professor Paulo de Carvalho and Professor Jorge Henriques for all the support and guidance. A special thanks to Professor Jorge Henriques for his dedication and motivation that were one of the keys for the realization of this project. His unconditional availability and fruitful discussions were essential to overcome the difficulties that appeared throughout the work and to its successful conclusion.

I would also like to express my gratitude to my co-advisor, Doctor Alberto Bonomi and his incredible team from Philips that shared their knowledge and ideas and provided me the support needed. I want to acknowledge the dedication, the availability and the team work of this group that made possible the realization of this project.

I want to thank Professor Lino Gonçalves for allowing CHUC to participate in the study, for providing access to the database from the Cardiac Intensive Care Unit of the Hospital dos Covões and for giving the driving idea of the thesis project (the evaluation of the SCORE risk ability to predict the patients' length of stay).

Finally, I am very thankful to Doctor José Pedro Sousa, cardiologist from CHUC working in the cardiac ICU, for guiding the work concerning the cardiac patients and for contributing with his clinical knowledge thorough the study. His insights were crucial to validate the clinical relevance of the developed work.

Resumo

A previsão do tempo de internamento dos pacientes é de grande importância para os hospitais, uma vez que pode determinar a utilização de recursos, melhorar o agendamento de futuros internamentos e cirurgias, e auxiliar no planeamento dos cuidados de saúde dos pacientes, desde a admissão até à alta. Consequentemente, uma melhor qualidade dos cuidados de saúde prestados pode ser proporcionada aos pacientes, sendo este o principal objetivo dos hospitais.

Neste projecto, quatro abordagens diferentes foram implementadas para desenvolver modelos de previsão de tempo de internamento: **i)** exploração de modelos de risco existentes (SCORE), **ii)** aplicação de modelos típicos de inteligência computacional (Random Forest, Support Vector Machine e Multilayer Perceptron), **iii)** desenvolvimento de um modelo interpretável e personalizável ao paciente com base em regras e **iv)** integração de dados dinâmicos (sinais vitais) nos modelos anteriores. Os dados clínicos usados neste trabalho foram fornecidos pelo CHUC (Centro Hospitalar e Universitário de Coimbra) e pela Philips Electronics Nederland B.V., compreendendo 1544 pacientes admitidos na unidade de cuidados intensivos de cardiologia do Hospital dos Covões (Coimbra) e 189 pacientes bariátricos admitidos para cirurgia no Catharina Hospital (Eindhoven), respetivamente.

O conjunto inicial de variáveis dos pacientes cardíacos foi obtido através de uma revisão da literatura e do conhecimento clínico de um cardiologista da unidade de cuidados intensivos de cardiologia do CHUC. Para os pacientes bariátricos, este conjunto resultou de uma revisão da literatura para a determinação das variáveis relevantes. Posteriormente, as variáveis de entrada dos modelos de previsão de tempo de internamento foram selecionadas desse conjunto inicial usando o coeficiente de correlação tau de Kendall. Adicionalmente, as variáveis de entrada selecionadas para os pacientes cardíacos foram também validadas pelo cardiologista. O desempenho dos modelos referidos, medido através da média geométrica (GE) e do F1 score, foi determinado aplicando este conjunto final de variáveis de entrada a cada um deles. Finalmente, através da aplicação do teste Friedman e do correspondente teste *post-hoc* Nemenyi, foi possível ordenar os modelos em função do seu desempenho.

A performance do modelo baseado no SCORE foi significativamente baixa, obtendo uma GE de 0.50. Assim, apesar deste modelo de risco ser de grande importância na prática

cardiológica europeia, não é adequado para estimar o tempo de internamento hospitalar. A segunda abordagem (modelo Black-box) superou o modelo anterior. Os melhores resultados foram obtidos pelo Multilayer perceptron com uma GE de 0.62 ± 0.03 para os pacientes cardíacos e 0.64 ± 0.08 para os bariátricos, respetivamente. Verificou-se ainda que o desempenho do modelo interpretável e personalizável foi superior ao modelo Black-box, para os dois tipos de pacientes, com uma GE de 0.66 ± 0.02 para os pacientes cardíacos e 0.83 ± 0.05 para os pacientes bariátricos. Adicionalmente, a inclusão de sinais vitais aos modelos de previsão mostrou-se vantajosa por levar a um aumento da performance em todos os classificadores. Estes resultados sugerem que a incorporação de dados dinâmicos em modelos de previsão de tempo de internamento deve ser explorada de forma aprofundada em estudos posteriores.

A análise dos resultados permitiu-nos concluir que, apesar de aceitável, a performance dos modelos desenvolvidos não parece ser adequada para o seu uso na prática clínica (GE máxima de 0.66 e 0.83 para os pacientes cardíacos e bariátricos, respetivamente). Este facto pode-se justificar pela dificuldade e complexidade que o problema apresenta. O estudo de outras variáveis, não só determinadas aquando a admissão, mas durante as primeiras horas ou no primeiro dia de internamento do doente, poderia ser uma estratégia a explorar no futuro.

Palavras-chave: Tempo de internamento hospitalar, Previsão, Modelos de inteligência computacional, Interpretabilidade, Modelos prognósticos de risco, Sinais vitais

Abstract

Predicting the patients' length of stay (LOS) is of major importance for hospitals, since it can determine the resource utilization, improve the scheduling of admissions and surgeries and helping in the development of effective clinical pathways. Consequently, a better quality of care can be provided to the patients, which is the main goal of the hospitals.

In this project, four different approaches were implemented to develop LOS prediction models: **i)** exploration of available risk tools (SCORE), **ii)** application of typical computational intelligence models (Random Forest, Support Vector Machine and Multilayer Perceptron), **iii)** development of an interpretable and patient customized model based on rules and **iv)** integration of dynamic data (vital signs) in the previous models. The clinical data used in this work was provided by the CHUC (Coimbra Hospital and University Center) and by Philips Electronics Nederland B.V., comprising 1544 patients admitted in the cardiac intensive care unit of Hospital dos Covões (Coimbra) and 189 bariatric patients admitted to surgery in Catharina Hospital (Eindhoven), respectively.

The initial set of features of the cardiac patients was obtained through a literature review and the clinical knowledge of an ICU cardiologist of CHUC. For the bariatric patients, this set resulted from a literature review for the determination of the relevant features. Then, the input features of the LOS prediction models were selected from this initial set using the Kendall's tau coefficient correlation. Moreover, the selected input features for the cardiac patients were also validated by the cardiologist. The performance of the referred models, measured in terms of the geometric mean (GE) and F1 score, was determined by employing this final set of input variables to each one of them. Finally, through the application of the Friedman test and the corresponding post-hoc Nemenyi test, it was possible to order the models according to their performance.

The SCORE model performance was significantly low, achieving a geometric mean (GE) of 0.50. Thus, although this risk tool is of high importance in the European cardiology practice, it isn't sufficiently accurate to estimate the actual LOS. The second approach (Black-box model) outperformed the previous model. The best results were achieved by the multilayer perceptron with a GE of 0.62 ± 0.03 for the cardiac patients and 0.64 ± 0.08 for the bariatric ones. Furthermore, we verified that the performance of the interpretable and customized model was higher than the Black-box model, for both types of patients,

obtaining a GE of 0.66 ± 0.02 for the cardiac patients and 0.83 ± 0.05 for the bariatric patients. Moreover, the addition of the vital signs to the prediction models was proved to be advantageous since it leaded to an increase of performance in all the classifiers. These results suggest that the incorporation of dynamic data in LOS prediction models is worthy of further exploratory studies.

The analysis of the results allowed us to conclude that, although acceptable, the performance of the developed models does not seem to be adequate for their use in clinical practice (maximum GE of 0.66 and 0.83 for the cardiac and bariatric patients, respectively). This fact may be justified by the difficulty and complexity that the problem presents. The study of other variables, not only determined at admission time, but during the first hours or on the first day of the patient's stay, could be a strategy to explore in the future.

Keywords: Length of stay, Prediction, Computational intelligence models, Interpretability, Prognostic risk tools, Vital signs

Acronyms

ACS Acute Coronary Syndrome.
ANN Artificial Neural Network.

CABG Coronary Artery Bypass Graft.
CHUC Coimbra Hospital and University Center.
CISUC Center of Informatics and Systems of the University of Coimbra.
CKD Chronic Kidney Disease.
CVD Cardiovascular Disease.

ECG electrocardiogram.

GE Geometric Mean.

ICU Intensive Care Unit.

LOS Length of Stay.

MEWS Modified Early Warning Score.
MLP Multilayer Perceptron.

NSTEMI Non-ST-elevation Myocardial Infarction.

RF Random Forest.
ROC Receiver Operating Characteristic.

SCORE Systematic Coronary Risk Evaluation.
SE sensitivity.
SMOTE Synthetic Minority Oversampling Technique.
SP specificity.
STEMI ST-elevation Myocardial Infarction.
SVM Support Vector Machine.

VS Vital Signs.

WMC ward medication of corticosteroids.
WMS ward medication of sedatives.

List of Figures

List of Tables

Contents

1

Introduction

1.1 Context of the Project

The project of the present thesis was developed from a partnership between the University of Coimbra and Philips Electronics Nederland B.V. (Eindhoven), together with the collaboration of the Coimbra Hospital and University Center (CHUC) and the Catharina Hospital, located in Eindhoven. The development of the work took place at the Center of Informatics and Systems of the University of Coimbra (CISUC) and Philips facilities and the Length of Stay (LOS) was studied and implemented in two contexts: hospitalisation in a cardiac Intensive Care Unit (ICU), in collaboration with CHUC, and hospital admission for bariatric surgeries, in collaboration with Catharina Hospital.

1.2 Motivation

The major goal of hospitals is to provide quality healthcare and a good indicator of efficiency of care and hospital performance is the length of stay, as it can determine the hospital resource utilization [1] allowing an effective management of them.

Predicting the patients' LOS can give an overview of upcoming availability of beds, leading to improved scheduling of future admissions and surgeries and cost containment [1]. The LOS prediction is particularly important in Intensive Care Units because they admit a large number of patients and the resources are limited. Moreover, they are one of the most costly units of an hospital [2].

Some models have been developed specifically for LOS prediction, however, its number is limited when compared with other generic models, such as, for predicting complications, re-admissions or mortality. Therefore, although particular models for LOS prediction can and have been developed, the potential of models currently available for other purposes, such as the risk score tools, already used to aid in the clinical decision and validated in large sets of populations, may be explored in this context. This hypothesis arises from the evidence of the existing correlation between the mortality/re-admission risk and the patients' length of hospital stay [3][4]. However, despite the usefulness of these general models, the development of specific models for LOS is of major interest.

Moreover, the models previously mentioned are generally based on static data collected at admission time, corresponding to the patients' demographics and medical history. This kind of information doesn't represent the patients' health evolution (trends), contrarily to dynamic data collected over time (like Vital Signs). These trends are essential when the clinicians evaluate patients [5], suggesting that these dynamic parameters can be used to integrate and potentially optimize existing LOS prediction models.

On the other hand, current specific models developed for the LOS prediction task are mostly data-driven models, commonly based on data-mining techniques due to their high performance consequent of their ability to find patterns in the data [6]. However, they are usually not interpretable (black-box models) and not customized to the patient since they are applied equally to all patients. In this sense, an alternative capable of improving the interpretation and customization of the models would be beneficial in order to increase the clinical relevance and the clinicians' confidence, and so their acceptance [7].

1.3 Contextualization

1.3.1 Length of stay prediction

The length of hospital stay refers to a single episode of hospitalization and is typically defined as the number of days that the patient is hospitalized in a medical facility [1][8]. This parameter is frequently used to assess the planning and management of the hospital resources and the consequent health cost. By determining the usage's scheduling of wards and other hospital resources, it allows an effective scheduling of upcoming admissions and surgeries. In addition, a proper prediction of LOS can also indicate the patient's severity of illness and be used to help in the development of an effective clinical pathway [1]. Therefore, LOS prediction models are essential to assist hospital administrators as it can be used to optimize their long term strategic planning, and to support clinicians by providing a decision support tool. Patients are also benefited considering that an adequate LOS prediction allows long-term care and discharge activities planning, contributing to their quality of care [9].

1.3.2 Risk Models for LOS prediction

There are specific prognostic scores available in the clinical practice that have been developed and validated in the cardiovascular context. These scores address the primary, as well as the secondary prevention domain being risk assessment models for, respectively, long-term (years) and short-term (months) prediction periods [10].

Some studies attempted to predict the length of stay based on those prognostic scores [11][12][13]. However, there is a wide variety of short term prognostic scores such as: **i)** The Global Registry of Acute Coronary Events (GRACE) risk model [14] that calculates the probability of dying of an ACS while in hospital and at 6 months after admission,

ii) the Acute Decompensated Heart Failure National Registry (ADHERE) Algorithm [15] that estimates the in-hospital mortality in admitted patients with acute decompensated Heart Failure, **iii)** the Emergency Heart Failure Mortality Risk Grade (EHMRG) [16] that estimates a 7-day mortality of emergency Chronic Heart Failure patients and **iv)** Simplified Acute Physiology Score (SAPS II) that estimates the probability of mortality for ICU patients [17]. There are also long term risk assessment scores like **v)** FRAMINGHAM [18], that estimates a 10 year risk of heart attack, **vi)** QRISK [19], that calculates a 10 year risk of developing a heart attack or stroke and **vii)** the SCORE risk model [20] that establish the 10 year risk of fatal Cardiovascular Disease (CVD). These risk models were already used for risk stratification in outcome comparisons in previous literature studies [12][21][22] and the clinical guidelines recommend the use of such prognostic risk scores to support the clinical decision. Therefore, although they have not been developed specifically for LOS prediction, this may indicate the potential of this type of prognostic scores for LOS prediction models.

1.3.3 Specific Models for LOS prediction

An effective prediction of the patients' LOS is of high importance in the healthcare and can be achieved by prediction models [8].

Different approaches and methods have been used with the aim to predict LOS. Arithmetic methods are based on the calculation of the average or the median length of stay but, they assume that the LOS is normally distributed and, due to the skewness of the LOS, that is not adequate. Statistic methods are based in the covariate's analysis (representing the patient's characteristics and external factors that can possibly predict LOS) which have been implemented through linear regression and logistic regression techniques [23]. More recently, numerous studies have been predicting LOS with a good performance using data mining techniques, such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) [24][1][25], due to their ability to discover patterns amongst the data.

1.3.4 Vital signs

It has been demonstrated by several studies that hospital patient's vital signs and its progression are linked with the patient's recovery. Vital signs scores (usually known as Early Warning Scores) like the National Early Warning Score of patients were demonstrated to be negatively correlated with the length of hospital stay [26][27]. These scores are used worldwide as tools to identify patients with clinical deterioration or with critical illness and are based on physiological readings, as systolic blood pressure, heart rate and body temperature [28].

Additionally, the analysis of vital signs trends is also an approach taken by some researchers to predict LOS, since they can demonstrate the patients health evolution, alerting for potential deterioration or recovery [29].

1.4 Goals

The goals of this thesis are organised in two sections: the general goals to be addressed in this work, and the specific goals, which defines the particular goals for each dataset.

1.4.1 General goals

In this study the main goal is to predict the patients' length of stay at hospital admission through the implementation of four different approaches:

1. Exploration of available risk tools (SCORE risk)
2. Application of typical computational intelligence models available
3. Development of an interpretable and patient customized model
4. Integration of dynamic data (vital signs) in the previous models

The performance of all models will be compared applying the adequate statistical tests, validated with two datasets: the ICU dataset (provided by CHUC from Coimbra) and the TRICA dataset (provided by Philips from the Catharina Hospital in Eindhoven).

In the first approach, the ability of the SCORE to estimate the LOS will be evaluated. The SCORE is the European cardiovascular disease risk assessment model that establishes the total 10 year risk for fatal CVD. This risk model was selected to be explored, as a suggestion of Prof. Lino Gonçalves, given his relevance in the European cardiology practice.

In a second approach, a model based on computational intelligence techniques will be developed (Black-box model). For the ICU case, the feature baseline, i.e, the initial set of variables to be analysed for the models, will be supported on the clinical knowledge of an ICU cardiologist (Dr. José Pedro Sousa), combined with a literature review. In this way, all the selected features will be confidently validated as having clinical relevance. For the TRICA patients, the feature baseline will be based on a literature review and exploration of the data, since we don't have the clinical knowledge of the doctors in this context. In particular, due to their good performance in the LOS prediction literature, three specific machine learning algorithms will be explored: Random Forest (RF), SVM and MLP.

Furthermore, in order to evaluate the influence of the inclusion of vital signs parameters in the LOS prediction, the same model will be applied in a different dataset provided by Philips (TRICA dataset) containing information of bariatric patients. First, a model using only demographic features (static data) will be trained. In a second stage, the integration of the vital signs measurements (dynamic features) in representative features will be added to evaluate if this addition is beneficial and results in an improvement of the models' performance. Since the dataset of the cardiac ICU does not contain dynamic data it was not possible to validate this approach using it.

On the other hand, black-box models are usually not interpretable. Therefore, a new model based on rules created from the features will be developed (Rules model), allowing a better understanding of the results and contributing to an increase of its clinical relevance and acceptance. This approach was applied to both ICU and TRICA datasets.

1.4.2 Specific goals

With concern to the main objective which is to predict patients LOS, and consequently improve hospital workflow and patient healthcare quality, the goals of this thesis are the following:

1. LOS prediction in a Cardiac Intensive Care Unit

- Exploitation of the Coimbra's ICU database, comprised of 1544 patients admitted in the cardiac ICU at Hospital dos Covões (in Coimbra, Portugal), by developing prediction models using its data.
- Determination of the SCORE risk tool ability to assess the patients' length of stay
- Development of a Black-box model using different algorithms (SVM, MLP, RF) and with feature baseline obtained mainly through clinical knowledge
- Performance comparison between the different algorithms applied in the Black-box model and with the SCORE based model.
- Development of an interpretable and patient customized model based on rules, and performance comparison with the previous models.

2. LOS prediction in bariatric surgeries and influence of the incorporation of dynamic data

- Application of the Black-box model in the TRICA database, containing information of 189 bariatric patients submitted to surgery (gastric sleeve or gastric bypass), using only the patients demographic (DEM) features and secondly, with the Vital Signs (VS) features added, to evaluate if these dynamic data are a beneficial and relevant input to the model.
- Implementation of the interpretable model to evaluate if a similar or better performance is reached by comparing it with the previous model with adequate statistical tests.

1.5 Research Contributions

This project also contributed with one paper accepted in an international conference:

- Cláudia Lopes, Jorge Henriques, Paulo de Carvalho, Lino Gonçalves, Carolina Négrier, José Pedro Sousa and Alberto Bonomi, "The use of SCORE and GRACE risk tools to assess the length of stay in a cardiac intensive care unit", EMBEC 2020 - 8th European Medical and Biological Engineering Conference, Portorož, Slovenia (Accepted on 11th September to be presented in November, 2020).

1.6 Organization of the thesis

This document is structured in 6 Chapters. In Chapter 2 a background knowledge necessary for a better understanding of the work developed in the study, focusing both the clinical and the technical components, is presented. Chapter 3 presents the state of the art regarding the determination of the LOS predictors and the developed models for LOS prediction. Moreover, some studies regarding the correlation between Vital signs and LOS are reported. Then, in Chapter 4, the methodologies employed for the different LOS prediction models are represented, including the data processing, the discretization of both LOS and SCORE risk, and the feature design, followed by the procedures integrating the Black-box model and the Rules model. Afterwards, the Chapter 5 describes and discuss the results of the experiments. Finally, in Chapter 6 the main conclusions of the work are presented.

2

Background concepts

The background concepts presented below are organized in two sections: clinical component and technical component. The clinical component focuses the major clinical aspects that are relevant for this work: the SCORE risk tool, acute coronary syndromes and bariatric surgeries. The technical component presents the main concepts and techniques that support the development and validation of the prediction models.

2.1 Clinical component

2.1.1 SCORE risk tool

The implemented SCORE risk model [30] stands for Systematic Coronary Risk Evaluation and is a primary prevention tool applicable to the general population that estimates a 10 year risk for fatal CVD using 5 risk factors: gender, age, smoking status, systolic blood pressure and total cholesterol. This risk tool was endorsed by several European societies, including the European Society of Cardiology and the European Society of General Practice, and divided in two charts, the high and the low risk chart, respectively for the high and the low risk regions of Europe according to their background risk for fatal CVD.

In this work, since the data was collected from the Portuguese population, the low risk chart was employed [20], providing a quantitative score value in the range from 0 to 26. This risk value was then converted to a qualitative risk category (percentage of risk), comprising seven distinct classes determining the 10 year risk of fatal CVD: less than 1 %, 1%, 2%, 3% - 4%, 5% - 9%, 10% - 14% and 15% and over.

2.1.2 Acute Coronary Syndromes

As previously mentioned, we intend to estimate the length of stay of patients admitted to a cardiac ICU, in particular due to an Acute Coronary Syndrome (ACS). Acute Coronary Syndrome is the term that describes the conditions in which the blood supply to the heart is restricted or totally blocked [31][32]. This condition typically results from the accumulation of atherosclerotic plaques on the walls of coronary arteries, which are responsible for the distribution of oxygen and nutrients to the heart muscle (myocardium). The posterior

rupture of this deposit leads to the formation of blood clots that block the blood flow and can consequently cause the heart muscle cell death (myocardial infarction). Even without the rupture of the deposit, it may still reduce the blood flow, so although there is no cell death, the decrease in oxygen still results in cardiac distress, for instance under the form of chest pain (stable angina).

An ACS is a medical emergency that requires a fast diagnosis and adequate care. The leading symptoms of this condition are: chest pain (most common), namely pain or discomfort spreading from the chest to one or both arms, jaw, neck, back or stomach, shortness of breath, nausea, sweating and dizziness [32]. When patients arrive at the hospital with these complaints, they undergo medical history collection and physical examination, and two tests are performed, an electrocardiogram (ECG) and a blood test. The ECG measures the heart's electrical activity and the blood test can determine if there is cardiomyocyte death. The results of these tests allow the correct classification of the patients in one of the 3 types of ACS: ST-elevation Myocardial Infarction (STEMI), Non-ST-elevation Myocardial Infarction (NSTEMI) and Unstable Angina [33].

The rule-out diagnosis starts by an inspection of the ECG, that must be performed within 10 min after the first medical contact, and detection of two categories of patients. If a specific alteration, namely ST-segment elevation, particularly if encompassing a typical coronary artery territory, is identified, the patient is diagnosed with STEMI [33][34]. Contrarily, if no ST-elevation is observed, the patient is classified according to the complementary information given by the blood test results, more specifically, by the plasmatic measurement of a cardiac troponin, since they reflect myocardial cellular damage (necrosis of the cardiac tissue). An elevated value indicates that damage has, indeed, occurred to the heart muscle, therefore allowing the diagnosis of a NSTEMI. When a normal level is measured the diagnosis of Unstable Angina is, likewise, confirmed [35].

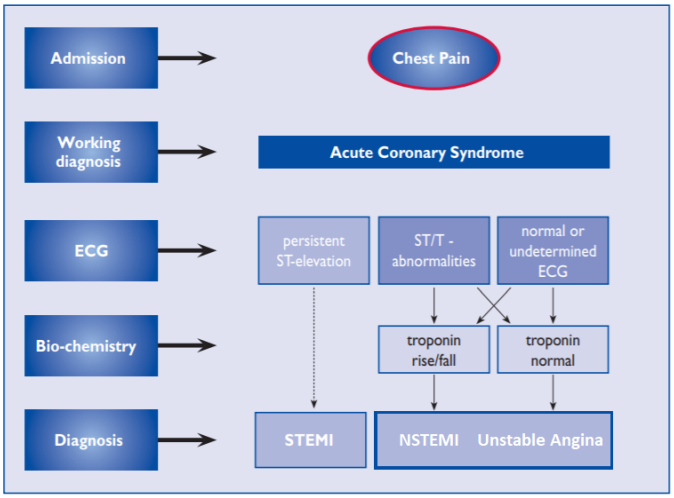


Figure 2.1: The spectrum of ACS and diagnosis steps. Figure adapted from figure 1 of [33].

The treatments are implemented according to the latest European Society of Cardiology guidelines [35][34]. The goals are to restore coronary blood flow and prevent or manage eventual complications, particularly heart failure and malignant arrhythmia.

In case of a STEMI, medical action must be immediate and should include performing an emergent cardiac catheterization. This procedure allows for the direct observation and treatment of the obstructed arteries consisting of a coronary angiography, to detect the obstruction location (diagnostic), typically followed by an angioplasty (treatment). In severe cases, as in the presence of obstruction of all three coronary arteries or an obstruction of the left main coronary artery, the patient is, instead, submitted to a cardiac surgery (Coronary Artery Bypass Surgery), frequently requiring further exams to rule out surgery contraindications. This preparation and the longer recovery process from the surgery typically lead to a LOS of at least 1 week.

When a NSTEMI is detected, the patient will, almost always, also qualify for invasive management, under the form of a cardiac cateterization. However, in this case, it is not emergent but urgent, being normally performed in 24 to 48 hours.

Before a diagnosis of unstable angina, a pharmacological anti-ischaemic therapy is given to the patient in order to decrease myocardial oxygen demand, or to increase myocardial oxygen supply [35][34]. If the symptoms and ischemic signs persist after this treatment the patient is selected for cardiac catheterization but, since it is not an urgent situation, it can be performed days later.

Perhaps more than in any other heart disease, an ACS is linked to several risk factors so effective prevention involves healthy lifestyle habits and the avoidance of those same risk factors (smoking, high blood pressure, high cholesterol, lack of physical activity, obesity, diabetes, etc) [32].

2.1.3 Bariatric surgeries

Bariatric surgeries are a solution for weight loss that works by reducing the stomach capacity to hold food [36].

Nowadays, the weight loss surgeries are performed using minimally invasive techniques. The most frequent bariatric surgeries are the Laparoscopic Roux-en-Y Gastric Bypass and Laparoscopic Sleeve Gastrectomy.

The Roux-en-Y Gastric Bypass consists in two procedures: stapling off the upper section of the stomach to create a small pouch that will be the 'new' stomach, and division of the small intestine. The bottom end of the small intestine is then connected to the newly created stomach pouch and the top portion connected further down to allow the stomach acids and digestive enzymes from the bypassed stomach to reach the food [36]. This way, the food goes into the new stomach and then directly into the small intestine that was sewn to it. This surgery leads to weight loss because the new stomach created can hold

smaller quantities of food and consequently, less calories are consumed. Additionally, it also promotes changes in the gut hormones reducing hunger and, by bypassing most of the stomach, the calories and nutrients absorption also decreases significantly.

In the Laparoscopic Sleeve Gastrectomy, approximately 80 percent of the stomach is removed resulting in a tubular pouch. Similarly to the gastric bypass, the amount of food and correspondent calories are also restricted and the hunger is reduced by changes in gut hormones [36].

The recent improvements of technology in medical procedures leaded to a decrease of the recommended length of stay. Nowadays, post-operative protocols aim to discharge patients 1 day after the surgery (first post-operative day) with studies showing reduction in length of stay without increasing complications. However, the predictors for early discharge after this type of surgery have not been clearly identified [37].

