UNIVERSIDADE D
COIMBRA

Sónia Cristina Santos Sousa

Development of Clinical Risk Models: Aspects of Interpretablity and Personalization

Dissertation in the context of Master in Biomedical Engineering, Specialization in Clinical Informatics and Bioinformatics, supervised by Prof. Dr. Jorge Henriques and Prof. Dr. Simão Paredes and Dr. José Pedro Sousa and presented to the Faculty of Science and Technology

September of 2022

This page is intentionally left blank.

Faculty of Sciences and Technology
Department of Physics

Sónia Cristina Santos Sousa

Development of Clinical Risk Models: Aspects of Interpretability and Personalization

Dissertation in the context of the Master in Biomedical Engineering
Specialization in Clinical Informatics and Bioinformatics

Supervisors:
Prof. Dr. Jorge Henriques (DEI - CISUC)
Prof. Dr. Simão Paredes (ISEC-IPC; CISUC)
Dr. José Pedro Sousa (CHUC)

Coimbra, 2022

This page is intentionally left blank.

Acknowledgements

First, I would like to thank my advisors, Jorge Henriques and Simão Paredes, for all the availability and support throughout this year. I'm very grateful for all the wisdom you shared with me. Furthermore, I want to thank Centro Hospitalar e Universitário de Coimbra and, in particular, Dr.José Pedro for the provided clinical dataset. I also want to express my gratitude for the clinical advice that Dr.José Pedro provided.

On a more personal note, I want to thank my parents and my sister for always being there for me, and for all the support and effort made during my academic path. I also want to thank my extended family for always giving me the motivation to keep going. To my friends in Coimbra and at home, thank you for being a safe harbor and making me laugh when I need the most. Last but definitely not least, I want to thank Miguel for always believing in me and for being my biggest supporter.

This page is intentionally left blank.

”If not me, who? If not now, when?”

Emma Watson

This page is intentionally left blank.

Abstract

In Europe, Cardiovascular Diseases (CVDs) are among the main causes of death, with 3.9 million deaths annually. For Acute Coronary Syndrome (ACS) patients, risk stratification at hospital admission enables the physician to tailor the therapeutic strategy. The Global Registry of Acute Coronary Events (GRACE) score is the most used risk assessment tool in Portugal. In parallel, Machine Learning (ML) models have shown notable performance. However, in healthcare, their deployment is still limited since "black-box" models don't provide explanations for their predictions.

Our goal was to develop ML models to predict 6-month mortality for ACS patients and compare them with the GRACE score regarding performance and interpretability. To obtain explanations, we used algorithms that due to their properties create interpretable ML models: logistic regression, naive Bayes, and decision trees. Furthermore, we proposed an interpretable approach based on decision rules. This method, besides interpretability, also addresses personalization, without impairing the model's performance. First, a global set of interpretable rules is generated based on risk factors of ACS. Then, an ML model is trained to predict the probability that each rule is correct for a given patient. In this work, we evaluate interpretability quantitatively through a stability measure, with a range of [-1,1], the 95% Confidence Interval (CI) interval on the geometric mean, and the Spearman correlation between the features' rank by importance attributed by each of the ML models and features' rank considered by the GRACE score.

Centro Hospitalar e Universitário de Coimbra (CHUC) provided the clinical dataset used for our methodology validation. Our proposed approach achieved the best performance, with 74.72% geometric mean. Furthermore, it has the highest Spearman correlation (0.83), a narrow CI (11.2%) and the same stability as the GRACE score (0.506).

Keywords: Machine Learning, Interpretability, Personalization, Cardiovascular Diseases, Acute Coronary Syndrome, GRACE Risk Score

This page is intentionally left blank.

Resumo

Na Europa, as doenças cardiovasculares estão entre as maiores causas de morte, com 3.9 milhões de mortes anualmente. Em relação a pacientes com Síndrome Coronária Aguda, uma estratificação de risco no momento de admissão no hospital permite que o médico adapte a estratégica terapêutica. O modelo de risco GRACE é a ferramenta de avaliação de risco mais usada em Portugal. Em paralelo, os modelos de inteligência computacional têm demonstrado uma performance notável. No entanto, nos cuidados de saúde, o seu uso ainda é limitado devido aos modelos "caixa-preta" não fornecerem explicações para as suas previsões.

O nosso objetivo era desenvolver modelos de inteligência computacional para prever a mortalidade, num período de 6 meses, para pacientes com Síndrome Coronária Aguda e comparar os modelos com o GRACE, quanto à performance e interpretabilidade. De forma a obter explicações, usamos algoritmos que devido às suas propriedades criam modelos de inteligência computacional interpretáveis: regressão logística, naive Bayes e arvores de decisão. Além disso, propusemos uma abordagem interpretável baseada em regras. Este método, além da interpretabilidade, também aborda a personalização, sem prejudicar a performance do modelo. Primeiramente, é gerado um conjunto global de regras interpretáveis baseado em fatores de risco da Síndrome Coronária Aguda. De seguida, um modelo de inteligência computacional é treinado para prever a probabilidade de cada regra estar correta para um determinado paciente. Neste trabalho, avaliamos quantitativamente a interpretabilidade através de uma medida de estabilidade, com uma variação entre [-1,1], do intervalo de confiança de 95% para a média geométrica e da correlação de Spearman entre as variavéis ordenadas pela importância atribuída por cada um dos modelos e pela importância considerada pelo GRACE.

O Centro Hospitalar e Universitário de Coimbra (CHUC) forneceu o conjunto de dados clínicos que foram usados na validação da nossa metodologia. A nossa abordagem alcançou a melhor performance, com 74.72% de média geométrica.

Ademais, possui a maior correlação de Spearman (0.83), um intervalo de confiança com uma variabilidade restrita (11.2%) e a mesma estabilidade que o modelo de risco GRACE (0.506).

Palavras-chave: Inteligência Computacional, Interpretabilidade, Personalização, Doenças Cardiovasculares, Síndrome Coronária Aguda, Modelo de Risco GRACE

Contents

List of Figures xii

List of Tables xiv

List of Abbreviations xix

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Main goal . . . 3
- 1.3 Structure . . . 4

2 Background 5
- 2.1 Clinical Background . . . 5
  - 2.1.1 Acute Coronary Syndrome (ACS) . . . 5
  - 2.1.2 Risk Scores . . . 8
- 2.2 Machine Learning Algorithms . . . 12
  - 2.2.1 Logistic Regression . . . 13
  - 2.2.2 Naive Bayes . . . 14
  - 2.2.3 Decision Rules . . . 15
  - 2.2.4 Decision Trees . . . 16
  - 2.2.5 Clustering . . . 17
  - 2.2.6 Nearest Neighbors . . . 18
- 2.3 Interpretability . . . 19
  - 2.3.1 Definitions of Related Terms . . . 20
  - 2.3.2 Taxonomy of Interpretability . . . 22
  - 2.3.3 Shapley Values . . . 24
  - 2.3.4 Evaluation of Interpretability . . . 26
- 2.4 Data Pre-processing and Data Validation . . . 27
  - 2.4.1 Statistical Tests . . . 28
  - 2.4.2 Validation Strategies . . . 29
  - 2.4.3 Random Sampling . . . 30
- 2.5 Performance Assessment Metrics . . . 32
- 2.6 Confidence Intervals . . . 34
- 2.7 Conclusions . . . 35

3 State of the Art 37

3.1 Machine Learning Studies on Cardiovascular Disease . . . 37
3.2 The Case for Interpretability . . . 38
3.3 The Case for Personalization . . . 40
3.4 Evaluation of Interpretability . . . 41
3.4.1 Application-Grounded and Human-Grounded Evaluation of Interpretability . . . 41
3.4.2 Functionally-Grounded Evaluation of Interpretability . . . 42
3.5 Conclusions . . . 46

4 Methodology **49**
4.1 Data Pre-processing and Data Validation . . . 50
4.1.1 Treatment of Missing Values . . . 50
4.1.2 Statistical Tests . . . 50
4.1.3 Preliminary Analysis . . . 51
4.1.4 Validation Strategy . . . 51
4.1.5 Handling Data Imbalance . . . 51
4.2 Implementation of the Clinical Reference and Interpretable Models . 51
4.2.1 Clinical Reference: GRACE Risk Score . . . 52
4.2.2 Intrinsic Interpretable Models . . . 52
4.3 Evaluation of the Performance of the Models . . . 60
4.4 Evaluation of the Interpretability of the Models . . . 60
4.4.1 Application-Grounded Evaluation . . . 61
4.4.2 Functionally-Grounded Evaluation . . . 61

5 Results and Discussion **65**
5.1 Dataset . . . 65
5.2 Data Pre-processing . . . 67
5.2.1 Computation of Variables . . . 67
5.2.2 Missing Values and Computation of Class Label . . . 67
5.2.3 Statistical Tests and Preliminary Analysis . . . 69
5.3 Implementation and Evaluation of the Models . . . 73
5.3.1 GRACE . . . 74
5.3.2 Logistic Regression . . . 79
5.3.3 Naive Bayes . . . 85
5.3.4 Decision Trees . . . 90
5.3.5 Our Approach . . . 96
5.4 Results Overview . . . 101
5.4.1 Performance Evaluation . . . 101
5.4.2 Interpretability: Functionally-Grounded Evaluation . . . 102

6 Conclusions **105**

Bibliography **107**

Appendices **117**
A SHAP Summary Plots . . . 119

List of Figures

1.1 Citation count for research articles (left) and google search trends (right) with keywords "Interpretable Machine Learning" and "Explainable AI". From Molnar et al. (2020). 3

2.1 ACS related concepts and their relationships. 6
2.2 Summary on the Unstable Angina, NSTEMI and STEMI conditions. Adapted from CanadiEM (2018). 7
2.3 A typical Electrocardiogram. From Alivecor (2022). 7
2.4 Bayes' theorem. From Prawtama (2021). 14
2.5 Conversion of a decision tree into decision rules. From Freitas et al. (2010). 17
2.6 Representation of 3 clusters obtained by the k-means clustering method with k=3. It is possible to see the mean values of the elements of the cluster being represented as the cluster centers. From Zhang et al. (2017). 18
2.7 Illustration of the K-Nearest Neighbor method for unsupervised learning with k=3 and k=6. Adapted from Italo (2018). 19
2.8 Function h_x maps a coalition to a valid data instance. From Molnar (2022). 25
2.9 Stratified k-fold method with k=5. From Müller (2020). 30
2.10 SMOTE-NC algorithm for numerical data. From Imbalanced-learn (2014). 32
2.11 Confusion Matrix. From Aras (2020). 32
2.12 Leave-one-out bootstrap. Adapted from Raschka (2020). 35

4.1 Summary of the proposed methodology. 49
4.2 Centroids for feature X. Each centroid represents a class and they are obtained by calculating the means of the values that belong to each cluster. Adapted from Valente et al. (2021b). 55
4.3 Methodology related to the creation of rules and followed application to the dataset. Adapted from Valente et al. (2021b). 56
4.4 Methodology related to the training of rules and the prediction of their acceptance degree. From Valente et al. (2021b). 57
4.5 Summary of the main phases of the proposed approach. Adapted from Valente et al. (2022). 60

5.1 Class distribution in each of the ACS diagnoses. 66
5.2 Class distribution for each gender. 67
5.3 Boxplots of the numerical variables used in our work: age, heart rate, systolic blood pressure, and creatinine. 70
5.4 Distributions of the categorical variables used in our work: Killip class, troponin, STEMI and cardiac arrest. 72
5.5 Correlation between the numerical and ordinal variables. 73
5.6 Correlation between the binary variables. 73
5.7 Class distribution according to risk stratification. 75
5.8 SHAP force plot for the GRACE risk score. 76
5.9 SHAP summary plot for the GRACE risk score. 79
5.10 SHAP force plot for logistic regression. 84
5.11 SHAP force plot for naive Bayes. 90
5.12 Decision tree for a randomly selected train partition of our dataset. . 92
5.13 SHAP force plot for decision tree. 95
5.14 Means of the clusters for the troponin variable representing both patient's classes. 97
5.15 Means of the clusters for the STEMI variable representing both patient's classes. 97
5.16 SHAP force plot for our proposed approach. 100

A.1 SHAP summary plot for logistic regression. 119
A.2 SHAP summary plot for naive Bayes. 119
A.3 SHAP summary plot for the decision tree model. 119
A.4 SHAP summary plot for our proposed approach. 119

List of Tables

2.1 Variables used and points attributed in the GRACE risk score. Adapted from Araújo Gonçalves de et al. (2005). 9
2.2 Heart rate values (bpm) and different conditions associated. Based on Meek (2002). 10
2.3 Values of systolic blood pressure (mmHg) and different conditions associated. Adapted from Hussain, Fadel (2020). 10
2.4 Creatinine values (μmol/L) for women and men and different conditions associated. Based on UCSF Health (n.d.a). 11
2.5 Killip Class. Adapted from Gjesdal et al. (2018). 11
2.6 ML algorithms and respective properties that allow them to create interpretable models. 23
2.7 Post-model/post hoc/model-agnostic methods and the type of explanations they provide. Adapted from Carvalho et al. (2019). . . . 24
2.8 Range of correlation values and their interpretation. 29
2.9 Goal, null hypothesis, and type of variables used in different statistical tests. Based on Hindle, Childs (2021); SciPy (2022a,b,c); Sheskin (2020); StatsTest.com (2020a,b,c). 29
2.10 Common metrics used to evaluate the performance of classifiers. Based on Brownlee (2020); Sunasra (2017). 33

4.1 Calculation of the training rules acceptance. Adapted from Valente et al. (2022). 58
4.2 Example of the computation of the predicted mortality risk and the reliability measure. Adapted from Valente et al. (2021b). 59

5.1 Range or values taken by each of the variables that were used throughout our work. 66
5.2 Percentage of missing values for each of the variables used throughout our work. 68
5.3 Results for the Chi-Square test (performed in the categorical variables), the Kolmogorov-Smirnov test, and Mann-Whitney U test (performed in the numerical variables). The p-values that aren't lower than 0.05 are marked in bold. 69
5.4 Mean values for the numerical variables used in our work separated by class. 70

5.5 Percentage of individuals that have each value in the categorical variables and mortality rate for each value. 72
5.6 Performance metrics for the GRACE risk score. 75
5.7 Functionally-grounded evaluation of interpretability for the GRACE risk score. 76
5.8 Ranks of the used features considering the features' importance returned by SHAP in our implementation of the GRACE risk score. . 76
5.9 Performance metrics for the logistic regression model. 80
5.10 Coefficients values and their 95% Confidence Interval returned by the logistic regression model. The LLR p-value, the standard errors and p-values of the coefficients are also represented in the table. The represented values are the averaged values on the 10 runs performed. 81
5.11 Values of the odds ratio and their 95% Confidence Interval. For the numerical variables, it is also represented the mean and standard deviation values averaged on the 10 runs performed. 82
5.12 Functionally-grounded evaluation of interpretability for the logistic regression model. 83
5.13 Ranks of the used features considering the features' importance returned by SHAP for the logistic regression model and for the GRACE risk score. 84
5.14 Categories for the numerical variables used in the implementation of the naive Bayes model. 85
5.15 Performance metrics for the naive Bayes model. 86
5.16 Probabilities for each category in the age variable given class 0 or 1. . 86
5.17 Probabilities for each category in the heart rate variable given class 0 or 1. 86
5.18 Probabilities for each category in the systolic blood pressure variable given class 0 or 1. 87
5.19 Probabilities for each category in the creatinine variable given class 0 or 1. 87
5.20 Probabilities for each category in the Killip class variable given class 0 or 1. 88
5.21 Probabilities for each category in the troponin variable given class 0 or 1. 88
5.22 Probabilities for each category in the STEMI variable given class 0 or 1. 88
5.23 Probabilities for each category in the cardiac arrest variable given class 0 or 1. 89
5.24 Functionally-grounded evaluation of interpretability for the naive Bayes model. 89
5.25 Ranks of the used features considering the features' importance returned by SHAP for the naive Bayes model and for the GRACE risk score. 90
5.26 Performance metrics for the decision tree model. 91
5.27 Features' importance returned for the decision tree represented in figure 5.12. The values different from 0 are marked in bold. 94

5.28 Functionally-grounded evaluation of interpretability for the decision tree model. 94
5.29 Ranks of the used features considering the features' importance returned by SHAP for the decision tree model and for the GRACE risk score. 95
5.30 Performance metrics for our proposed approach. 97
5.31 Representation of feature values, rule outputs, and predicted rules acceptance for a specific patient. Furthermore, we present the average acceptance of positive and negative rules, the mortality risk, and the reliability measure. 99
5.32 Functionally-grounded evaluation of interpretability for our proposed approach. 99
5.33 Ranks of the used features considering the features' importance returned by SHAP for our approach and for the GRACE risk score. . 100
5.34 Overview of the performance of the different models. 101
5.35 Overview of the functionally-grounded evaluation of the model's interpretability. 102

This page is intentionally left blank.