2.2 Technical component

To develop the proposed LOS prediction models some fundamental procedures need to be implemented. The first step consists in finding the relevant input features, that means, the features that are more correlated with the patients LOS (output). For that aim, the feature selection was performed with the implementation of the Kendall tau coefficient statistic. Afterwards, given the inputs with high potential, we need to compute the LOS output, through the development of the Black-box and Rules models that employ classification models (SVM, MLP, RF).

The performance of the classification models should then be evaluated using the most adequate metrics, that in this context, we determined to be the Geometric Mean (GE) and the F1 score.

Finally, in order to compare the models and conclude which one achieved the best performance, and therefore, is the most suitable for the LOS prediction of the analysed patients, two statistical test were employed: Friedman test and Nemenyi test.

2.2.1 Kendall's Tau coefficient statistic

In this work, the Kendall's tau correlation statistic will be employed to perform the feature selection for both cardiac and bariatric patients. This method was selected taking into account the type of variables involved in both datasets and the discretization of the length of stay in two ordinal categorical classes (Short and Long) [38]. The Kendall's test was preferred over the Spearman's Rank Correlation Coefficient (also employed in this type of variables) for being less sensitive to error and discrepancies in the data, and for having a more direct interpretation of the statistic [39][40].

The Kendall's Tau statistic is a non parametric test used to determine the degree of

association between two variables measured on at least an ordinal scale. It is also desirable the existence of a linear relationship between the variables when this test is implemented, however, this is not a strict assumption [41].

To the application of this correlation statistic, the data must be first ranked for further testing of the similarities in the ordering of the ranked data [42]. The null hypothesis states that 'There is no statistically significant relationship between both variables' and can be rejected at a given level of significance [43].

For this statistic two types of pairs are defined: concordant pairs (n_c), where the pair ranks are in the same order, following the same direction, and discordant pairs (n_d), when they are ranked in opposite directions.

The correlation coefficient comprises values from -1 to 1 being these limits obtained, respectively, when one order is the exact reverse, or the identical of the other order respectively [42].

There are three variations of the Kendall's Tau coefficient: Tau-a, Tau-b and Tau-c. In this work the Tau-b was selected since it takes ties into account by giving to tied observations the mean of the ranks they would have if they weren't a tie. This coefficient is calculated according to:

$$\tau_{\boldsymbol{B}} = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{2.1}$$

Where:

$$n_0 = n\,(n-1)\,/2 \tag{2.2}$$

$$n_1 = \sum_i \frac{t_i(t_i - 1)}{2} \tag{2.3}$$

$$n_2 = \sum_j \frac{u_j(u_j - 1)}{2} \tag{2.4}$$

Being n the number of items, t_i the number o tied ranks in the i-th group of ties for the first quantity and u_j the number o tied ranks in the j-th group of ties for the second quantity [44].

2.2.2 Classification models

Machine learning algorithms are being increasingly used for LOS prediction due to their capacity to learn how to predict the class of an entity based on a set of examples by searching for patterns. They are a subgroup of Artificial Intelligence, and can be categorized as supervised or unsupervised. Supervised learning models learn on labeled data, contrarily to unsupervised models, to which no corresponding output of each input variable is given

for the learning process. These models are heavily used in the health and medical domain in applications such as diagnosis and patient management [45] and more recently in LOS prediction [45] [24][1].

In this work three supervised algorithms were chosen, Random Forest, Support Vector Machine and Artificial Neural Network due to their good performance demonstrated in recent studies [24][1][46].

2.2.2.1 Random Forest

The Random Forest is introduced by Leo Breiman [47] as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest". It can be seen as an ensemble of randomly trained decision trees that are trained independently in the training phase. For the predictions during testing, it uses the Decrease of Gini Impurity as a splitting criterion and the final decision is obtained by aggregating all the predictions of each tree and selecting the class with most votes.

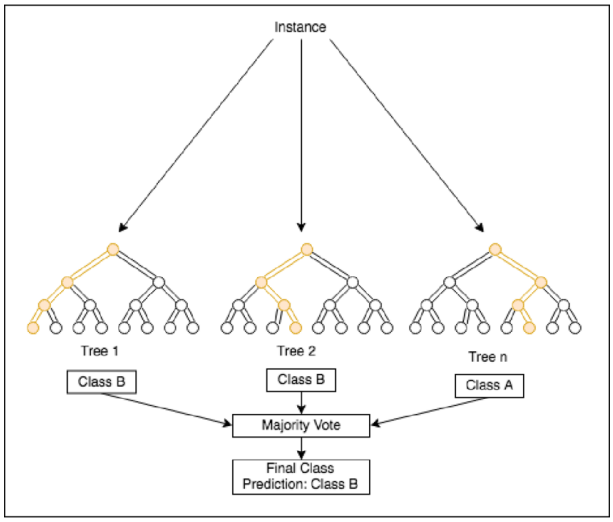


Figure 2.2: Random Forest diagram. Figure adapted from figure 4 of [48].

It has been proved that this method provides high accuracy, robustness to noise and stability [47]. Moreover, it can handle big data and missing values (replacing each one by the variable appearing the most in a particular node) and is a good technique to implement in imbalanced data [46].

2.2.2.2 Support Vector Machine

The Support Vector Machine is a supervised algorithm that can be used for classification or regression problems and is able to present good performance in classification of both linear and non-linear data, due to its capacity to capture complex patterns [49].

Each data point is plotted in a n-dimensional space where n represents the number of features, in which each feature value is regarded as a coordinate.

The aim of the SVM is to fit a hyperplane (decision boundary) that distinctly separates the data points in two classes by an optimization criterion possible due to a transformation called kernel trick. With this trick, the data points are projected into a higher dimensional space, where it might be easier to separate in two classes than using the original dimensional space. Then, the optimal hyperplane is the one which has maximum margin from the nearest points (support vectors) of all the classes [50].

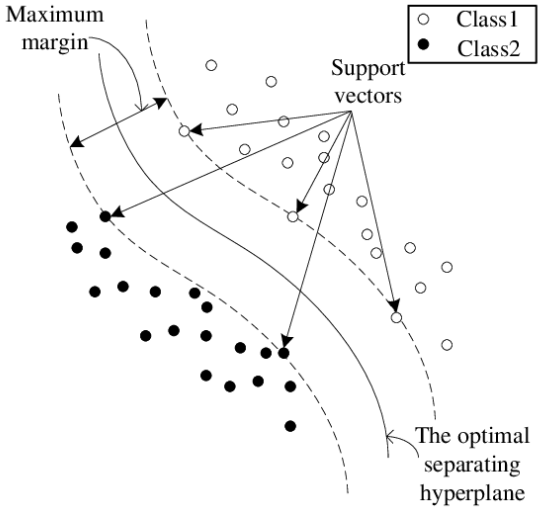


Figure 2.3: SVM diagram. Figure adapted from figure 1 of [51].

When the data is not linearly separable, non-linear kernels, that best fit the data, can be applied, such as: polynomial, radial basis function and sigmoid.

This is an algorithm of interest for the many advantages it presents. The classification task depends only on the number of support vectors rather than the input space domain (computationally efficient). This algorithm is also widely used when the distribution of the data is unknown because of the several alternatives for the choice of the separation threshold form (kernel types). In addition, the SVM solution determined is unique, since the optimization problem is convex. This is an advantage compared to other methods, such as Neural Networks (which have multiple solutions associated with local minima), thus, being more robust over different samples [52].

2.2.2.3 Artificial Neural Network

Artificial Neural Networks achieve great success with both linear and non-linear data and are one of the most popular methods used in medicine and many other fields [53].

They emerged through the combination of several units named perceptrons, that are a popular machine learning algorithms developed in the 1950s and 1960s by the scientist

Frank Rosenblatt [54].

As the name suggests, they try to simulate the behaviour of the brain neurons system by interconnecting layers of them, with each neuron having a respective weight, bias and activation function [5]. The activation function captures the non-linear relationship of the inputs and converts them in a more useful output. In the training phase the initial weights are chosen randomly and posteriorly adjusted in order to produce the desired output.

The ANN are a great tool for the prediction task because of the many advantages associated: robustness due to the use of weights even in noisy environments, low error rate, high degree of accuracy and improvement of performance with the learning ability in the training phase [53].

The most widely used type of ANN are the feedforward multilayer networks, also called multilayer perceptrons since they are effective for non-linear problems. They are composed of: one input layer, one output layer and at least one hidden layer, where the output of each element is computed layer by layer. Then, the difference between the output of the output layer and the target is back-propagated to the previous layers, with the adjustment of the connection weights [53].

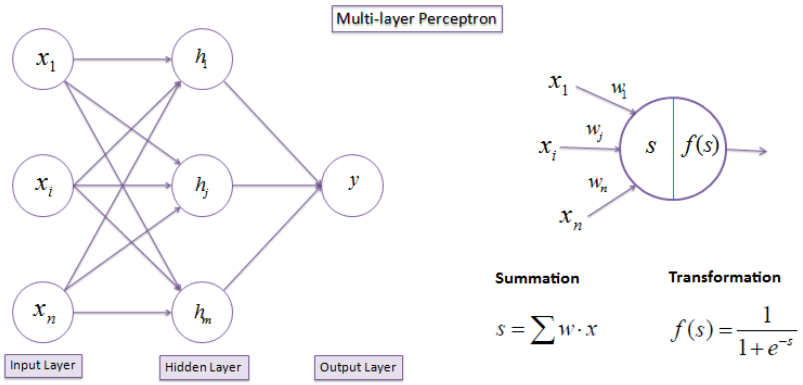


Figure 2.4: ANN and perceptron diagrams. Figure extracted from [55].

2.2.3 Performance assessment

Length of stay prediction models have been evaluated using different assessment metrics and, to select the most adequate ones, we must take into account the major goals of the health care centers and hospitals, in the context where the models will be implemented. It is also crucial to explore and understand the data. The LOS distribution is highly skewed and does not present a normal distribution, sometimes resulting in an class imbalance. Therefore, the metrics chosen must be suitable for a proper assessment in this context.

From the clinical perspective, clinicians are interested in knowing which patients will have the recommended LOS and which will stay longer. In this sense, and following the suggestion of our clinical partners, two classes were assumed $LOS = \{Short, Long\}$,

where *Short* represents the recommended LOS and *Long* an unexpected prolonged length of stay. Moreover, a misclassification of both (sensitivity and specificity) leads to serious consequences in terms of hospital workflow and costs.

With respect to specific metrics, the geometric mean (GE) is typically used as a "figure of merit" in the sense that it combines the sensitivity and the specificity, while trying to maximize the accuracy on each of the classes, thus, being adequate for non-balanced data. This metric is commonly used when both classes are equally important.

The GE consists in the squared root of the product of the sensitivity (SE) and specificity (SP) as demonstrated in equation (2.7).

$$SE = \frac{TP}{TP + FN} \tag{2.5}$$

$$SP = \frac{TN}{TN + FP} \tag{2.6}$$

$$GE = \sqrt{SE \times SP} \tag{2.7}$$

In the previous equations the TP (true positive) represents patients with extended LOS that the model classified as extended/long and the FP (false positive) is the recommended/short LOS incorrectly classified by the model as long. Similarly, the TN (true negative) identifies short LOS predicted by the model that are indeed short and the FN (false negative) represents the patients with long hospital stay that were misclassified as short.

The sensitivity is also called true positive rate since it measures how often the model made a correct positive prediction for people with long length of stay. The ability to correctly identify people with a short LOS is represented by the specificity, as it is the number of people classified with short LOS divided by the total number of people with short LOS. To measure the balance between those two metrics the GE aggregates them, making possible to capture the classes predictions correctness with the same importance.

On the other hand, when the classes distribution is highly unbalanced and the minority is the positive class, the most adequate metric is the F1 score. In the presence of an extremely low number of positive cases (extended LOS) the F1 score must be evaluated to determine if these few cases are detected (sensitivity) and, if the long LOS predicted by the model are correct (precision).

The F1 score is interpreted as a weighted average of the precision (PR) and recall, also

called sensitivity (SE), according to:

$$PR = \frac{TP}{TP + FP} \tag{2.8}$$

$$F1\ score = 2 \times \frac{PR \times SE}{PR + SE}. \tag{2.9}$$

The precision is the ratio of prolonged stays (positive cases) correctly predicted divided by the total number of the model's positive predictions.

2.2.4 Statistical Validation

Given the performance values of different models, the expected final goal is to determine, with confidence, which is the best model. For that, several statistical tests evaluate if there is a statistically significant difference between the performance of the models.

The tests for multiple comparison, like repeated-measures ANOVA (parametric) and Friedman (non parametric) tests are the most appropriate for the comparison of several models. Nevertheless, when working with machine learning algorithms, the assumptions of the ANOVA are highly probable to be violated, or not easy to meet, as mentioned by Demsar et al.[56] and Garcia et al.[57] respectively, therefore, the non parametric alternative is the most suitable.

On the other hand, whatever sampling method selected to obtain the train and tests sets from a dataset with a limited size (facing the problem of train and test proportion, as we need to use as much data as possible for the model training, and reserve enough data for the model evaluation, to obtain reliable performance estimations [58]), the sampling independence, fundamental assumption of statistical tests, will be violated, since there can be an overlap between the training and/or test sets. Therefore, the statistical tests implemented cannot be viewed as rigorously correct, but only approximate, heuristic tests, as referred by Pizarro et al.[59] and Dietterich et al.[60].

In this work we decided to test two different sampling strategies, the random sampling and the 30 repeated permutations followed by two-fold cross-validation employed by Pizarro et al.[59]. The first is presented in the following document and the second one, in Appendix A.

2.2.4.1 The Friedman Test

It is essential to perform a correct validation when comparing the results of different models since a minimum requirement must be overcome to assume that a model achieved a better performance than others.

The Friedman test is frequently used for a proper validation since it is a nonparametric test that is equivalent to the repeated-measures ANOVA (parametric test) when the assumption of independency, normality and homoscedasticity are not verified [61].

This statistical test is a rank based test that enables the comparison of several models by determining if there is a statistically significant difference among the several methods' performance. It starts by ranking the models according to their performances results (using the metric desired) in each sample: the algorithm with the best performance gets the rank of 1, the second best rank 2, etc. If two or more models have the same metric value, they receive an equal rank which is the mean of their ranks if they were ordered consecutively each by other [56]. Then, the average rank (computed from all the samples) of each model is compared according to:

$$R_j = \frac{1}{n} \sum^{i} r_i^j \tag{2.10}$$

$$X_F^2 = \frac{12n}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{2.11}$$

The variable n is the number of observed samples, k denotes the number of algorithms and r_i^j describes the rank of algorithm j on sample i. The Friedman statistic test null-hypothesis states that all the algorithms are equivalent and so their ranks R_j should be equal, and it is distributed according to a chi-square distribution X_F^2 with $(k-1)$ degrees of freedom when n and k are big enough (as a rule of thumb, $n > 10$ and $k > 5$). For a smaller number of algorithms and samples, there are tables of critical values defined that can be found in statistical books [56][62].

The Table 2.1 presents the critical values for the Chi-Square test.

Table 2.1: Critical values for Friedman's two-way analysis

Level of significance α		**0.10**	**0.05**
Degrees of freedom *df*	**1**	2.71	3.84
	2	4.61	5.99
	3	6.25	7.82
	4	7.78	9.49
	5	9.24	11.07

If the null hypothesis is rejected at the selected significance level (commonly $\alpha = 0.05$), the post-hoc Nemenyi test is implemented to compare all classifiers to each other, testing all pairs of them [63].

2.2.4.2 The Nemenyi Test

With the rejection of Friedman's null hypothesis it can be concluded that there is a difference between the algorithms, therefore, the next step is to do a pairwise comparison to identify which models outperformed the others. The Nemenyi test is adequate for this task and the comparison is made based on the average ranks computed in the Friedman test [61].

This test states that two methods are significantly different at a determined significance level ($\alpha = 0.05$ for example) if their average ranks differ at least the critical distance CD [56]. In other words, the CD defines the threshold to determine whether the performance between the algorithms is significantly different and is computed as:

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6n}} \tag{2.12}$$

The q_α is based on the Studentized range statistic divided by $\sqrt{2}$ and some of the values are given in Table 2.2:

Table 2.2: Critical values of the two-tailed Nemenyi test

Number of classifiers	2	3	4	5	6	7	8
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780

The results of the Nemenyi test can also be represented by means of critical difference diagrams [63].

3

Related work

As mentioned previously, LOS predictive models are an essential tool for hospital management and health care systems that help in the decision-making process. Therefore, several studies have been proposing different methods aiming to: **i)** define the LOS predictors (factors affecting LOS) using descriptive analysis - the analyse of data through statistical methods to find patterns or summarize the data in a meaningful way, and **ii)** estimate the LOS with predictive models.

Additionally, the association between LOS and vital signals was also addressed by some studies.

3.1 Length of stay predictors

Several methods have been proposed to determine the LOS predictors.

Sosyal Güvence et al. used a multinomial logistic regression to determine the factors affecting the LOS of patients with several diseases admitted in multiple hospitals in Turkey. The average length of stay had a statistically significant difference according to all independent variables used in the study including: age, gender, type of disease, type of hospitalization, presence of comorbidity, type and number of surgeries, season of hospitalization and geographic region of the hospital [64]. Arab et al. determined the factors affecting LOS in public hospitals in Lorestan Province, Iran. The researchers used the t-test, one-way ANOVA, and multifactor regression for the descriptive analysis. Age, gender, reason of admission, type of insurance and discharge status were proved to have impact on the average length of stay [65]. Aghajani et al.[66] determined the factors affecting the LOS in a general surgery department by extracting useful features found in literature. The set of features was then given to experts to be endorsed.

3.1.1 Cardiac patients

In the study of Wright et al.[67] peripheral congestion, concomitant acute medical problems requiring specific treatment, the development of renal impairment and the presence of social problems were found to be related to a longer than average LOS of heart failure

patients.

Pei-Fang et al.[25] performed the statistical analysis to study the relationship between the cardiac patient's characteristics and their LOS predictors using the Pearson's correlation coefficients. They found that the highest correlated risk factors were: history of heart failure, age and gender.

From a National Health Service hospital in Portugal, Teresa Magalhães et al.[68] developed a model (multiple logistic regression) to find the predictive factors on extended LOS of patients with Acute Myocardial Infarction using clinical and administrative data. In this study, patients with comorbidities were found to have increased risk of extended LOS. Wasfy et al.[69] selected the potential features for LOS predictors of patients with acute myocardial infarction based on clinical knowledge, ensuring that all the variables where known to clinicians at admission and available at admission time. The candidate variables included: age, gender, type of ACS, race, systolic blood pressure, race, smoking status and diabetes and other comorbidities.

3.1.2 Bariatric patients

In a multivariate analysis performed by Jonathan et al. [70] the longer hospitalization was predicted by diabetes, chronic obstructive pulmonary disease, bleeding diathesis, renal insufficiency, hypoalbuminemia, prolonged operating time, and resident involvement with the procedure, but not by patient age, sex, body mass index, and other comorbidities.

More recently, by a retrospective review of a Single UK Tertiary Centre Experience, Mahmood et al. [37] determined the factors predicting LOS following bariatric surgery. The patients integrating their study were discharged on post-operative day 1 and they found that undergoing Sleeve Gastrectomy and a BMI higher than 50 were associated with longer LOS (greater than day 1 post-operation discharge).

3.2 Predictive models

Regression methods like linear and logistic regression have been used to predict LOS [71]. However they were not considered successful methods as they don't take into account the skewness of the data. Verbug et al.[72] compared the predictive performance of 8 regression models (ordinary least squares regression on untransformed and on truncated at 3 days LOS , a generalized linear model with a Gaussian distribution and a logarithmic link function, Poisson regression, negative binomial regression, the original and recalibrated APACHE IV model and Gamma regression with a logarithmic link function) in an Intensive Care Unit and concluded that it is difficult to predict the length of stay using only patient characteristics at ICU admission time. The poor performance was also result of the skewness of the data and the high mortality rate of ICU patients.

Recent studies are using data mining techniques to predict LOS. In an Indian hospital,

Tanuja et al.[24] investigated the LOS prediction performance of four methods: a multi-layer neural network (MLP), a Naive Bayes classifier, the k-nearest neighbors algorithm (KNN) and a C4.5 decision tree, using elderly hospital eletronic discharge data. The MLP algorithm had better performance than the other three techniques. Aiming to predict the LOS of critically ill patients in the intensive care unit, Houthooft et al.[73] compared different machine learning algorithms and concluded that the best model was a support vector regression, achieving a mean absolute error of 1.79 days for those patients surviving a non-prolonged stay.

3.2.1 Cardiac patients

Lior Turgeman et al. applied a regression tree model (Cubist) using static inputs at admission time that do not change during hospitalization. The LOS was predicted for congestive heart failure patients with high interpretability but the Cubist error was not uniformly distributed [74]. They suggested to use a SVM to help separate the cases by their level of Cubist error.

In their study Pei-Fang et al.[25] developed an artificial neural network model to predict LOS for cardiac inpatients with one of the diagnoses: coronary atherosclerosis (CAS), heart failure (HF) and acute myocardial infarction (AMI). Pearson's correlation coefficients were used in the statistical analysis to explore the relationships between LOS and the patient's characteristics. The ANN predicted the continuous LOS values (0 to 35 days), and, for CAS patients the model accuracy was from 88.31% to 91.53% and for HF and AMI it ranged from 63.69% to 67.47% at the preadmission stage. The ANN was also proposed by Rowan et al.[75] to predict the LOS of postoperative cardiac patients due to its good discriminating ability (AUC of 0.819).

Rezaei et al.[1] divided the LOS in three classes and applied three classification algorithms: decision tree, support vector machines and artificial neural network. The SVM obtained a better performance, furthermore, the LOS of the cardiac patients tended to be longer in patients with lung or respiratory disorders and high blood pressure.

In late studies, Daghistani et al.[46] found that the variables with highest impact on the prediction of in-hospital LOS were on admission heart rate, on admission systolic and diastolic blood pressure, age and insurance status and the Random Forest outperformed the other machine learning models with a sensitivity of 0.80, accuracy of 0.80 and AUROC of 0.94.

On the other hand, some interpretable models have been developed to predict the outcomes in several domains of the cardiac scope, using different approaches, as fuzzy systems, decision trees and clinical prediction rules [76][77][78]. Recently, Simão et al. [10] created an interpretable approach for CVD risk assessment based on the identification of clinically relevant rules with encouraging results, reaching a geometric mean of 0.86 for all patients, and also appearing to have the potential to be implemented to predict the LOS of cardiac

patients.

3.2.2 Bariatric patients

Few studies were developed aiming to predict LOS of bariatric patients, and the ones found on literature, explored the ability of already used mortality risk scores for the LOS prediction. Gilhooly et al. [79] performed an evaluation of risk prediction models in predicting outcomes (including LOS) after bariatric surgery. They found that POSSUM (the Physiological and Operative Severity Score for the enumeration of Morbidity and Mortality) may be used to predict patients who will have prolonged postoperative LOS after bariatric surgery due to morbidity or poor mobility. Additionally, Miguel et al.[80] also tested the Obesity Surgery Mortality Risk Score (OS-MRS) to predict the length of stay of patients submitted to a gastric bypass surgery. They observed an association between the OS-MRS score and the LOS and they also verified that longer length of stay was slightly associated with longer surgery duration.

3.3 Vital signs and LOS

From a literature review, some studies have proved a relationship between the patients' vital signs parameters and scores, and their length of hospital stay. Paterson et al.[26] identified a relationship between the SEWS (Standardized Early Warning Score) and the length of stay where the median length of stay extended significantly for higher values of SEWS score. The National Early Warning Scores (NEWS) were also used in S. Barth et al. study [27] as a tool to determine patient's outcomes including mortality and length of stay. They suggested that there is an association between the NEWS at admission time and the patient's length of stay. Additionally, the NEWS measurements of different time points were proved by N. Alam et al.[81] to be good predictors of patient outcomes including LOS.

Moreover, vital signs' trends can also represent the evolution of the patients health state and help predicting their hospital stay. Brekke et al.[82] studied the respiratory rate as a predictor using, the current value and adding trend models, and this it was proved to be the most accurate when comparing to other vital signs. The improvements using trend were considered minor despite the statistically increased model accuracy.

3.4 Conclusion

From the findings of the state of the art we can verify the existence of several studies aiming to determine the LOS predictors for both cardiac and bariatric patients, using statistical methods and other strategies, such as the doctors' clinical knowledge. Comorbidities were found to be highly associated with the patients LOS, for both cardiac and bariatric patients.

On the other hand, the predictive models developed for the cardiac context largely exceed the ones targeting the bariatric patients (and the few studies found were based on risk scores). Moreover, from the explored classification algorithms, the best performances were reached by the MLP [24] [25], the SVM [1] [73] and the RF [46], outperforming other data mining techniques.

Furthermore, interpretable models have also been developed to predict the outcomes of cardiac patients with satisfactory results and providing an explanation of the reasoning behind predictions [7].

In addition, the relationship between the patients' vital signs and their length of stay was also proved to exist, being worth more exploration, and motivating the inclusion of vital signs in LOS prediction models.

4

Experimental setup

4.1 Data

In the present work two datasets were used, the Cardiac ICU dataset and the TRICA dataset, both preprocessed by removing the patients that died during hospitalization.

4.1.1 Cardiac ICU Data

The ICU dataset is comprised of 1544 patients admitted between 2009 and 2016 at the Cardiac Intensive Care Unit of the Hospital dos Covões located in Coimbra with one type of Acute Coronary Syndrome. The table with the baseline characteristics of CHUC dataset used in this work is presented in Appendix D. It contains information about the patients including demographics, comorbidities, complications induced by the ACS and laboratory data. The average LOS is 4,92 days, however, patients typically stay 3 days (mode).

Figure 4.1 depicts the histogram of the patient's LOS and it can be observed its asymmetric distribution: a relatively small number of patients with long LOS exists on the right side of the graph (particularly for periods greater than 15 days) and a high concentration of patients is observed on the left side.

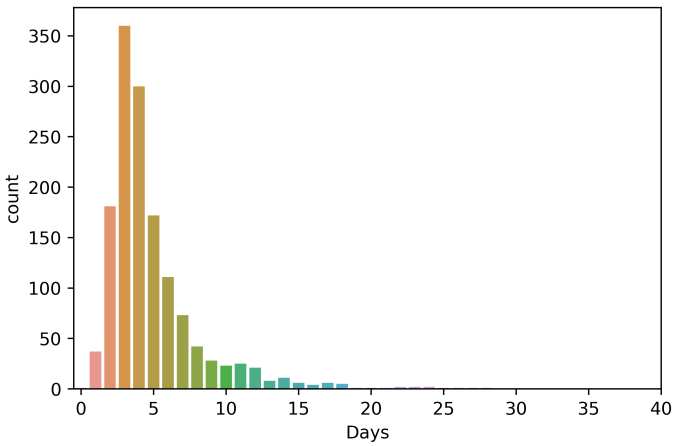


Figure 4.1: Length of stay distribution of the cardiac ICU patients.

4.1.2 TRICA Data

The TRICA study dataset was provided by Philips, resulting from a non-randomized, observational, single-center study conducted at the Department of Anesthesiology of Catharina Hospital in Eindhoven, in the Netherlands. The primary objective was a technical validation of offline metrics, using data collected from the ELAN device in the perioperative period as input (photoplethysmography and accelerometer data). The validation was performed in different groups of patients submitted to surgery requiring anesthesia, such as bariatric surgery, esophagectomy and pancreatectomy.

The dataset contains data from 350 patients, however, in this work only the bariatric patients were selected to be explored. Therefore, 190 bariatric patients that were admitted to surgery: gastric bypass or gastric sleeve, between May and December of 2019 were employed in this study (baseline characteristics presented in Appendix E). Besides the patients' demographics, comorbidities, ward medication and other static data, this dataset also contains periodic vital signs measurements, starting at the surgery day and collected until the patient is discharged. These signs consist of 23 variables and are obtained in a frequency of 30 seconds during the surgery time and less periodically after that.

The LOS distribution is represented in Figure 4.2 demonstrating the heavy imbalance in the data, since the majority of the patients stay only 1 day (162 patients corresponding to 85 %).

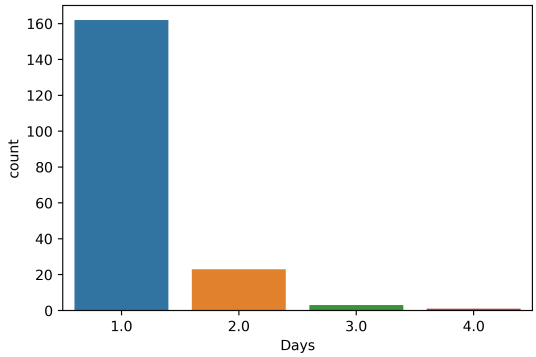


Figure 4.2: Length of stay distribution of the bariatric patients.

4.2 Discretization

The data discretization is the process of transforming continuous data to a discrete one, reducing large domains of numeric values to a smaller set of classes.

4.2.1 Discretization of the LOS

From the clinical perspective it is not realistic the exact estimation of the length of stay since it is hard to trust and rely in a model which predicts the exact number of days a

patient will stay when he arrives at the hospital. It is more relevant to know if a patient will have the recommended length of stay in the context it is admitted, or if the hospital stay will be extended and the adequate consequent measures should be implemented. This perspective was confirmed by the ICU specialist who provided the clinical support for this study. In this sense, the strategy considered was to assume only two categorical classes $LOS = \{Short, Long\}$ where *Short* represents the recommended LOS and *Long* an extended LOS. The threshold day for the discrimination between the two classes was chosen accordingly to the clinical context the patients were admitted.

In the Cardiac Intensive Care Unit, the patients were hospitalized with one type of Acute Coronary Syndrome: STEMI, NSTEMI or unstable Angina. The European recommendation for the length of stay of patients admitted with this condition is 3 days, therefore, this was the threshold defined for the distinction between *Short* and *Long* stays, equation (4.1). Usually, patient stays that exceed this value are the most serious/critical cases.

$$LOS = \begin{cases} Long \text{ if number of days in ICU } > 3 \\ Short \text{ if number of days in ICU} \leq 3 \end{cases} \tag{4.1}$$

On the other hand, the bariatric patients are admitted in the hospital to be submitted to a bariatric surgery and the normal LOS for this type of patients is one day, since they are usually discharge the day after surgery. Hence, the selected threshold day is 1 day.

$$LOS = \begin{cases} Long \text{ if post-operative discharge day } > 1 \\ Short \text{ if post-operative discharge day } = 1 \end{cases} \tag{4.2}$$

4.2.2 SCORE risk tool for LOS estimation

The initial approach implemented to estimate the LOS was the implementation of the SCORE risk tool. The first step was the calculation of the *SCORE value* using the information collected from patients at hospital admission, consisting in 5 risk factors: gender, age, smoking status, systolic blood pressure and total cholesterol. By implementing these risk factors in the SCORE risk tables (presented in Appendix C) the *SCORE value* was obtained and the respective *SCORE class* was determined according to:

$$SCORE\,class = \begin{cases} 0 \text{ if } SCORE \text{ value} = 0 \\ 1 \text{ if } SCORE \text{ value} = 1 \\ 2 \text{ if } SCORE \text{ value} = 2 \\ 3 \text{ if } SCORE \text{ value} \in \{3,4\} \\ 4 \text{ if } SCORE \text{ value} \in \{5,6,7,8,9\} \\ 5 \text{ if } SCORE \text{ value} \in \{10,11,12,13,14\} \\ 6 \text{ if } SCORE \text{ value} \geq 15 \end{cases} \tag{4.3}$$

Each *SCORE class* establishes the 10-year risk level of developing fatal CVD. To perform the correlation assessment with the LOS, the *SCORE class* was divided in two classes $Risk = \{Low, High\}$ defined as:

$$Risk = \begin{cases} High\ if\ SCORE\ class \geq C_{threshold} \\ Low\ if\ SCORE\ class\ < C_{threshold} \end{cases} \tag{4.4}$$

Then, a simple rule could be used to estimate the LOS, defined as:

$$\text{IF}\ \ Risk = High\ \ \text{THEN}\ \ LOS > 3\ days \tag{4.5}$$

In order to find the best $C_{threshold}$ (the optimal value for the discrimination of the *SCORE class*), an exhaustive search was performed testing all the possible $C_{threshold}$ values {1,...,6} in order to maximize the GE value.

This first analysis relied on the assessment of the capacity to distinguish the LOS of all the patients integrating the dataset, therefore, no usual process of validation/testing was carried out.

Later, we also evaluate the SCORE ability to distinguish the LOS of different test subsets, in order to compare with the other developed models.

4.3 Features Design

The features design involves two steps: identification of all the potential features (feature baseline) and the posterior selection of the most relevant ones (feature selection).

4.3.1 Feature baseline

4.3.1.1 ICU Dataset

Aiming the development of the predictive models based on computational intelligence, the features baseline was defined through the clinical knowledge of the ICU cardiologist who guided this work in the cardiac context. In addition to the variables proposed by the doctor, a literature research was performed and the selected features were also validated by him. The final feature baseline was divided in 6 groups: demographic variables, co-morbidities, type of acute coronary syndrome, presence/absence of surgery, complications induced by the ACS, laboratory variables and systolic blood pressure, consisting in a total of 19 variables. The following table depicts the features that integrate each group and the clinical reason for their selection.

Table 4.1: Feature baseline divided by features groups, with the correspondent clinical reason of selection and variable type

Variable	Clinical reason	Variable type
Demographic variables		
Age	Older patients stay longer due to comorbidities associated with age	continuous
Gender	Women live longer than men, so they have ACS in a later stage, moreover they have higher probabilities of bleeding leading to longer LOS than male gender	binary
BMI	Overweight patients have more complications associated and consequent extended LOS, additionally, underweight represents fragility due to low physiological reserve	continuous
Comorbidities		
Myocardial infarction		binary
Congestive heart failure		binary
Peripheral vascular disease	Several associated complications	binary
Cerebrovascular Accident		binary
Diabetes		binary
Chronic kidney disease		binary

Variable	Clinical reason	Variable type
Type of ACS		
Type of ACS (STEMI, NSTEMI, Untable Angina)	Each type of ACS has its recommended LOS because of the respective procedures and complications	discrete
Presence/Absence of surgery		
Surgery (Yes/No)	Patients submitted to a Coronary Artery Bypass Graft (CABG) need to stay longer since the time required for the pre-surgical study and recovery is prolonged	binary
ACS induced complications		
Killip class	Quantifies the severity of the heart failure	discrete
Laboratory variables		
Albumin	Fragility marker	continuous
Creatinine	Renal function marker	continuous
Troponin	Indicates the severity of the infarction	continuous
C-reactive protein	Inflammation marker	continuous
Glycose	Important for the patients with diabetes since both low and high values lead to bad prognostic	continuous
Hemoglobin	Indicates possible existence of anemia	continuous

Variable	Clinical reason	Variable type
Systolic Blood Pressure		
SBP	Patients with hypotension or hypertension tend to have prolonged LOS	continuous

4.3.1.2 TRICA Dataset

For the bariatric patients of the TRICA dataset it wasn't possible to have the doctors' clinical knowledge on the bariatric surgery department. Therefore, the strategy was to perform a literature review and explore the data (containing 260 parameters) to define the feature baseline by focusing on the variables that could give relevant information for the problem and removing the ones that weren't related with the problem.

With that aim, the features with more than 75% of patients with missing values were removed and the remaining variables selected were related to the patient's demographics, medical history, comorbidities, medication administered during hospitalization and complications during surgery. As a result, a total of 58 features were identified as the baseline.

4.3.1.3 TRICA - Vital Signs features

The TRICA dataset was also composed by vital signs measurements of the patients, that were collected during surgery (with a frequency of 30 seconds) and in the post operative days (less frequently and not periodically) until the patient was discharge.

To evaluate the influence of adding the vital signs information in length of stay prediction models, 4 of the 23 vital signs variables were selected: heart rate, respiration rate, oxygen saturation (SpO2) and the Modified Early Warning Score (MEWS), since their measurements were collected more frequently and in more patients than the remaining signs. The MEWS is not a vital sign, however it is a score representative of 4 physiological readings (systolic blood pressure, heart rate, respiration rate and temperature) and the patient's level of consciousness, that identifies patients who are at risk of clinical deterioration and who may require a higher level of care. The higher the score the more serious the patient's condition [28].

From the time-series data of each vital sign parameter, we only focus on the first 24 hours after the end of the surgery, since the goal was to integrate these features in a LOS prediction model that uses data from the admission day and first post operative day to perform the estimations.

Then, for each selected vital signs, a set of 4 features were extracted: variance, mean value, slope and a binary feature, indicating if the mean value was in the normal range (according to values found on literature). The variance was calculated in order to verify how stable the vital parameter was during the hospital stay and the slope extracted to represent its evolution (trend). The mean value represents the overall value during the 24 hours. Concerning the MEWS, only 2 features were extracted: mean and maximum value.

With this procedure a total of 14 new vital signs features were added to the initial set (performing a total of 72 features as baseline) and are represented in the next table.

Table 4.2: Vital signs features developed from the heart rate, respiration rate and SpO2 periodic measurements

Feature	Variable type
Heart rate features	
Mean heart rate	continuous
Heart rate slope	continuous
Heart rate variance	continuous
Mean heart rate in/not in normal range	binary
Respiration rate features	
Mean respiration rate	continuous
Respiration rate slope	continuous
Respiration rate variance	continuous
Mean respiration in/not in normal range	binary
SpO2 features	
Mean SpO2	continuous
SpO2 slope	continuous
SpO2 variance	continuous
Mean SpO2 in/not in normal range	binary
MEWS features	
Mean MEWS	discrete
Maximum MEWS	discrete

4.3.2 Feature selection

In order to evaluate the correlation between the variables composing the feature baseline and the patients' length of stay, the Kendall's Tau correlation statistic was applied. This strategy was employed to select the most relevant features of both cardiac and bariatric patients. In this way, the final set of features to be given as input for the models will contain only the best LOS predictors. The selected features needed to achieve an absolute value of the tau coefficient higher than 0.1 and a p value < 0.01.

The ICU dataset was divided in two sets: the train set, corresponding to 70 % of the data, and the test set, composed of the remaining data. The Kendall's tau statistic was then applied in the train set and the final set of selected features was validated by the clinical specialist who supervised the work.

Concerning the TRICA dataset, due to the nonexistence of the clinical knowledge in the context of the bariatric patients, two additional feature selection methods were carried out: Random Forest and Logistic Regression, to verify if they were a best alternative to select the features (however we later concluded that they weren't advantageous).

Moreover, for the bariatric patients, due to the high data imbalance we couldn't selected the adequate features only in one train set of patients. For that reason, 20 different train sets were generated and the features selected were the most common ones obtained in all the train sets.

4.3.3 Feature discretization

In this work, the feature discretization was employed in the continuous features to take advantage of the interpretability gain since it leaded to the creation of rules sets that could be understood by the clinicians and serve as foundation for the LOS prediction results.

4.3.3.1 Cardiac patients

With this purpose 3 different discretization strategies were employed when working with the cardiac patients.

The first discretization was made by applying the normal ranges' limits of each feature, according to the standard reference intervals of the Clinical Pathology Service of the CHUC, and conforming the values found in the literature. For example, according to the European Society of Cardiology, the systolic blood pressure (SBP) of elderly people is considered normal [83] if:

$$90 \text{ mmHg} \leq \text{SBP} < 140 \text{ mmHg} \tag{4.6}$$

On the other hand, the knowledge and experience of the doctor, that interacts with the ICU patients daily and evaluates their characteristics and health state (such as demographic and laboratory data) could be seen as an essential insight, for a discretization aligned with the decisions that are taken in the clinical practice. For that reason, the second discretization of the features was done using the thresholds suggested by the doctor who guided this work.

Finally, the third approach uses the Kendall's tau results. Since there was a possibility that the Kendall's tau coefficients obtained with the two previous strategies could be low, a last approach was adopted in order to determine if higher coefficients values could be found taking advantage of the data (data-driven approach), using computational methods. In this approach the variables were discretized by performing an exhaustive search, which tested all the possible features values as cut off points, aiming to maximize the Kendall's Tau coefficient.

The best discretization approach (the one achieving the highest Kendall's tau coefficient values) was then selected to be implemented in the features for the posterior procedures.

4.3.3.2 Bariatric patients

Regarding the TRICA patients, the feature discretization strategy employed was the one with the best results in the ICU dataset, the data-driven method. Nonetheless, some modifications were applied as a result of the strategies implemented to overcome the high data imbalance problem. Since the features were selected using different training sets (selecting the most common ones) if we only used one train set to find the optimal thresholds for the selected features, not all of them would have got a p value < 0.05, rejecting the null hypothesis, in that specific set.

For that reason, 100 different train sets were generated, considering only the features selected previously, and the Kendall's tau coefficient maximization procedure was employed only on the features that were good predictors (rejected the null hypothesis) in each train set. Then, the cut off value (optimal threshold) of each parameter was obtained as the mean value of all the threshold values achieved in each one of the train sets.

4.4 Black-box model

4.4.1 Methodology

The next figure depicts the steps involved in the Black-box model after the feature selection and discretization: classification models hyperparameters' tuning, determination of the optimal cut point of the Receiver Operating Characteristic (ROC) curve and LOS assessment.

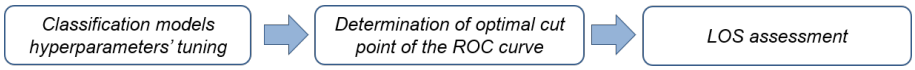


Figure 4.3: Steps integrating the Black-box model

4.4.2 Classification models hyperparameters' tuning

With the adequate features selected and discretized, the next step was to predict the patients LOS by means of classification models.

We implemented three classification models intending to determine the best one for the LOS prediction task. The selected algorithms were the Random Forest, the SVM (with a regularization parameter of C=1) and the Multi-layer Perceptron (that trains using Back-propagation) since they stood out in the literature review for achieving better performances in comparison to other models.

Hyperparameter tuning or optimization is the process of choosing a set of optimal hyperparameters that maximizes the model's performance. With this aim, a brief tuning of the 3 classification models was performed to obtain the adequate hyperparameters for the input data and for the purpose of the models application. This tuning was performed through the implementation of a cross-validated grid-search during the training phase of the algorithms.

The grid search consists in an exhaustive search through a manually specified subset of the hyperparameter options of a selected algorithm. By iterating through every parameter option and storing a model for each one, when it finishes testing all the options, it provides the combination which produced the highest value of the scoring metric that was previously specified.

The goal of the Black-box model is to predict the patients' LOS with the maximum GE to ensure that both the long and short LOS are correctly detected. Thus, as mentioned in Chapter 2.2.3, the scoring parameter defined for the metric which will be evaluated for the models optimization is the GE.

The following table depicts the hyperparameter settings of each algorithm to be optimized: the kernel type of the SVM, the MLP hidden layer size and the number of estimators of the RF.

Table 4.3: Algorithms hyperparameters settings for grid search

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer size	number of estimators
Settings	poly, linear, RBF	6, 8, 10, 12, 14	10, 50, 100

Moreover, to deal with the data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed in each training dataset prior to fitting a model, and, as the name suggests, it generates new examples from the minority class.

The SMOTE selects a random example from the minority class and finds n of the nearest neighbors, usually n=5. Then, a randomly selected neighbor is chosen and a new synthetic example is created by selecting a random point from a line segment defined between the two examples in the feature space. This technique was chosen because is one of the most widely used approach to synthesize new examples [84].

4.4.3 Determination of the optimal cut point of the ROC curve

With the proper models hyperparameters defined, the ROC curve of the test set was calculated.

The ROC curve is a probability curve that works as a performance measurement for classification models, at various thresholds settings for the predicted probabilities of the algorithms, plotting the true positive rate (y-axis) against the false positive rate (x-axis).

For each observation in the data, a binary classification is given (short or long LOS) that results from the transformation of the predicted probability (a continuous value between 0 and 1), outputted from the classification algorithms, into a binary response, using a cut-off point. If the predicted probability exceeds the chosen cut-off point, the classification is positive (long LOS), otherwise, the predicted class is a short LOS [85].

In order to find the point that gives the maximum correct classification, we determined the optimal cut-point, by maximizing the Youden function which is the difference between true positive rate, also called sensitivity, and the false positive rate (1-specificity) over all possible cut-point values of the curve[86]. The Youden index is defined as:

$$J(c) = \{SE(c) + SP(c) - 1\} = \{SE(c) - (1 - SP(c))\} \tag{4.7}$$

Where SE denotes the sensitivity and SP the specificity.

With this function, the point on the ROC curve which is farthest from line of equality (diagonal line) is considered the optimal cut-point because it is the point that optimizes the model's differentiating ability when equal weight is given to sensitivity and specificity.

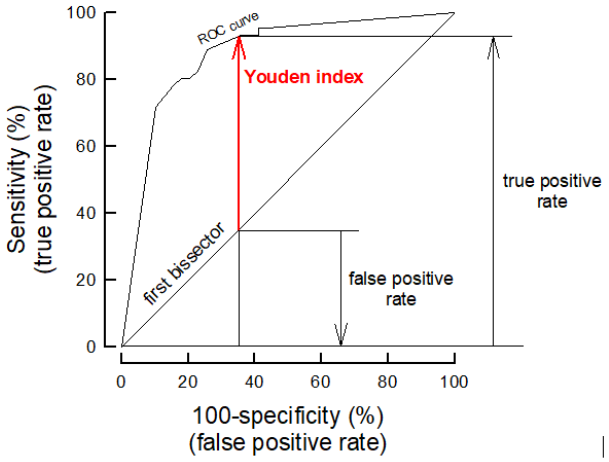


Figure 4.4: Youden index representation in ROC curve. Adapted from Figure 3.4 of [87]

4.4.4 LOS classification

Having the optimal cut point to discriminate between the short and long LOS it was possible to execute a proper LOS classification of the test set patients using this value.

In order to evaluate the model's performance, it was employed in 20 different train and test sets obtained by randomly splitting the data (70% of the patients for the train set and the remaining for the test set). Additionally, we also implemented the 30 repeated permutation followed by two-fold cross-validation, employed by Pizarro et al.[59] (that also implemented non parametric tests to compare multiple models in a single dataset), presented in Appendix A, and the obtained results and conclusions were equivalent.

4.5 Rules model

One major disadvantage of black-box models is their lack of interpretability. This is of extreme importance for the clinical environment since interpretable models allow a better understanding of the results and a higher clinical relevance and acceptance, by providing an explanation of the reasoning behind predictions [7][10]. In addition, patient customized models would also be beneficial in the decision making process and it is not provided by the previous Black-box model. Regarding these motivations, an additional model was developed through the creation of rules for each feature. This rule based approach was recently developed by Simão et al. [10] integrating the CISUC group to which we belong. The aim of their study was the CVD risk assessment of acute coronary syndrome patients, however, the encouraging results, motivated the implementation and validation of an approach similar to theirs (that we defined as Rules Model) in this context of LOS prediction.

4.5.1 Methodology

The next figure depicts the steps involved in the Rules model after the feature selection and discretization: derivation of rules, classification models hyperparameters' tuning, selection of rules, determination of the optimal cut point of the Precision-Recall curve and LOS assessment.

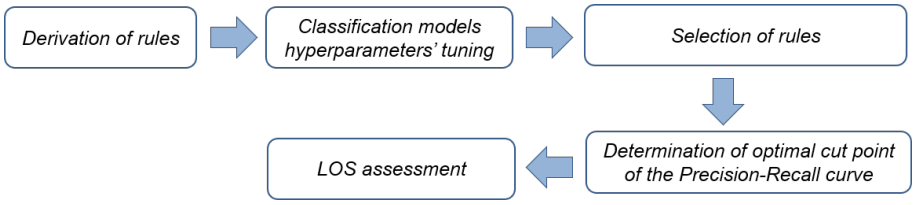


Figure 4.5: Steps integrating the Rules model

4.5.2 Derivation of Rules

The first step for the derivation of rules is to obtain, for each feature, the optimal discriminative value ($F_{threshold}$) capable of achieving the best separation between the two LOS classes (short/ long). This is accomplished by applying the data driven discretization, obtained by maximization of the Kendall's tau coefficient previously referred in chapter 4.4.2. Each feature best cut-off value establishes the threshold for the correspondent decision rule according to:

$$\text{IF } \textit{Feature value} \geq F_{threshold} \text{ THEN } \textit{LOS} = \textit{Long} \tag{4.8}$$

This approach was applied to both continuous and categorical variables and allows to maximize the interpretability of the model. For the binary features the decision rule was directly defined based on the clinical consequences originated by the presence/absence (1/0) of the feature (risk factor). Giving a matrix $Xinputs_{(N\times M)}$ representing the features values (M) for all patients (N), the rules were derived with the application of the best separation threshold of each feature as described in equation (4.8). The joining of all rules for all the patients generated the $RulesOutcome_{(N\times M)}$ matrix, composed by binary values where 0 and 1 represent Short and Long LOS respectively (step 1 of Figure 4.6).

4.5.3 Selection of Rules

The combination of all the rules derived may not be advantageous because not all of them can be adequate for each patient and contribute positively for the patient's LOS classification. For this reason, only the best rules for classifying each patient LOS were applied. With that purpose, an intermediate step was performed, aiming the identification of the adequate rules (correct rules) to classify each patient.

The comparison between the $RulesOutcome_{(N\times M)}$ matrix and the actual length of stay of the patients $C_{(N\times 1)}$ enables to conclude which rules were correct for each of them. If the rule outcome corresponded to the actual value of the patient LOS it was considered as correct, otherwise, classified as incorrect. With this operation a matrix containing the rules that classified correctly and incorrectly each patient's LOS ($RulesClassification_{(N\times M)}$ matrix) was obtained (step 2 of Figure 4.6). In this sense, for a specific patient, a value of 0 represented an incorrect rule and 1 a correct one.

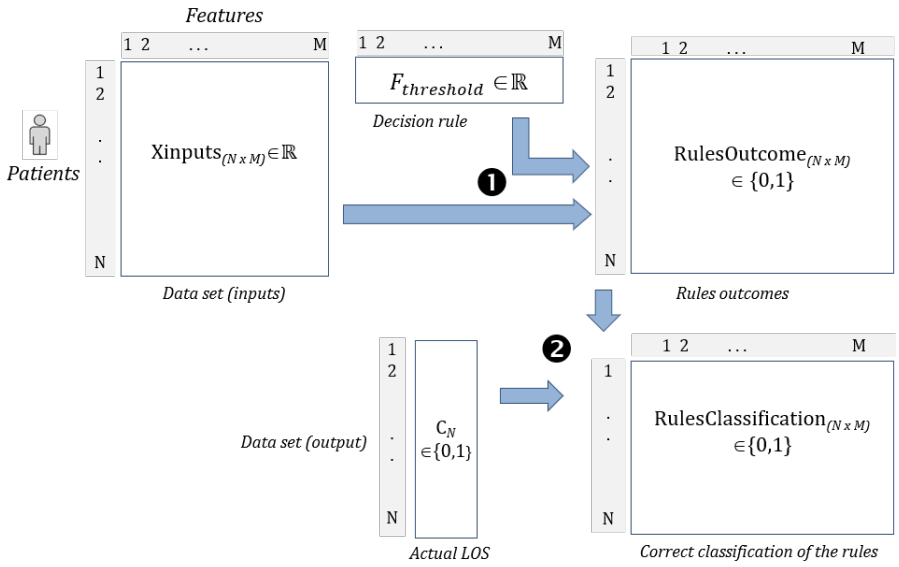


Figure 4.6: Methodology for the determination of the correct rules. Figure adapted from Figure 3 of [10].

Having the correct classification of the rules (target) we implemented different supervised strategies (the same three classification algorithms of the Black-box model): the SVM, the MLP and the RF, to perform the selection of rules, by predicting the correct ones for each patient. They were given the $Xinputs_{(N\times M)}$ matrix containing all the features original values of the N patients and the respective target, the $RulesClassification_{(N\times M)}$ matrix. The three different classifiers were then trained to predict the patients' correct rules, originating the $SelectedRules_{(N\times M)}$ matrix.

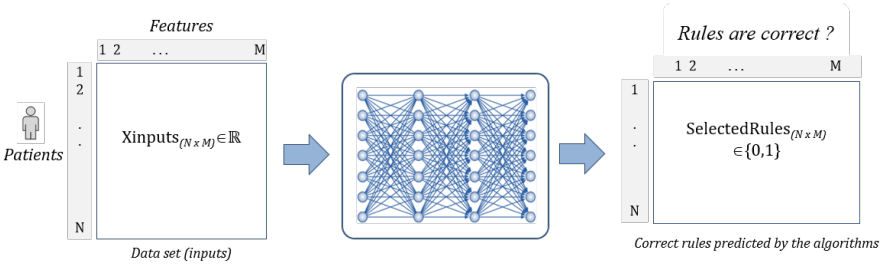


Figure 4.7: Selection of rules by the classification algorithms. Figure adapted from Figure 4 of [10].

4.5.4 Classification models hyperparameters' tuning

In order to determine the best hyperparameter settings of the algorithms, to predict the rules correctness, here again, a brief tuning was previously executed through the implementation of a cross-validated grid search (a procedure similar to the one executed for the black-box model, described in chapter 4.4.3).

The score parameter selected to be optimized was the F1 score, since we wanted to maximize the number of correct rules. This is important because the determination of the LOS will be made using only the rules that were considered correct by the classification models.

4.5.5 Determination of the optimal cut point of the Precision-Recall curve

In line with the search of the hyperparameters settings that achieved the maximum F1 score in the previous step, the following procedure was the calculation of the precision-recall curve. This curve plots the positive predictive value, or precision, (y-axis) against the true positive rate (x-axis) also called recall.

The aim was to find the optimal cut point of the curve (the point that achieves the best precision and recall values) that, in this case, represents the highest F1 score value, according to equation (2.9). This value was found by calculating the F1 score of all the curve's points and determination of the maximum value.

4.5.6 LOS classification

After the identification of each patient's most appropriate rules through the classification models, the determination of the respective LOS was directly obtained with an ensemble method. The strategy selected was the calculation of the rules outcome mode, employing only the patient's correct rules. Since only the rules that are correct for a specific patient are selected, a customization strategy is achieved. In this sense, the LOS classification of a patient L_i is a voting system where the final LOS is the mode of the rules outcome of all the patient's correct rules, according to:

$$L_i = mode\left(RulesOutcome_{\ i \times M}\right) \text{ if } SelectedRules_{\ i \times M} = 1 \tag{4.9}$$

Since the number of correct rules was not controlled, two distinct situations could occur: a draft or zero correct rules identified. In both cases, a long LOS was assumed to give more importance to the detection of extended stays.