List of Abbreviations

ACS Acute Coronary Syndrome

AI Artificial Intelligence

bpm beats per minute

CART Classification and Regression Trees

CHUC Centro Hospitalar e Universitário de Coimbra

CI Confidence Interval

COMPAS Correctional Offender Management Profiling for Alternative Sanctions

CVDs Cardiovascular Diseases

DSS Decision Support System

ECG Electrocardiogram

FN False Negative

FP False Positive

GDPR General Data Protection Regulation

GRACE Global Registry of Acute Coronary Events

ICM Intuitive Confidence Measure

ID3 Iterative Dichotomiser 3

ITR Information Transfer Rate

KNN K-Nearest Neighbor

LIME Local Interpretable Model-Agnostic Explanations

LLR Log-Likelihood Ratio

ML Machine Learning

NSTE-ACS Non-ST-Elevation Acute Coronary Syndrome

NSTEMI Non-ST-Elevation Myocardial Infarction

PCA Principal Component Analysis

PDR Predictive, Descriptive, Relevant

PURSUIT Platelet Glycoprotein IIb/IIIa in Unstable Agina: Receptor Suppression Using Integrilin

SHAP SHapley Additive exPlanations

SMOTE Synthetic Minority Over-sampling Technique

SMOTE-NC Synthetic Minority Over-sampling Technique for Nominal and Continuous

STEMI ST-Elevation Myocardial Infarction

TIMI Thrombolysis In Myocardial Infarction

TN True Negative

TP True Positive

UA Unstable Angina

XAI eXplainable Artificial Intelligence

1

Introduction

Cardiovascular Diseases (CVDs) are the leading cause of morbidity and mortality in the world with an estimated 17.9 million deaths from CVDs in 2019 (Weng et al., 2017; World Health Organization, n.d.). Coronary heart disease is one group of cardiovascular diseases and they are responsible for 1 out of 5 deaths in the United States of America, being the most common cause of death (Viera, Sheridan, 2010). Acute Coronary Syndrome (ACS) is a term used to describe a range of conditions including: ST-Elevation Myocardial Infarction (STEMI), Non-ST-Elevation Myocardial Infarction (NSTEMI) and Unstable Angina (UA) (Smith et al., 2015).

For ACS patients, early risk stratification at hospital admission is very important as it allows for a tailored therapeutic strategy. The risk stratification is performed using risk scores based on initial clinical history, Electrocardiogram (ECG), and laboratory tests (Araújo Gonçalves de et al., 2005). The Thrombolysis In Myocardial Infarction (TIMI) (Antman et al., 2000), the Platelet Glycoprotein IIb/IIIa in Unstable Agina: Receptor Suppression Using Integrilin (PURSUIT) (Boersma et al., 2000) and Global Registry of Acute Coronary Events (GRACE) (Granger et al., 2003) are example of risk scores. The latter is the most used risk score in clinical settings.

These risk scores oversimplify complex relationships like risk factors with non-linear interactions, assuming that each risk factor has a direct relationship with the outcome (Weng et al., 2017). Furthermore, score models have been found to perform well at the population level but worse at the individual level (Valente et al., 2021b).

1.1 Motivation

Due to recent research progress on Machine Learning (ML), which has been demonstrating remarkable performances in many different tasks, approaches based

on ML methods have been proposed for the implementation of risk score models (Valente et al., 2021b, 2022).

Although ML has the potential to create personalized models, most Machine Learning models follow the principle "one fits all" since the model is applied in the same way to all individuals. However, in healthcare, physicians do not apply their complete expertise in the same way to all patients, but instead, they perform a personalized diagnosis, considering the particular characteristics of each one (Valente et al., 2022). Despite ML models usually having better predictive performance, physicians are doubtful to approve them, not only due to the personalization issue mentioned above but also because they lack interpretability. Since ML models behave like "black-box" systems, as they do not provide explanations for their conclusions, it can be hard to trust the predictions made by the models in high-stakes domains, where crucial decisions are made. (Linardatos et al., 2020). These issues are preventing the widespread adoption of ML models in healthcare, and in particular, in the implementation of risk score models (Valente et al., 2022).

The need for interpretable high-performing models for real-world applications led to the current popularity of the eXplainable Artificial Intelligence (XAI) field. It focuses on the understanding and interpretation of the behavior of Artificial Intelligence (AI) systems, making it more comprehensible to humans (Abedin, 2022; Linardatos et al., 2020; Valente et al., 2022). The field had lost the attention of the scientific community since most of the research focused on the predictive power of algorithms. The work on interpretability had existed for many years, for example, with the built-in feature importance measure of random features, but it was not as investigated and reported as it is nowadays. The popularity of the search terms "Explainable AI" and "Interpretable Machine Learning" and the number of papers published with these terms throughout the years is illustrated in figure 1.1, where it is noticeable the increase in recent years (Linardatos et al., 2020; Molnar et al., 2020). The number of review papers that have been written on the the topic of Machine Learning (ML) interpretability over the past years shows the room for improvement that exists in this field (Linardatos et al., 2020; Valente et al., 2022).

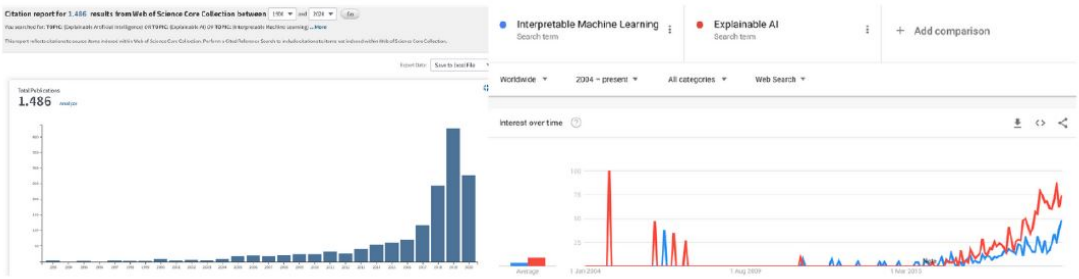

Figure 1.1: Citation count for research articles (left) and google search trends (right) with keywords "Interpretable Machine Learning" and "Explainable AI". From Molnar et al. (2020).

There is also a need to develop measures to quantify the interpretability of different ML models in order to compare them (Schmidt, Biessmann, 2019). However, little research has been done on this topic, and the work found in the literature is very recent.

1.2 Main goal

Due to the limitations of the risk score models, our main goal is to develop different ML models to solve the problem of mortality prediction in patients with ACS. Therefore, we took advantage of the predictive power of ML models. Furthermore, we explored and addressed the interpretability and personalization problems.

A real Portuguese dataset provided by the Centro Hospitalar e Universitário de Coimbra (CHUC) hospital was used in this work. The dataset contains information on 1544 patients admitted to CHUC that experienced an ACS episode. The code behind this project was developed in Python language, using Jupyter Notebook and resorting to several libraries, namely: Scikit-Learn, Numpy, Pandas, SciPy, and Matplotlib.

Objectives

Considering our main goal, we can define several objectives for this work.

- Implementation of interpretable ML models to evaluate the 6-month mortality risk of patients diagnosed with ACS.
- Implementation of an ML model developed by our work group to solve both the personalization and interpretability issues.
- Development of measures to quantify the interpretability of the different models.

- Comparison of the ML models to the GRACE risk score regarding predictive power and interpretability.

1.3 Structure

This document is structured in six chapters. In the chapter 2 we introduce pertinent background concepts related to the clinical aspect of our work, the ML models used, the concepts related to interpretability, the pre-processing methods, and data validation strategies. Finally, we introduce the concept of data validation metrics and confidence intervals. In chapter 3, we present state of the art in the development of ML models in the cardiovascular field. We also present the need for personalization along with prior studies of interpretability as a concept and studies related to the evaluation of interpretability. The proposed methodology is presented in chapter 4 and in chapter 5 the results obtained are discussed. Lastly, chapter 6 exposes the conclusions and enumerates ideas for future work.

2

Background

In this chapter, several essential concepts are explored in order to allow a complete understanding of the proposed work and respective results. In section 2.1, we introduce the clinical concepts, namely information about Acute Coronary Syndrome and risk scores. In section 2.2, information on the different interpretable models that is relevant to our work is given. Then, in section 2.3, interpretability is defined, along with other related terms often mentioned in the literature. A taxonomy of interpretability is presented based on the information found in the literature to provide a rigorous background to understand the work developed. In particular, the Shapley values method is introduced. Furthermore, formal concepts on interpretability evaluation are mentioned. Statistical tests used in this work are mentioned in section 2.4, along with data validation strategies and random sampling. In section 2.5, we mention the typical performance metrics used to evaluate models and in section 2.6, the concept of confidence intervals is introduced. Lastly, in section 2.7, considering the information presented in this chapter, we mention a summary of the work that was done in this thesis.

2.1 Clinical Background

2.1.1 Acute Coronary Syndrome (ACS)

Cardiovascular Diseases are a group of disorders of the heart and blood vessels, including cerebrovascular disease, rheumatic heart disease, peripheral arterial disease, and coronary heart disease (International Diabetes Federation, 2021; World Health Organization, n.d.).

Coronary Heart Diseases are a condition of the blood vessels supplying oxygen and blood to the heart muscle where atherosclerosis (accumulation of cholesterol plaques) occurs in the vessels (International Diabetes Federation, 2021; Mayo Clinic, 2021b).

Acute Coronary Syndrome is the term for suddenly reduced blood flow to the heart, which is an immediate cause for medical emergency care. It happens after the rupture of a coronary arterial plaque, given that a thrombus (blood clot) forms and blocks the flow of blood to heart muscles, a condition called myocardial ischemia. (Mayo Clinic, 2021a, n.d.). Acute Coronary Syndrome includes Unstable Angina, ST-Elevation Myocardial Infarction and Non-ST-Elevation Myocardial Infarction. A summary of the different concepts mentioned and their relationships is presented in figure 2.1.

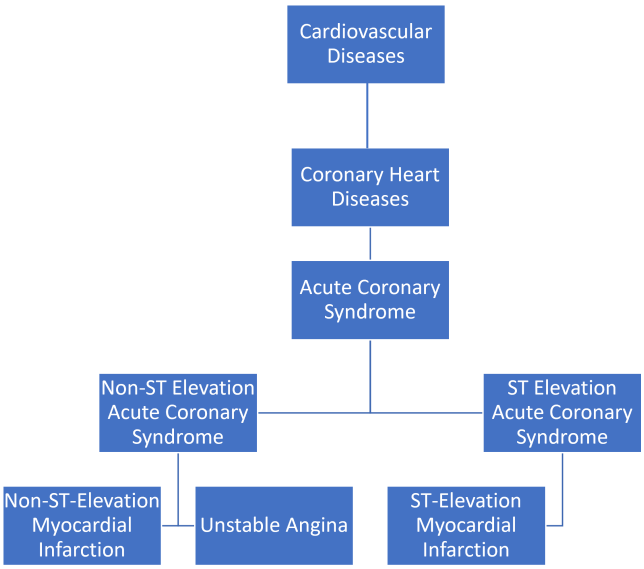

Figure 2.1: ACS related concepts and their relationships.

If the supply of oxygen to heart muscle cells is too low those cells can die, resulting in damage to muscle tissues. In this case, a myocardial infarction, commonly designated as a heart attack occurs. However, even when there is no cell death, there is still a decrease in oxygen. In this situation, Unstable Angina occurs. Therefore, in the STEMI and NSTEMI diagnosis, cell death occurs (measured by the level of troponin) and in Unstable Angina, those levels remain the same (Mayo Clinic, 2021a). We can distinguish between the 2 types of myocardial infarction, based on the fact that in the presence of a STEMI diagnosis, an ST-segment elevation is visible on the ECG (Smith et al., 2015). In figure 2.2, a summary of the different ACS conditions mentioned above is presented.

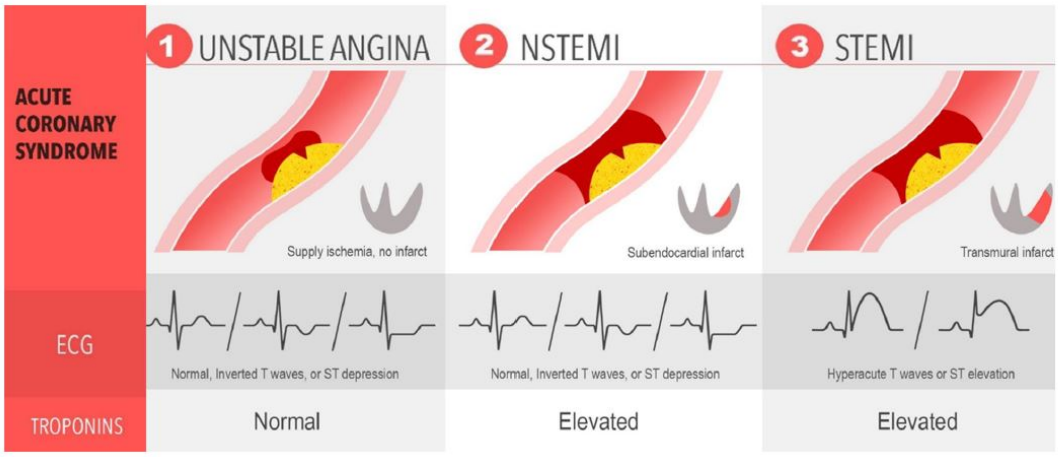

Figure 2.2: Summary on the Unstable Angina, NSTEMI and STEMI conditions. Adapted from CanadiEM (2018).

The levels of troponin and the presence of an ST-elevation in the ECG are determinant factors for the diagnosis of a patient since they allow to differentiate between the 3 different types of ACS.

ECG and ST-Segment Elevation

An Electrocardiogram allows checking the heart's rhythm, electrical activity, and the resulting waveforms by placing electrodes on the skin. The number of electrodes that are placed varies, the most popular choice being the 12-lead ECG which records the electrical activity as seen from 12 different points around the heart. A typical ECG morphology (figure 2.3) includes a P-wave that consists of atrial depolarization. The P-wave is followed by the QRS complex (ventricular depolarization), and finally, ventricular repolarization occurs (T-wave). These waveforms occur in a repeating rhythm called sinus rhythm (Sampson, McGrath, 2015).

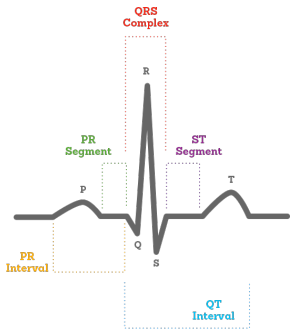

Figure 2.3: A typical Electrocardiogram. From Alivecor (2022).

In the case of a STEMI diagnosis, the ST-segment has an elevation superior of 2mm in men and 1.5mm in women on at least two ECG contiguous leads (Hwang, Levis, 2014).

Troponin Levels

Cardiac biomarkers, also known as troponins are a group of proteins (troponin C, troponin I, and troponin T) from the cardiac muscle fibers that regulate muscular contractions. Therefore, troponin levels in the blood are typically very low but in the case of myocardial infarction, troponin is sent into the bloodstream, indicating necrosis in myocardial cells and causing the troponin levels to increase significantly. As heart damage increases, greater amounts of troponin are released in the blood (Medical News Today, n.d.; MedlinePlus, n.d.). Troponin levels may remain high for 1 to 2 weeks after the myocardial infarction event (UCSF Health, n.d.b). Troponin tests typically measure the levels of troponin I or troponin T in the blood and use a threshold of 0.40 ng/ml for myocardial infarction (Medical News Today, n.d.).

2.1.2 Risk Scores

Risk stratification in ACS enables the physician to triage the patients and decide the best course of therapy. Besides that, when a patient knows their risk level they are more likely to initiate risk-reducing actions and therapies (Araújo Gonçalves de et al., 2005; Viera, Sheridan, 2010). In this context, the most acknowledged risk scores models are the TIMI, PURSUIT and GRACE risk scores. They were all developed for a short-term risk assessment (occurrence of myocardial infarction event or death), after hospital admission with ACS diagnosis. The GRACE was developed for a prognosis within 6 months after hospital admission, the TIMI after 14 days, and in the PURSUIT risk score, 30 days are considered. Furthermore, the PURSUIT and TIMI risk scores were developed with the databases from large clinical trials of Non-ST-Elevation Acute Coronary Syndrome (NSTE-ACS) which includes the NSTEMI and Unstable Angina conditions (figure 2.1). On the other side, the GRACE score is the latest and was developed from an international registry of patients across the entire spectrum of ACS (Araújo Gonçalves de et al., 2005). Therefore, nowadays, the GRACE score is the most used risk model in Portugal.

In the GRACE risk score, the final score of a patient (ranging between 2 and 383) is the sum of all variables' scores. Each variable score is attributed depending on the value that the patient presents for that variable. The variables used in the GRACE score and the different points attributed are represented in table 2.1.

Variable	Values	Points
Age (years)	<40	0
	40-49	18
	50-59	36
	60-69	55
	70-79	73
	≥80	91
Heart Rate (beats per minute-bpm)	<70	0
	70-89	7
	90-109	13
	110-149	23
	150-199	36
	>200	46
Systolic Blood Pressure (mmHg)	<80	63
	80-99	58
	110-119	47
	120-139	37
	140-159	26
	160-199	11
	>200	0
Creatinine (mg/dL)	0-0.39	2
	0.4-0.79	5
	0.8-1.19	8
	1.2-1.59	11
	1.6-1.99	14
	2-3.99	23
	>4	31
Killip Class	Class I	0
	Class II	21
	Class III	43
	Class IV	64
Cardiac Arrest at Admission		43
Elevated Cardic Markers		15
ST-segment Deviation		30

Table 2.1: Variables used and points attributed in the GRACE risk score. Adapted from Araújo Gonçalves de et al. (2005).