Additionally, as a strategy to verify if we could reach a similar or better performance using only a subset of the initial set of rules (features), instead of the emsemble of all, we

performed an exhaustive search, testing all the possible combinations of them. Then, the subset reaching the highest GE value was selected, which we called 'best features subset'. It is important to mention that also in this approach, only the rules that were correct for each patient were employed. The model's performance was then compared with the one employing the totality of the rules.

Similarly to the Black-box model, this model was also applied to 20 different sets of patients obtained through splits of the data in random train and test sets, in order to have a better assessment of the model's performance.

4.6 Assessment metrics and Statistical validation

To quantify the performance of all the models, the geometric mean was computed for a proper evaluation of the length of stay predictions in the clinical context. Moreover, the F1 score was also evaluated in the bariatric patients due to its importance, specially for data with an extremely low number of the positive class, which is the case of these patients' LOS. For the developed models to be considered accurate they should identify the majority of these rare positive cases.

To validate the results of all the models and perform a correct comparison between them, two statistical tests were employed: the Friedman test, to analyse if there is a statistical significant difference between the models' performance and, if that is concluded, the Nemenyi test for a pairwise comparison.

These tests were employed, first, in the cardiac patients, and secondly in the bariatric ones. For each type of patient, we conducted performance comparisons between: **i)** the 3 classification algorithms of the Black-box model and **ii)** the 3 classification algorithms of the Rules model, to determine the best classification algorithms of each model. Then, we compared the performance of **iii)** the best algorithms of the Black-box and the Rules models, to observe if one model outperformed the other.

5

Results and Discussion

This chapter presents the main results of this research obtained for the cardiac and bariatric patients (work developed using Python).

The results of the SCORE risk tool for the assessment of the cardiac patients' LOS are presented.

Then, the feature baseline and feature selection, as well as the feature discretization for each group of patients (cardiac and bariatric) are depicted.

Regarding the Black-box model, the results of the grid search for the best hyperparameters are reported and the mean ROC curve of the test sets of each classification algorithm are represented. Moreover, the mean values and standard deviations of the evaluation metrics are plotted and discussed.

Additionally, the results of the Black-box model applied in the bariatric patients before and after the addition of the vital signs features are described.

Concerning the Rules model, the grid search results and the mean values and standard deviations of the GE and F1 score are represented for both types of patients.

Finally, the comparison between the Black-box and the Rules models is demonstrated, using the metrics GE and F1 Score and the appropriate statistical tests for ranking the models.

5.1 SCORE correlation

In a first approach, the SCORE risk model capacity to estimate the LOS was evaluated using all the patients integrating the dataset.

The GE values obtained with the variation of the $C_{Threshold}$ for the LOS assessment, equations (4.4) and (4.5), are depicted in Table 5.1.

Table 5.1: SCORE risk: GE values correspondent to the $C_{Threshold}$ variation

SE	SP	GE	$C_{Threshold}$
0.95	0.05	0.22	1
0.78	0.23	0.43	2
0.56	**0.45**	**0.50**	**3**
0.16	0.80	0.36	4
0.05	0.94	0.21	5
0.01	0.98	0.11	6

According to the obtained results, the maximum GE achieved was $GE = 0.50$ for a $C_{Threshold} = 3$, resulting in the following rule:

$$\text{IF } SCORE\ class \geq 3 \text{ THEN } LOS > 3\ days \tag{5.1}$$

This value ($GE = 0.50$) demonstrates that this prognostic risk tool isn't adequate for the LOS prediction and the development of a specific model for that aim is recommended in order to obtain a more accurate prediction.

5.2 Feature design and discretization

5.2.1 ICU patients

The next table depicts the Kendall's tau results for all the non continuous features of the ICU patients' feature baseline.

Table 5.2: Kendall's tau results for the non continuous features

Feature	p value	tau coefficient
Type of ACS	$\mathbf{6.8 \times 10^{-06}}$	**-0.14**
Gender	2.4×10^{-01}	0.04
Killip class	$\mathbf{6.1 \times 10^{-12}}$	**0.22**
CABG	7.5×10^{-03}	0.09
History of coronary artery disease	3.8×10^{-01}	-0.03
History of congestive heart failure	1.2×10^{-03}	0.10
History of CVA	$\mathbf{2.1 \times 10^{-04}}$	**0.12**
Diabetes	3.5×10^{-02}	0.07
CKD	$\mathbf{6.6 \times 10^{-07}}$	**0.16**
Peripheral vascular disease	5.1×10^{-02}	0.06

From the table observation, we selected 4 features: type of ACS, Killip class, History of CVA and Chronic Kidney Disease (CKD), since they meet the requirements of having a coefficient with an absolute value superior to 0.1 and a p value lower than 0.01 (the

remaining 6 were excluded). As previously mentioned, aiming to increase the model's interpretability, the continuous features were discretized, in two or three classes, originating discrete features that could be interpreted as rules (which facilitates the understanding of future results). For that, three different discretization approaches were employed to the 7 continuous features present in the baseline set: **i)** using the normal ranges found in literature and according to the standard reference intervals of the Clinical Pathology Service of the hospital center [83][88][89][90], **ii)** applying the values suggested by the ICU cardiologist and **iii)** performing an exhaustive search, testing all the possible features values as threshold to find the optimal value. The Kendall's tau test was then applied to the discretized features. The tau coefficients and respective thresholds are represented in the next Table.

Table 5.3: Kendall's tau results for the 3 types of discretization

Doctor's discretization

Feature	p value	tau coefficient	Thresholds
Age	4.0×10^{-06}	0.14	70
SBP	5.0×10^{-02}	0.04	90
C-reactive protein	1.0×10^{-04}	0.13	1.5
Hemoglobin	7.7×10^{-01}	0.01	8
Creatinine	2.6×10^{-05}	0.14	174.8
Troponin	5.5×10^{-04}	0.11	0.4-100
Glycose	4.5×10^{-05}	0.13	11

Discretization from normal ranges

Feature	p value	tau coefficient	Thresholds
Age	4.6×10^{-04}	0.11	60 - 80
SBP	3.8×10^{-01}	- 0.03	90 - 140
C-reactive protein	7.9×10^{-02}	0.06	0.5
Hemoglobin	1.3×10^{-08}	-0.18	12/13*
Creatinine	1.0×10^{-05}	0.14	63.65 - 104.31
Troponin	7.8×10^{-04}	0.11	0.4
Glycose	1.3×10^{-05}	0.14	3.3 - 6

Data driven's discretization

Feature	p value	tau coefficient	Thresholds
Age	2.1×10^{-06}	0.15	71
SBP	7.7×10^{-01}	0.09	101
C-reactive protein	5.6×10^{-08}	0.18	3.1
Hemoglobin	1.3×10^{-08}	-0.18	13
Creatinine	5.2×10^{-09}	0.19	89
Troponin	2.9×10^{-04}	0.12	0.4-168
Glycose	3.8×10^{-06}	0.15	6

The superscript * denotes a cut-off value of 12 for women and 13 for men.

The features presenting only one value in the 'Thresholds' column were divided in two classes, and the ones with two threshold values, separated by an en dash, were discretized in three classes.

The discretization method with worst results was the one using the thresholds from the normal range values (with the exception of the Hemoglobin and Glycose which achieved a coefficient higher than using the doctor's discretization). The reason for the smaller coefficients in this approach can be that the thresholds defined by the normal ranges may not be sufficiently low or high, to identify health conditions that promote prolonged LOS (poor discriminative power).

As expected, the data driven's discretization (based on the maximization of the Kendall's tau) was the one with the best results, determining 6 features as good LOS predictors: age, c-reactive protein, hemoglobin, creatinine, troponin and glycose that, together with the 4 features previously selected, makes a total of 10 features integrating the final set. The systolic blood pressure was removed as we couldn't reject the null hypothesis for this feature in any of the discretization strategies.

However, although the coefficients obtained using this method were higher than in the previous two, the overall increase wasn't sharp. In some variables there was little difference in both tau coefficient and threshold, when comparing with the doctor's discretization, which is the case of the Age and the Troponin.

From these findings, we can already verify that the association between the selected features and the LOS isn't very strong since the highest coefficient value found was 0.22 from the Killip class, therefore, this suggests that it will be difficult to obtain a LOS prediction model with a satisfactory performance.

Moreover, the features selected were already demonstrated as good LOS predictors in previous studies. The age has been found to predict LOS in cardiac patients [68][46]. Pei-Fang et al. [25] also determined that both age and the history of cerebrovascular disease were correlated with the cardiac patients' LOS. In addition, Teresa et al. [68] verified that comorbidities (namely cardiovascular diseases) and some laboratory variables (including troponin, creatinine, c-reactive protein and hemoglobin) could also be potential covariates. Still in her study, the type of acute myocardial infarction (STEMI, NSTEMI) was also associated with the hospitalization time.

5.2.2 Bariatric patients

The feature selection method used for the bariatric patients was also the Kendall's tau statistic and the discretization method employed was the data driven approach since it achieved the best results in the cardiac ICU patients. The mean cut off values obtained for the non binary selected features are presented in table 5.4.

Table 5.4: Mean cut off values for the non binary selected features

Feature	Mean cut off value
Surgery duration	60
Mean MEWS	1
Mean respiration rate	14
Heart rate slope	4
Heart rate variance	26

A total of 9 features was selected consisting in 5 static variables (ward medication of Corticosteroids, ward medication of Sedatives, Type of surgery, Cardiovascular comorbidity and Surgery duration) and 4 vital signs variables collected in the first 24 hours after surgery (mean MEWS, mean respiration rate, heart rate slope and the heart rate variance). The surgery type and surgery duration were tested for correlation for suspicion that the first depended on the second variable, but were proved to be uncorrelated.

The features selected for the bariatric patients are consistent with studies that refer them as good LOS predictors. The surgery duration was found to be correlated with the hospitalization time by Miguel et al.[80] and Jonathan et al.[70]. Furthermore, in Jonathan et al. study, the type of surgery was associated with greater than 1 day post-operation discharge [70].

In addition, the vital signs features selected also appeared in previous literature works as having a relationship with the length of stay. The National Early Warning Scores were proved by Barth et al.[27] to be associated with LOS, where higher values of NEWS leaded to longer hospitalization times. Brekke et al.[82] obtained an increase in their model's accuracy using the trend models of the respiration rate suggesting that this feature could be a good input feature for the LOS prediction.

5.3 Black-box model implemented in the cardiac patients

5.3.1 Grid search results

The grid search was employed to find the best hyperparameter of each classifier. Although we could have explored more deeply the optimal models' parameters, that isn't the goal of the study. The brief tuning was only performed in order to make sure that the parameters weren't random and were chosen with a minimum confidence.

Table 5.5 presents the results of the grid search performed for the 3 classifiers of the Black-box model, applied in the cardiac patients. In this table we can observe the most common parameter option of the different train sets employed.

Table 5.5: Grid search results for the Black-box model (cardiac patients)

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer size	number of estimators
Most common option	RBF	10	100

For the SVM classifier, the most common kernel type selected was the RBF (radial basis function) also called Gaussian kernel. This is a non linear kernel that is widely used when there is no prior knowledge about the data, outperforming other kernels [91]. Here, the RBF was also proved to be the most fitted for the LOS prediction task.

The most common hidden layer size selected for the MLP was 10 neurons and the number of estimators of the RF was 100.

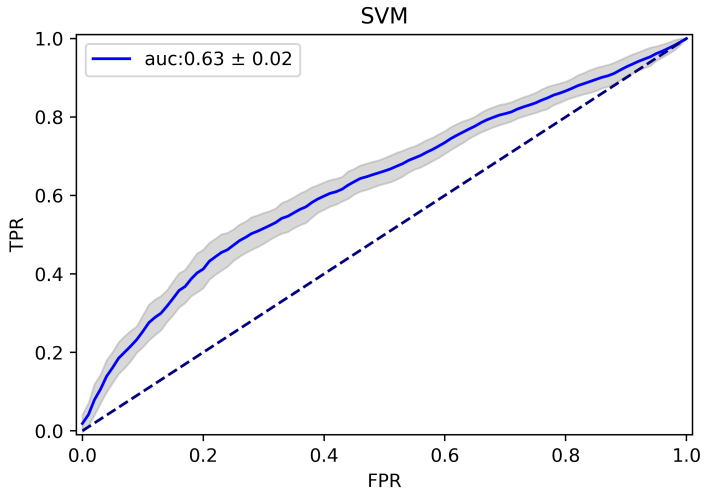
Naturally, for the various train sets, different hyperparameter options were selected, however, if we want to set a final design for each classifier, the most appropriate options are the most common values determined.

5.3.2 ROC Curves of the test sets

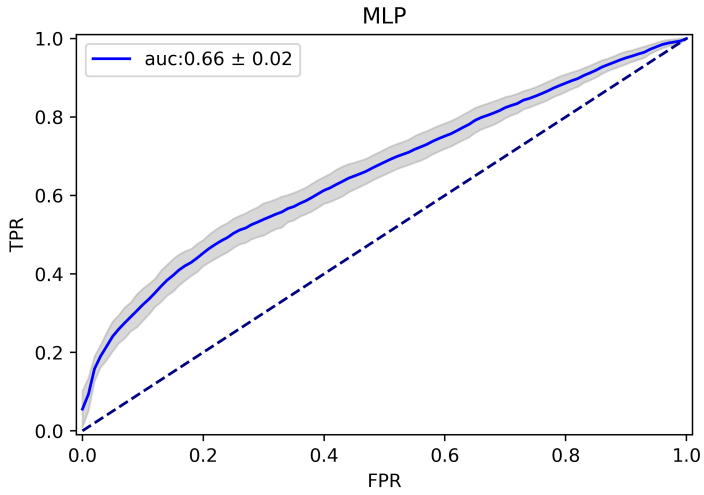
Figure 5.1 represents the ROC curves of all 20 patients' test sets, where the blue bold curve illustrates the mean ROC curve. In addition, the mean AUC value and respective standard deviation is depicted.

The area under the ROC curve is a performance measurement for classification problems at various thresholds settings [92] that can evaluate the models' capacity to distinguish between classes. In this sense, a higher AUC is indicative of a better distinction between the long and short lengths of stay.

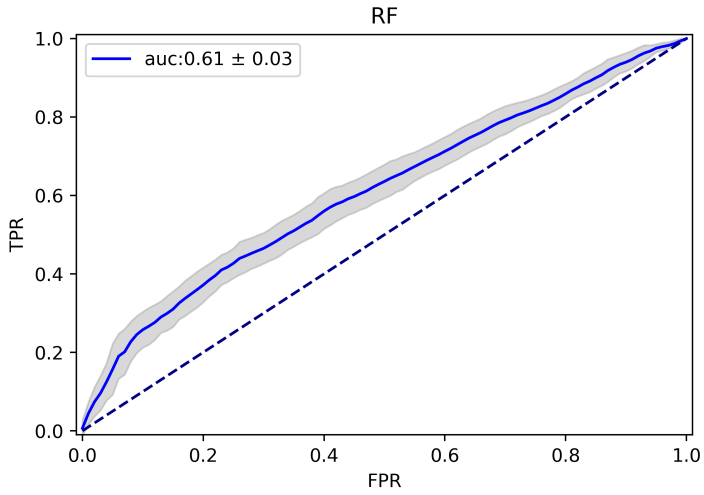
With the observation of these curves we can expect a better performance, in terms of GE, for the MLP and SVM since they achieved higher values of area under the ROC curve (0.66 ± 0.02 and 0.63 ± 0.02, respectively).



(a) Mean SVM ROC Curve.



(b) Mean MLP ROC Curve.



(c) Mean RF ROC Curve.

Figure 5.1: Mean ROC Curves of the Black-box model algorithms: cardiac patients

5.3.3 LOS classification

The mean values and standard deviations obtained from the LOS prediction of the three classification algorithms, used in the Black-box model, applied to the different test sets, are illustrated in Figure 5.2. The highest GE mean values were achieved by the MLP (0.62 ± 0.03) and SVM (0.60 ± 0.04) classifiers (Table 5.6). The RF had a slightly lower performance of $GE = 0.58 \pm 0.03$. These findings are in line with the observed in the ROC curves figure which revealed a higher LOS discrimination capacity for the MLP, that resulted in higher GE values.

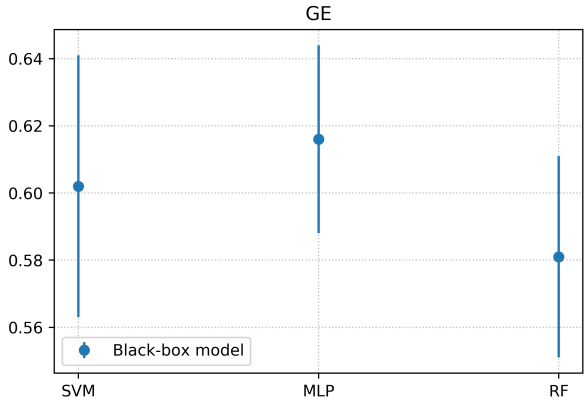


Figure 5.2: GE evaluation of the Black-box model algorithms: mean values and standard deviations (cardiac patients)

Table 5.6: Geometric mean evaluation of the Black-box model classification algorithms: mean values and standard deviations (cardiac patients)

Algorithm	SVM	MLP	RF
GE	0.60 ± 0.04	0.62 ± 0.03	0.58 ± 0.03

The present results also demonstrate the difficulty of predicting the patients' hospitalization time in the ICU with a GE good enough for the clinical application. However, this performance level was expected since the Kendall's tau coefficients obtained for the selected features (even after optimization) didn't exceed 0.2, with the exception of the Killip class (τ_B=0.22).

5.3.4 Statistical validation

To perform an adequate statistical validation for the comparison of all the models, the Friedman test was first employed to determine if the null hypothesis, which states that all the models perform similarly and therefore their ranks should be the same, can be rejected. If that was verified, the Nemenyi test was also implemented for the pairwise comparison.

The Friedman statistic was calculated from the average ranks of the algorithms, presented in Table 5.7. Since the comparison was made between 3 classifiers (k=3), the respective critical value, obtained from the Friedman's two-way analysis table (df = k - 1) with a significance level of 5 %, is 5.99, Table 2.1.

Table 5.7: Average ranks of the Black-box model classifiers using the GE metric - cardiac patients

GE	SVM	MLP	RF
Average rank	1.700	1.625	2.675

The value calculated from the average ranks was $X_F^2 = 13.90$, equation (2.11), overcoming the critical value of 5.99. Therefore, the null hypothesis was rejected and the Nemenyi test was performed for a pairwise comparison of the classification algorithms, based on the computed average ranks.

The critical value of the two-tailed Nemenyi test, for a significance level of 5%, considering that 3 classifiers are compared (k=3) is 2.344, according to Table 2.2. Using this value, the computed critical distance, equation (2.12), was 0.74. Hence, for a classifier to be determined better than other, the absolute value of their average ranks difference must be at least 0.74.

The Table 5.8 presents the Nemenyi test results. The difference between the average ranks of the MLP and the RF of $D = 1.05$ exceeded the critical distance, so we can conclude that the MLP performance was better than the Random Forest one. Similarly, the SVM also outperformed the RF with a difference in the average ranks of $D = 0.98$. On the other hand, the SVM and the MLP, should be ranked equally since their average ranks difference of $D = -0.08$ didn't reached the critical distance.

Table 5.8: Nemenyi test of the Black-box model classification algorithms for the GE (cardiac patients)

Algorithms	MLP	RF
SVM	-0.08	0.98
MLP		1.05

These results are represented in the Nemenyi diagram (Figure 5.3), where we can observe that the classifier with the best average rank was the MLP followed by the SVM and the RF. Additionally, we can observe that the classifiers with a statistically significant difference in their performances are not linked, contrarily to the MLP and SVM, that are connected with a bold line, demonstrating that they should be ranked equally. These findings are consistent with the observed in figure 5.2.

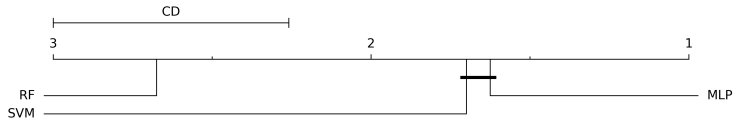


Figure 5.3: Nemenyi diagram of the Black-box model classification algorithms for the GE (cardiac patients)

Overall, we can conclude that, the MLP and SVM behaved similarly and both outperformed the Random Forest, thus, these classifiers should be the preferable algorithms to be implemented as the Black-box model. However, their GE mean values of 0.62 and 0.60 respectively, don't seem to be sufficiently high for the model to be adequate for an implementation in clinical institutions, and confirm the difficulty of LOS prediction in the ICU context. However, these results were expected since the tau coefficients of the selected features were low, with a maximum value of 0.22 for the 'Killip class'. This weak degree of association between the variables and the LOS was suggestive of a LOS prediction without a very high performance.

5.4 Rules model implemented in the cardiac patients

The Rules model was also implemented in the 20 different patients' sets to determine if a similar or better performance could be reached. In this approach the same 3 classifiers (SVM, MLP and RF) were employed to assess which was the most adequate for the task.

5.4.1 Derivation of rules

The derivation of the 10 rules from the 10 risk factors (M): Age, C-reactive protein (CRP), Hemoglobin, Creatinine, Troponin, Glycose, type of ACS, Killip class, History of CVA and Chronic Kidney Disease (CKD), by implementing the thresholds from the data driven approach of Table 5.3 (according to equation 4.8), originated the $RulesOutcome_{\,i\times M}$ matrix defined as:

$$\text{IF} \begin{cases} M{=}AGE \geq 71 \\ M{=}CRP \geq 3.1 \\ M{=}HEMOGLOBIN < 13 \\ M{=}CREATININE \geq 89 \\ M{=}TROPONIN \geq 0.4 \\ M{=}GLYCOSE \geq 6 \\ M{=}TYPE\ OF\ ACS = \text{STEMI} \\ M{=}KILLIP\ CLASS \geq 2 \\ M{=}HISTORY\ OF\ CVA = \text{'Yes'} \\ M{=}CKD = \text{'Yes'} \end{cases} \text{THEN } RulesOutcome_{\,i\times M} = 1 \tag{5.2}$$

5.4.2 Grid search implementation and selection of rules

After the derivation of the rules, the SVM, MLP and RF classifiers were then trained to classify the correctness of the several rules (M) for each patient (N), returning as output the $SelectedRules_{(N\times M)}$. As a result, for each patient it was possible to identify the particular rules that were later combined in the ensemble scheme.

With that aim, we first performed a grid search in order to find the classifiers best hyperparameters design which optimized the identification of the correct rules (F1 score metric). Table 5.9 depicts the search results, where we can observe that the most common options were, the RBF kernel, a size of 12 neurons in the hidden layer and 100 estimators for the SVM, MLP and RF hyperparameters, respectively.

Table 5.9: Grid search results for the Rules model (cardiac patients)

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer size	number of estimators
Most common option	RBF	12	100

With the best hyperparameters found for each classifier, and their identification of each patient's correct rules ($SelectedRules_i$), we obtained a patient customized scheme that uses only a particular set of rules that best classifies its length of stay according to its characteristics. As a result, it was possible to define a simple rule that could be used to classify the length of stay of a patient, according to equation (4.9), by employing the $RulesOutcome\ _{i\times M}$ matrix obtained in (5.2).

5.4.3 LOS classification

From the observation of Figure 5.4 and Table 5.10, we can verify that the overall performance of the classifiers - SVM, MLP, RF was the same, with all of them achieving a mean GE of 0.63.

On the other hand, when the best combination of rules (best features subset) was employed, instead of using all them, the mean GE of the same classifiers - SVMb, MLPb, RFb increased to approximately 0.66. This demonstrates that even using only the correct rules for each patient, the use of a more adequate subset of rules can reduce wrong predictions, and that a simpler model, that needs less information of the patients, can be implemented with a more satisfactory performance.

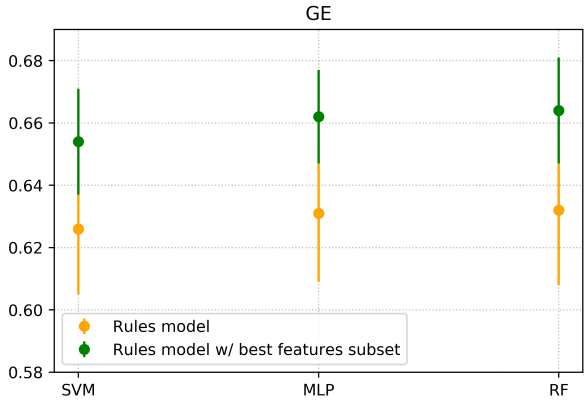


Figure 5.4: GE evaluation of the Rules model classification algorithms: mean values and standard deviations (cardiac patients)

Table 5.10: Geometric mean evaluation of the Rules model classification algorithms: mean values and standard deviations (cardiac patients)

Algorithm	SVM	MLP	RF	SVMb	MLPb	RFb
GE	0.63±0.02	0.63±0.02	0.63±0.02	0.65±0.02	0.66±0.02	0.66±0.02

Moreover we determined that, for each patients' set, all classifiers selected more or less the same subset of rules (with small variations between). The rules that were present in the majority of the best subsets were: Type of ACS, Age, Glicose, Killip class, C-reactive protein and History of CVA.

5.4.4 Statistical validation

For the comparison of the Rules model classifiers, two sets were evaluated: **i)** the SVM, MLP and RF with all the rules and **ii)** the same three algorithms using only the best subset of rules (SVMb, MLPb and RFb).

By observing the Friedman's critical values table, Table 2.1, we verify that, for 6 classifiers (k=6) and a significance level of 5%, the critical value is 11.07. This is the value that needs to be overcome to reject the Friedman's null hypothesis.

From the average of ranks, represented in Table 5.11, the value $X_F^2 = 77.92$ was obtained, equation (2.11). As a result the null hypothesis which states: 'all methods behave similarly', could be rejected and the Nemenyi test implemented.

Table 5.11: Average ranks of the Rules model classifiers using the GE metric - cardiac patients

GE	SVM	MLP	RF	SVMb	MLPb	RFb
Average rank	5.325	4.850	4.625	2.825	1.925	1.450

The SVMb, MLPb, RFb denotes the SVM, MLP, RF employing only the best rules subset.

Having six classifiers for comparison, the critical value of the two-tailed Nemenyi test for a significance level of 5%, obtained from Table 2.2, is now 2.850. Thus, the critical distance that the classifiers need to overcome to have a statistically significant difference is $CD = 1.69$, equation (2.12). From the differences between the average ranks of the classifiers, demonstrated in Table 5.12, it can be verified that the performance of the algorithms is only significantly different between the set of classifiers **i)** SVM, MLP and RF, and the set **ii)** SVMb, MLPb and RFb. Within each set all the classifiers behaved similarly, as the critical distance was not reached. Therefore, they are linked with a bold line in the Nemenyi diagram, Figure 5.5.