Below we explain the meaning of the variables used in the GRACE risk score, excluding the age variable and the STEMI and troponin variables, that were already explained.

Heart Rate

The heart rate is the number of times that the heart beats in a minute, which is normally between 60 and 100 beats per minute (bpm) for adults (Cleveland Clinic, n.d.). In table 2.2, it is possible to see the different conditions associated with abnormal heart rhythms.

Values (bpm)	Condition
<60	Bradycardia
60-100	Healthy Adult
>100	Tachycardia

Table 2.2: Heart rate values (bpm) and different conditions associated. Based on Meek (2002).

Systolic Blood Pressure

Systolic blood pressure indicates how much pressure the blood is doing against the artery walls when the heart beats (American Heart Association, n.d.). In table 2.3, different blood pressure categories are defined with the associated systolic blood pressure values.

Blood Pressure Category	Systolic Blood Pressure (mmHg)
Normal	<120
Elevated	120-129
High Blood Pressure (Hypertension) Stage 1	130-139
High Blood Pressure (Hypertension) Stage 2	140-180
Hypertensive Crisis (consult your doctor immediately)	>180

Table 2.3: Values of systolic blood pressure (mmHg) and different conditions associated. Adapted from Hussain, Fadel (2020).

Creatinine

Creatinine is a waste product made by the muscles as part of everyday activity. Normally, the kidneys filter creatinine from the blood and it is excluded from the body in the urine. If there is kidney disease, creatinine can build up in the blood, and less will be released in urine. Therefore, in those cases, are found high levels of creatinine in blood and low levels in urine. The normal values for creatinine in the blood are represented in table 2.4.

Women	Men	Condition
Values (μmol/L[1])	Values (μmol/L[1])	
<53.0	<61.9	Low Creatinine Values
53.0 - 97.2	61.9 - 114.9	Healthy Adult
>97.2	>114.9	High Creatinine Values

Table 2.4: Creatinine values (μmol/L) for women and men and different conditions associated. Based on UCSF Health (n.d.a).

Killip Class

The Killip class was introduced for clinical assessment of patients with acute myocardial infarction, providing effective stratification of long-term and short-term outcomes. It stratifies individuals according to the severity of their heart failure after the myocardial infarction event (Hashmi et al., 2020). In table 2.5, the different classes for Killip can be seen along with a description of the state that the patient is in for each class.

Killip class I:	Individuals with no clinical signs of heart failure.
Killip class II:	Individuals with rales or crackles in the lungs, an S3, and elevated jugular venous pressure.
Killip class III:	Individuals with frank acute pulmonary edema.
Killip class IV:	Individuals in cardiogenic shock or hypotension (measured as systolic blood pressure lower than 90 mmHg) and evidence of peripheral vasoconstriction (oliguria, cyanosis or sweating).

Table 2.5: Killip Class. Adapted from Gjesdal et al. (2018).

Cardiac Arrest

Cardiac arrest occurs when the heart suddenly stops pumping blood around the body, causing a lack of oxygen in the brain. This causes the person to fall unconscious and stop breathing and without immediate medical attention, the patient will die (British Heart Foundation, 2021).

[1] 1μmol/L=0.0113 mg/dL

In this work, the GRACE risk score was used as our clinical reference, given that it is the most used risk assessment tool in Portugal. The developed models were compared to it in terms of performance and interpretability.

2.2 Machine Learning Algorithms

Machine Learning is an application of Artificial Intelligence that uses data and algorithms to imitate the way that humans learn (IBM Cloud Education, 2020). Machine Learning can be divided into four approaches: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Supervised and unsupervised learning are the most common methods.

Supervised learning

Supervised learning uses labeled datasets (inputs and outputs) to train algorithms. In this approach, the model learns over time. There are two groups of problems inside supervised learning techniques: classification and regression.

Classification problems use an algorithm to predict the outcome of the test data, assigning it to specific categories, such as separating death/survival outcomes. Examples of algorithms that are used in classification are logistic regression, decision trees, and naive Bayes.

Regression methods use an algorithm to explain a relationship between dependent and independent variables. Regression models can help predict numerical values based on different data points, such as the risk score for a given patient. An example of an algorithm used for regression is linear regression.

Unsupervised learning

Unsupervised learning uses ML algorithms to interpret and cluster unlabelled datasets. These algorithms discover hidden patterns in the data without the need for human interference. Clustering is an example of a well-known method that consists in an unsupervised learning approach.

Out of all the supervised learning methods, we will address the ones used in this work: logistic regression, naive Bayes, decision trees, and decision rules. Furthermore, regarding unsupervised learning, we will mention clustering, in particular, k-means clustering, since we used this method in the development of our proposed approach. Nearest neighbors will also be mentioned since it was employed to develop the stability metric used to evaluate interpretability.

2.2.1 Logistic Regression

Logistic regression computes the probabilities for classification problems with two possible outcomes. The probabilities are modeled using the logistic function to force the output to assume only values between 0 and 1. The relationship between the output and the features is represented in equation 2.1, in which i represents a specific instance, p the number of features, x the different features, and β the learned feature weights/coefficients.

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + ... + \beta_p x_p^{(i)}))} \tag{2.1}$$

By considering a threshold and the estimated probability, a data point is classified into one of the two classes. A clear advantage of logistic regression is that it also gives probabilities, so it isn't just a classification model.

The probability of an event divided by the probability of no event is called the "odds". In equation 2.2, we can see the relationship between the odds and the logistic regression coefficients.

$$\frac{P(y = 1)}{1 - P(y = 1)} = odds = exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \tag{2.2}$$

The "odds ratio" is the ratio between two odds, for example, the odds when a numerical feature is changed by one unit by the odds when the numerical feature remains unchanged. In equation 2.3, we can see the relationship between the "odds ratio" and the coefficient of the feature that was changed by one unit.

$$\frac{odds_{x_j+1}}{odds} = exp(\beta_j) \tag{2.3}$$

Interpretation of the odds ratio

A change in a feature by one unit changes the odds ratio by a factor of the exponential of the coefficient of that feature. If the coefficient is positive (negative), the change in the odds ratio will be bigger (less) than 1, meaning it will represent an increase (decrease) in the odds ratio (Molnar, 2022).

If the numerical feature is standardized (z-score normalization), the interpretation changes: An increase in 1 standard deviation in the numerical feature

is associated with an increase/decrease in the odds ratio. The increase/decrease has a magnitude of the exponential of the coefficient of that feature (Choueiry, n.d.).

In the case of binary categorical features, for the interpretation of the odds ratio, we need to assume a reference category (for example, the value 0 of the feature). Changing the feature from the reference category to the other category changes the estimated odds by a factor of the exponential of the coefficient of that feature (Molnar, 2022).

Feature Importance

The values for the odds ratios of continuous variables are not directly comparable with one another or to the odds ratios of binary variables in terms of their relative importance to the outcome because the numerical variables are not measured on the same scale. However, because the binary variables assume the same values (0 and 1 values), their values for the odds ratios are comparable and indicate feature importance (Anderson et al., 2003). If we standardize (z-score normalization) the numerical variables and obtain the standardized coefficients, then we can compare the magnitude of the coefficients between the numerical variables (Choueiry, n.d.; Menard, 2011).

2.2.2 Naive Bayes

The naive Bayes classifier uses the Bayes' theorem (figure 2.4) of conditional probabilities (probability of an event occurring given the probability of another event that has already occurred).

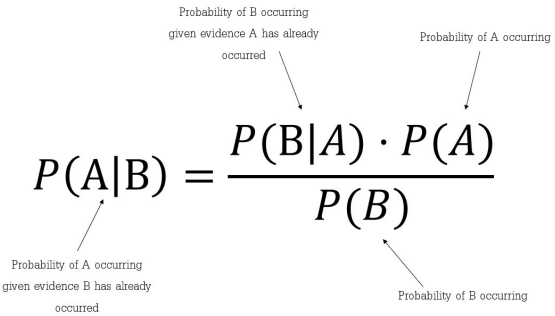

Figure 2.4: Bayes' theorem. From Prawtama (2021).

We can rewrite Bayes' theorem as represented in equation 2.4, with features x_i and class y. Furthermore, n represents the number of features, $P(y)$ represents the prior probability of class y (the relative frequency of the class in the training set),

and $P(y|x_1,...,x_n)$ the posterior probabilities (the class with the highest posterior probability corresponds to the outcome).

$$P(y|x_1,...,x_n) = \frac{P(y)P(x_1,...,x_n|y)}{P(x_1,...,x_n)} \tag{2.4}$$

The naive Bayes classifier assumes conditional independence of the features x_i given the value of class y (Molnar, 2022; Scikit-learn, n.d.-a). This naive assumption represented in 2.5 reduces the complexity of the problem. (Burkart, Huber, 2021). $P(x_i|y)$ are the conditional probabilities that depict the probabilistic relationship between the features x_i and the class y.

$$P(x_1,...,x_n|y) = \prod_{i=1}^{n} P(x_i|y) \tag{2.5}$$

Considering the independence assumption and that the denominator of equation 2.4 remains constant for a given input, we can rewrite that equation to:

$$P(y|x_1,...,x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.6}$$

Lastly, to obtain the class for a given input, we find the class value with maximum probability (equation 2.7).

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.7}$$

Assuming categorical features, we can obtain the conditional probabilities $P(x_i = t \mid y = c)$, meaning, probability of the category t in feature i given class value c. (Scikit-learn, n.d.-a).

2.2.3 Decision Rules

If-then rules have the following structure (Burkart, Huber, 2021):

IF condition AND/OR condition THEN label (prediction) ELSE other label (prediction).

Each condition is composed of a feature, an operator, and a value. In the case of more than one condition aggregated by 'AND' (union), all conditions must be true for the rule to apply. On the other hand, in the case of conditions aggregated by 'OR' (conjunction), only one condition needs to be true for the rule to apply.

One decision rule or a combination of multiple rules can be used to make predictions (Molnar, 2022). Decision rules have the advantage that domain knowledge in the form of cause-effect relationships can be extracted from them (Valente et al., 2021a).

2.2.4 Decision Trees

Decision trees can be used for classification or regression (Molnar, 2022). The model works by learning simple decision rules inferred from the data features (Scikit-learn, n.d.-b). Logistic regression fails when features interact with each other. In those situations, a decision tree model may be better suited (Molnar, 2022). Decision trees are simple to understand given their visualization property (Molnar, 2022; Scikit-learn, n.d.-b; Valente et al., 2021a). However, they have the disadvantage of being unstable, given that, small variations in the data may change the entire tree structure changes, which affects the confidence in the model (Molnar, 2022; Scikit-learn, n.d.-b).

There are various algorithms to construct decision trees, namely Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), and C4.5. We developed this work in Python, that implements CART (Molnar, 2022; Scikit-learn, n.d.-b).

Classification and Regression Trees (CART) Algorithm

The CART algorithm takes a feature and determines which cut-off point minimizes the Gini index (or other criteria). The Gini index tells us how "impure" a node is, impure meaning when the data points in that node have very different values for the class. As a consequence, the best cut-off point makes the two resulting subsets as distinct as possible regarding the target outcome. After the best cutoff for each feature has been calculated, the algorithm selects the feature for splitting that would result in the best partition in terms of the Gini index and adds this split to the tree, creating different subsets of the dataset.

The final subsets are called terminal/leaf nodes, the intermediate subsets are called internal/split/decision nodes, and the first node is the root node (Molnar, 2022).

The algorithm continues recursively until a stop criterion is reached, for example, the minimum number of samples required to split an internal node (Molnar, 2022; Scikit-learn, n.d.-b). In the end, each instance belongs to one terminal node that tells us the predicted outcome (Molnar, 2022).

Decision Trees to Decision Rules

Every decision tree can be transformed into a rule-based model (Burkart, Huber, 2021; Margot, Luta, 2021; Valente et al., 2021a). In figure 2.5, it is possible to see an example. By incorporating a set of decision rules, decision trees mimic human reasoning. (Valente et al., 2021a)

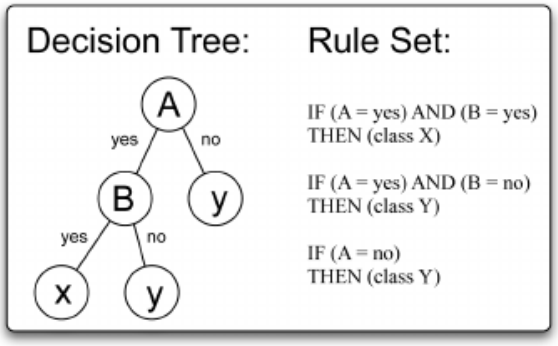

Figure 2.5: Conversion of a decision tree into decision rules. From Freitas et al. (2010).

Feature Importance

Feature Importance is determined by going through all the splits for which the feature was used and measuring how much it has reduced the Gini index compared to the parent node (node above). Therefore, feature importance tells us how much a feature helped to improve the "purity" of all nodes (Molnar, 2022).

2.2.5 Clustering

Clustering is a technique used for grouping unlabeled data based on maximizing within-group-object similarity and between-group-object dissimilarity. (Delua, 2021; Liao, 2005). The clustering technique can be divided into different methods, for example, partitioning methods, that construct k-independent partitions of the data. Each partition is a cluster and the k number can be chosen. An example of a partitioning method is the k-means clustering (Liao, 2005).

K-Means Clustering

In the k-means clustering method, each cluster is represented by the mean value of all elements in the cluster (centroid) (Liao, 2005). A representation of the method can be seen in figure 2.6.

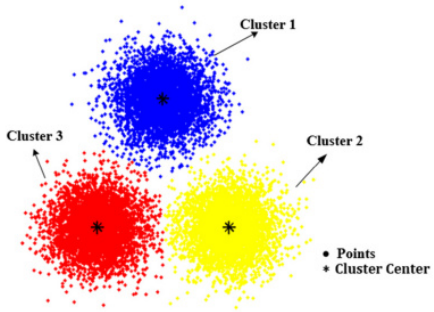

Figure 2.6: Representation of 3 clusters obtained by the k-means clustering method with k=3. It is possible to see the mean values of the elements of the cluster being represented as the cluster centers. From Zhang et al. (2017).

Initially, k objects are randomly selected as the initial centroids, the most simple method being to choose the objects from the original dataset. Secondly, for each object in the dataset, the Euclidean distance with each centroid is calculated and each object is assigned to the nearest cluster. Thirdly, the centroids are updated by calculating the mean of the cluster. This process is repeated until the difference between the old and the new centroids is less than a defined threshold (the centroids don't change significantly) (Scikit-learn, 2019; Zhang et al., 2017).

2.2.6 Nearest Neighbors

The nearest neighbors method can be used for unsupervised and supervised learning. The principle behind the method is to find a predefined number of training samples closest in distance to a certain point. If the nearest neighbors method is being used for supervised learning, the method also predicts the label (in case of classification) or the predicted value for the dependent variable (in the case of regression) for a certain point from the label/values of its nearest neighbors (Burkart, Huber, 2021; Scikit-learn, 2022).

The predefined number of samples can be a user-defined integer constant (k) as is the case in K-Nearest Neighbor (KNN) learning, the most used method. In radius-based neighbor learning, a method based on the local density of points, the number of neighbors can vary depending on the fixed radius of each training point.

In figure 2.7, it is possible to see a representation of the K-Nearest Neighbor method for unsupervised learning. The method simply founds the training samples closest in distance to a certain point. Any metric measure can be used to find the nearest neighbors, but the standard Euclidean distance is the most common choice (Scikit-learn, 2022).

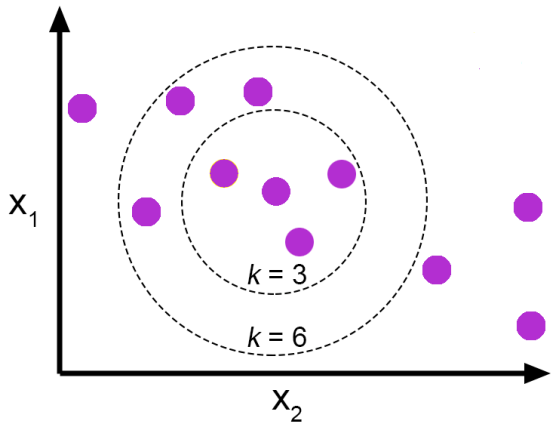

Figure 2.7: Illustration of the K-Nearest Neighbor method for unsupervised learning with k=3 and k=6. Adapted from Italo (2018).

2.3 Interpretability

There is a lack of formality in the XAI field, as there aren't rigorous and universal definitions that can be accepted by the scientific community (Lahav et al., 2018; Linardatos et al., 2020; Molnar et al., 2020; Schmidt, Biessmann, 2019).

However, there is a general understanding of interpretability, with Qayyum et al. (2020) defining it as "the ability to describe the internal processes of an ML system in a human-understandable manner". Therefore, interpretability allows humans to understand how ML models make decisions (Abedin, 2022). The goal of interpretability is to increase the user's trust and willingness to utilize ML systems (Lahav et al., 2018). Explanations and interpretability go hand-in-hand, with Carvalho et al. (2019) stating "Interpretability is the end-goal that we want to achieve and explanations are the tools to reach interpretability". Interpretability is inherently a subjective field and by giving explanations we need to take into account the application domain and the type of audience that the explanations are being given to (Carvalho et al., 2019).

Although some authors distinguish between interpretability and explainability, those definitions are sometimes contradictory (Burkart, Huber, 2021; Linardatos et al., 2020). Therefore, like many authors, in this work we will use both terms interchangeably (Carvalho et al., 2019; Doran et al (2017, as cited in, Burkart, Huber, 2021)).

2.3.1 Definitions of Related Terms

Besides the use of explainability to refer to interpretability, we can also find in the literature other synonyms like intelligibility, transparency, understandability, comprehensibility, and model trust (Abedin, 2022; Ahmad et al., 2018; Burkart, Huber, 2021; Murdoch et al., 2019). All these terms refer to solve the problem of the "black box" nature of ML for better human understanding of the decision-making process (Burkart, Huber, 2021). Trust in ML systems can be optimized through interpretability, due to being "easier for humans to trust a system that explains its decisions rather than a black box that just outputs the decision itself" (Carvalho et al., 2019). The trust component is essential to convince non-technical experts like clinicians to adopt ML solutions (Burkart, Huber, 2021; Lahav et al., 2018). Transparency of ML models can be defined as the opposite of "black-box" (Carrington et al., 2018). Both understandability and comprehensibility are goals for which interpretability aims (Carvalho et al., 2019) and Carrington et al. (2018) defines understandability as "how likely we are able to provide an interpretation". Comprehensibility is related to how well humans understand the explanations and depends on the audience and context (Carvalho et al., 2019).

In the literature, it is possible to find definitions of other related terms of interpretability.

- **Representativeness / Completeness / Broadness / Degree of integration**: Although found with different names in the literature, this concept refers to the generalization of a given explanation. It is related to the number of instances of the dataset that are covered by the explanation provided (Burkart, Huber, 2021; Carvalho et al., 2019; Nguyen, Martínez, 2020). Explanations can cover the entire model (Burkart, Huber, 2021; Carvalho et al., 2019).
- **Simulatability:** Defined as the user's ability to "run" mentally a model on a given input, meaning that a person can mentally simulate and reason about the decision-making process of a model (how a trained model produces an output for an arbitrary input) in reasonable time (Carrington et al., 2018; Molnar et al., 2020; Murdoch et al., 2019).
- **Decomposability:** Being able to see and understand the parts of the model (parameters) and the parts of the data (features and instances) and how they contribute to an output of the model (Carrington et al., 2018).
- **Relevancy:** An explanation is relevant if it provides insight for a particular audience in a particular domain (Murdoch et al., 2019).

- **Contrastiveness / Counterfactual Faithfulness:** There is a tendency for humans to think in counterfactual propositions, meaning that usually, they do not ask why a certain prediction was made but rather why some prediction was made instead of another prediction. They are interested in the factors that need to change (in the input) so that the ML prediction/decision (output) would also change (Carvalho et al., 2019).
- **Fairness / Unbiasedness:** There have been cases of discrimination made by ML models due to biases in the training data. In this case, the model favors certain cases over others, in particular, often minorities are explicit or implicitly discriminated against. Therefore, interpretability aims at improving fairness, ensuring that predictions are unbiased. This is done through explanations that help understand if the decision of the ML model is based on a learned demographic (for example, racial) bias (Carvalho et al., 2019; Doshi-Velez, Kim, 2017; Qayyum et al., 2020).
- **Accountability:** Panch et al. (2018) mentions that in "the type of problems that are experienced in clinical practice, where the objectives are not always clear and there is a high likelihood of external factors, explanation is necessary for accountability".
- **Privacy:** Interpretability aims at addressing privacy, that in this context means that sensitive information in the data is protected (Abedin, 2022; Carvalho et al., 2019; Doshi-Velez, Kim, 2017).
- **Causality:** Causal inference is related, but distinct, from interpretable machine learning. Causal inference is associated with extracting causal relationships from data (statements that altering one variable will cause a change in another). On the other hand, interpretable ML is used to describe general relationships (correlations), causal or not. Therefore, causality shouldn't be attributed to an explanation, but the results of an explanation can inform future experimental studies to investigate causal associations (given that the goal of learning predictive models is to use them as guides to action) (Ahmad et al., 2018; Domingos, 2012; Murdoch et al., 2019).

  Ideally, in healthcare applications, an ML model should reflect the true causal relations of its underlying phenomena because most of the crucial healthcare problems require causal reasoning ("what if?"). Some authors predict that causal explanations are going to be the next frontier of machine learning research (Ahmad et al., 2018; Burkart, Huber, 2021; Qayyum et al., 2020; Zhou et al., 2021).

2.3.2 Taxonomy of Interpretability

Interpretability methods and techniques can be classified according to different criteria. One of those is regarding when these methods are applicable: before (pre-model), during (in-model), or after (post-model) building the ML model.

Pre-model interpretability techniques are independent of the model, as they are only applicable to the data itself (data interpretability). It consists of exploratory data analysis techniques, for example, Principal Component Analysis (PCA) and clustering techniques. Data visualization and intuitive features are properties that help to achieve pre-model interpretability. This work does not address this type of interpretability.

Another criterion to classify interpretability is distinguishing whether it is achieved through using interpretable ML models (intrinsic interpretability) or by applying methods that analyze the model after training (post hoc interpretability). Finally, we can divide the methods into model-specific and model-agnostic (Carvalho et al., 2019).

A. In-Model / Intrinsic / Model-Specific

In this section we consider ML models that have inherent interpretability (through constraints imposed on the complexity or not), meaning they are intrinsically interpretable. The constraints can be related to sparsity (low number of features), monotonicity, causality, or can come from domain knowledge. These type of models are transparent and called "white-box" (Carvalho et al., 2019; Linardatos et al., 2020). We consider these methods model-specific, meaning they are limited to specific model classes. This is because each method is based on some specific model's internals and the interpretation of the explanation is tied to the structure of the model, for example, the interpretation of weights in a linear model (Carvalho et al., 2019; Molnar et al., 2020). These model-based interpretability methods generally use simpler models, which can sometimes result in lower predictive accuracy on complex datasets (Murdoch et al., 2019).

There is a subset of algorithms that create interpretable models, including linear regression, logistic regression, decision trees, naive Bayes, and rule-based classifiers since they have meaningful parameters that we can use to explain predictions. In table 2.6, some of the ML algorithms referred in section 2.2 that create interpretable models are mentioned along with their respective parameters (Carvalho et al., 2019). These algorithms were the ones implemented in our work. However, it is important to mention that these algorithms are only interpretable up to a certain dimension.

If we use hundreds of features in logistic regression or increase too much the depth of decision trees (how many splits a tree can make before coming to a prediction [Galarnyk, 2019]), these methods aren't interpretable anymore (Molnar et al., 2020).

ML Algorithm	Parameters
Logistic Regression	Feature coefficients; odds ratio
Naive Bayes	Probability of obtaining a certain value for a feature, given a class.
Decision Trees	Natural visualization of the tree; decision rules extracted; feature importance
Decision Rules	Domain knowledge extracted from the rules

Table 2.6: ML algorithms and respective properties that allow them to create interpretable models.

B. Post-Model / Post Hoc / Model-Agnostic

Post-model and post hoc interpretability refers to improve interpretability after building a model, therefore these methods are applied after training. It is important to mention that, there are post hoc methods that can be applied to intrinsically interpretable models, for example, SHapley Additive exPlanations (SHAP). There are also model-specific methods that are post hoc, for example, methods for deep neural networks (Carvalho et al., 2019). However, most Post hoc methods are model-agnostic since they can be applied to any ML model (black box or transparent/intrinsic interpretable). They are decoupled from the model and don't have access to their inner workings, such as weights. Another property of these methods is that models are interpreted without sacrificing their predictive power, as they are applied after training (Carvalho et al., 2019; Murdoch et al., 2019).

We can divide these explanation methods according to the type of explanation they provide: feature summary, example-based, and surrogate model.

Feature summary consists of feature statistics for each feature, such as feature importance, the most studied method in interpretable ML. It assigns an importance value to each feature depending on its contribution to the prediction. Other feature summary methods use feature effects (how a change in a feature changes the predicted outcome) like partial dependence plots and accumulated local effects plots.

An example-based method explains a prediction of interest either directly, providing representative examples (data points) with the same prediction, or counterfactually, by providing examples with a different prediction. The data points can be already existent or not.

Finally, with surrogate models, we approximate the original black-box model (either globally or locally meaning using all the data instances or some) with an

intrinsically interpretable model, simpler and easier to understand. In order to train the surrogate model, it is only needed the input and output data of the original model. The explanations returned by the surrogate model will then provide insights into the inner workings of the original model. Surrogate model approaches vary in the target black-box ML model and the interpretable model that is used (Burkart, Huber, 2021; Carvalho et al., 2019; Lahav et al., 2018; Molnar et al., 2020; Nguyen, Martínez, 2020; Zhou et al., 2021). Although these methods have the advantage of applying to any model, most of them do not consider the intrinsic properties of specific types of models in order to generate explanations (Carvalho et al., 2019).

In table 2.7, are mentioned examples of post-model/post hoc/model-agnostic methods and the type of explanations they provide.

Explanation Method	**Type of explanation**
Partial Dependence Plots	Feature summary
Individual Condition Expectation	Feature summary
Accumulated Local Effects Plots	Feature summary
Feature Interaction	Feature summary
Feature Importance/ Feature Attribution	Feature summary
Local Interpretable Model-Agnostic Explanations (LIME)	Surrogate interpretable model
SHAP	Feature summary
BreakDown	Feature summary
Anchors	Feature summary
Counterfactual Explanations	(new) Data point
Prototypes and Criticisms	(existent) Data point
Influence Functions	(existent) Data point

Table 2.7: Post-model/post hoc/model-agnostic methods and the type of explanations they provide. Adapted from Carvalho et al. (2019).

In the following section, we will explore the explanation method SHAP, and the theory behind it: Shapley values. In this work, we used SHAP to understand the different importance that each model we implemented attributes to each feature.

2.3.3 Shapley Values

Shapley values are originated from the collaborative game theory field. A prediction can be explained by assuming that each feature value of a dataset instance is a “player” in a ”game” that is the prediction task for a single instance and the ”payout” is the actual prediction for this instance subtracting the average prediction for all instances. The different feature values collaborate to the payout, and the

feature contribution is the Shapley value of the feature. The definition of Shapley value is the average marginal contribution of a feature value across all possible coalitions (combinations) of features. To obtain the Shapley value for a specific feature, we compute the prediction for each of the features coalitions with and without the value of the feature in question. The difference between the prediction with and without the feature is the marginal contribution. Finally, we average the marginal contribution across all coalitions to obtain the Shapley value. We replace the feature values of features that are not in a coalition with random feature values from other instances of the dataset (sampling values from the feature's marginal distribution) to get a prediction from the machine learning model (López, 2021; Molnar, 2022; Molnar et al., 2020).

In figure 2.8, we can see this replacement of feature values in action: we have 3 features (age, weight, and color). In a coalition where all the features are present (x'), the mapping from coalition to feature value is direct: we use the feature values of instance x. However, considering a coalition where we only have the age feature present (z'), the weight and color feature values are replaced with those feature values from another random dataset instance, while for the age feature we use the feature value of the instance x.

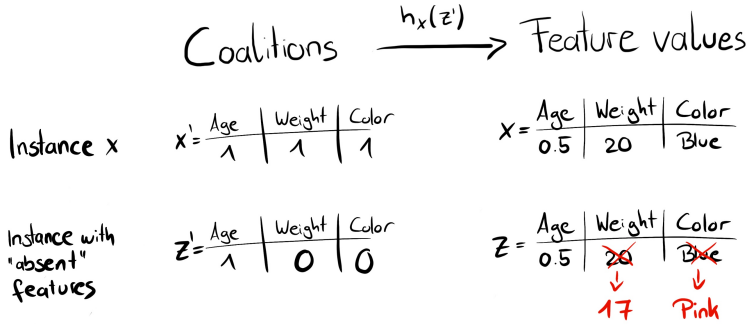

Figure 2.8: Function h_x maps a coalition to a valid data instance. From Molnar (2022).

Shapley values can be computed for both regression and classification tasks. It is important to mention some disadvantages of Shapley values, namely the fact that they require a lot of computation time, due to calculating the marginal contributions for all coalitions possible. Furthermore, access to all data instances is needed due to sampling values from the feature's marginal distribution. Marginalizing the feature also has negative consequences if the features are correlated since the Shapley value obtained for a correlated feature might not be reliable (Molnar, 2022).

SHAP

SHAP is a method that uses the kernel SHAP to calculate Shapley values with much fewer coalitions, solving the disadvantage of computation time mentioned earlier. It is based on a weighted linear regression model that is built with coalitions, predictions, and weights. Once optimized, the coefficients of the solution are the Shapley values. Going into detail, we sample coalitions and for each coalition, the prediction of the model is obtained and the weight for that coalition is calculated. Coalitions with few features and coalitions with a lot of features get the biggest weights (López, 2021). Kernel SHAP is a model-agnostic method, but there are other variants of SHAP for the calculation of Shapley values for specific types of models, for example, tree SHAP for decision trees, random forests, and gradient boosted trees.

There are available software tools where we can visualize Shapley values as "forces" in the force plot. The Shapley value for each feature is a force that pushes to increase (positive Shapley value) or decrease (negative Shapley value) the prediction for a specific data instance. In this plot, the base value is also represented and consists of the average of all predictions. The interpretation of the Shapley value for a feature value is the contribution to the prediction for this particular instance compared to the average prediction for the dataset. Furthermore, the implementations of SHAP come with many global interpretation methods based on the aggregation of Shapley values. We can interpret the entire ML model by analyzing the Shapley values returned for every instance in the dataset. The SHAP summary plot combines feature importance with feature effects.

2.3.4 Evaluation of Interpretability

Besides the lack of formality in the definition of interpretability, there is also ambiguity regarding interpretability measurement (Carvalho et al., 2019). Therefore, in this section, we define a taxonomy of evaluation approaches for interpretability: application-grounded, human-grounded, and functionally-grounded evaluation (Carvalho et al., 2019; Doshi-Velez, Kim, 2017).

Application-Grounded Evaluation

Consists of evaluating interpretability by conducting human experiments with domain experts within a real-world application (the exact application task). An example of a real application for a model is helping doctors diagnose patients with a specific disease. The quality of the explanation can be evaluated as to whether it

results in better identification of errors, new facts, or less discrimination. Since the system is tested directly for the end task that is built for, good performance in this evaluation is an evidence of success (Carvalho et al., 2019; Doshi-Velez, Kim, 2017).

Human-Grounded Evaluation

In this case, we maintain the target application but use unskilled people in the experiments instead of domain experts, disregarding the domain in which the assessed interpretability would be applied. This is useful when experiments with the target community are challenging due to a small subject poll and high expenses. This type of evaluation can be used to test more general notions of the quality of an explanation, for example, study what kinds of explanations are best understood under limited time (Carvalho et al., 2019; Doshi-Velez, Kim, 2017).

Functionally-Grounded Evaluation

While human evaluation is important to access interpretability, designing a human experiment is not an easy task (Doshi-Velez, Kim, 2017). Besides that, with application-grounded evaluation is difficult to compare results in different domains (Carvalho et al., 2019). Therefore, in functionally-grounded evaluation, no humans are required, and a formal definition of interpretability is used as a proxy for evaluating interpretability. The challenge is to determine which metrics/proxies to use and this remains an open problem (Doshi-Velez, Kim, 2017; Linardatos et al., 2020). This type of evaluation has the least costs and its results can be compared across domains, however, they present less validity since the proxies chosen are not real measures of interpretability (Carvalho et al., 2019).

In this work, we evaluated interpretability by using a set of measures (functionally-grounded evaluation). Some of the conclusions were also shared with our clinical partner, so we considered our work to be partially validated through application-grounded evaluation.

2.4 Data Pre-processing and Data Validation

In this section, some common statistical tests are mentioned and the concept of random sampling is presented. This background information will be essential to understand the pre-processing performed in our work, where among other steps like handling missing values, we performed statistical tests and handled data imbalance. Furthermore, we present validation strategies that were used for data partition.

2.4.1 Statistical Tests

A statistical test is a procedure for deciding whether a hypothesis is true. They can be used to estimate the difference between two or more groups, to determine whether a predictor variable has a statistically significant relationship with an outcome variable, or to assess the relationship between two variables (Ille, Milic, 2008). A statistical test always has a null hypothesis referred to as H0 (defined, for example, by the absence of a relationship between two study variables) and an alternative hypothesis referred to as H1 (defined by the existence of a relationship between two study variables) (Han et al., 2011; XLSTAT, 2022).

In statistical tests, two values are calculated: a test statistic (a number that describes how much the relationship between variables in the test differs from the null hypothesis of no relationship) and the p-value. The p-value estimates how likely it is to obtain this test statistic if the null hypothesis of no relationship was true (Hindle, Childs, 2021). Therefore, the smaller the p-value the stronger the evidence against the null hypothesis. (Sheskin, 2020).

It is necessary to define a threshold above which we reject the null hypothesis, this is referred to as the significance level alpha (XLSTAT, 2022). This value is usually 5%, therefore if p - $value < 0.05$, we reject H0 and accept H1 with a 5% risk. For example, if H0 is that the means of two groups are equal and we accept the H1 (the means of two groups are different), there is a 5% likelihood that the difference is due to chance (Ille, Milic, 2008).