Table 5.12: Nemenyi test of the Rules model classification algorithms for the GE (cardiac patients)

Algorithms	MLP	RF	SVMb	MLPb	RFb
SVM	-0.48	-0.70	-2.50	-3.40	-3.88
MLP		-0.23	-2.03	-2.93	-3.40
RF			-1.80	-2.70	-3.18
SVMb				-0.90	-1.38
MLPb					-0.48

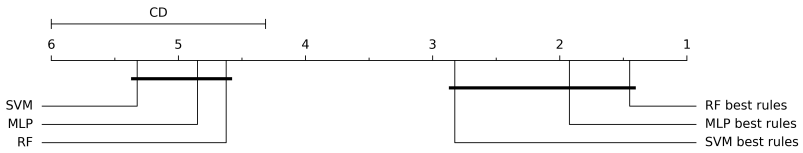


Figure 5.5: Nemenyi diagram of the Rules model classification algorithms for the GE (cardiac patients)

These results are in concordance with the observed in Figure 5.4, confirming the similar performance between the classifiers using all the rules (SVM, MLP, RF) with a mean GE of 0.63, and also between the classifiers using the best subset of rules (SVMb, MLPb, RFb) achieving a GE of approximately $GE = 0.66$. This increase of performance was proved to be statistically significant.

5.5 Comparison of the Black-box and Rules models implemented in the cardiac patients

The analyse of Figure 5.6 allows to observe and compare the Black-box and the Rules models performance.

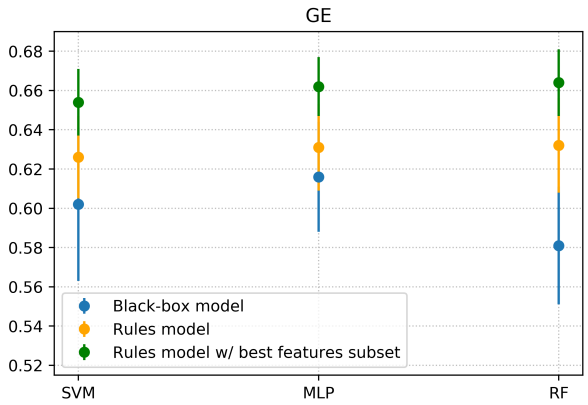


Figure 5.6: GE evaluation of all the classification algorithms of the two models (Black-box and Rules): mean values and standard deviations - cardiac patients

We can visualize an increase of the mean GE when comparing the Rules model in relation to the Black-box model, specially in the Random Forest and, as previously mentioned, the results are even better when the best subset of rules is employed.

5.5.1 Statistical validation

To determine if we can conclude that the difference between the GE of the Black-box and the Rules models is statistically significant, the classification algorithms with the best average rank of each one were selected to be compared using the Friedman and Nemenyi tests.

From the observation of the previous results, the classifier of the Black-box model with the best average rank was the MLP and, withing the groups of the Rules model i) using all the rules and ii) using the best subset of rules, we selected the Random Forest for both.

The resulting average ranks of this new comparison are presented in the next table.

Table 5.13: Average ranks of each model's best classifier using the Geometric mean metric - cardiac patients

GE	MLP	RF	RF best subset
Average rank	2.85	2.15	1.00

The Friedman statistic obtained in this case was $X_F^2 = 34.9$, so the null hypothesis could again be rejected, since it overcame the critical value of 5.99, Table 2.1. The Nemenyi test was then implemented with a critical distance of 0.74 as the conditions were: three classifiers (k=3) and a significance level of 5%, equation (2.12).

As it can be observed in Figure 5.7 and Table 5.14, only the RF with the best subset of rules achieved the critical distance when comparing with the other classifiers. The distance between the RF with all the rules and the MLP from the Black-box model (D = 0.70) was very close to the critical distance but not sufficiently high. The best classifier design was the RF using the best subset of rules, followed by the RF of the Rules model and the MLP from the Black-box model.

Table 5.14: Nemenyi test of the best classification algorithms of each model using the GE metric - cardiac patients

Algorithms	RF (rules)	RF best subset
MLP (black-box)	-0.70	-1.85
RF (rules)		-1.15

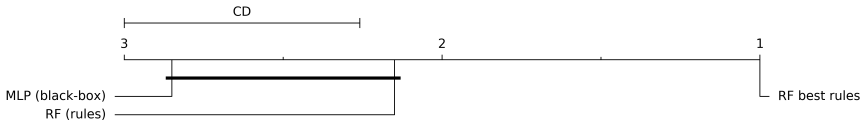


Figure 5.7: Nemenyi diagram of the best classification algorithms of each model for the GE metric - cardiac patients.

With these findings we can conclude that, although the GE values achieved by the Rules model still don't seem sufficient for the model to be implemented in clinical institutions, this model is more accurate and adequate to predict the patient's LOS than the Black-box model with the use of the best subset of correct rules, that leads to a significant increase of performance.

In addition, the Rules model also has the advantages of being interpretable, allowing to observe which rules were correct for each patient, and patient customized, by using only the rules that were correct for each of them. These are important aspects for the clinicians, since they can give an insight of the more relevant risk factors and serve as one of the inputs that can contribute to the decision making.

5.6 Results of the alternative sampling technique: 30 repeated permutations followed by two-fold cross-validation

As mentioned in chapter 2.2.4, we also tested the sampling technique suggested by Pizarro et al.[59] - 30 repeated permutations followed by two-fold cross-validation (demonstrated in appendix A), since they also addressed the problem of the violation of the sampling independence when comparing the performance of different models in a single dataset.

The results obtained were equivalent to the present ones. The algorithm of the Black-box model with the best average rank was again the MLP, and it achieved a mean GE of 0.60 $\pm$ 0.02. The Rules model also outperformed the Black-box one with a maximum GE mean value of 0.65 (using the best subset of features) and this value was achieved, again, by the Random Forest. When the classifiers employed all the rules, they obtained the same mean GE of the presented sampling technique (GE = 0.63). This performance increase was proved to be statistically significant by the Friedman and Nemenyi statistical tests.

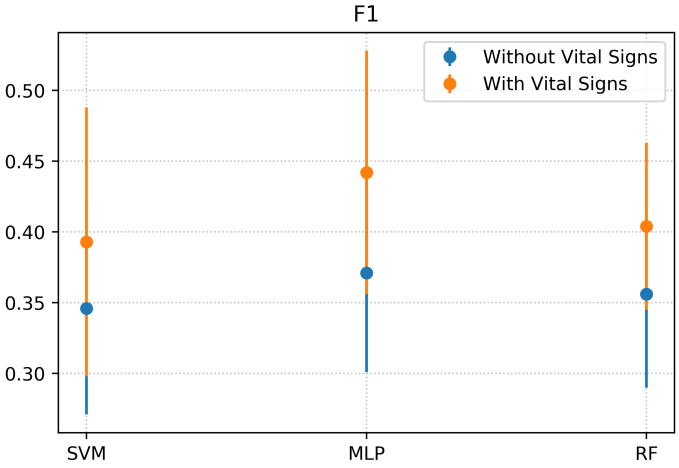
5.7 Bariatric patients' LOS prediction before and after the Vital Signs features addition

The influence of the Vital Signs addition in the LOS prediction was tested with a performance comparison of two models: one containing only the: demographic and ward features and comorbidities, and other containing the same features with the vital signs features added (Mean respiration rate, Heart rate slope, Heart rate variance and mean MEWS).

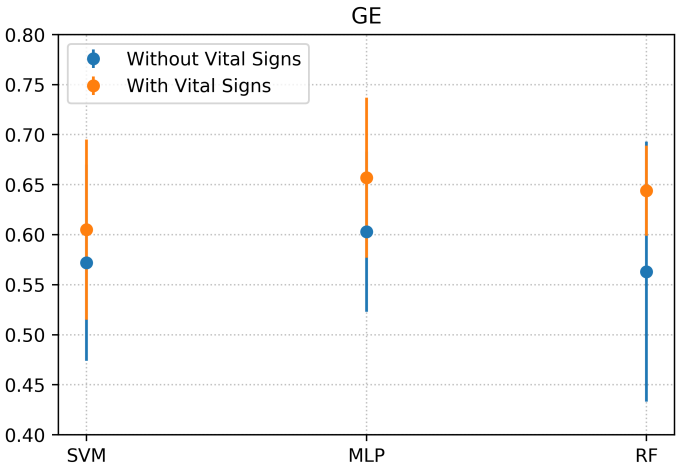
The goal was to verify if the inclusion of the dynamic data could improve the LOS prediction. Therefore, the two models were applied in 20 different patients' sets and the F1 score and GE mean results obtained are observed in the next tables and plotted in Figure 5.8.

Table 5.15: F1 score and GE mean values and respective standard deviations of the algorithms without and with the vital signs features addition

		SVM	MLP	RF
F1 score	Without vital signs	0.35±0.08	0.37±0.07	0.36±0.07
	With vital signs	0.39±0.09	0.44±0.09	0.40±0.06
GE	Without vital signs	0.57±0.10	0.60±0.08	0.56±0.13
	With vital signs	0.60±0.09	0.66±0.08	0.64±0.04



(a) F1 score results.



(b) GE results.

Figure 5.8: Comparison between the Black-box algorithms before and after the addition of the vital signs features.

From the observation of the figure results, we can verify an increase of the GE and F1 score mean values, for all classifiers, when the model includes the vital signs features. It can be observed an increase of the F1 score mean value of 0.04 for the SVM and RF classifiers, and 0.07 for the MLP. In terms of GE, the increase is more accentuated for the MLP and the RF algorithms with a difference of 0.06 and 0.08, respectively. The standard deviation of the Random Forest also reduced significantly.

We can then conclude that the addition of vital signs features improves the performance of the LOS prediction models. Therefore, they were added to the input features set of the future models presented.

5.8 Black-box model implemented in the bariatric patients

The Black-box model was then employed in the bariatric patients using the final set of features defined previously and its performance was evaluated using the same statistical tests implemented in the cardiac patients.

5.8.1 Grid search results

We performed a grid search for the 3 classifiers in order to obtain the best hyperparameter options for the LOS prediction task (similarly to the Black-box model of the cardiac patients).

In Table 5.16, we observe that, for the SVM and the RF classifiers, the most common kernel type and number of estimators selected were, respectively, the RBF and 100 estimators (also selected for the cardiac patients). The most common hidden layer size selected for the MLP was 14 neurons. Thus, if final models would be defined, in order to best fit all patients, they could be a SVM with a RBF kernel, a RF with 100 estimators and a MLP with a hidden layer size of 14 neurons.

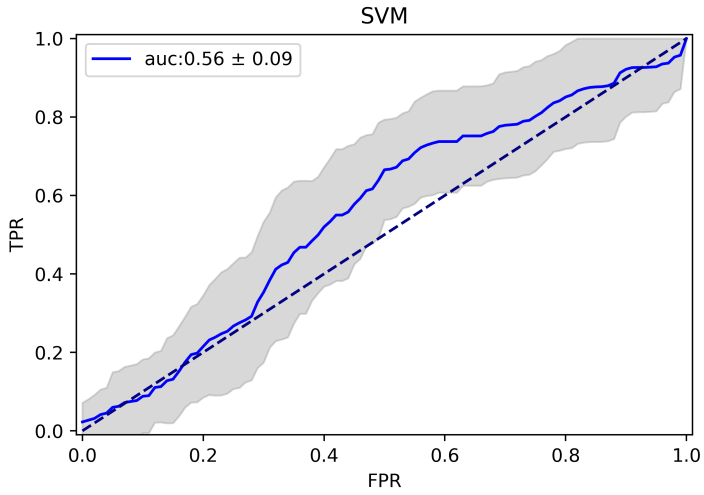
Table 5.16: Grid search results for the Black-box model (bariatric patients)

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer sizes	number of estimators
Most common option	RBF	14	100

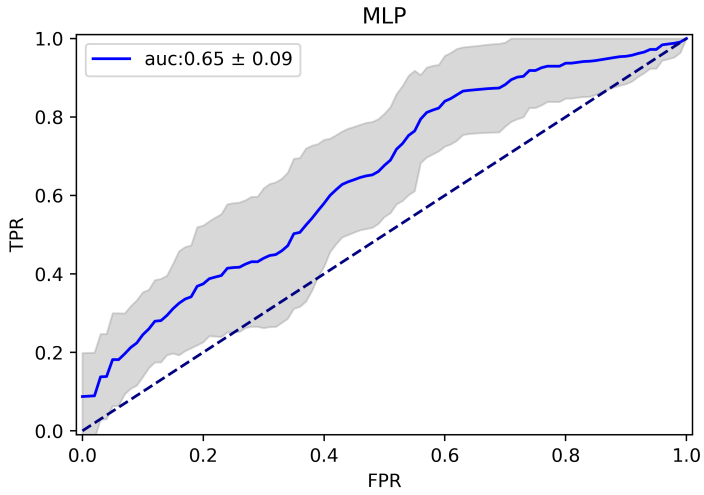
5.8.2 ROC Curves of the test sets

The ROC curves of the test sets are plotted in Figure 5.9. For each classifier it is represented the mean ROC curve and also the mean AUC and correspondent standard deviation. We observe that the highest mean AUC was obtained for the MLP with a value of $AUC = 0.65 \pm 0.09$ followed by the RF ($AUC = 0.63 \pm 0.06$) and the SVM ($AUC = 0.56 \pm 0.09$).

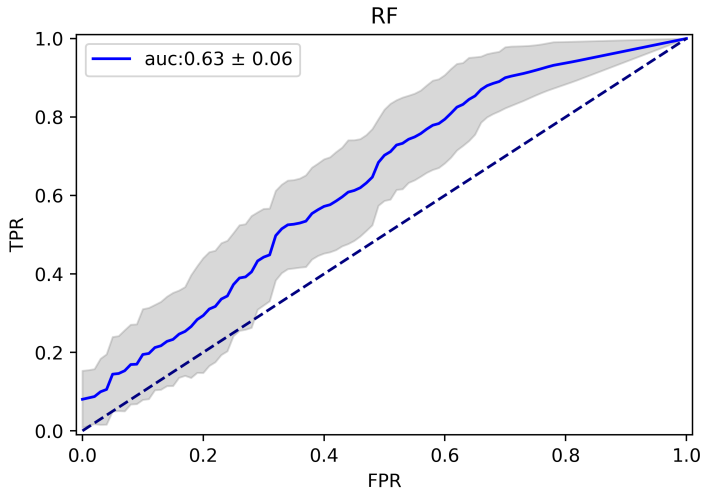
Similar to the cardiac patients, the MLP was again the classifier that reached a higher mean value but, with the elevated standard deviation value, its discrimination ability can be similar to the RF, therefore, we can't directly determine that the MLP was the Black-box model classifier with the best capacity to distinguish the two LOS classes. For that, the statistical validation was later implemented.



(a) Mean SVM ROC Curve.



(b) Mean MLP ROC Curve.



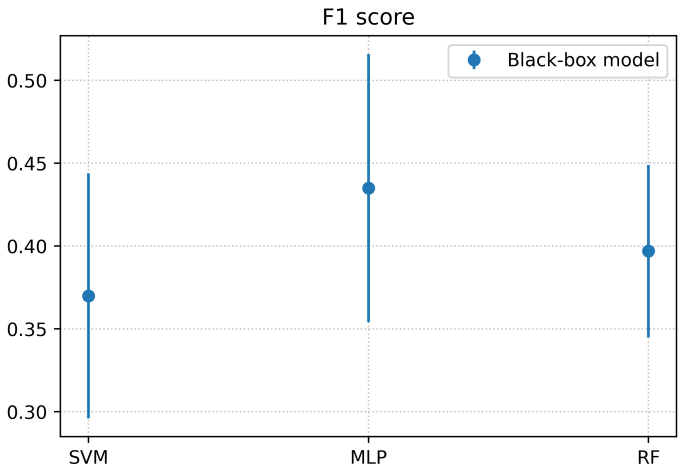
(c) Mean RF ROC Curve.

Figure 5.9: Mean ROC Curves of the Black-box model algorithms: bariatric patients.

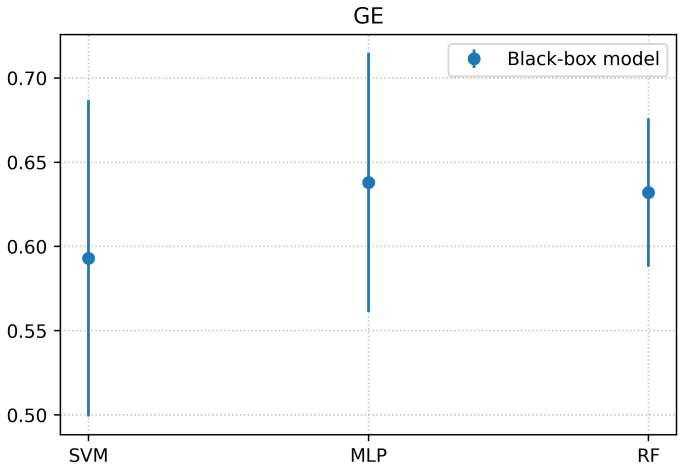
5.8.3 LOS classification

The performance of the three classification algorithms of the Black-box model in terms of F1 score and GE is plotted in Figure 5.10 and the respective values represented in Table 5.17.

The MLP achieved the highest mean values of both metrics and the SVM the lowest. However, when we compare the GE of the MLP and the RF, we observe that although the mean value of the first classifier is higher ($GE = 0.64 \pm 0.08$), the standard deviation is also bigger than the RF one ($GE = 0.63 \pm 0.04$), making it harder to define which algorithm performed better. To execute a proper performance comparison between the three classifiers a statistical validation was implemented.



(a) F1 scores results.



(b) GE results.

Figure 5.10: F1 score and GE evaluation of the Black-box model classification algorithms: mean values and standard deviations (bariatric patients)

Table 5.17: Geometric mean and F1 score evaluation of the Black-box model classification algorithms: mean values and standard deviations (bariatric patients)

Algorithm	SVM	MLP	RF
F1	0.37 ± 0.07	0.44 ± 0.08	0.40 ± 0.05
GE	0.59 ± 0.09	0.64 ± 0.08	0.63 ± 0.04

From these findings, we verify an overall low performance in terms of F1 score that can be a result of the high data imbalance (extremely low number of extended LOS), making it harder to predict those cases. When evaluating the GE, the best classifier achieved values of approximately $GE = 0.64$ which doesn't appear to be sufficiently accurate for the clinical implementation.

5.8.4 Statistical validation

To identify the best algorithm of the model, a statistical validation was implemented.

From the average ranks of the classifiers, for the F1 score and GE, the Friedman statistic values obtained were $X_F^2 = 10.80$ and $X_F^2 = 3.769$ respectively. Only the value correspondent to the F1 score was higher than the critical value of 5.99 (Table 2.1), for that reason, the Friedman null hypothesis could only be rejected for the F1 score. Therefore, when evaluating the GE, we can conclude that the behaviour of all classifiers was similar and thus they can have the same rank. In terms of F1 score there was at least one classifier with a performance statistically different than the others.

Table 5.18: Average ranks of the Black-box model classifiers using the F1 score metric - bariatric patients

F1 score	SVM	MLP	RF
Average rank	2.50	1.48	2.03

The Nemenyi test was then implemented only for the F1 score metric, with a critical value of the Nemenyi test (significance level of 5%) of 2.344 as observed in Table 2.2, and a consequent critical distance of 0.74, equation (2.12).

The differences of the average ranks calculated for the Nemenyi test are represented in Table 5.19. We can verify that the MLP outperformed the SVM with a difference of $D = -1.03$, overcoming the critical distance. However, the critical distance wasn't achieved when comparing the MLP with the RF performance, and also, the RF with the SVM, so they can be equally ranked. This is represented by the connecting bold lines in the Nemenyi diagram (Figure 5.11).

Table 5.19: Nemenyi test for the Black-box model classifiers for the F1 score metric

Algorithms	MLP	RF
SVM	-1.02	-0.47
MLP		0.55

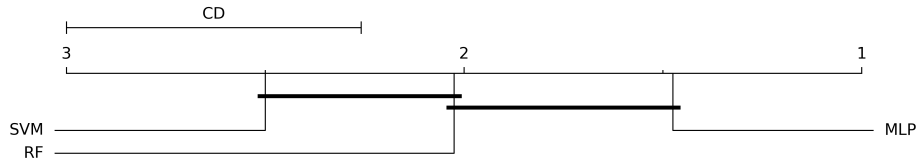


Figure 5.11: Nemenyi diagram of Black-box model classification algorithms for the F1 score (bariatric patients)

Overall, we can conclude that the classifier implemented in the bariatric patients with the best rank was the MLP, however, its performance is only significantly better than the SVM. Concerning the GE, all classifiers behaved similarly. Therefore, the preferable algorithm to be selected as the Black-box model is the MLP (also selected as the preferred Black-box model for the cardiac patients).

5.9 Rules model implemented in the bariatric patients

5.9.1 Derivation of rules

The risk factors (M) selected for the bariatric patients: Type of surgery, Surgery duration, Cardiovascular comorbidity, ward medication of corticosteroids (WMC) and Sedatives (WMS), Mean MEWS, Mean respiration rate (Mean RESP), Heart rate slope, Heart rate variance, were also transformed in 9 rules, by implementing the thresholds depicted in Table 5.4 originating the $RulesOutcome_{\ i \times M}$ matrix according to:

$$\text{IF} \left\{ \begin{array}{l} M{=}Surgery\ type = \text{Gastric sleeve} \\ M{=}Surgery\ duration \geq 60 \\ M{=}WMC = \text{'Yes'} \\ M{=}WMS = \text{'Yes'} \\ M{=}Cardiovascular\ comorbidity = \text{'Yes'} \\ M{=}Mean\ MEWS \geq 1 \\ M{=}Mean\ RESP \geq 14 \\ M{=}Heart\ rate\ slope \geq 4 \\ M{=}Heart\ rate\ variance \geq 26 \end{array} \right\} \text{THEN } RulesOutcome_{\ i \times M} = 1 \quad (5.3)$$

5.9.2 Grid search implementation and selection of rules

After the derivation of the rules, we implemented the classifiers (SVM, MLP, RF) to determine the correctness of the several rules (M) for each patient (N), but first, the grid search was again employed to find the classifiers best hyperparameters for the identification of the correct rules.

Table 5.20 depicts the search results. The most common options were, the RBF kernel, a size of 14 and 100 estimators for the SVM, MLP and RF hyperparameters respectively.

Table 5.20: Grid search results for the Rules model (bariatric patients)

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer size	number of estimators
Most common option	RBF	14	100

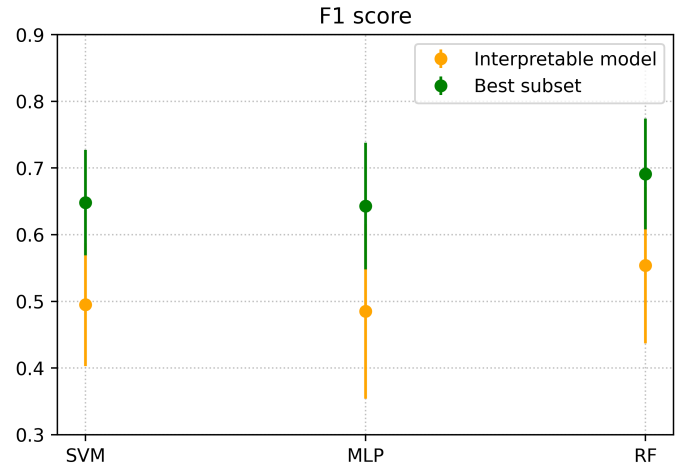
Having the best hyperparameters option defined for the classifiers, they were trained to predict the correctness of the rules. The rules that were determined as correct by the classifiers ($SelectedRules_{\,i\,\mathrm{x}M}$) were then applied to equations (5.3) and (4.9) for the determination of each patient' LOS.

5.9.3 LOS classification

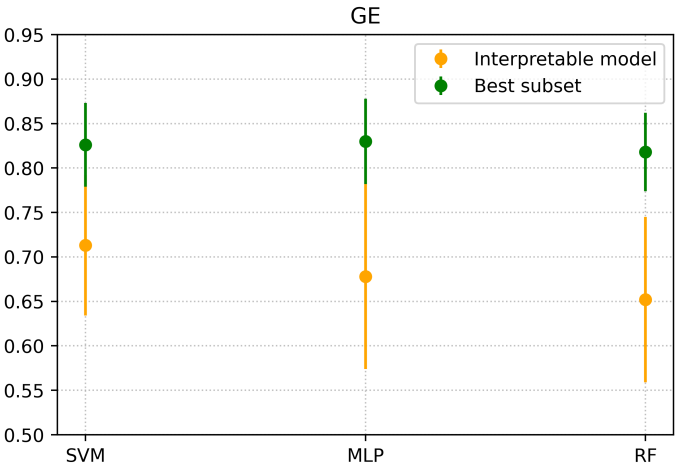
From the observation of Figure 5.12, we can verify that the classifiers performed differently when evaluating both the F1 score and the GE.

The SVM reached a higher GE mean value ($GE = 71 \pm 0.08$) followed by the MLP ($GE = 0.68 \pm 0.10$) and, in last, the RF with $GE = 0.65 \pm 0.09$. On the other hand, the best F1 score was achieved by the RF with $F1 = 0.55 \pm 0.12$.

Furthermore, the use of the best combination of rules for each classifier, resulted in a performance increase for both F1 score and GE, achieving a maximum mean value of 0.69 and 0.83 respectively (Table 5.21). We can again verify that, although only the correct rules for each patient are considered, the use of a more adequate subset of rules, reducing the patients' information needed for the prediction, can be implemented with a satisfactory performance.



(a) F1 score results.



(b) GE results.

Figure 5.12: F1 score and GE evaluation of the Rules model classification algorithms: mean values and standard deviations (bariatric patients)

Table 5.21: F1 score and GE evaluation of the Rules model classification algorithms: mean values and standard deviations (bariatric patients)

Algorithm	SVM	MLP	RF	SVMb	MLPb	RFb
F1	0.50±0.09	0.49±0.13	0.55±0.12	0.65±0.08	0.64±0.10	0.69±0.08
GE	0.71±0.08	0.68±0.10	0.65±0.09	0.82±0.05	0.83±0.05	0.82±0.04

Moreover, the rules that were present in the majority of the best subsets were: type of surgery, ward medication of Corticosteroids, mean MEWS, mean respiration rate, heart rate slope and heart rate variance. We can verify that all the vital signs features integrate the majority of the best subsets, demonstrating their relevance in this context.

5.9.4 Statistical validation

The two sets: **i)** SVM, MLP and RF with all the selected features and, **ii)** the same three algorithms using the best subset of features (SVMb, MLPb and RFb), were evaluated with the statistical tests.

From the average of ranks, Table 5.22, the values $X_F^2 = 54.34$ and $X_F^2 = 76.73$ were obtained, for the F1 score and GE respectively, equation (2.11). As a result, the null hypothesis "Ho: all the methods behave similarly" was rejected for the two metrics, since the critical value for 6 classifiers (k=6) and a significance level of 5% is 11.07 (Table 2.1).

Table 5.22: Average ranks of the Rules model classifiers using the F1 score and GE metrics - bariatric patients

	SVM	MLP	RF	SVMb	MLPb	RFb
Average rank using F1 score	5.050	5.225	4.275	2.575	2.450	1.425
Average rank using GE	4.650	5.025	5.250	2.150	1.750	2.175

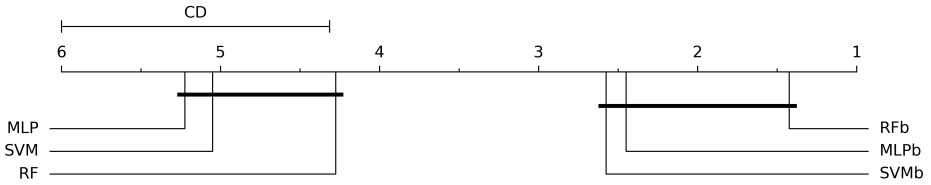
Therefore, the Nemenyi test was employed considering a critical distance of $CD = 1.69$, as we performed a comparison between 6 methods, equation (2.12). The results are represented in the next tables and the respective diagrams are illustrated in Figure 5.13.