Parametric vs. Non-parametric Tests

Parametric tests make specific assumptions regarding one or more of the population parameters that characterize the underlying distributions of the data that is being tested. On the other hand, non-parametric tests don't make assumptions about population parameters, meaning data can be collected from a sample that does not follow a specific distribution, such as normal distribution. In conclusion, if the data doesn't follow a normal distribution and we don't know the distribution's parameters, a non-parametric test should be used (Ille, Milic, 2008).

Correlation Statistical Tests

Correlation is defined as a relation existing between statistical variables which tend to be associated in a way not expected by chance alone (Mukaka, 2012). Furthermore, correlation doesn't assume any cause-and-effect relationship (Han et al., 2011).

In table 2.8, it is possible to see the rule of thumb to interpret correlation coefficients. They can vary from -1 (perfect negative correlation) through 0 (no correlation) to +1 (perfect positive correlation) (Mukaka, 2012).

Correlation Value	Interpretation
0.9 to 1 (-0.9 to -1)	Very high positive (negative) correlation
0.7 to 0.9 (-0.7 to -0.9)	High positive (negative) correlation
0.5 to 0.7 (-0.5 to -0.7)	Moderate positive (negative) correlation
0.3 to 0.5 (-0.3 to -0.5)	Low positive (negative) correlation
0.0 to 0.3 (0.0 to -0.3)	Negligible correlation

Table 2.8: Range of correlation values and their interpretation.

Overview of Common Statistical Tests

Different statistical tests can be used depending on our goal, the distribution of the data (parametric or non-parametric test), and the type of data involved. In table 2.9, we can see an overview of common statistical tests, the goal for their implementation, the type of variables used, and their null and alternative hypothesis. In this work, we used the Kolmogorov-Smirnov test and the non-parametric tests mentioned in table 2.9, since our data didn't follow a normal distribution.

Statistical Test	Goal	Parametric/Non-Parametric	Type of variables	H0/H1
Kolmogorov-Smirnov	Comparing sample distributions		Continuous	H0: The two distributions are identical
Two-Sample T-Test	Compare two unpaired groups	Parametric	Continuous	H0: The means of the groups are equal
Mann-Whitney U Test	Compare two unpaired groups	Non-Parametric	Continuous/Ordinal	H0: The groups follow the same distribution
Chi-Squared Test		Non-Parametric	Categorical	H0: Observed frequencies for the variables match the frequencies that we would get by chance
Spearman Correlation	Correlation between 2 variables	Non-Parametric	Continuous/Ordinal	
Phi Coefficient		Non-Parametric	Binary	

Table 2.9: Goal, null hypothesis, and type of variables used in different statistical tests. Based on Hindle, Childs (2021); SciPy (2022a,b,c); Sheskin (2020); StatsTest.com (2020a,b,c).

2.4.2 Validation Strategies

To avoid overfitting the model and to test its generalization capabilities with unseen data, the dataset needs to be divided into train and test partitions. That can be done using different methods. Below, we present the ones used in our work: holdout and stratified k-fold.

A. Holdout

In the holdout method, 70% of our dataset is used to train the model. The remaining 30% is be used to test the model. It is possible to implement, the stratified version of holdout, in which the class frequencies of the dataset are preserved in the train and test partitions.

B. Stratified K-Fold

In the k-fold method, the data is divided into k folds (groups of samples of equal sizes), and all but one fold are used to train the model. The remaining fold is used to test the model. This process is repeated k times and the fold that is used to test the model is always changed.

The stratified k-fold is a variation of the k-fold method in which the folds are made by preserving the class frequencies of the complete dataset. A representation of this method using 5 folds can be seen in figure 2.10 (Scikit-learn, 2013).

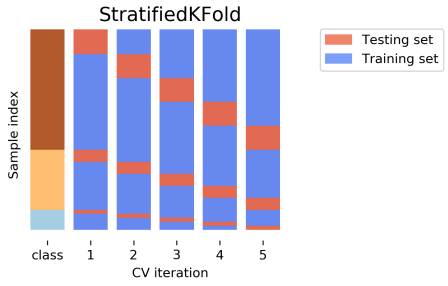

Figure 2.9: Stratified k-fold method with k=5. From Müller (2020).

2.4.3 Random Sampling

The learning and prediction phase of Machine Learning algorithms can be affected by an imbalanced dataset. In the case of a binary problem, this corresponds to a significant difference in the number of samples in the positive and negative classes. Therefore, we have one under-represented class and another over-represented one. Usually, the under-represented class is the one we are more interested in (positive class) (Brownlee, 2020; Imbalanced-learn, 2014). The techniques used to solve this problem and achieve equal or almost equal representation of both classes in the dataset, are undersampling, oversampling, or a combination of both. The goal is to improve the performance of the models.

Undersampling

This method consists in deleting samples from the majority class. It has the disadvantage that since vast quantities of data are discarded, there is a loss in classification performance (Analytics Vidhya, 2020; Imbalanced-learn, 2014).

Oversampling

This method consists in randomly duplicating samples from the minority class. In those cases, overfitting may occur. To overcome that, we can generate synthetic samples instead of just duplicating existing ones. The augmented dataset is then used instead of the original dataset to train a classifier (Imbalanced-learn, 2014; Weiss, 2007).

Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) is a technique that generates new samples by interpolation. The Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) is an extension of the SMOTE algorithm for which categorical data are treated differently (Imbalanced-learn, 2014). In our work, we applied SMOTE-NC in order to deal with the data imbalance issue.

SMOTE-NC

This technique is used when the data is a mixture of numerical and categorical features. Furthermore, it uses the KNN method to construct synthetic samples by randomly selecting one or more neighbors for each instance. Typically, the number of neighbors selected can be chosen by the user, along with the final number of samples that will belong to the minority and majority classes.

The algorithm used to generate new samples is represented in figure 2.10 and equation 2.8. A new sample x_{new} will be generated considering the sample x_i and the chosen number of nearest neighbors (in the figure that number corresponds to 3). Further clarifying the notation, x_{zi} is one of the nearest neighbors and λ is a random number between 0 and 1.

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \tag{2.8}$$

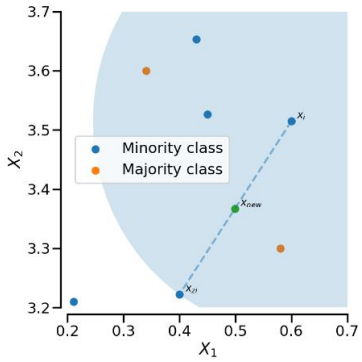

Figure 2.10: SMOTE-NC algorithm for numerical data. From Imbalanced-learn (2014).

To deal with the categorical data, we specify which of the variables are categorical ones. This is because the new data samples added can only have specific values in those categorical variables, so they belong to the same categories originally presented without any other extra interpolation. The categories of a newly generated sample correspond to the most common category of the nearest neighbors (Imbalanced-learn, 2014).

2.5 Performance Assessment Metrics

In this section, we present common performance assessment metrics that were used to evaluate the implemented ML models.

Confusion Matrix

The performance of a classifier is usually evaluated taking into account the confusion matrix. An example of a confusion matrix for a binary classification problem is represented in figure 2.11, where the predicted class (by the model) and true class (real labels) are displayed. The positive class (class 1) can be, for example, having a disease, and the negative class (class 0) not having a disease.

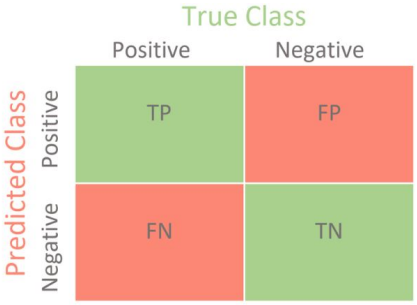

Figure 2.11: Confusion Matrix. From Aras (2020).

Below is an explanation of the confusion matrix:

- **True Positive (TP):** The predicted class is positive and the actual class is also positive.
- **True Negative (TN):** The predicted class is negative and the actual class is also negative.
- **False Positive (FP):** The predicted class is positive and the actual class is negative.
- **False Negative (FN):** The predicted class is negative and the actual class is positive.

Common Metrics

In table 2.10, the most common metrics used for performance assessment of classifiers are represented.

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FN+FP}$	Percentage of instances correctly classified
Sensitivity	$\frac{TP}{TP+FN}$	Percentage of positive instances classified as positive
Specificity	$\frac{TN}{TN+FP}$	Percentage of negative instances classified as negative
Precision	$\frac{TP}{TP+FP}$	Percentage of correct positive predictions
G-Mean	$\sqrt{sensitivity * specificity}$	Geometric mean of sensitivity and specificity
F1-Score	$2 * \frac{precision*sensitivity}{precision+sensitivity}$	Harmonic average of precision and sensitivity

Table 2.10: Common metrics used to evaluate the performance of classifiers. Based on Brownlee (2020); Sunasra (2017).

It is important to mention, that when the datasets are imbalanced, accuracy is not the best indicator of model performance. This is due to the fact that high accuracy is achievable by a model that only predicts the majority class. In those cases, the G-mean and/or the F1-score should be used (Brownlee, 2020; Sunasra, 2017). Therefore, the report metrics in our results will be sensitivity, specificity, and the G-mean.

2.6 Confidence Intervals

Nandeshwar (2006) defines a Confidence Interval (CI) as a range of values (with upper and lower confidence bounds) that has a specified probability of containing the parameter being estimated from a given set of sample data (Nandeshwar, 2006; Zhang, 2019). The width of the Confidence Interval gives us an idea of how uncertain we are about the estimated parameter. The most widely used confidence intervals are the 95% and 99% confidence intervals (Nandeshwar, 2006).

Often we want to calculate the CI of a parameter but we do not know its distribution. In those cases, we need to calculate a non-parametric Confidence Interval, that does not any assumption about the functional form of the distribution of the parameter (Brownlee, 2017, 2018). Bootstrapping is a commonly used statistical technique used for calculating a non-parametric Confidence Interval. We can use bootstrap to calculate, for example, confidence intervals for model coefficients in logistic regression or the Confidence Interval of a performance measure like prediction accuracy (Folkman, 2019; Zhang, 2019).

Bootstrap

The general bootstrap method for computing a CI works by generating bootstrap samples by resampling with replacement from the original dataset in random order. Each bootstrap sample has the same size as the original data. Since we are resampling with replacement, certain data samples may appear more than once in a bootstrap sample and some not at all. An estimate of the parameter that we are trying to calculate the CI for, is computed for each bootstrap sample, yielding various estimates of the parameter. In the case of the 95% CI, the 2.5 and 97.5 percentile values of the parameter estimates are used to compute the Confidence Interval (Nandeshwar, 2006; Raschka, 2020; Zhang, 2019).

Leave-one-out bootstrap is a variation of bootstrap that is better suited for when it is being used for the evaluation of predictive models. In this case, the dataset is split in 70% for the training set and the remaining 30% into a testing set (out-of-bag samples). The bootstrap samples are generated from the training set. On each bootstrap sample, we train an ML algorithm and use the results of the training to predict on the testing set and obtain a parameter like accuracy. Lastly, the proceeding is the same: we obtain certain percentiles depending on the used Confidence Interval to calculate its range (Raschka, 2020; Zhang, 2019). A representation of this variation of the bootstrapping method can be seen in figure 2.12.

Figure 2.12: Leave-one-out bootstrap. Adapted from Raschka (2020).

In our work, we implemented the 95% Confidence Interval on the G-Mean as a metric for interpretability evaluation. The confidence intervals were obtained by using the leave-one-out bootstrap method.

2.7 Conclusions

In this chapter, we presented background information that is be essential to understanding the pre-processing performed in our work. Regarding statistical tests, we presented an overview of common ones, and in our work, we used the Kolmogorov-Smirnov test and the non-parametric tests. We mentioned data validation strategies that were used for data partition in our work: holdout and stratified k-fold. Furthermore, we presented SMOTE-NC that was used as our random sampling strategy in order to deal with the data imbalance issue.

In addition, we explored ACS and the risk scores applied in patient risk stratification, in particular, the GRACE risk score (the most used method in Portugal). In this work, the GRACE risk score is our clinical reference and the developed ML models were compared to it in terms of performance and interpretability. The implemented ML models are supervised learning methods based on algorithms mentioned in this chapter, namely logistic regression, naive Bayes, decision trees, and decision rules. One of the implemented models is our proposed approach, which employs k-means clustering, an unsupervised learning method mentioned in this chapter.

From the mentioned performance assessment metrics, we used sensitivity, specificity, and the G-mean to compare the results of the models with the GRACE score. In this work, we evaluated interpretability resorting to a set of metrics (functionally-grounded evaluation). Shapley values, and in particular, SHAP were

essential to develop one of those metrics. The nearest neighbors method was used in the stability metric. Finally, we implemented confidence intervals (using the leave-one-out bootstrap method) also as a metric for interpretability evaluation.

3

State of the Art

Since our goal is to implement ML models in the context of mortality prediction in ACS, a Machine Learning study applied to Cardiovascular Diseases is mentioned in section 3.1. The need for interpretability and therefore, to implement interpretable ML models, is clarified in section 3.2. The problem of personalization that was addressed in one of our implemented models is presented in section 3.3. Furthermore, we present the state of art in interpretability evaluation in section 3.4, since one of our objectives is to develop measures to quantify the interpretability of the implemented models. Several properties of interpretability (descriptive accuracy, simplicity, and stability/confidence) that can be used as proxies for functionally-grounded evaluation are described. In section 3.5, our conclusions after analysing the literature, are presented. A summary of the work that was done is also given.

3.1 Machine Learning Studies on Cardiovascular Disease

As some authors have stated (Carvalho et al., 2019; Doshi-Velez, Kim, 2017; Schmidt, Biessmann, 2019), Machine Learning (ML) systems have become ubiquitous, leading to their widespread. This fact can be justified by ML systems having, in many cases, higher predictive power than statistical methods. Therefore, these systems have begun to be applied in the support of critical decision-making in many domains, including healthcare (Lahav et al., 2018).

As Qayyum et al. (2020) stated, the "early prediction and diagnosis of diseases from medical data are one of the exciting applications of ML". In Weng et al. (2017), the authors investigated whether Machine Learning can improve cardiovascular risk prediction.

Contrary to standard risk score models that assume linear relationships, ML methods "can model more complex relations between the predictors and the output" and "can easily incorporate new features detected as important ones and be applied to incomplete datasets".

Weng et al. (2017) used 4 different Machine Learning algorithms (random forest, logistic regression, gradient boosting, and neural networks) and compared them to an established algorithm (American College of Cardiology guidelines) to predict the first cardiovascular event in 10 years using a database from the United Kingdom population. The authors concluded that Machine Learning significantly improves the accuracy of cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others.

Given their potential, we developed different ML models to predict 6-month mortality of ACS patients.

3.2 The Case for Interpretability

Despite the high predictive performance (Carrington et al., 2018), ML systems, in particular, deep neural networks, are "black boxes", meaning that the internal process of generating an inference from data is not transparent and they do not provide explanations about their complex learning process. (Qayyum et al., 2020).

Interpretability has become increasingly important over the years, even more in certain high-stake domains like healthcare where accountability is crucial, and the lack of it has been the biggest adversary in the widespread adoption of ML by this sector (Panch et al., 2018). Carvalho et al. (2019) also argued for the importance of interpretability, given that it is indispensable to understand and trust Machine Learning (ML) systems. Ahmad et al. (2018) stated that Interpretable ML allows "the end-user to interrogate, understand, debug and even improve the ML system".

Going one step further, there are now regulatory incentives for interpretability, like the European Union's General Data Protection Regulation (GDPR), enforceable since May 2018, that gives to citizens the right to an explanation of algorithmic decisions made about them (Ahmad et al., 2018; Carrington et al., 2018; Carvalho et al., 2019; Panch et al., 2018). In Portugal, the National Initiative on Digital Skills published in the document AI Portugal 2030, that transparent AI is one of the fundamental research lines in the future (República Portuguesa, 2022).

Accordingly to Doshi-Velez, Kim (2017), there are also a series of desiderata of ML systems that can be optimized through interpretability, namely fairness, privacy,

causality, and trust (concepts described in chapter 2). Regarding fairness, some examples of biased decisions made by ML models in contexts with high impact were identified, e.g. with a widely used criminal risk assessment algorithmic tool (Correctional Offender Management Profiling for Alternative Sanctions- COMPAS) that was proved to perform unreliable decisions harming minority groups. Therefore, explanations are of "uttermost importance to ensure algorithmic fairness" (Carvalho et al., 2019). As for privacy, Qayyum et al. (2020) described how to use interpretability as a countermeasure to adversarial attacks (external parties try to manipulate an algorithm after learning how it operates) (Abedin, 2022).

There are studies in the healthcare context based on ML, namely to provide risk assessments for heart failure patients, that present interpretability modules and other forms of evidence about the underlying Ml models to the user. It is the case of Lahav et al. (2018), that trained a neural network to predict 1-year mortality risk and then designs a reinforcement learning-based Decision Support System (DSS) around the neural network model. The system can learn from its interactions with users and presents information about the dataset, training methodology, and model accuracy, in addition to interpretability modules including linear approximations coefficients and decision-tree approximations.