Table 5.23: Nemenyi test for the Rules model classifiers using the F1 score metric

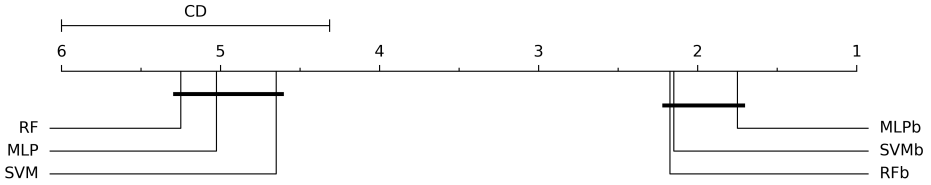
Algorithms	MLP	RF	SVMb	MLPb	RFb
SVM	0.18	-0.78	-2.48	-2.60	-3.63
MLP		-0.95	-2.65	-2.78	-3.80
RF			-1.70	-1.83	-2.85
SVMb				-0.13	-1.15
MLPb					-1.03

Table 5.24: Nemenyi test for the Rules model classifiers using the GE metric

Algorithms	MLP	RF	SVMb	MLPb	RFb
SVM	0.38	0.60	-2.50	-2.90	-2.48
MLP		0.23	-2.88	-3.28	-2.85
RF			-3.10	-3.50	-3.08
SVMb				-0.40	0.03
MLPb					0.43



(a) F1 scores results.



(b) GE results.

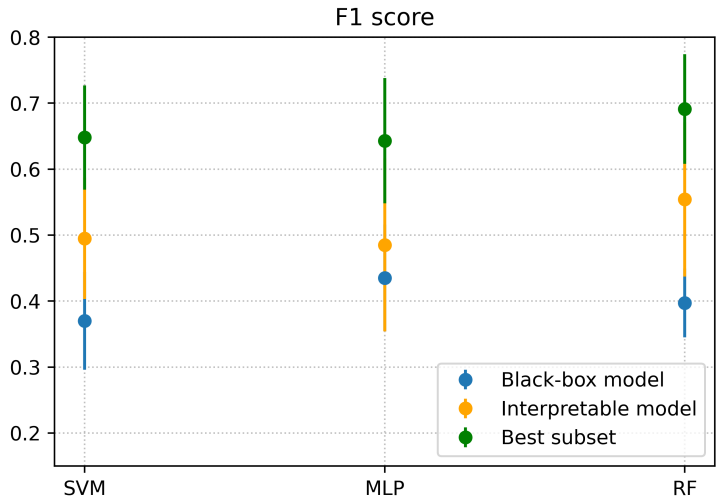
Figure 5.13: Nemenyi diagrams of the Rules model classification algorithms for the F1 score (a) and GE (b) - bariatric patients

The differences between the classifiers average ranks presented in the previous tables demonstrate that, for both metrics, there is a significantly difference between the performance of the classifiers with all the rules and with the best subset of them.

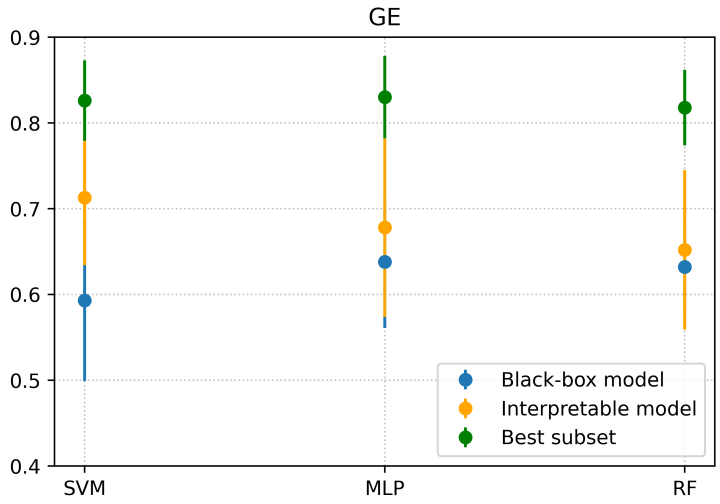
On the other hand, within the group of classifiers: SVM, MLP and RF, the critical distance wasn't reached. The same occurred with the set of the classifiers: SVMb, MLPb and RFb. From these findings, we can conclude that the selection of the best subset of rules for each model is a strategy that can improve the model's performance.

5.10 Comparison of the Black-box and Rules models implemented in the bariatric patients

Finally, the performance of all the models classifiers, using the two metrics, was compared and the results are illustrated in Figure 5.14. It can be observed an increment of the mean values of both metrics for the Rules model, in relation to the Black-box model. Nevertheless, for an adequate comparison of these findings we proceeded with the same statistical validation process implemented for the cardiac patients.



(a) F1 scores results.



(b) GE results.

Figure 5.14: F1 score and GE evaluation of all the classification algorithms of the two models (Black-box and Rules): mean values and standard deviations - bariatric patients

5.10.1 Statistical validation

For the statistical validation, the classifiers with the best average rank of the Black-box and Rules models were selected in order to compare the performance of each model. In this sense, the MLP was the classifier selected from the Black-box model (best ranked in terms of F1 score and GE). When evaluating the Rules model, the RF and the RFb were selected for the F1 score, and the SVM and MLPb for the GE metric.

The three classifiers selected for each metric were then compared and the resulting average ranks, in terms of GE and F1 score, are presented in Table 5.25.

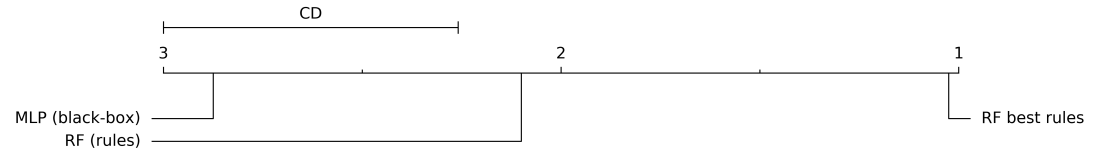
Table 5.25: Average ranks of each model's best classifier using the F1 score and GE metrics - bariatric patients

F1 score	MLP	RF	RFb
Average rank	2.88	2.10	1.03
GE	MLP	SVM	MLPb
Average rank	2.75	2.25	1.00

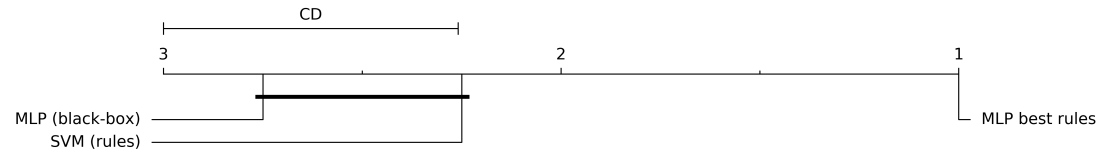
The critical value for the 3 classifiers (k=3) was 5.99. From the average of ranks, the Friedman statistic values of $X_F^2 = 35.41$ and $X_F^2 = 32.50$ were obtained for the F1 score and GE respectively, equation (2.11), and the null hypothesis could be rejected for both metrics. Thus, the Nemenyi test was performed with a respective critical distance of 0.74.

The Nemenyi test results represented in the Table 5.26 allow us to conclude that, when evaluating the F1 score, the Rules model achieved a performance significantly better than the Black-box model ($D = -0.78$). Moreover, the RBb also outperformed the RF of the Rules model ($D = -1.07$). Therefore, they are plotted in the Nemenyi diagram of Figure 5.15 a) without any bold line connecting them.

When evaluating the GE (Table 5.27), only the MLPb achieved the critical distance in comparison with the other classifiers, with $D = -1.75$ and $D = -1.25$, when comparing with the MLP (Black-box) and SVM (Rules) respectively.



(a) F1 scores results.



(b) GE results.

Figure 5.15: Nemenyi diagrams of the best classification algorithms of each model for the F1 score (a) and GE (b) metric - bariatric patients

Table 5.26: Nemenyi test of the best classification algorithms of each model using the F1 score metric - bariatric patients

Algorithms	RF (rules)	RFb
MLP (black-box)	-0.78	-1.85
RF (rules)		-1.07

Table 5.27: Nemenyi test of the best classification algorithms of each model using the GE metric - bariatric patients

Algorithms	SVM (rules)	MLPb
MLP (black-box)	-0.50	-1.75
SVM (rules)		-1.25

In sum, it can be concluded that the Rules model was better than the Black-box one in predicting the LOS. The Nemenyi test validated that, the performance of the Rules model, in terms of GE, is significantly better than the Black-box model when using the best subset of rules, and the performance increase is also statistically significant when evaluating the F1 score (using all the rules and also the best subset of them). Therefore, the Rules model should be the preferable model to implement for the LOS prediction task, providing also higher interpretability and patient customization.

5.11 LOS classification performance of the SCORE model

The performance of the SCORE based model was also evaluated in the same 20 test sets of the cardiac patients (ICU dataset) and compared with the Black-box model. The mean GE obtained by the SCORE model was 0.50 with a standard deviation of 0.02 which is clearly lower than the worst Black-box classifier GE mean value of 0.58 ± 0.03, therefore, no statistic validation was performed since we could easily conclude that the SCORE model was less accurate to estimate the actual LOS.

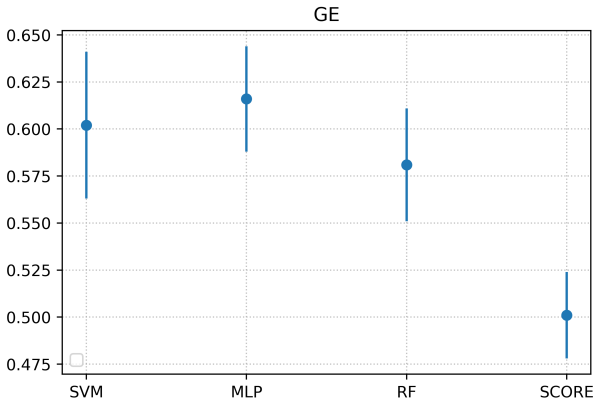


Figure 5.16: Performance comparison between the GE of the Black-box and SCORE models - cardiac patients

These results are in line with the findings of Chapter 5.1. Although this prognostic risk tool is highly used for long term predictions in the context of Portuguese hospitals, for a more adequate prediction of the LOS the development of a specific models for that task is preferable.

6

Conclusion and Future work

In this work, different approaches for LOS prediction of two types of patients, cardiac and bariatric, were developed and validated. The obtained results enabled us to achieved the initial goals, that addressed four main research lines: **1.** Exploration of available risk tools (SCORE risk), **2.** Application of typical computational intelligence models, **3.** Development of an interpretable and patient customized model, and **4.** Integration of dynamic data (vital signs) in the previous models.

Firstly, the features were selected using the Kendall's tau coefficient correlation, and, for the cardiac patients, they were also validated by the ICU cardiologist. The features selected for both types of patients are consistent with studies found in literature that refer them as good predictors.

In the first approach we implemented the SCORE risk tool, a prognostic risk tool widely used in the European cardiology practice, to assess its capacity to estimate the LOS of the cardiac patients. We concluded that, although this risk tool is of high importance in the cardiac context for the Portuguese hospitals, it isn't adequate for the LOS prediction task and, for a more adequate prediction, the development of a specific model for this task is preferable. On the other hand, despite not being within the scope of this thesis, the application of a short-term risk score seemed to us that could have more potential. In fact, this study was carried out as part of a scientific publication, using the GRACE model for this purpose, a model normally used in Portuguese hospitals (Appendix B).

Secondly, typical computational intelligence models were employed in the cardiac patients, due to their demonstrated good performance in previous literature studies. We verified that, this model clearly outperformed the previous one, achieving a $GE = 0.62 \pm 0.03$ with the MLP. The value achieved doesn't seem satisfactory enough for the implementation of the model in clinical institutions and demonstrates the difficulty to predict the length of stay using only the patient's characteristics at ICU admission time, as concluded also by Verbug et al.[72]. This performance level was expected since we verified that the strength of association between the variables and the LOS was very low (lower than 0.2 for the majority).

Furthermore, an interpretable and patient customized model was developed, based on

rules. These are major advantages for a LOS prediction model, contributing to an increase of its clinical relevance and acceptance. It was demonstrated a similar performance of this model in comparison to the Black-box one, obtaining a slightly higher GE ($GE = 0.63 \pm 0.02$), but the use of the best subset of rules was statistically proved to reached even better results ($GE = 0.66 \pm 0.02$), outperforming the previous models.

Concerning the bariatric patients, we started by exploring the influence of the vital signs addition to the model and verified that it leaded to a performance increase. However, more experiments need to be carried out for a more deep exploration.

Finally, with the vital signs added to the input features, both Black-box and Rules models were also employed in these patients, and their performances compared. Here again, we concluded that the Rules model was preferable, since it reached a similar performance in terms of GE, using all the rules, outperforming the Black-box model when employing the best subset of them, and when evaluating with the F1 score. The Black-box classifier with the highest mean values of the evaluation metrics was also the MLP (as for the cardiac patients) with a GE of 0.64 ± 0.08 and a F1 score of 0.44 ± 0.08. The interpretable model reached a mean GE of 0.71 ± 0.08, using the SVM with all the selected features (rules), and 0.83 ± 0.05 with the MLP employing the best subset of them.

In conclusion, the Rules model was the preferable LOS prediction model for both types of patients. Although the values achieved still don't seem sufficiently high for an implementation of the model in clinical institutions, its interpretability and patient customization can give an insight of the more relevant risk factors and serve as one of the inputs that can contribute to the decision making of the clinicians.

As future work, different improvements and experiments can be done. The two models should be applied in datasets with more balanced data, since it was the major problem, specially, for the TRICA patients. The implementation of these models in bigger datasets would also be beneficial to improve the learning of the classifiers applied.

Moreover, the study of other variables acquired, not only at admission time, but during the first hours or on the first day of the patient's stay in the ICU, could be a strategy to explore in the future that may lead to a better performance.

Furthermore, an alternative exploration may be the validation of the Rules model in other clinical problems, for example, in the cardiovascular risk prediction, and comparison with existing risk models, as the GRACE and SCORE, and also with other data-based models (MLP, SVM ...).

On the other hand, the first results of the integration of dynamic data were optimistic, therefore, motivating an extended exploration: i) other dynamic data, besides the vital signs, can be analyzed, ii) the time series data (continuous measurements) can be used instead of the features created from them (mean, variance...) and iii) the use of dynamic signals should be tested in different types of patients.

Bibliography

[1] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare Informatics Research*, vol. 19, no. 2, pp. 121–129, 2013.

[2] H. Maharlou, S. R. Niakan Kalhori, S. Shahbazi, and R. Ravangard, "Predicting length of stay in intensive care units after cardiac surgery: Comparison of artificial neural networks and adaptive neuro-fuzzy system," *Healthcare Informatics Research*, vol. 24, no. 2, pp. 109–117, 2018.

[3] M. Sud, B. Yu, H. C. Wijeysundera, P. C. Austin, D. T. Ko, J. Braga, P. Cram, J. A. Spertus, M. Domanski, and D. S. Lee, "Associations Between Short or Long Length of Stay and 30-Day Readmission and Mortality in Hospitalized Patients With Heart Failure," *JACC: Heart Failure*, vol. 5, no. 8, pp. 578–588, 2017.

[4] H. F. Lingsma, A. Bottle, S. Middleton, J. Kievit, E. W. Steyerberg, and P. J. Marang-Van De Mheen, "Evaluation of hospital outcomes: The relation between length-of-stay, readmission, and mortality in a large international administrative database," *BMC Health Services Research*, vol. 18, no. 1, pp. 1–10, 2018.

[5] O. Andersson, "Predicting Patient Length Of Stay at Time of Admission Using Machine Learning," 2019.

[6] T. A. Jilani, H. Yasin, M. Yasin, and C. Ardil, "Acute coronary syndrome prediction using data mining techniques- An application," *World Academy of Science, Engineering and Technology*, vol. 59, pp. 474–478, 2009.

[7] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable Machine Learning in Healthcare," pp. 559–560, 2018.

[8] V. Stangenberger, "Predicting the length of stay for optimizing hospital level capacity planning in Dutch hospitals," 2018.

[9] L. Lella, A. Di Giorgio, and A. F. Dragoni, "Length of stay prediction and analysis through a growing neural gas model," *CEUR Workshop Proceedings*, vol. 1389, pp. 11–21, 2015.

[10] S. Paredes, J. Henriques, T. Rocha, P. de Carvalho, and J. Morais, "Identification of Clinically Relevant Rules: An Interpretable Approach for CVD Risk Assessment," 2020.

[11] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for today's critically ill patients," *Critical Care Medicine*, vol. 34, pp. 1297–1310, 2006.

[12] M. Ghorbani, H. Ghaem, A. Rezaianzadeh, Z. Shayan, F. Zand, and R. Nikandish, "A study on the efficacy of APACHE-IV for predicting mortality and length of stay in an intensive care unit in Iran," *F1000Research*, vol. 6, no. May, pp. 1–9, 2017.

[13] S. T. Rocha, D. F. Pizzol, C. Ritter, C. M. Fraga, D. C. Tamiozo, and V. H. P. Ricci, "Desempenho do escore SAPS II em uma unidade de terapia intensiva," *Arquivos Catarinenses de Medicina*, vol. 41, no. 4, pp. 26–31, 2012.

[14] E. W. Tang, C.-k. Wong, P. Herbison, and N. Zealand, "Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome," 2007.

[15] K. F. Adams, W. T. Abraham, C. W. Yancy, and W. J. Boscardin, "Risk Stratification for In-Hospital Mortality in Acutely Decompensated Heart Failure Classification and Regression Tree Analysis," vol. 293, no. 5, 2005.

[16] P. D. Levy, T. O. Neill, P. C. Austin, and J. V. Tu, "Prospective Validation of the Emergency Heart Failure Mortality Risk Grade for Acute," pp. 1146–1156, 2019.

[17] J. Allyn, C. Ferdynus, M. Bohrer, C. Dalban, D. Valance, and N. Allou, "Simplified acute physiology score II as predictor of mortality in intensive care units: A decision curve analysis," *PLoS ONE*, vol. 11, no. 10, 2016.

[18] R. B. D. Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel, "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study," 2008.

[19] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study," *British Medical Journal*, vol. 335, no. 7611, pp. 136–141, 2007.

[20] "Score risk charts, the european cardiovascular disease risk assessment model." European Society of Cardiology. Available at `https://www.escardio.org/Education/Practice-Tools/CVD-prevention-toolbox/SCORE-Risk-Charts` [Accessed: 09.11.2020].

[21] A. Godinjak, A. Iglica, A. Rama, I. Tančica, and S. Jusufović, "Predictive value of SAPS II and APACHE II scoring systems for patient outcome in a medical intensive care unit," vol. 45, no. 2, pp. 97–103, 2016.

[22] A. Junior, A. Mocelin, F. Andrade, L. Brauer, F. Giannini, A. Nunes, and C. Dias, "ORIGINAL ARTICLE SAPS 3 , APACHE IV or GRACE : which score to choose for acute coronary syndrome patients in intensive care units ? SAPS 3 , APACHE IV ou GRACE : qual escore escolher para pacientes," vol. 131, no. 3, pp. 173–178, 2013.

[23] A. Awad, M. Bader–El–Den, and J. McNicholas, "Modeling and Predicting Patient Length of Stay: A Survey," *Journal of Advanced Scientific Research and Management*, vol. 1, no. 8, pp. 90–101, 2016.

[24] S. Tanuja, D. U. Acharya, and K. R. Shailesh, "Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay," *Journal of Pharmaceutical and Biomedical Sciences*, vol. 7, no. 7, pp. 1–4, 2011.

[25] P. F. J. Tsai, P. C. Chen, Y. Y. Chen, H. Y. Song, H. M. Lin, F. M. Lin, and Q. P. Huang, "Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network," *Journal of Healthcare Engineering*, vol. 2016, no. April, 2016.

[26] R. Paterson, D. C. MacLeod, D. Thetford, A. Beattie, C. Graham, S. Lam, and D. Bell, "Prediction of in-hospital mortality and length of stay using an early warning scoring system: Clinical audit," *Clinical Medicine, Journal of the Royal College of Physicians of London*, vol. 6, no. 3, pp. 281–284, 2006.

[27] S. Barth, K. Buell, A. Nanavati, S. Wijetilleka, J. F. D. Wolff, and R. Tennant, "Using National Early Warning Scores to Predict Length of Stay and Appropriate Use of Short Stay Medical Teams," 2016.

[28] Medtronic, "Clinical evidence guide - early warning scores." Available at https://www.medtronic.com/content/dam/covidien/library/us/en/product/health-informatics-and-monitoring/vital-sync-remote-monitoring-early-warning-score-clinical-evidence-guide.pdf [Accessed: 28.03.2020].

[29] N. Zealand and D. Health, "Recognising and Responding to Clinical Deterioration : Background paper June 2008," *Signs*, no. June, pp. 1–16, 2008.

[30] R. M. Conroy, K. Pyörälä, A. P. Fitzgerald, S. Sans, A. Menotti, G. De Backer, D. De Bacquer, P. Ducimetière, P. Jousilahti, U. Keil, I. Njølstad, R. G. Oganov, T. Thomsen, H. Tunstall-Pedoe, A. Tverdal, H. Wedel, P. Whincup, L. Witheimsen, and I. M. Graham, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, 2003.

[31] "Acute coronary syndrome." American Heart Association, Available at `https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks/acute-coronary-syndrome` [Accessed: 14.08.2020].

[32] "Acute coronary syndrome." mayo Clinic, Available at `https://www.mayoclinic.org/diseases-conditions/acute-coronary-syndrome/symptoms-causes/syc-20352136` [Accessed: 07.02.2020].

[33] C. W. Hamm, J. P. Bassand, S. Agewall, J. Bax, E. Boersma, H. Bueno, P. Caso, D. Dudek, S. Gielen, K. Huber, M. Ohman, M. C. Petrie, F. Sonntag, M. S. Uva, R. F. Storey, W. Wijns, and D. Zahger, "ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation," *European Heart Journal*, vol. 32, pp. 2999–3054, 2011.

[34] B. Ibanez, S. James, S. Agewall, M. J. Antunes, C. Bucciarelli-Ducci, H. Bueno, A. L. Caforio, F. Crea, J. A. Goudevenos, S. Halvorsen, G. Hindricks, A. Kastrati, M. J. Lenzen, E. Prescott, M. Roffi, M. Valgimigli, C. Varenhorst, P. Vranckx, and P. Widimský, "2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation," *European Heart Journal*, vol. 39, no. 2, pp. 119–177, 2017.

[35] M. Roffi, C. Patrono, J. P. Collet, C. Mueller, M. Valgimigli, F. Andreotti, J. J. Bax, M. A. Borger, C. Brotons, D. P. Chew, B. Gencer, G. Hasenfuss, K. Kjeldsen, P. Lancellotti, U. Landmesser, J. Mehilli, D. Mukherjee, R. F. Storey, and S. Windecker, "2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent st-segment elevation: Task force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the european society of cardiology (ESC)," *European Heart Journal*, vol. 37, no. 3, pp. 267–315, 2015.

[36] "Bariatric surgery procedures." American Society for Metabolic and Bariatric Surgery, Available at https://asmbs.org/patients/bariatric-surgery-procedures [Accessed: 04.08.2020].

[37] F. Mahmood, A. J. Sharples, A. Rotundo, N. Balaji, and V. S. Rao, "Factors predicting length of stay following bariatric surgery: Retrospective review of a single UK tertiary centre experience," *Obesity Surgery*, vol. 28, no. 7, pp. 1924–1930, 2018.

[38] D. Normando, "Tutorial para a escolha do teste estatístico para a pesquisa científica." [Accessed: 26.05.2020].

[39] "Kendall's tau and spearman's rank correlation coefficient." Available at https://www.statisticssolutions.com/kendalls-tau-and-spearmans-rank-correlation-coefficient/ [Accessed: 15.12.2019].

[40] "Spearman's rho and kendall's tau." Posted on July 8, 2019, Statistical Odds Ends, Available at https://statisticaloddsandends.wordpress.com/2019/07/08/spearmans-rho-and-kendalls-tau/ [Accessed: 11.11.2019].

[41] "Kendall's tau-b using spss statistics." Laerd statistics, Available at https://statistics.laerd.com/spss-tutorials/kendalls-tau-b-using-spss-statistics.php [Accessed: 01.09.2020].

[42] E. Marshall and E. Boggis, "The Statistics Tutor' s Quick Guide to Commonly Used Statistical Tests,"

[43] D. Sheskin, "Handbook of Parametric and Nonparametric statistical procedures second edition," 2000. Chapman Hall/CRC.

[44] "Kendall rank correlation coefficient." Available at https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient [Accessed: 21.12.2019].

[45] P. Liu, L. Lei, J. Yin, W. Zhang, W. Naijun, and E. El-Darzi, "Healthcare data mining: Prediction inpatient length of stay," *IEEE Intelligent Systems*, pp. 832–837, 2006.

[46] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, and M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: A machine learning approach," *International Journal of Cardiology*, vol. 288, pp. 140–147, 2019.

[47] L. Breiman, "ST4_Method_Random_Forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[48] D. Murphy, "Using Random Forest Machine Learning Methods to Identify Spatiotemporal Patterns of Cheatgrass Invasion through Landsat Land Cover Classification in the Great Basin from 1984 - 2011," no. May, 2019.

[49] C. Burges and C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, January 1998.

[50] S. Ray, "Understanding support vector machine (svm) algorithm from examples (along with code)." Available at https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [Accessed:11.06.2020].

[51] Z. Li, R. Outbib, S. Giurgea, D. Hissel, and S. Jeme, "Online implementation of SVM based fault diagnosis strategy for PEMFC Online implementation of SVM based fault diagnosis strategy for PEMFC systems," no. February, 2015.

[52] L. Auria and R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis," *SSRN Electronic Journal*, no. August, 2011.

[53] K. U. Rani, "Analysis of heart diseases dataset using neural network approach," *International Journal of Data Mining & Knowledge Management Process*, vol. 1, no. 5, pp. 1–8, 2011.

[54] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[55] "Artificial neural network - perceptron." Available at `https://www.saedsayad.com/artificial_neural_network_bkp.htm` [Accessed: 02.09.2020].

[56] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[57] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining : Experimental analysis of power," *Information Sciences*, vol. 180, pp. 2044–2064, 2010.

[58] J. Gardner and C. Brooks, "Statistical approaches to the model comparison task in learning analytics," *CEUR Workshop Proceedings*, vol. 1915, 2017.

[59] J. Pizarro, E. Guerrero, and P. L. Galindo, "Multiple comparison procedures applied to model selection," *Neurocomputing*, vol. 48, no. 1-4, pp. 155–173, 2002.

[60] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[61] T. Rocha, "Similarity based approaches for the analysis and prediction of physiologic time series," 2012.