According to Ahmad et al. (2018), there is a trade-off between interpretability and performance. Intrinsic interpretable models like regression models and decision trees often perform worse on prediction tasks compared to black-box models like gradient boosting, deep learning models, and others. Despite being a challenge, Machine Learning in the context of healthcare is expected to be highly accurate and understandable at the same time (Murdoch et al., 2019; Qayyum et al., 2020), like Lahav et al. (2018) stated: "users wishing to leverage predictive models for critical decision making such as medical risk prognosis, must be professionally and ethically able to justify their medical actions". There has been some research on models which exhibit high performance as well as interpretability, like the model-agnostic method Anchors (mentioned in table 2.7). The applicability of these models in healthcare has not been convincingly demonstrated due to their infrequent application, the difficulty to understand the involved concepts from the clinical perspective and the existence of doubts concerning the reliability of the obtained explanations (Ahmad et al., 2018; Valente et al., 2022).

Considering the problems of model-agnostic methods mentioned above, and that "the easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models" (algorithms mentioned in table 2.6), in our work, we implemented intrinsic interpretable models (Molnar, 2022) with the goal to achieve higher performance than the clinical method currently in place in Portugal (GRACE risk score).

3.3 The Case for Personalization

ML has the potential to create personalized models, for example, by creating an ensemble of different models and varying the weights given to the models accordingly to the characteristics of each patient. Another example, even simpler, is to create a different model for each gender, therefore, the model applied to each patient would vary if they are a man or a woman. Moreover, some ML models give more importance to the training of patients that are similar to the patient that is being evaluated, therefore, are more related to personalized medicine. On the contrary, risk score models "can be seen as a more generalized point system" (Valente et al., 2021b).

Personalization is a conflicting topic but, in this work, we consider personalization in the sense that the model varies according to the characteristics of the patients. Considering the output of the logistic regression model (represented in equation 2.1), we would deem it a personalized model if the weights (β) would change according to the patient's features ($x^{(i)}$). Therefore, instead of having the model's weights β_0,β_1,..,β_p , we would have the different weights $\beta_0^{(i)}$,$\beta_1^{(i)}$,..,$\beta_p^{(i)}$ for each instance (represented by i). It is important to note, that we do not consider personalization the fact that the output of the model will be different for each patient because the values of the patient's features vary, even if the same weights are applied to all patients.

One of the implemented models was developed by our research group and consists of a new risk assessment methodology that besides being interpretable, also addresses the personalization issue mentioned above (Valente et al., 2021b). Furthermore, the proposed approach has several steps, including using risk factors of ACS to create several decision rules and train those rules with an Machine Learning (ML) classifier to predict the probability that each rule is correct for each patient. The information is then used to compute the risk of mortality. Therefore, the goal of personalization for ML is fulfilled, since the model selects the most relevant rules for each patient, mirroring the actions that a physician would take when doing a prognostic for a ACS patient (Valente et al., 2021b).

3.4 Evaluation of Interpretability

The need for interpretability measurement is stated clearly by Schmidt, Biessmann (2019): it provides a way for interpretability methods to be "directly compared across studies in unified benchmark tests" and other authors (Doshi-Velez, Kim, 2017; Nguyen, Martínez, 2020) also mention the importance to be able to compare methods to move forward the field of interpretability.

There are well-defined and widely used metrics of performance, like the ones mentioned in section 2.5. On the contrary, as stated in Carvalho et al. (2019) and the same year, by Schmidt, Biessmann, the quality of explanations and trust in ML predictions are difficult to measure. Developing measures for interpretability remains a challenge, which validates the need for research in the field (Molnar et al., 2020). In the following sections, we will present a few of the measures developed to assess interpretability, either by using humans (field experts or unskilled people) or proxies (qualitative or quantitative metrics).

3.4.1 Application-Grounded and Human-Grounded Evaluation of Interpretability

The work presented in this section can be considered quantitative, however, it requires task-specific user studies. Accordingly to the taxonomy presented in Zhou et al. (2021), the experiments with humans can be divided into objective metrics, such as task time length and task performance, and subjective metrics, for example, user trust, satisfaction, and understanding.

A. Objective metrics

In Schmidt, Biessmann (2019), experiments with unskilled people are performed, and using a set of equations, the quality of the explanations and the trust in ML models are evaluated. One of the metrics defined is the Information Transfer Rate (ITR), which measures the rate at which humans replicate model predictions, considering that actual model decisions are not shown. They measured the ITR in two scenarios, given that in both humans must make a decision: when no explanation of the ML predictions is provided and when it is. In their experiments, by providing explanations, humans replicated faster and more accurately the predictions of the ML model. Therefore, the study validates the advantages of explanations in the context of using ML to assist human decision-making. Moreover, the ITR measure is used to determine the trust in ML models, dividing it by the mutual information

between human decisions and true labels. It identifies when human decisions are overly biased towards the ML models and humans are trusting blindly in the models and not relying on their judgment as when model predictions are wrong and humans still agree with them.

B. Subjective metrics

In Lahav et al. (2018) doctors and Machine Learning (ML) experts were asked to rate how useful they find each piece of model evidence to quantify user trust in different types of explanations. Linear approximations coefficients and decision-tree approximations were some of the types of explanations evaluated and the decision-tree approximation proved to be the method that both doctors and ML experts trusted the most.

The survey written by Zhou et al. (2021) concluded that "subjective measures, such as trust and confidence, have been embraced as the focal point for the human-centered evaluation of explainable systems." Therefore, in this work, the explanations were partially validated by our clinical partner, since our domain end-users and field experts are clinicians. Other properties of interpretability can be quantitatively evaluated and in the next section, examples of different developed measures are presented.

3.4.2 Functionally-Grounded Evaluation of Interpretability

As Carrington et al. (2018) pointed out, the literature has few definitions for quantitative measures of model interpretability. Therefore, we also considered in this section qualitative and subject concepts that support the mentioned quantitative measures.

By reviewing the literature it is possible to conclude that to evaluate interpretability, we must evaluate a set of properties that are considered important for a model with high interpretability to have. This is due, as stated in Margot, Luta (2021), to the absence of a strict mathematical definition of interpretability since it involves many concepts. Additionally, as stated in Molnar et al. (2020)., "various quantifiable aspects of interpretability are emerging."

Either by presenting exact formulas to evaluate these properties or by given theoretical definitions of desiderata (Carvalho et al., 2019), the work of many authors falls under three pillars: evaluate the quality of the explanations (descriptive accuracy), evaluate if the model providing the explanations is stable and can be trusted (stability and confidence) and evaluate if the model is simple (simplicity). In

the following sections, we will present the work developed in the context of evaluating interpretability in an objective manner that can be applied to intrinsic interpretable models, given that those are the ones being developed in this work.

However, it is important to note the advances in research in the field of measuring interpretability in the context of different methods. Nguyen, Martínez (2020) developed several measures to evaluate methods that use feature importance and example-based methods. Arya et al. (2019) developed, among others, a measure, for feature-based local explanations, that tries to evaluate if the features' importance returned by an explainability method are the correct ones. Finally, Carrington et al. (2018) developed several measures of interpretability to be used with support vector machines.

A. Descriptive Accuracy

In Carvalho et al. (2019) accuracy and fidelity are considered goals that interpretability aims for. To understand their definition of accuracy as predictive accuracy, we must look at the work of Murdoch et al. (2019), who created a framework for rating the interpretability methods - the Predictive, Descriptive, Relevant (PDR) framework.

Murdoch et al. (2019) defined predictive accuracy as approximating the underlying data relationships with a model, meaning evaluating the quality of a model's fit through measures such as test-set accuracy, presented in section 2.5.

On the other side, the authors define descriptive accuracy as the approximation of what the model has learned using an interpretation method. Descriptive accuracy is a synonym of fidelity for other authors such as Molnar et al. (2020), who define it as "How well an explanation approximates the ML model". Moreover, fidelity is also mentioned as faithfulness in Burkart, Huber (2021) and in Nguyen, Martínez (2020). Descriptive accuracy is also related to the quality of explanations, a concept defined as soundness in Burkart, Huber (2021).

Finally, in Murdoch et al. (2019), relevancy is judged relative to a specific human audience and domain problem and can help choose, in a particular context, what type of accuracy - descriptive or predictive - is more important.

Many authors mention the difficulty in measuring descriptive accuracy, namely Murdoch et al. (2019): "descriptive accuracy is generally very challenging to measure or quantify" and Molnar et al. (2020) who stated that we don't have a way of knowing "how correct an explanation is". However, the authors Murdoch et al. (2019) propose a solution: "when an underlying scientific problem has been previously studied, prior experimental findings can serve as a partial ground truth to retrospectively

validate interpretations". Ahmad et al. (2018), also proposed "that the concordance in explanations as well as how well the explanations align with what is already known in the domain will determine explanation model preference".

For the specific problem addressed in this work, we conclude that the quality of the explanations should be evaluated considering the GRACE. In our work we compared the features' importance that each ML model gives to the different risk factors with the features' importance set by GRACE. In order to do this, we used the kernel SHAP method, described in section 2.3.3 to obtain the Shapley Values.

B. Simplicity

Carrington et al. (2018) mentioned that the simpler a model is, the easier it is to understand, interpret and describe it. Margot, Luta (2021) defined a measure for simplicity and stated that this property of interpretability ensures understandability for humans. Their measure is based on the sum of the length of all the rules of the prediction, considering that the measures of these authors are to be used in rule and tree-based algorithms. Simplicity in this context is related to sparsity, Molnar et al. (2020) and Wilson et al. (2018, as cited in, Carvalho et al., 2019) mentioned compactness as one of the "quantitative interpretability indicators" and defined it specifically as the "number of conditions in the decision rule".

In the literature, we can find examples of metrics based on model size for decision trees: the number of rules, length of rules, depth of the tree, and number of features used as splitting features (Craven & Shavlik, 1996, as cited in, Burkart, Huber (2021); Zhou et al., 2021). Other authors, mention measures for sparsity that are concerned with the total number of features used (Su et al., 2015, as cited in, Burkart, Huber, 2021). Murdoch et al. (2019) referred to the importance of simplicity, given that increasing the complexity of the model by increasing, for example, the depth of a decision tree, makes the model more difficult for a human to internally simulate. Therefore, it affects the simulatability aspect of interpretability discussed in section 2.3.1. Valente et al. (2021b) also emphasized the phenomenon that even if every single element of a model can be interpretable if the model becomes too large, the interpretability of the global model is affected, and so, models which were "a priori interpretable may become hard to explain".

The measure of simplicity, although easy to understand may be difficult to implement, considering the existence of different properties that make models interpretable (table 2.6). For example, in order to evaluate which model is simpler, logistic regression or decision trees, we would need, respectively, to compare a model sparse in features to a model sparse in rules. Nguyen, Martínez (2020) and Zhou

et al. (2021), mentioned the fact that it is not possible to define measures that can be applied to all interpretability methods, given the contextual nature of explanations.

C. Stability / Confidence

In Margot, Luta (2021) a measure for stability is presented and the authors believe that the stability property of interpretability ensures robustness (Carvalho et al., 2019; Doshi-Velez, Kim, 2017; Murdoch et al., 2019), which is also defined as reliability by a few authors (Carvalho et al., 2019; Doshi-Velez, Kim, 2017). Given that the measures presented by Margot, Luta (2021) are to be applied to rule and tree-based algorithms, their measure of stability (q-stability score) is the ratio of common rules between two sets of rules generated by an algorithm based on two independent samples (observations drawn from the same distribution). The definition of reliability and robustness is presented in Carvalho et al. (2019) - "Ensure that small changes in the input do not cause large changes in the prediction". Murdoch et al. (2019) stated that stability is a "prerequisite for trustworthy interpretations" and that "one should not interpret parts of a model which are not stable to appropriate perturbations to the model and data". In conclusion, the meaning behind stability stays the same - similar instances should be attributed the same label by the model.

Waa et al. (2018b,2018a,2020) developed a Intuitive Confidence Measure (ICM) that, according to the authors, is easy to understand and can predict how likely a given model output is correct. The measure is independent of its underlying ML model, meaning it is model-agnostic. In equation 3.1, the measure mentioned in their later work (Waa et al., 2020) can be observed. The measure consists in, for a single data point $\overrightarrow{x} \in \mathbb{R}^n$, selecting the k neighbors using a distance function (e.g. the Euclidean distance). After that step and considering only the neighbors that the ML system correctly classified, the neighbors (set S) are divided into two sets: S^+ and S^-. The S^+ consists of the neighbors where the decision of the ML model was the same as the decision for point $\overrightarrow{x}$ and that decision was correct (concerning the ground truth label). The S^- consists of the neighbors where the decision of the ML model was different from the decision for point $\overrightarrow{x}$ and that decision was correct. The ICM weights each neighbor with a kernel based on its similarity to $\overrightarrow{x}$ according to a distance function. In this case, the radial basis function is used as the kernel. Therefore, the ICM depends not only on the number of points belonging to S^+ and S^- but also on their similarity to $\overrightarrow{x}$. The standard deviation (σ) is the average similarity between the k neighbors.

The measure is bounded in the interval [-1,1] and a value of -0.5 can be interpreted as there is 50% evidence that the output of the ML model will be incorrect.

$$C\left(\overrightarrow{x}|S^{+},S^{-},d\right) = \frac{1}{|S^{+}|}\sum_{\overrightarrow{x_i}\in S^{+}} exp\left[-\left(\frac{1}{\sigma}||\overrightarrow{x}-\overrightarrow{x_i}||^2\right)^2\right] - \frac{1}{|S^{-}|}\sum_{\overrightarrow{x_j}\in S^{-}} exp\left[-\left(\frac{1}{\sigma}||\overrightarrow{x}-\overrightarrow{x_j}||^2\right)^2\right] \tag{3.1}$$

Another way of verifying that "slight changes to the input data only lead to slight changes in the output" is by computing confidence intervals (Folkman, 2019). The narrower the range of the CI, the more confidence we have in the measure since it is more precise (Brownlee, 2018).

In Zhang (2019), the author derived 95% confidence intervals on the accuracy of different Machine Learning algorithms to rank them against each other. Furthermore, in Nandeshwar (2006), confidence intervals on the results of neural networks are calculated to quantify the reliability of their performance. Both authors used several different methods to compute the confidence intervals, bootstrap being one of those (described in section 2.6).

In our work, we implemented a measure for stability based on the work of Waa et al. (2020), but adapt it to fit the interpretability issue. Instead of using ICM as a confidence measure in the ML output, we implemented it as a measure of stability, allowing us to evaluate interpretability in a variety of intrinsic interpretable models. Our measure of confidence is the 95% Confidence Interval on the geometric mean (G-Mean) of the ML models. A lot of confidence intervals are reported on accuracy, but since we are dealing with an imbalanced dataset, we used the G-Mean.

3.5 Conclusions

To conclude, after reviewing the literature on the Machine Learning studies on CVDs, we decided to implement ML models to predict the mortality of patients with ACS, given the advantages of ML compared to risk scores.

Furthermore, two problems arise in the context of implementing ML to be used in a domain like healthcare, interpretability, and personalization. To address the first problem, we implemented intrinsically interpretable models and one of them (our approach) addresses the personalization issue, mirroring the reasoning of a physician.

After analyzing the need to evaluate interpretability, we decided to compare the different implemented models not only in terms of performance but also regarding interpretability. Following the taxonomy presented in chapter 2, an application-grounded evaluation (using domain experts) is more reliable than an human-grounded evaluation (using unskilled people) (Carvalho et al., 2019; Doshi-Velez, Kim, 2017). Since the human-centered evaluation of interpretability has been linked to subjective measures like trust (Zhou et al., 2021), in our work, the explanations returned by some of the implemented models were validated by our clinical partner.

Regarding proxies measures to evaluate interpretability, three properties stand out in the literature: descriptive accuracy (quality of explanations), simplicity, and stability/confidence. After analyzing the literature on the first property, we concluded that the quality of explanations should be evaluated considering what is already known in the domain. Therefore, descriptive accuracy was evaluated by comparing the features' importance attributed by the different models with the features' importance attributed by the GRACE risk score. The features' importance were obtained using the Shapley Values returned by kernel SHAP method. Considering the existence of different properties that make models interpretable, is difficult to implement a measure of simplicity to compare the different models. Therefore, we didn't implement a measure for this property, however, this will be an important line of work for the future. Finally, for the stability property, we are interested in understanding if small changes in the input may originate significant changes in the output. To evaluate that, we adapted a measure developed by Waa et al., that evaluates if similar instances are attributed the same label by the model. Furthermore, we computed the 95% CI on the geometric mean of the models, in order to determine which models have a narrower interval and therefore, are more trustworthy.

This page is intentionally left blank.

4

Methodology

This chapter describes the proposed methodology, which consists of a pre-processing and validation stage (described in section 4.1) where the missing values were handled and various statistical tests to characterize the dataset were applied. Furthermore, we visualized the features in boxplots and histograms to perform a preliminary analysis. The dataset was divided for validation and the data imbalance issue was addressed. Following that, we implemented different interpretable models that output the prognostic of a patient with ACS (described in section 4.2), given that the two distinct outcomes are that the patient survives or dies. Finally, we explain how the models are going to be compared to the clinical reference (GRACE risk score) with regard to performance and interpretability in sections 4.3 and section 4.4, respectively. In figure 4.1, a summary of the proposed methodology is represented.

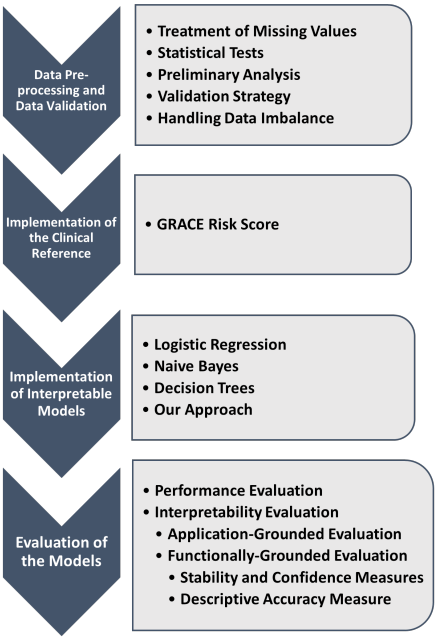

Figure 4.1: Summary of the proposed methodology.

4.1 Data Pre-processing and Data Validation

The initial stage of our work consisted of data pre-processing and validation. This phase comprised several steps namely: treatment of missing values, statistical tests, preliminary analysis, validation strategy, and handling of data imbalance.

4.1.1 Treatment of Missing Values

Concerning the treatment of missing values, the patients that don't have the class label information can't be used in our study, so we eliminated them. Furthermore, if the patient has the class label information but has missing values on the variables that were used in our work, we also decided to drop those patients.

4.1.2 Statistical Tests

Statistical tests are important to obtain more information about our dataset, namely normality, correlation, and discriminating power of variables. This information will later be useful to understand the results of the implemented models. Taking that into account, after the pre-processing, we performed on the dataset some of the statistical tests mentioned in table 2.9.

- **Kolmogorov-Smirnov Test**: used to understand if our numerical variables follow a normal distribution. We used this test to decide whether to apply parametric tests (normally distributed variables) or non-parametric tests.
- **Mann-Whitney Test**: used in each of the numerical variables to assess the difference between two distributions (the two outcomes). In our work, these two outcomes were death and survival of the patients.
- **Chi-Squared Test**: used in the categorical variables to assess the relationship between two categorical variables (each one of the categorical variables and the outcome of the patients). If the variables are independent, the observed frequencies for the categorical variables match the frequencies that we would get by chance.

Correlation

To analyze the correlation between the variables of our dataset, we computed the correlation matrix using the Spearman correlation coefficient for the numerical and ordinal variables and the Phi coefficient for the binary variables.

4.1.3 Preliminary Analysis

In order to do a preliminary analysis, we visualized our features, resorting to boxplots in the numerical variables and histograms in the categorical variables. The goal was to analyze the distributions of the features, in particular, the median of the numerical features and the absolute frequencies (for each class and feature value) of the categorical features. Furthermore, we also computed other information such as the mean and standard deviation of each numerical variable, in the two different groups corresponding to the two outcomes.

4.1.4 Validation Strategy

Before, implementing our models, we divided our data into training and test sets. We implemented the holdout method with the stratified version and the stratified k-fold method with 10 folds.

4.1.5 Handling Data Imbalance

Given that in cardiovascular risk death analysis the problem is usually imbalanced, the next step was to use random sampling in the training portion of the data. We decided to use oversampling since with undersampling there is a loss in classification performance. The oversampling technique is only used in the training portion of the data since we want the test data to represent the real world to see how our models would perform if used in a clinical setting. In our specific case, we used SMOTE-NC since we had categorical variables, and this method allows us to use oversampling in datasets that contain both categorical and numerical variables. It is important to mention that oversampling was not used in the implementation of the GRACE risk score since there weren't train and test steps and there was no need to determine parameters.

4.2 Implementation of the Clinical Reference and Interpretable Models

After the pre-processing phase described in the previous section, we implemented the GRACE risk score and the interpretable models.

4.2.1 Clinical Reference: GRACE Risk Score

We implemented the GRACE risk score by attributing to each patient the scores represented in table 2.1 depending on the values taken by each of the variables. Following that, each patient had a total score given by the sum of the scores attributed in each variable and the patients were divided into three groups (Elbarouni et al., 2009).

- **Low-risk Group** : score $\leq$ 108 for NSTEMI and Unstable Angina diagnoses and score $\leq$ 125 for STEMI.
- **Intermediate-risk Group**: score 109-140 for NSTEMI and Unstable Angina diagnoses and score 126-154 for STEMI.
- **High-risk Group**: score $\geq$141 for NSTEMI and Unstable Angina diagnoses and score $\geq$ 155 for STEMI.

Since our problem is binary (class 1 or 0), and the GRACE divides the patients into 3 groups, to attribute a class to the patients, we considered 2 scenarios.

- **Scenario 1**- Separate the patients into the following 2 groups:
  - Low-risk+Intermediate-risk
  - High-risk
- **Scenario 2**- Separate the patients into the following 2 groups:
  - Low-risk
  - Intermediate-risk+High-risk

4.2.2 Intrinsic Interpretable Models

Since we are using a labeled dataset and we aim at creating ML models to predict 6-month mortality of ACS patients, we used supervised learning, and in particular, classification methods. To address the problem of interpretability, that arises in the context of implementing ML in a critical domain like healthcare, we need to implement interpretable models. As mentioned in section 3.2, considering the problems of obtaining explanations through model-agnostic methods, in our work, we implemented intrinsic interpretable models by using the algorithms mentioned in table 2.6.

A. Logistic Regression

We implemented the logistic regression model and analyzed its interpretable properties: the coefficients and the odds ratio.

Logistic Regression Coefficients

Regarding the magnitude of the coefficients, we can only compare the coefficients of the variables that follow the same distribution. Since we standardized (z-score normalization) the numerical variables, we can compare the coefficients of the numerical variables (Choueiry, n.d.; Menard, 2011). The coefficients of the categorical binary variables can also be compared with each other (Anderson et al., 2003). In this work, we also standardized the categorical ordinal variables and compared the coefficients returned with the other coefficients returned by the numerical variables.

The logistic regression model returns a set of information that can be analyzed.

- **Log-Likelihood Ratio (LLR) Test Value**: The null hypothesis of the LLR test is that a restricted model (using only some of the variables) performs better than a complete model with all the variables that were used (Jansen, 2018).
- **P-values of the Coefficients**: If the p-values are greater than 0.05, they aren't statistically significant and this means we can't reject the null hypothesis which states that the coefficient is equal to zero (Jansen, 2018).
- **95% CI of the Coefficients**: If the interval contains the zero value, the coefficient isn't statistically significant.
- **Standard Errors**: If the standard errors are bigger than half of the magnitude of the coefficients, the coefficients aren't statistically significant (Anderson et al., 2003).

Odds Ratio

Subsequently, we calculated the exponential of the coefficients to obtain the odds ratio and we calculated the 95% CI for the odds ratio, which was obtained by calculating the exponential of the CI of the coefficients (Gonçalves, 2013).

B. Naive Bayes

We decided to implement the categorical naive Bayes algorithm and since our dataset contained numerical and categorical variables, we transformed the numerical variables into categorical ones. Our intention was for the categories to

have scientifically meaning, therefore for the variables that can be associated with clinical conditions, we used the range of values that are linked to the conditions (section 2.1.2). In the remaining variables, the categories adopted in the GRACE risk score were applied.

Probabilities Interpretation

The naive Bayes algorithm is based on $P(x_i = t \mid y = c)$: the probability of each category (for example t) in each feature (for example i), given class c (in our case, class 0 or 1) (Scikit-learn, n.d.-a). Since the naive Bayes algorithm creates interpretable models, those probabilities can be assessed to understand if they are compatible with the clinical knowledge.

C. Decision Trees

We implemented the decision tree model using the CART algorithm with the Gini index as the criterion to evaluate the quality of a split (Scikit-learn, n.d.-b). Furthermore, we analyzed the properties that make decision trees interpretable: the decision trees themselves, and the decision rules and feature importance extracted from them.

Feature Importance

The decision tree algorithm returns the features' importance that can be analyzed.

Visualization of the Tree and Decision Rules

The decision tree can be visualized, and that allows for the decision rules to be derived. In addition, we can assess if the domain knowledge extracted from the decision rules matches the actual ground truth (the clinical knowledge).

D. Our approach

Our proposed approach addresses interpretability and personalization while maintaining the goal of obtaining good performance. It involves several steps: the creation of decision rules, implementation of the personalization mechanism, computation of the prediction, and computation of a reliability measure. Furthermore, the computation of the reliability measure is an additional step and provides information that usually is not available to clinicians: the trustfulness of the algorithm for a particular prediction (Valente et al., 2021b, 2022).

Creation of Decision Rules

The first step was the creation of several rules through the dichotomization of the risk factors used in the GRACE risk score. An example of such a rule can be seen below for the age risk factor, considering a binary outcome ($\hat{t_i}$).

$$IFAGE \geq 80 \; THEN \; \hat{t_i} = 1 \tag{4.1}$$

We used the decision rules that were created by Valente et al. (2022) for some of the variables used in our work. For the remaining variables, the methodology explained next was applied. We considered "virtual patients", obtained through clustering, assuming that patients with similar characteristics can be grouped. Therefore, we obtained two centroids for each risk factor, and each centroid represents a class. In our specific case, we used k-means clustering, so the centroids are computed using the mean of the values of the corresponding group. In figure 4.2, we can see an example for feature X (Valente et al., 2022). According to the guidelines of Valente et al. (2022), for the computation of the centroids, we treated the binary variables as continuous ones.

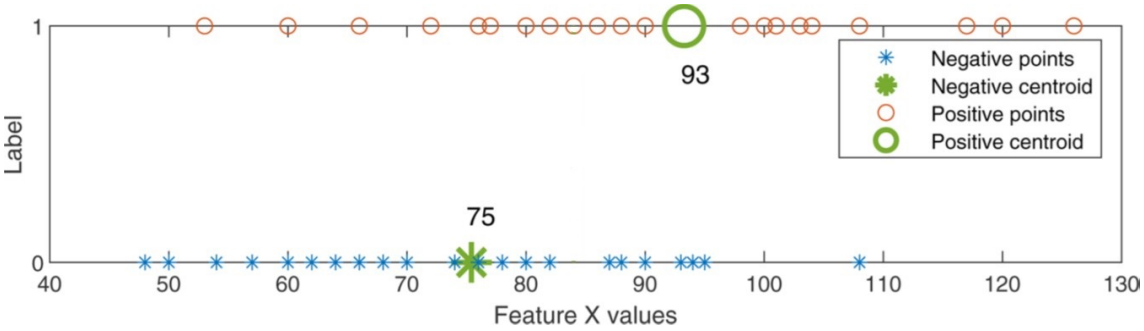

Figure 4.2: Centroids for feature X. Each centroid represents a class and they are obtained by calculating the means of the values that belong to each cluster. Adapted from Valente et al. (2021b).

To generalize the methodology for all the variables, we considered the normalized distance d_n, defined in equation 4.2, where d_1 represents the Euclidean distance from each patient to the centroid representing class 1 and d_0 represents the Euclidean distance from each patient to the centroid representing class 0 (Valente et al., 2021b, 2022).

$$d_n = 1 - \frac{d_1}{d_1 + d_0} \tag{4.2}$$

Therefore, $d_n = 0$ when the patient coincides with the "virtual patient" that represents class 0 and $d_n = 1$ when the patient matches the "virtual patient"

that represents class 1. Having that in mind, the decision rules have the following notation, where L represents a threshold, with the standard value being $L = 0.5$ (represents the mean between the 2 centroids).

$$IF\ d_n \geq L\ THEN\ \hat{t}_i = 1 \tag{4.3}$$

We tested different thresholds, to assess which one maximized the model's performance (Valente et al., 2022). After defining the decision rule for each feature, we applied them to the train and test partitions of our data. This process is represented in figure 4.3 where we consider 5 features, and therefore the creation of 5 rules. In the example, m is the number of patients in the training set and n is the number of patients in the test set.

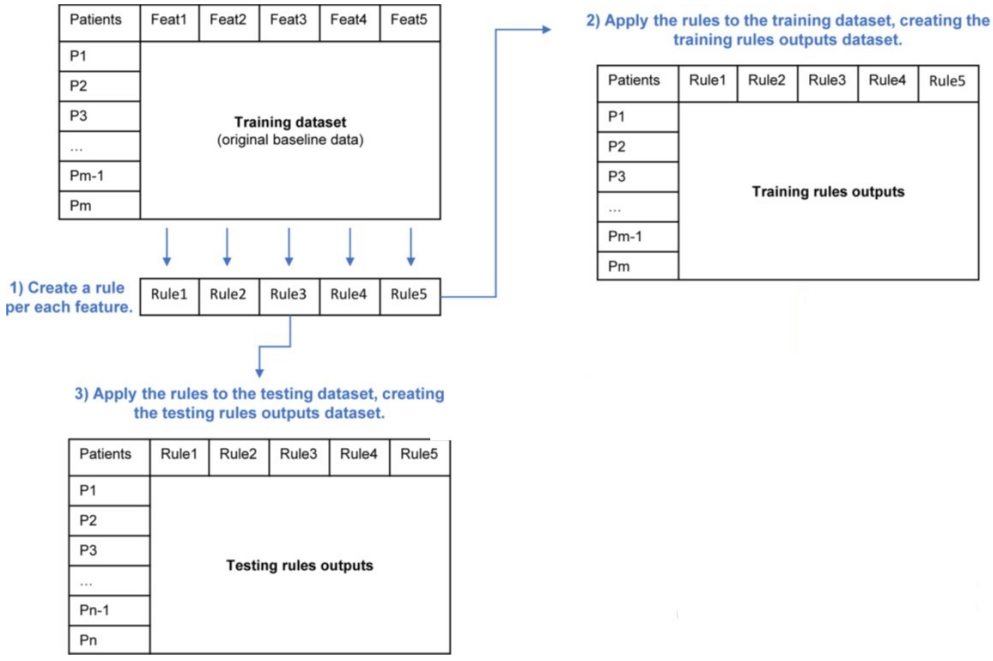

Figure 4.3: Methodology related to the creation of rules and followed application to the dataset. Adapted from Valente et al. (2021b).

The proposed approach can be considered an intrinsic interpretable model since we use decision rules that are considered the most interpretable prediction method (Molnar, 2022; Valente et al., 2021b). That is a result of their if-then structure that "semantically resembles natural language and the way we think" (Molnar, 2022). By using decision rules in our approach, the incorporation of clinical knowledge is straightforward and done in an interpretable manner.

Personalization Mechanism

In this phase, the goal was to train the rules defined in the previous phase with an ML technique to learn which rules are more appropriate for a specific patient. The interpretability of the rules is preserved, since this phase does not modify the baseline rules, only attributes an acceptance probability to each rule. This allows for personalized explanations that are simple. It is important to note that this phase, mimics the reasoning of physicians, that based on their clinical expertise (the global model's set of rules) and facing the specific characteristics of a patient, they select only the rules that are more appropriate to that particular patient. The step-by-step methodology of this phase is represented in figure 4.4. We trained a different model per rule, using all of the features as the training data, and the target corresponded to the rule acceptance.

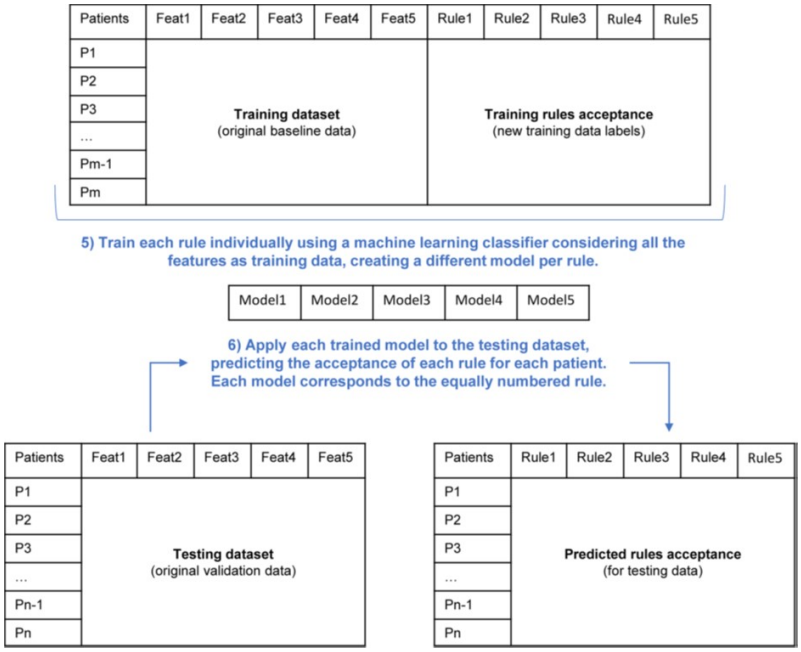

Figure 4.4: Methodology related to the training of rules and the prediction of their acceptance degree. From Valente et al. (2021b).

To obtain the training rules acceptance, we compared each rule output (training rules output) with the true output (the original training data labels) and if they are the same the training rule acceptance is 1. On the contrary, if the rule output is different from the true output, the rule acceptance is 0. An example of the computation of the rules acceptance for a patient belonging to class 0 is represented in table 4.1, where we considered the existence of 4 features, and therefore, 4 rules (Valente et al., 2022).