[62] "Friedman's Two-way Analysis of Variance by Ranks –Analysis of k-Within-Group Data with a Quantitative Response Variable,"

[63] J. Gardner, A. Arbor, and A. Arbor, "A Statistical Framework for Predictive Model Evaluation in MOOCs," pp. 269–272, 2017.

[64] F. YILMAZ, H. K. İLTER, S. MALHAN, and R. NUMANOĞLU TEKİN, "Determination of Factors Affecting Length of Stay With Multinomial Logistic Regression in Turkey," *Sosyal Guvence*, no. 10, pp. 74–74, 2016.

[65] M. Arab, A. Zarei, F. Rahimi, A.and Rezaiean, and F. Akbari, "Analysis of factors affecting length of stay in public hospitals in lorestan province, iran," *Hakim Health Systems Research Journal*, vol. 12, no. 4, 2010.

[66] S. Aghajani and M. Kargari, "Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department," *Hospital Practices and Research*, vol. 1, no. 2, pp. 53–58, 2016.

[67] L. Penzenstadler, A. Chatton, G. Thorens, D. Zullino, and Y. Khazaal, "Factors influencing the length of hospital stay of patients with substance use disorders," *Journal of Substance Use*, vol. 5, pp. 201–209, 2020.

[68] T. Magalhães, S. Lopes, J. Gomes, and F. Seixo, "The Predictive Factors on Extended Hospital Length of Stay in Patients with AMI: Laboratory and Administrative Data," *Journal of Medical Systems*, vol. 40, no. 1, 2016.

[69] J. H. Wasfy, K. F. Kennedy, F. A. Masoudi, T. G. Ferris, S. V. Arnold, V. Kini, P. Peterson, J. P. Curtis, A. P. Amin, S. M. Bradley, W. J. French, J. Messenger, P. M. Ho, and J. A. Spertus, "Predicting Length of Stay and the Need for Postacute Care After Acute Myocardial Infarction to Improve Healthcare Efficiency," *Circulation. Cardiovascular quality and outcomes*, vol. 11, no. 9, p. e004635, 2018.

[70] J. Carter, S. Elliott, J. Kaplan, M. Lin, A. Posselt, and S. Rogers, "Predictors of hospital stay following laparoscopic gastric bypass: Analysis of 9,593 patients from the National Surgical Quality Improvement Program," *Surgery for Obesity and Related Diseases*, vol. 11, no. 2, pp. 288–294, 2015.

[71] D. Mahadevan, C. Challand, and J. Keenan, "Revision total hip replacement: Predictors of blood loss, transfusion requirements, and length of hospitalisation," *Journal of Orthopaedics and Traumatology*, vol. 11, no. 3, pp. 159–165, 2010.

[72] I. W. Verburg, N. F. De Keizer, E. De Jonge, and N. Peek, "Comparison of regression methods for modeling intensive care length of stay," *PLoS ONE*, vol. 9, no. 10, 2014.

[73] R. Houthooft, J. Ruyssinck, J. van der Herten, S. Stijven, I. Couckuyt, B. Gadeyne, F. Ongenae, K. Colpaert, J. Decruyenaere, T. Dhaene, and F. De Turck, "Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores," *Artificial Intelligence in Medicine*, vol. 63, no. 3, pp. 191–207, 2015.

[74] L. Turgeman, J. H. May, and R. Sciulli, "Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission," *Expert Systems with Applications*, vol. 78, pp. 376–385, 2017.

[75] M. Rowan, T. Ryan, F. Hegarty, and N. O'Hare, "The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors," *Artificial Intelligence in Medicine*, vol. 40, no. 3, pp. 211–221, 2007.

[76] K. K. Oad, X. Dezhi, and P. K. Butt, "A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease," vol. 14, no. 3, 2014.

[77] S. Merja, R. H. Lilien, and H. F. Ryder, "Clinical Prediction Rule for Patient Outcome after In-Hospital CPR : A New Model , Using Characteristics Present at Hospital Ad-

mission , to Identify Patients Unlikely to Benefit from CPR after In-Hospital Cardiac Arrest," pp. 19–27, 2015.

[78] E. P. Hess, R. J. Brison, J. J. Perry, and L. A. Calder, "Development of a Clinical Prediction Rule for 30-Day Cardiac Events in Emergency Department Patients With Chest Pain and Possible Acute Coronary Syndrome," *YMEM*, vol. 59, no. 2, pp. 115–125.e1, 2012.

[79] D. A. Gilhooly, M. Cole, and S. R. Moonesinghe, "The evaluation of risk prediction models in predicting outcomes after bariatric surgery: a prospective observational cohort pilot study," *Perioperative Medicine*, vol. 7, no. 1, pp. 1–9, 2018.

[80] D. S. Miguel, P. Ramos, J. Oliveira, C. Ferreira, and F. Cruz, "OS-MRS as a predictor of hospital length of stay – a retrospective audit of patients submitted to elective gastric bypass surgery," *Anaesthesia, Pain and Intensive Care*, vol. 21, no. 1, pp. 54–58, 2020.

[81] N. Alam, I. L. Vegting, E. Houben, B. van Berkel, L. Vaughan, M. H. Kramer, and P. W. Nanayakkara, "Exploring the performance of the National Early Warning Score (NEWS) in a European emergency department," *Resuscitation*, vol. 90, no. March, pp. 111–115, 2015.

[82] I. J. Brekke, L. H. Puntervoll, P. B. Pedersen, J. Kellett, and M. Brabrand, "The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review," *PLoS ONE*, vol. 14, no. 1, pp. 1–13, 2019.

[83] I. Ramzy, "Definition of hypertension and pressure goals during treatment (esc-esh guidelines 2018)." Vol. 17, N 18 - 14 Aug 2019, European Society of Cardiology, Available at https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-17/definition-of-hypertension-and-pressure-goals-during-treatment-esc-esh-guidelin [Accessed: 09.01.2020].

[84] J. Brownlee, "SMOTE for Imbalanced Classification with Python," Machine Learning Mastery, Available at `https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/` [Accessed: 16.01.2020].

[85] M. Pandey and A. Jain, "ROC Curve : Making way for correct diagnosis," *PharmaSUG*, pp. 1–12, 2016.

[86] A. Michils, R. Louis, R. Peché, S. Baldassarre, and A. Van Muylem, "Exhaled nitric oxide as a marker of asthma control in smoking patients," *European Respiratory Journal*, vol. 33, no. 6, pp. 1295–1301, 2009.

[87] I. T. Hsiao and L. Gao, "Models of Bankruptcy Prediction Since the Recent Financial Crisis: KMV, Naïve, and Altman's Z-score," 2016.

[88] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. D. Ferranti, J.-p. Després, H. J. Fullerton, V. J. Howard, M. D. Huffman, C. R. Isasi, M. C. Jiménez, S. E. Judd, B. M. Kissela, J. H. Lichtman, L. D. Lisabeth, S. Liu, R. H. Mackey, D. J. Magid, D. K. Mcguire, E. R. M. Iii, C. S. Moy, P. Muntner, M. E. Mussolino, K. Nasir, R. W. Neumar, G. Nichol, L. Palaniappan, D. K. Pandey, M. J. Reeves, C. J. Rodriguez, W. Rosamond, P. D. Sorlie, J. Stein, A. Towfighi, T. N. Turan, S. S. Virani, D. Woo, R. W. Yeh, and M. B. Turner, *AHA Statistical Update Heart Disease and Stroke Statistics — 2016 Update A Report From the American Heart Association WRITING GROUP MEMBERS*. 2016.

[89] I. T. Hsiao and L. Gao, "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity ," Vitamin and Mineral Nutrition Information System. Geneva, World Health Organization, 2011 (WHO/NMH/NHD/MNM/11.1). Available at `http://www.who.int/vmnis/indicators/haemoglobin.pdf`.

[90] L. K. Newby, R. L. Jesse, J. D. Babb, R. H. Christenson, T. M. D. Fer, G. A. Diamond, F. M. Fesmire, S. A. Geraci, G. C. Larsen, S. Kaul, C. R. Mckay, G. J. Philippides, and W. S. Weintraub, "EXPERT CONSENSUS DOCUMENT ACCF 2012 Expert Consensus Document on Practical Clinical Considerations in the Interpretation of Troponin Elevations," vol. 60, no. 23, 2012.

[91] D. Kancherla, J. D. Bodapati, and N. Veeranjaneyulu *International Journal of Recent Technology and Engineering*.

[92] S. Narkhede, "Understanding auc - roc curve." Towards data science, Available at https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Accessed: 13.07.2020].

Appendices

A

Results of the 30 repeated permutations followed by two-fold cross-validation sampling technique

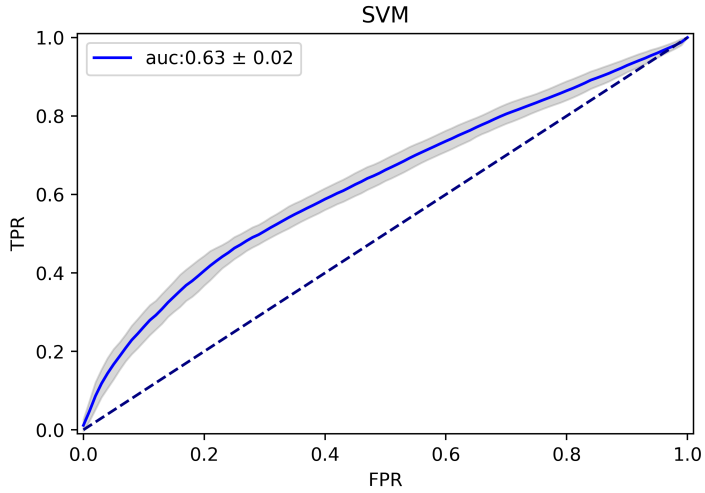
A.1 Black-box model

Grid search results

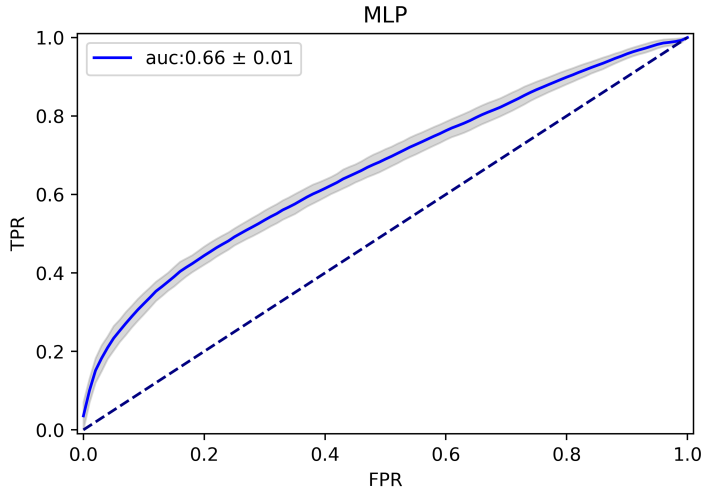
Table A.1: Grid search results for the Black-box model (cardiac patients)

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer size	number of estimators
Most common option	RBF	10	100

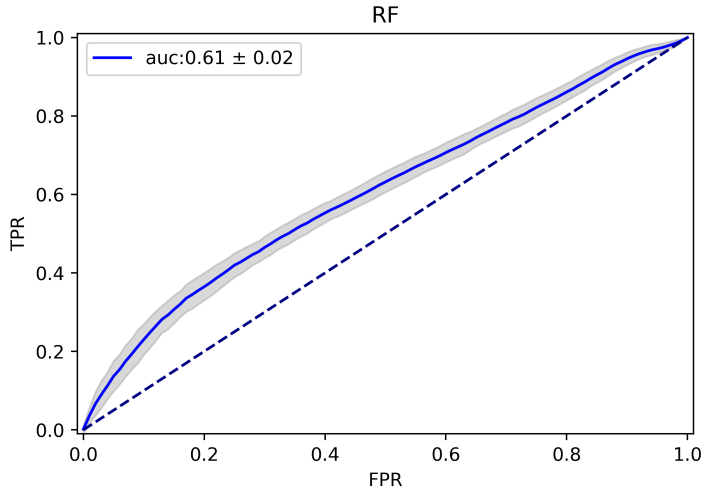
ROC Curves



(a) Mean SVM ROC Curve.



(b) Mean MLP ROC Curve.



(c) Mean RF ROC Curve.

Figure A.1: Mean ROC Curves of Black-box algorithms: cardiac patients

LOS Classification

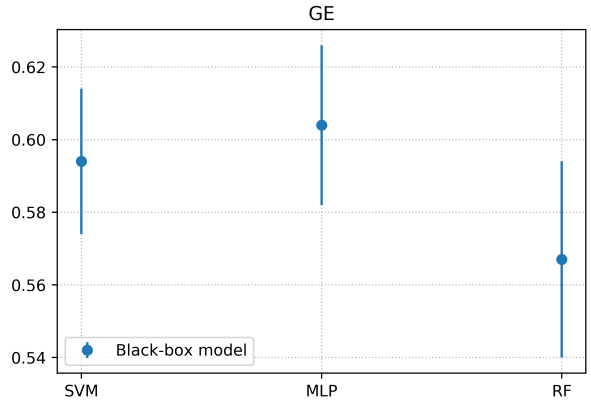


Figure A.2: Mean values and standard deviation of the GE for the classification algorithms of the Black-box model

Table A.2: Geometric mean evaluation of the Black-box model classification algorithms: mean values and standard deviations

Algorithm	SVM	MLP	RF
GE	0.59 ± 0.02	0.60 ± 0.02	0.57 ± 0.03

Statistical validation

Table A.3: Average ranks using the Geometric mean metric - Black-box model

	SVM	MLP	RF
Average rank	1.800	1.458	2.741

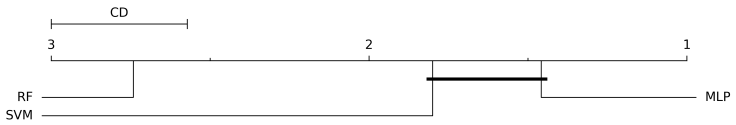


Figure A.3: Nemenyi diagram of the Black-box model classification algorithms for the GE

Table A.4: Nemenyi test

Algorithms	MLP	RF
SVM	-0.34	0.94
MLP		1.28

A.2 Rules model

Grid search

Table A.5: Grid search results for the Rules model

Algorithm	SVM	MLP	RF
Hyperparameter	kernel type	hidden layer size	number of estimators
Most common option	RBF	12	100

LOS Classification

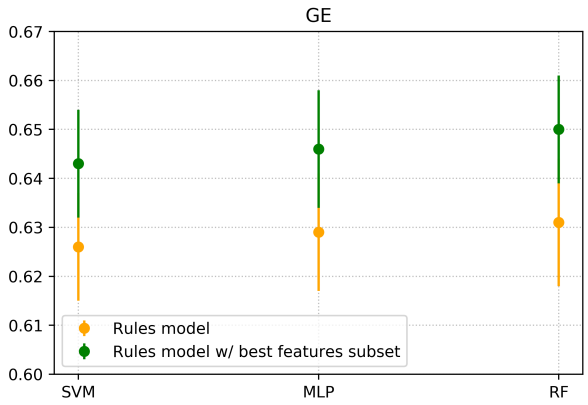


Figure A.4: Mean values and standard deviation of GE for the classification algorithms of the Rules model

Table A.6: Geometric mean evaluation of the Rules model classification algorithms: mean values and standard deviations

Algorithm	SVM	MLP	RF	SVMb	MLPb	RFb
GE	0.63±0.01	0.63±0.01	0.63±0.01	0.64±0.01	0.65±0.01	0.65±0.01

Statistical validation

Table A.7: Average ranks using the Geometric mean metric - Rules model

	SVM	MLP	RF	SVMb	MLPb	RFb
Average rank	5.350	4.808	4.700	2.617	2.158	1.367

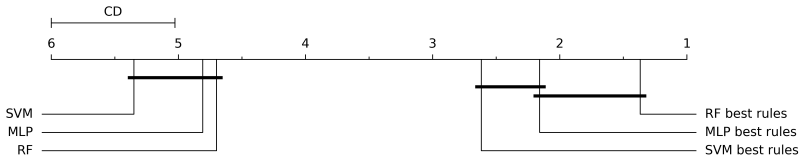


Figure A.5: Nemenyi diagram of the Rules model classification algorithms for the GE

Table A.8: Nemenyi test

Algorithms	MLP	RF	SVMb	MLPb	RFb
SVM	-0.54	-0.65	-2.73	-3.19	-3.98
MLP		-0.11	-2.19	-2.65	-3.44
RF			-2.08	-2.54	-3.33
SVMb				-0.46	-1.25
MLPb					-0.79

Comparison of the Black-box and Rules models implemented in the cardiac patients

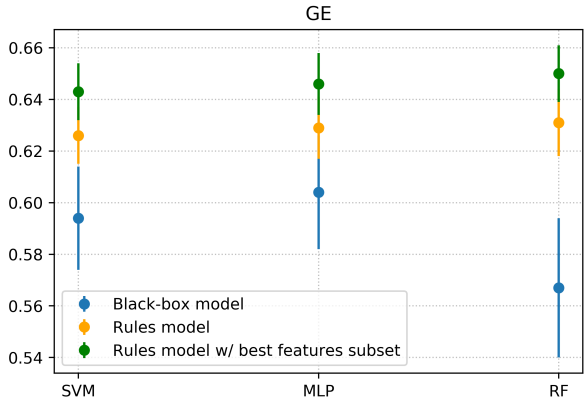


Figure A.6: Mean values and standard deviation of GE for all the classification algorithms of the 2 models (Black-box and Rules)

Statistical validation

Table A.9: Average ranks of each model's best classifier using the Geometric mean metric

	MLP	RF	RF best subset
Average rank	2.95	2.05	1.00

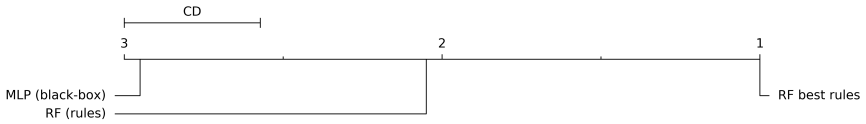


Figure A.7: Nemenyi diagram of the best classification algorithms of the two models for the GE metric

Table A.10: Nemenyi test of best classification algorithms of the two models for the GE metric

Algorithms	RF	RF best subset
MLP	-0.90	-1.95
RF		-1.05
RF best subset		

B

The use of SCORE and GRACE risk tools to assess the Length of Stay in a Cardiac Intensive Care Unit

The use of SCORE and GRACE risk tools to assess the Length of Stay in a Cardiac Intensive Care Unit

Cláudia Lopes[1], Jorge Henriques[1], Paulo de Carvalho[1], Lino Gonçalvez[2], Carolina Négrier[2], José Pedro Sousa[2], Alberto Bonomi[3]

[1] Center for Informatics and Systems of University of Coimbra – CISUC, Coimbra, Portugal; lopes.cr97@gmail.com, jh@dei.uc.pt, carvalho@dei.uc.pt
[2] Centro Hospitalar Universitário de Coimbra, Coimbra, Portugal. lgoncalv@ci.uc.pt, carolinanegrier@gmail.com, zpedro_14@hotmail.com
[3] Patient Care and Measurements department, Philips Research, Eindhoven, The Netherlands alberto.bonomi@philips.com

Abstract. The possibility of using simple and effective models to estimate the patient's length of stay in intensive care units is decisive to support the clinical professional decisions. These models can help professionals in the stratification process and, particularly, in the identification of the necessary intervention plan to improve the patient's health condition. In clinical practice specific prognostic scores are available and validated in the cardiovascular context. These risk tools address the primary prevention domain, as well as the secondary prevention domain, usually involving long-term (years) and short-term (months) prediction periods, respectively.

The aim of this study is to investigate the capacity of available prognosis risk tools, in particular SCORE (primary tool) and GRACE (secondary tool), to estimate the length of stay in a cardiac intensive care unit. For validation purposes a dataset collected by the Centro Hospitalar Universitário de Coimbra was used, consisting of approximately 1400 patients that have been admitted into the cardiology intensive care unit. The obtained results suggested that SCORE and GRACE models are not sufficiently accurate to estimate the actual length of stay. Moreover, GRACE presents better results than SCORE, which can be justified by the employed risk factors, more specific for short-term prediction periods.

Keywords: Length of Stay, Prediction, Cardiovascular risk scores, Intensive care units.

2

1 Introduction

The major goal of hospitals is to provide quality healthcare, while implementing effective management of resources. The Length of stay (LoS) refers to a single episode of hospitalization, and is defined as the number of days that the patient is hospitalized in a clinical facility. LoS is a good indicator of efficiency of care and hospital performance, as it can be an indicator of hospital resource utilization [1]. This is especially important in intensive care units (ICU), where the clinical condition of the patient is most of the times critical and the resources are limited.

In addition, a proper prediction of LoS can also indicate the patient's severity of illness [2] and be used to help the definition of efficient clinical pathways [3], as well as in the establishment of adequate intervention plans. Consequently, LoS prediction models are essential to assist hospital administrators, as it can be used to optimize their strategic planning, and to assist professionals by providing a clinical decision support tool. Patients are also benefited considering that LoS prediction allows long term care and discharge activities planning, contributing to the quality of care.

There are available different approaches and methods used for LoS prediction. The arithmetic and statistical approaches are the simplest techniques. Basically, arithmetic methods use the average or the median LoS as the prediction value. This is, however, a biased value since LoS distribution is highly skewed and does not presents a normal distribution, as it is assumed by those methods. Statistic methods are based in covariate's analysis (using the patient's characteristics and external factors that correlate with LoS and can be used to predict LoS), typically implemented through linear regression (LR) and logistic regression techniques [4]. In parallel, data-mining techniques have been also employed, in order to find unknown relationships among variables [4]. Using data-mining approaches Hachesu et al. [2] developed accurate models to predict the LoS of heart patients using decision trees, support vector machines and artificial neural networks (ANNs). Rowan et al. [5] demonstrated that ANNs could be used as an effective LoS stratification instrument in postoperative cardiac patients. Pei-Fang et al. [6] attempt to predict LoS of in-patients with one of the three primary diagnoses: coronary atherosclerosis, heart failure and acute myocardial infarction in a cardiovascular unit using an ANN model. Furthermore, the ANN results were compared to the ones using LR model, concluding that the LR model had a slightly better accuracy. Mohammad et al. [7] conducted a study on the efficacy of APACHE-IV for predicting mortality and LoS in an intensive care unit, showing that APACHE-IV model underestimated LOS.

On the other hand, in the context of cardiovascular diseases management, there are available several prognostic risk tools, usually validated in large sets of populations, allowing physicians to evaluate the probability of an individual developing an event based on a set of risk factors. These tools can be divided in two main categories: long term (years), specific for primary prevention (e.g. SCORE [8], FRAMINGHAM [9] and QRISK [10]) and short term (months), specific for secondary prevention (e.g. APACHE-IV [11], GRACE [12], TIMI [13]). Moreover, other classifications are possible, such as the disease (coronary artery disease, heart failure, etc) or patient's conditions (ambulatory patients, hospitalized patients, etc.). Therefore, although they

have not been developed specifically for LoS prediction, the potential of such models may be exploited in this context. Actually, clinical guidelines recommend the use of such prognostic risk scores to support the clinical decision. As result, its extension to LoS prediction can be straightforward, without the need of development additional models.

The main goal of this work is to investigate the capacity of some prognosis risk scores, in particular SCORE (primary tool) and GRACE (secondary tool), to estimate the LoS in a cardiac ICU. Besides the importance to investigate long and short term predictions, these scores are recommended by European Society of Cardiology and have a major relevance in the context of Portugal hospitals. The paper is organized as follows: section 2 provides an overview of the developed methodologies and section 3 presents the main achieved results. Some final considerations are drawn in section 4.

2 Materials and Methods

2.1 Dataset

A dataset provided by the Centro Hospitalar e Universitário de Coimbra (CHUC) was inspected in this work for validation purposes. It contains information (N=1414) collected in the context of the admission of cardiac patients to the ICU at Hospital de Covões, gathered between 2009 and 2016. Although a vast set of parameters was collected, only a subset was employed. In effect, just the necessary inputs (risk factors) for SCORE and GRACE risk scores were considered (**Table 1**).

Table 1. Risk factors – baseline characteristics of CHUC dataset used in this work.

Variable		Value
GEN	Gender {Man/Woman}	{989/425}
AGE	Age (Years)	67.79 ± 13.27
SMO	Smoke status {No/Yes}	{907/507}
SBP	Systolic blood pressure (mmHg)	134.33±27.39
TCH	Total cholesterol (mmol/L)	4.64±1.22
HR	Hear rate (bpm)	76.36±18.15
KIL	Killip class {1/2/3/4}	{1089, 221, 78, 26}
CRT	Creatinine (μmol/L)	110.00±113.53
CAA	Cardiac arrest at admission*	n.a.
ECM	Elevated cardiac markers {No/Yes}**	{700/714}
STD	ST segment deviation {No/Yes}	{886/528}
LoS	Length of stay (Days)	4.92±0.48

* The information regarding CAA is not applicable in the context of study carried out by CHUC. A result, this input was considered as zero for all patients for the GRACE model.
** For computing the ECM the troponin value was used. A threshold of 34 μg/L or higher has been defined as a positive case.

4

2.2 Discretization of the LoS

The **Fig. 1** depicts the histogram of the patient's LoS for the available dataset. As can be observed, the LoS presents an asymmetric distribution: a relatively small number of patients with long LoS exist on the right of the graph (particularly for periods greater than 15 days) and a high concentration of patients is observed on the left of the graph. In fact, 1316 patients present a length of stay equal or less than 10 days, corresponding to 93% of all patients.

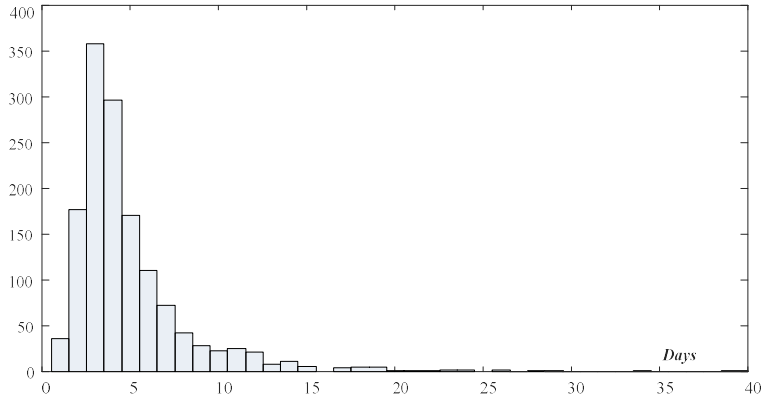


Fig. 1. Length of Stay – Days.

Although the average LoS is 4.92 days, the typical period of stay (mode) is three days (358 patients, corresponding to 25.3%) and the majority of patients were in ICU less or equal than four days (868 patients, corresponding to 61.4%).
Since from the clinical perspective it is not realistic the exact estimation of the LoS, a strategy considering only two categorical classes was assumed LoS={*Short*, *Long*}. As result, one of the decisive parameter to be determined is the threshold *NDays*, able the discrimination between *Short* and *Long* periods. This parameter *NDays* is defined according to equation (1), being its value considered in the interval {2,..,10}.

$$LoS = \begin{cases} Short & if \ \ number\ of\ days\ in\ ICU < NDays \\ Long & if \ \ number\ of\ days\ in\ ICU \geq NDays \end{cases} \quad (1)$$

As example, if $NDays = 5$, it means that if a patient has remained in ICU 4 days, his stay was considered a *Short* period. On the other hand, if he has stayed a period greater or equal than 5 days, this is considered as a *Long* period.

2.3 Discretization of the SCORE and GRACE models

The implemented risk scores, SCORE and GRACE, were calculated from the initial values collected on patient's admission (**Table 1**).
The SCORE risk model [9] is a primary prevention tool applicable to the general population. It estimates a 10 year risk of death due to a general cardiovascular event, making use of 5 risk factors {GD, AGE, SMO, SBP, TCH}. Two alternatives of

SCORE model are available, respectively for the high and for the low risk regions of Europe. In this work (Portugal population) the low risk chart was employed, providing a quantitative score value in the interval {0,..,26}. This risk value is then converted to a qualitative risk category, comprising seven distinct classes.
The GRACE risk model [13] is a secondary prognosis tool, commonly applied to patients admitted with a first episode of myocardial infarction. The main goal is to estimate the probability of death or of a new myocardial infarction event during a given period, usually the next month or during next six months. GRACE employs eight risk factors {AGE, HR, SBP, CRT, KIL, CAA, ECM, STD} to obtain a score value in the interval {0,..,258}, that is used to obtain a risk category, comprising three distinct classes: low, intermediate, and high.
To allow a practical correlation assessment with the LoS, only two categorical risk classes have been considered for both models, i.e., Risk={Low, High}. Thus, analogous to LoS, a cut-off point *CLevel* has to be determined in order to separate the risk category of both models in low and high risk classes, according to (2).

$$Risk = \begin{cases} Low \ \ if \ \ risk\,Score < CLevel \\ High \ if \ \ risk\,Score \geq CLevel \end{cases} \tag{2}$$

The parameter *CLevel* was inspected in the interval {1,..,26} and {1,..,258}, for the SCORE and GRACE models, respectively.

2.4 Assessment metrics

Having transformed the outputs of the both risk models as {*Low*, *High*}={0,1}, and the LoS as {*Short*, *Long*}={0,1} it is straightforward to assess the accuracy of each one of the models in the estimation of the LoS. In effect, in order to quantify the performance of SCORE and GRACE models the sensitivity (SE), specificity (SP), and geometric mean (GM) can be computed. The geometric mean is typically used as a "figure of merit" in the sense that it combines the sensitivity and the specificity, being also adequate to assess results in non-balanced datasets.

$$SE = \frac{TP}{TP + FN} \tag{3}$$

$$SP = \frac{TN}{TN + FP} \tag{4}$$

$$GM = \sqrt{SE \times SP} \tag{5}$$

The TP value identifies a patient that presents a LoS considered as {Long}={1} and the risk tool classified his risk as {High}={1}. Similarly, the TN value represents a patient with a {Short}={0} LoS, and the risk tool classified his risk as {Low }={0}. The FP value identifies a {Short} LoS patient that has been classified as {High} risk; the FN identifies a {Long} LoS patient that has been classified as {Low} risk.
As result, the main goal of this work is to determine the values of *NDays* and *CLevel*, respectively defined by equations (1) and (2), which maximize the GM, equation (5), for the SCORE and GRACE models.

6

3 Results

3.1 SCORE correlation

In a first phase the SCORE risk was applied to all the patients. The result is depicted in **Fig. 2**, for the different categorical classes {0,..,6} (from low to high risk). As defined by the SCORE, these categorical classes have been obtained from the discretization of the score values {0,..,26}, to seven classes, according to (6).

$$class = \begin{cases} 0 & if\ scoreValue = 0 \\ 1 & if\ scoreValue = 1 \\ 2 & if\ scoreValue = 2 \\ 3 & if\ scoreValue \in \{3,4\} \\ 4 & if\ scoreValue \in \{5,6,7,8,9\} \\ 5 & if\ scoreValue \in \{10,11,12,13,14\} \\ 6 & if\ scoreValue \geq 15 \end{cases} \quad (6)$$

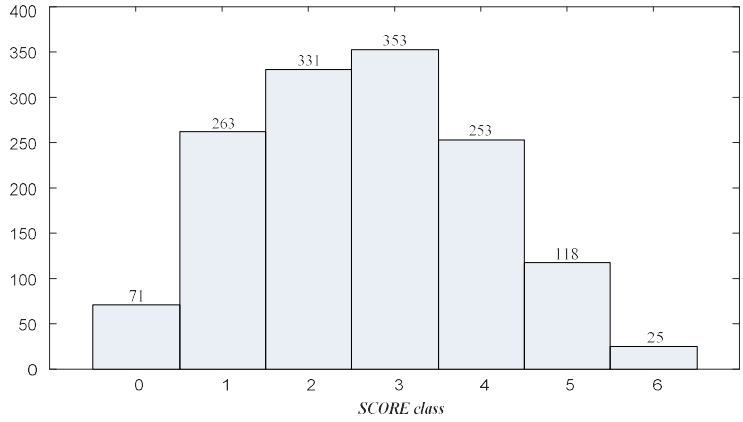


Fig. 2. SCORE class results.

To assess the performance of SCORE risk to estimate the LoS, an exhaustive search was performed. To this aim, the thresholds *NDdays*={2,..,10} and *CLevel*={2,..26} were tested in order to maximize the GM value. Since this analysis does not involve the development of a model, but only the assessment of the capacity to distinguish the LoS, no usual process of validation/testing was carried out. **Table 2** presents the maximum GM values for each one of the *NDays* values.

According to the obtained results, it is possible to achieve a maximum value of GM=0.5314, corresponding to a SP=0.5787 and a SP=0.4880, for the values of *NDays*=6 and *CLevel*=3. As conclusion, the best results for the LoS prediction can be achieved using a simple rule, given by (7).

IF (SCORE *scoreValue* ≥ 3) *THEN* (LoS is greater or equal than 6 days) (7)

An equivalent result can be obtained using the categorical classes, equation (8), using the categorization procedure defined in (6).

IF (SCORE *class*) ≥ 3 *THEN* (LoS is greater or equal than 6 days) (8)

Table 2. SCORE risk – SE, SP, GM values with respect to *NDays* and *CLevel* variation.

SE	SP	GM	NDays	CLevel
0.5298	0.4722	0.5002	2	3
0.5337	0.4930	0.5129	3	3
0.5362	0.4799	0.5072	4	3
0.5659	0.4931	0.5283	5	3
0.5787	**0.4880**	**0.5314**	**6**	**3**
0.5795	0.4817	0.5284	7	3
0.5654	0.4759	0.5187	8	3
0.5570	0.4735	0.5136	9	3
0.5785	0.4749	0.5241	10	3

3.2 GRACE correlation

Fig. 2 depicts for the different risk categorical classes (low, intermediate and high), obtained using GRACE score for the population under study. As defined by GRACE score, these categorical classes have been obtained from the discretization of the score values {0,..,248}, according to (9).

$$class = \begin{cases} 0 = low & if\ scoreValue < 109 \\ 1 = \text{int}\, ermediate & if\ 109 \le scoreValue < 140 \\ 2 = high & if\ scoreValue \ge 140 \end{cases} \quad (9)$$

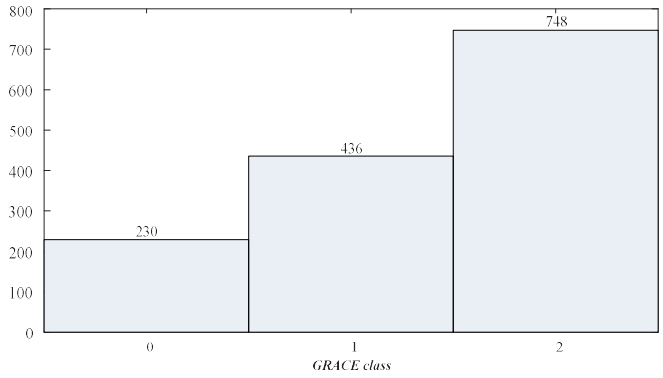


Fig. 3. GRACE risk results.

To determine the values of the thresholds an exhaustive search was carried out considering *NDdays*={2,..,10} and *CLevel*={2,..248}, in order to maximize GM. The obtained results are presented in **Table 3**, for the different *NDays* values.

8

Table 3. GRACE risk – SE, SP, GM values with respect to *NDays* and *CLevel* variation.

SE	SP	GM	NDays	Clevel
0.4274	0.6944	0.5448	2	151
0.6037	0.5540	0.5783	3	137
0.5184	0.7145	0.6086	4	151
0.6117	0.6935	0.6514	5	151
0.6667	0.6631	0.6649	6	151
0.7348	**0.6470**	**0.6895**	**7**	**151**
0.7435	0.6255	0.6819	8	151
0.6846	0.6767	0.6806	9	158
0.7355	0.6311	0.6813	10	154

The maximum value of GM=0.6895, resultant from a SE=0.7348 and a SP=0.6470, when considering *NDays*=7 and *CLevel*=151 values. As conclusion, the maximum geometric mean value was obtained considering the following rule (10).

IF (GRACE *scoreValue* ≥151) *THEN* (LoS is greater or equal than 7 days) (10)

In case the categorical level is used, since the risk score value 151 corresponds to a {high} risk, an analogous result can be obtained, equation (11).

IF (GRACE *class* = High) *THEN* (LoS is greater or equal than 7 days) (11)

However, using the categorical value {high} a slight decreasing in the obtained performance values was observed: SP=0.8106, SP=0.5357, GM=0.6589.
Moreover, from the analysis of **Table 3**, a further conclusion can be stated using the class {Intermediate}, corresponding to a *scoreValue*=137, equation (12).

IF (GRACE *class* ≥ *Intermediate*) *THEN* (LoS is equal or greater than 3 days) (12)

In this case a decreasing in the performance results was verified: SP=0.6037, SP=0.5540, GM=0.5783.

3.3 Risk factors correlation

Risk factors correlation
The next step involved the investigation of the relative importance of each one of the inputs (risk factors) (**Table 1**) with respect to the output (LoS). To this aim a simple Pearson correlation, equation (13), was computed for all inputs ($rFactor_i \;\; i = 1..10$). i.e., *rFactor*={GEN, AGE, SMO, SBP, TCH, HR, KIL, CRT, ECM, STD}.

$$Cf = \frac{\sum_{n=1}^{N} (LoS(n) - \overline{LoS})(rFactor_i(n) - \overline{rFactor_i})}{\sqrt{\sum_{n=1}^{N} (LoS(n) - \overline{LOS})^2} \sqrt{\sum_{i=1}^{N} (rFactor_i(n) - \overline{rFactor_i})^2}} \quad (13)$$

Regarding the LoS, two different alternatives were used: *i*) the raw data (*rawLoS*) comprising values in the interval {1,..,40} and *ii*) the optimal discretization of the length of stay (*dLoS*) obtained for SCORE model, considering *NDays*=6 (equivalent results were obtained with the optimal value for GRACE *NDays*=7).

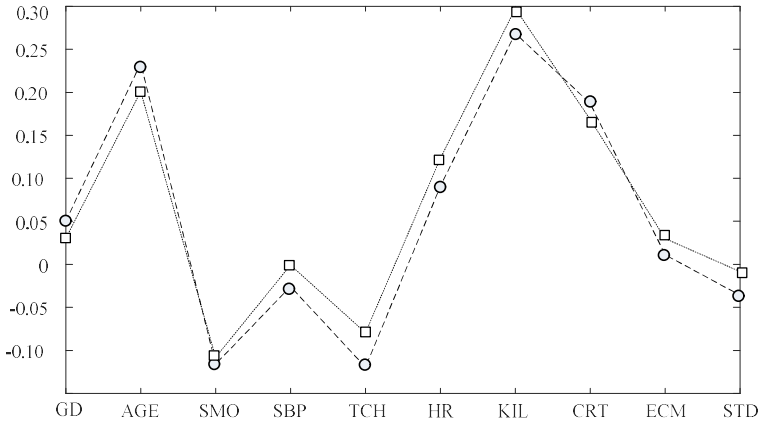


Fig. 4. Risk factors correlation with LoS.

As can be observed from **Fig. 4** the values are consistent in the sense that the *rawLoS* and the discretized version (*dLoS*) achieve approximately the same correlation values. Moreover, the highest values of correlation were verified for KIL (Killip class), AGE, CRT (creatinine) and HR (Heart rate).

3.4 Discussion

From the analysis of the SCORE and GRACE performances in the prediction of the LoS (Sections 3.1 and 3.2), it is possible to conclude that both models are not sufficiently accurate to estimate the actual LoS. In effect SCORE and GRACE presented a GM of 0.5314 and 0.6895, respectively. Additionally, SCORE and GRACE models were consistent in the optimal discretization of the LoS period: a value of *NDays* equal to 6 and 7, respectively. When comparing the performance of the two models, GRACE enabled to obtain better predictions of the LoS than SCORE, which can be justified by its short-term purposes.

From the correlation analysis (section 3.3) it can be concluded that the most relevant risk factors for prediction LoS are: Killip class, Age, Creatinine and Heart rate. This result confirms the superiority of GRACE: in effect, three of these four risk factors are exclusively used by GRACE (Killip class, Creatinine and Heart rate). Consequently, these inputs seem to be preferable to be used in short term contexts, such as ICUs. The age, Heart rate and Creatinine were also found to be variables with high impact on the prediction of LOS by previous literature studies as Daghistani et al. [14] and LaFaro et al. [15].

As conclusion, the obtained results indicate that for an accurate prediction of the LoS it is recommended the development of a specific model.

10

4 Conclusions

This work investigated the potential of some existing prognosis risk tools, in particular SCORE (primary tool) and GRACE (secondary tool), to estimate the LoS of stay in a cardiac ICU belonging to the Centro Hospitalar Universitário de Coimbra. Through a discretization process it was possible to assess this potential using sensitivity, specificity and geometric mean metrics. The obtained results suggested that SCORE and GRACE models are not sufficiently accurate to estimate the actual LoS. Moreover, GRACE presents superior results than SCORE, in accordance with the type of risk factors this model employs, which are more adequate for short-term prediction periods, such as the cardiac ICU.
Currently, a comparison study is being performed involving the implementation of some specific prediction models, based on computational intelligent methodologies, namely decision trees, ANNS and support vector machines.

References

1. Marshall, A., Vasilakis, C., El-Darzi, E.: Length of stay- based patient flow models: recent developments and future directions. Health Care Management Science, 8(3):213–220 (2005).
2. Hachesu, PR., Ahmadi, M., Alizadeh, S., Sadoughi F.: Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res;19(2):121-9 (2013).
3. Rowan, M., Ryan, T., Hegarty, F., O'Hare, N.: The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artif Intell Med;40(3):211–221 (2007).
4. Awad, A., Bader-El-Den, M., McNicholas, J.: Modeling and predicting patient length of stay: A survey. International Journal of Advanced Scientific Research and Management, 1(8) (2016).
5. Rowan, M., Ryan, T., Hegarty, F., OHare, N.: The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artificial Intelligence in Medicine.;40(3):211 – 221 (2007)
6. Tsai, P.F., Chen, P.C., Chen, Y.Y., Song, H.Y., Lin, H.M., Lin, F.M., Huang, Q.P.: Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. Journal of Healthcare Engineering (2016), 1–11 (2016)
7. Ghorbani, M., Ghaem, H., Rezaianzadeh, A., Shayan, Z., Zand, F., Nikandish, R.: A study on the efficacy of APACHE-IV for predicting mortality and length of stay in an intensive care unit in Iran. F1000Research 6:2032 (2017).
8. Conroy, R., Pyorala, K., Fitzgerald, A., Sans, S., Menotti, A., G. De Backer, G.,De Bacquer, D., Ducimetière, P., Jousilahti, P., Keil, U.,Njølstad, I. , Oganov, R.G.,Thomsen, T.,Tunstall-Pedoe, H.,Tverdal, A.,Wedel, H.,

Whincup, P., Wilhelmsen, L., Graham, I.M.: Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project; European Heart Journal Vol. 24, 987-1003 (2003).

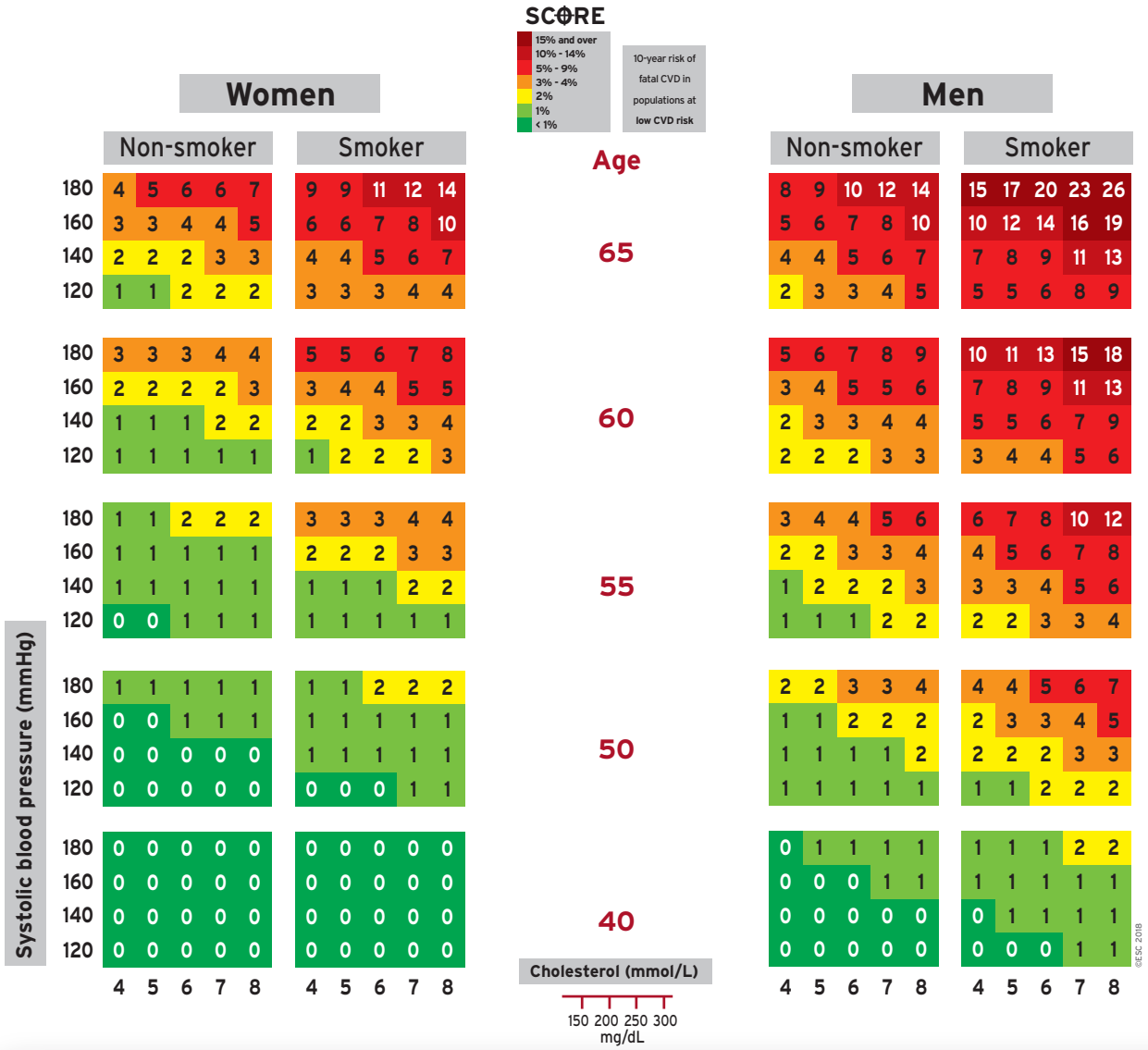
9. D'Agostino, R., Vasan, R., Pencina, M., Wolf, P., Cobain, M., Massaro, J., Kannel, W.: General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study, AHA (2008).
10. Cox, J., Coupland C., Robson J., May M., Brindle P.: Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study; BMJ (2007).
11. Ghorbani, M., Ghaem, H., Rezaianzadeh, A., Shayan, Z., Zand, F., Nikandish, R.: A study on the efficacy of APACHE-IV for predicting mortality and length of stay in an intensive care unit in Iran. F1000Research 6:2032 (2017)
12. Tang, E., Wong, C., Herbison, P.: Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long term mortality post-acute coronary syndrome, American Heart Journal, Vol. 153, nº1 (2007).
13. Antman E., Cohen, M., Bernink, P., McCabe, C., Horacek, T., Papuchis, G., Mautner, B., Corbalan, R., Radley, D., Braunwald, E.: The TIMI risk score for unstable angina/non-ST elevation MI- A method for Prognostication and Therapeutic Decision Making; JAMA, 284, 7, 835:842 (2000).
14. Daghistani, T., Elshawi, R., Sakr, S., Ahmed, A., Al-Thwayee, A., Al-Mallah, M.: Predictors of in-hospital length of stay among cardiac patients: A machine learning approach, Int. J. cardiology, vol. 288, pp. 140–147, Aug. 2019. doi: 10.1016/j.ijcard.2019.01.046 (2019).
15. LaFaro, R., Pothula, S., Kubal, K., Inchiosa, M., Pothula, V., Yuan, S., Maerz, D., Montes, L., Oleszkiewicz, S., Yusupov, A., Perline, R., Inchiosa, M.: Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre-Incision Variables. PLoS ONE 10(12): e0145395. doi:10.1371/journal.pone. 014539 (2015)

C

SCORE - European Low Risk Chart

SCORE - European Low Risk Chart

10 year risk of fatal CVD in low risk regions of Europe by gender, age, systolic blood pressure, total cholesterol and smoking status



How do I use the SCORE charts to assess CVD risk in asymptomatic persons?

1. Use the **low risk charts** in Andorra, Austria, Belgium*, Cyprus, Denmark, Finland, France, Germany, Greece*, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Monaco, The Netherlands*, Norway, Portugal, San Marino, Slovenia, Spain*, Sweden*, Switzerland and the United Kingdom.

Use the **high risk charts** in other European countries. Of these, some are **at very high risk** and the charts may underestimate risk in these. These include Albania, Algeria, Armenia, Azerbaijan, Belarus, Bulgaria, Egypt, Georgia, Kazakhstan, Kyrgyzstan, Latvia, FYR Macedonia, Moldova, Russian Federation, Syrian Arab Republic, Tajikistan, Turkmenistan, Ukraine and Uzbekistan.

*Updated, re-calibrated charts are now available for Belgium, Germany, Greece, The Netherlands, Spain, Sweden and Poland.

2. Find the cell nearest to the person's age, cholesterol and BP values, bearing in mind that risk will be higher as the person approaches the next age, cholesterol or BP category.

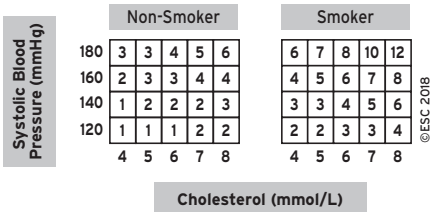
3. Check the qualifiers

4. Establish the total 10 year risk for fatal CVD.

Relative Risk Charts

Note that a low total cardiovascular risk in a young person may conceal a high relative risk; this may be explained to the person by using the relative risk chart. As the person ages, a high relative risk will translate into a high total risk. More intensive lifestyle advice will be needed in such persons. This chart refers to relative risk, not percentage risk, so that a person in the top right corner is at 12 times higher risk than a person in the bottom left corner.

Another approach to explaining risk to younger persons is to use cardiovascular **risk age**. For example, in the high risk chart, a 40 year old male hypertensive smoker has a risk of 4%, which is the same as a 65 year old with no risk factors, so that his risk age is 65. This can be reduced by reducing his risk factors.



Risk estimation using SCORE: Qualifiers

- The charts should be used in the light of the clinician's knowledge and judgement, especially with regard to local conditions.
- As with all risk estimation systems, risk will be over-estimated in countries with a falling CVD mortality rate, and under estimated if it is rising.
- At any given age, risk appears lower for women than men. However, inspection of the charts shows that their risk is merely deferred by 10 years, with a 60 year old woman resembling a 50 year old man in terms of risk.
- Risk may be higher than indicated in the chart in:
  - Sedentary or obese subjects, especially those with central obesity
  - Those with a strong family history of premature CVD
  - Socially deprived individuals and those from some ethnic minorities
  - Individuals with diabetes- the SCORE charts should only be used in those with type 1 diabetes without target-organ damage; Other diabetic subjects are already at high to very high risk
  - Those with low HDL cholesterol* or increased triglyceride, fibrinogen, apoB, Lp(a) levels and perhaps increased high-sensitivity CRP.
  - Asymptomatic subjects with evidence of pre-clinical atherosclerosis, for example plaque on ultrasonography.
  - Those with moderate to severe chronic kidney disease (GFR <60 mL/min/1.73 m^2)

*Note that HDL cholesterol impacts on risk in both sexes, at all ages, and at all level of risk. This effect can be estimated using the electronic version of SCORE, HeartScore, which has been updated to include HDL cholesterol level.

Visit www.heartscore.org
For the interactive version of the SCORE risk charts

Source: European Guidelines on CVD Prevention in Clinical Practice 2016
Eur J Prev Cardiol. 2016 Jul;23(11):NP1-NP96. doi: 10.1177/2047487316653709

EAPC
European Association of Preventive Cardiology
European Society of Cardiology

D

Baseline characteristics of the CHUC dataset

Although a vast set of parameters was available, only the necessary inputs (risk factors) for the developed models were employed.

Table D.1: Risk factors – baseline characteristics of the CHUC dataset used in this work

Variable	Value
Gender {Male/Female}	{916/373}
Age (Years)	67.07 ± 13.18
Type of ACS {Stemi/Nstemi/Unstable Angina}	{517/556/216}
Smoke status {No/Yes}	{813/476}
Systolic blood pressure (mmHg)	135.13 $\pm$27.36
Total cholesterol (mmol/L)	4.68 ± 1.21
Killip class {1/2/3/4}	{1024/183/64/18}
Creatinine (μmol/L)	105.53 ± 108.49
C-reactive protein (mg/dL)	2.44 ± 9.79
Chronic Kidney Disease {No/Yes}	{1070/219}
History of Cerebrovascular Accident {No/Yes}	{1179/110}
Troponin (ng/mL)	20.44 ± 63.52
Hemoglobin (g/dL)	13.61 ± 1.91
Glycose (mmol/L)	8.01 ± 4.32
Length of stay (Days)	4.92 ± 3.55

E

Baseline characteristics of the TRICA dataset

The TRICA dataset was composed by a vast set of parameters: static and dynamic data. The following table presents the main characteristics of the bariatric patients before any preprocessing of the data.

Table E.1: Risk factors – baseline characteristics of the TRICA dataset bariatric patients

Variable	Value
Gender {Male/Female}	{50/140}
Age (Years)	46.10 ± 11.54
Height (m)	1.70 ± 0.09
Weight (Kg)	118.11 ± 18.31
BMI (Kg/m^2)	40.44 ± 4.50
Surgery duration (Minutes)	75.43 ± 17.53
Surgery type {Gastric bypass/Gastric Sleeve}	{113/76}
Cardiovascular comorbidity {No/Yes}	{102/87}
Ward medication of Corticosteroids {No/Yes}	{187/2}
Ward medication of Sedatives {No/Yes}	{180/9}
Length of stay (Days)	1.17 ± 0.45