	Rule 1	Rule 2	Rule 3	Rule 4
Target - t_i		0		
Rule Output- $\hat{t}_i$	0	1	0	1
Rule Acceptance	1	0	1	0

Table 4.1: Calculation of the training rules acceptance. Adapted from Valente et al. (2022).

In this specific case, the ML method used was logistic regression and we standardized the numerical variables. The ML model was trained to produce a probability output, corresponding to the predicted probabilities of the rule acceptance that were obtained for each rule and each patient (Valente et al., 2021b).

Computation of the Prediction

We are applying this methodology to the problem of ACS, therefore are interested in predicting the 6-months mortality of the patients. To compute the mortality risk of each patient ($\hat{y}_i$), we used the outcomes of the rules ($\hat{t}_{ij}$) and their predicted probability acceptance ($\hat{r}_{ij}$) as can be seen in equation 4.4, where M represents the total number of rules.

$$\hat{y}_i = \frac{\sum_{j=1}^{M} \hat{t}_{ij}\hat{r}_{ij}}{M} \tag{4.4}$$

It is important to note that before applying equation 4.4, the rules that output 0 (negative rules) should be considered to have an output of -1. This is because if we considered an output of 0, that would lead to a contribution of 0 in the summation represented in equation 4.4, when dealing with the negative rules. By doing the transformation of 0 to -1 for the output of the negative rules, the risk of the patients will now be a value in the range [-1,1] (t_i). We converted this risk to the range [0,1] (s_i) through equation 4.5 (Valente et al., 2021b).

$$s_i = \frac{t_i + 1}{2} \tag{4.5}$$

To binarize the mortality risk and therefore attribute a class (0 or 1) to each one of the patients, we used equation 4.6, where P represents a threshold. The negative class (0) in this case corresponds to survival and the positive class (1) represents death.

$$\hat{t}_i = (\hat{y}_i \geq P) \tag{4.6}$$

The threshold P was obtained by finding the one that maximizes the performance of the model (Valente et al., 2022).

Reliability Measure

As the final step of our approach, we also computed the reliability measure, meaning, the degree of trust, in each patient's outcome. If a patient belongs to the positive class, then it is expected that the rules with an output of 1 have a greater predicted acceptance than the rules with an output of 0, and vice versa. In an ideal scenario, the rules that suggest the patient belongs to their true class would have 100% predicted acceptance and the rules of the opposite class would be rejected (0% predicted acceptance). Therefore, the bigger the difference between the acceptation of rules that suggest the patient belongs to class 0 and rules that suggest the patient belongs to class 1, the more confidence we have in the algorithms' output (Valente et al., 2022). The reliability measure of a patient's prediction ($\hat{w}_i$) was computed through equation 4.7, where p represents the number of rules that suggest the patient belongs to the positive class and q the number of rules that suggest the patient belongs to the negative class (Valente et al., 2021b)

$$\hat{w}_i = \left| \frac{1}{p} \sum_{j=1}^{p} \hat{r}_{ij} - \frac{1}{q} \sum_{j=1}^{q} \hat{r}_{ij} \right| \tag{4.7}$$

An example of a computation of this reliability measure for two patients, having into consideration their predicted rules acceptance and rule outputs is represented in table 4.2. The predicted mortality risk is computed through equation 4.4, followed by the transformation in equation 4.5. In the example, we considered 5 different rules.

	Patient 1		**Patient 2**	
	Predicted rule acceptance	**Rule output**	**Predicted rule acceptance**	**Rule output**
Rule 1	73%	0	88%	0
Rule 2	91%	0	95%	0
Rule 3	41%	1	24%	1
Rule 4	34%	1	11%	1
Rule 5	70%	0	91%	0
Predicted mortality	**0.34**		**0.26**	
Mean acceptance of the positive rules	37.50%		17.50%	
Mean acceptance of the negative rules	78.00%		91.33%	
Predicted reliability estimate	**40.50%**		**73.83%**	

Table 4.2: Example of the computation of the predicted mortality risk and the reliability measure. Adapted from Valente et al. (2021b).

Overview

In figure 4.5, is represented an overview of the proposed approach. First, we address interpretability through the creation of decision rules based on the risk factors used in the GRACE risk score. Then we train the rules with an ML model that gives the probability of each rule being correct for each patient (rule acceptance - $\hat{r}_{ij}$). We compute the mortality risk for each patient ($\hat{y}_i$) using the outcomes of the decision rules and their predicted acceptance. Finally, we binarize the mortality risk ($\hat{t}_i$) to obtain the class that each patient belongs to. The proposed approach also gives a reliability measure ($\hat{w}_i$) for each patient, so the clinicians are provided with a degree of trust in each patient's outcome.

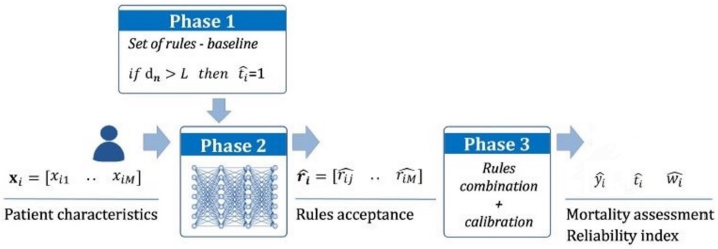

Figure 4.5: Summary of the main phases of the proposed approach. Adapted from Valente et al. (2022).

4.3 Evaluation of the Performance of the Models

To evaluate the performance of the implemented models, we used some of the performance assessment metrics described in section 2.5, namely sensitivity, specificity, and G-Mean will be presented. The reported performance of the models corresponds to the average values of the metrics after 10 runs.

4.4 Evaluation of the Interpretability of the Models

Interpretability evaluation was done based on an application-grounded evaluation and on a functionally-grounded evaluation (according to the taxonomy presented in section 2.3.4). In this manner, we have the information if the models can be used in a real scenario and have a way of quantifying objectively their interpretability.

4.4.1 Application-Grounded Evaluation

The application-grounded evaluation of the interpretability of the models consists of using a domain expert, in our case, the clinical partner, to assess if the explanations returned by the different models are trustworthy. Therefore, the information given by the domain expert answers the question of whether the physicians would accept that the models were implemented in the healthcare domain. As expected, the application-grounded evaluation doesn't apply to the clinical reference (GRACE risk score) since it is already the most widely used method in Portuguese hospitals.

4.4.2 Functionally-Grounded Evaluation

The functionally-grounded evaluation of interpretability uses a set of proxies to evaluate interpretability in a quantifiable manner. The proxies are properties of interpretability that can be found in the literature, namely stability/confidence and the quality of explanations provided by the interpretable methods (descriptive accuracy). The stability/confidence property relates to the fact that the models should be stable in a way to provide confidence in their results.

A. Stability and Confidence Measures

To evaluate if the models attribute the same label to similar instances, we computed a stability measure (equation 4.8) inspired by the work of Waa et al.

$$Stability\left(\overrightarrow{x}|S^{+},S^{-},d\right) = \frac{\sum_{\overrightarrow{x_i}\in S^{+}} exp\left(-\frac{||\overrightarrow{x}-\overrightarrow{x_i}||^2}{2\sigma^2}\right) - \sum_{\overrightarrow{x_j}\in S^{-}} exp\left(-\frac{||\overrightarrow{x}-\overrightarrow{x_j}||^2}{2\sigma^2}\right)}{k} \tag{4.8}$$

There were two types of alterations that were made on the measure developed by Waa et al. (equation 3.1).

- **Numerical alterations**: Some numerical alterations like adding constants were performed in order to obtain a metric that would fit the stability problem. We verified that the metric achieved greater stability when we increased the number of neighbors with the same label.
- **Considerations of the equation**: In our case, for the definition of the S set, all neighbors of a specific point were considered, instead of employing only the neighbors correctly classified like Waa et al. did. The goal is to evaluate stability in isolation, without considering the performance of the model.

Considering the alterations, let's clarify the meaning of the notation used in equation 3.1.

- $\overrightarrow{\boldsymbol{x}}$: The data point that we are calculating the stability for ($\overrightarrow{x}$ *in* $\mathbb{R}^n$).
- $\boldsymbol{k}$: The number of neighbors of point $\overrightarrow{x}$ considered. The k-nearest neighbors algorithm was used to find the neighbors and we tested different numbers of neighbors. We concluded that the best results were obtained by using 3 neighbors.
- $\boldsymbol{\sigma}$: The standard deviation was calculated the same way as in Waa et al. (2020):

$$\sigma = \frac{1}{k} \sum_{\overrightarrow{x}_i \in S} ||\overrightarrow{x} - \overrightarrow{x}_i||$$

- $\boldsymbol{S}$: The set that contains the k neighbors of $\overrightarrow{x}$.
- $\boldsymbol{S^+}$: Neighbors where the decision of the ML model was the same as the decision for point $\overrightarrow{x}$.
- $\boldsymbol{S^-}$: Neighbors where the decision of the ML model was different than the decision for point $\overrightarrow{x}$.

The stability measure returns values in the range [-1,1]. Below we provide some examples of possible values, in order to clarify their meaning.

- **-1:** All the k neighbors belong to a class other than the class of $\overrightarrow{x}$.
- **-0.5:** Around 25% of the neighbors belong to the same class as point $\overrightarrow{x}$.
- **0:** If k is an even number, half of the neighbors belong to a class other than the class of $\overrightarrow{x}$, and the other half belongs to the same class as point $\overrightarrow{x}$.
- **0.5:** Around 75% of the neighbors belong to the same class as point $\overrightarrow{x}$.
- **1:** All the k neighbors belong to the same class as point $\overrightarrow{x}$.

The stability measures calculates the stability for each prediction, therefore to obtain the general stability for each model we averaged the stability obtained for all patients in the dataset. Furthermore, the reported stability of the different implemented models is the average of the stability obtained in the 10 runs we performed. It is important to note that by using the stability measure defined in equation 4.8, the models that attribute the same label to all patients, would have a very high stability. Therefore, this measure should be evaluated combined with a confidence measure.

Our confidence measure consists in the 95% CI on the geometric mean of the predicative models. The Confidence Interval was computed using leave-one-out bootstrap (method describe in section 2.6). In our case, we used 1000 bootstrap samples. In each iteration, 70% of the dataset was used to generate the bootstrap

sample and train the model, and the remaining 30% to compute the predictions and calculate the geometric mean. After the 1000 iterations, we had several estimates of the geometric mean. Using the 2.5 and 97.5 percentile values of the geometric mean estimates, we computed the 95% CI. This way, we can evaluate if the classifier has a reasonable performance combined with a good stability, and the CI provides a measure of trust in the sense that the narrower the CI, the more confidence we can have in the model output.

B. Descriptive Accuracy Measure

Descriptive accuracy is related to the quality of explanations and was evaluated by comparing the features' importance attributed by the different models with the features' importance attributed by the GRACE risk score. To obtain the features' importance, we average the absolute Shapley values (returned by the kernel SHAP) per feature across the data. Features with Shapley values closer to zero have less impact (López, 2021; Molnar, 2022). It is important to note that some of our used models, already compute feature importance, namely logistic regression and decision trees. The advantage of using SHAP is to have a common way of analyzing feature importance for all the models.

In order to quantify the descriptive accuracy, we obtained the features' importance, and then, for the models and also for the GRACE risk score, we ranked the features by importance. Finally, for each intrinsic interpretable model, we computed the Spearman correlation, between the features' rank attributed by that model and the rank attributed by GRACE. We considered the Spearman correlation coefficient for the GRACE risk score to be 1 since the rank of the features attributed by each model was compared to the rank considered by GRACE.

Furthermore, by using the SHAP package in Python we also computed, for each model, the summary plot and the force plot. Besides feature importance, the SHAP summary plot allows us to analyze feature effects, to understand if increasing or decreasing the value of a feature has a positive (increase the prediction) or negative (decrease the prediction) impact on the output. The force plot for a random data instance analyzes the contribution of each feature value for the prediction of that instance compared to the average prediction for the dataset.

This page is intentionally left blank.

5

Results and Discussion

This chapter describes the used dataset (section 5.1) and the relevant results of our work. In section 5.2 the computation of the variables and class label is explained, alongside the description of how the missing values were handled. Besides that, the results of the different statistical tests performed are presented, and a preliminary analysis is performed on the dataset. In section 5.3 some notes are given on the implementation of the validation strategy and random sampling. Furthermore, it is explained the implementations of the models, the performance evaluation, and the properties that make the models interpretable are analyzed. Besides that, the results from the functionally-grounded evaluation of interpretability are also exhibited. Regarding application-grounded evaluation, our results were partially validated by our clinical partner. In section 5.4, is given an overview of the evaluation of the models regarding performance and functionally-grounded evaluation of interpretability.

5.1 Dataset

The dataset contains information on 1544 patients admitted between 2009 and 2016 at the Hospital dos Covões Cardiology ICU, containing patients from all spectrum of ACS diagnosis (STEMI, NSTEMI and UA). Of the 1544 patients, 70.08% are men and 29.92% are women. Moreover, it includes information such as the variables used in the GRACE risk score (**age, heart rate, systolic blood pressure, creatinine, Killip class, troponin, STEMI and cardiac arrest**), and the information if the patient died or survived. Besides that, it also includes medication information, the presence of comorbidities, records of prior medical procedures of the patients, and other pieces of information like height, weight, and smoking habits. The results from medical exams performed on the patients like blood tests and urinalysis are also registered. In the dataset, it is included the date of admission to the hospital and the date of death in the patients that died, as well as the cause of death.

It is important to note that prior approval was obtained to conduct this study, and the patient data were anonymized.

In this work, we used the variables from the GRACE risk score and their range is represented in table 5.1. It is important to note that in the binary variables, a value of 1 (0) means the presence (absence) of a specific feature (for example, a value of 1 (0) in the cardiac arrest feature indicates that the patient was (was not) in cardiac arrest at hospital admission).

	Range
Age (years)	30-101
Heart Rate (bpm)	32.00-180.00
Systolic Blood Pressure (mmHg)	50.00-248.00
Creatinine (μmol/L)	24.70-1934.40
Killip Class	1-4
Troponin	0/1
STEMI	0/1
Cardiac Arrest	0/1

Table 5.1: Range or values taken by each of the variables that were used throughout our work.

In figure 5.1 we can see the class distribution (survival or death) in each of the ACS diagnosis (NSTEMI, STEMI and Unstable Angina) from where we can conclude that our dataset has a higher percentage of NSTEMI cases (42%), following STEMI (37%) and lastly Unstable Angina (18%). Furthermore, almost 16% of people diagnosed with NSTEMI die and the mortality rate for STEMI diagnosis is close to that, with around 14% diagnosed people being deceased. For Unstable Angina cases, the mortality decreases to nearly 8%.

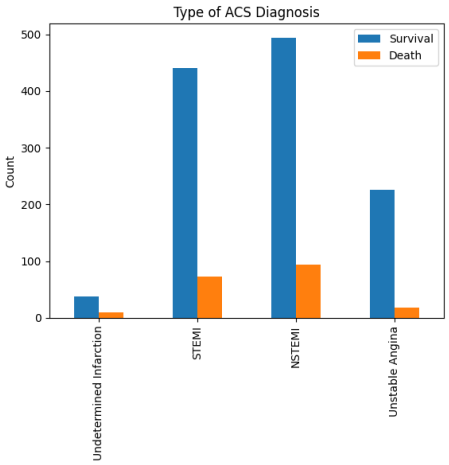

Figure 5.1: Class distribution in each of the ACS diagnoses.

In figure 5.2 we can see the class distribution for each gender: 12.34% of men died and 17.87% of women die. In our dataset, 71% of the patients are men and 29% are women.

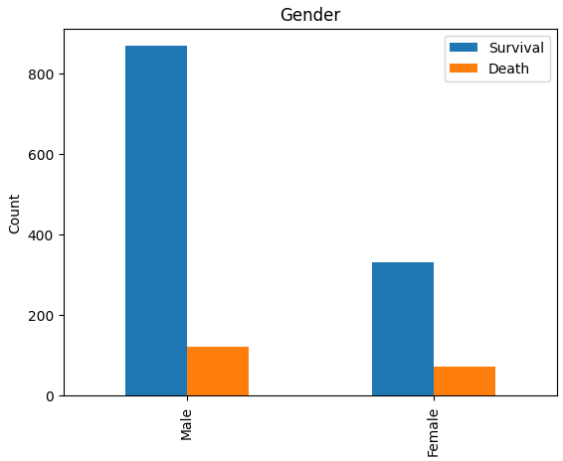

Figure 5.2: Class distribution for each gender.

5.2 Data Pre-processing

5.2.1 Computation of Variables

Concerning the GRACE variables, we extracted the following ones directly from the dataset: age, heart rate, systolic blood pressure, creatinine, and STEMI. The troponin variable was obtained using information contained in a variable from the dataset where the ACS diagnosis of the patient is registered. In that way, if a patient has STEMI or NSTEMI it means the troponin levels were elevated, so a value of 1 in the troponin variable is attributed to that patient and a value of 0 is attributed in the patients that have unstable angina. The cardiac arrest variable was calculated based on the Killip class of the patients, if they had a Killip of 4, a value of 1 was attributed to the cardiac arrest variable and a value of 0 was attributed in the remaining situations. Therefore, the Killip class and the cardiac arrest are highly correlated variables. That correlation played an important factor in logistic regression and the measures we took to overcome the high correlation are described in section 5.3.2.

5.2.2 Missing Values and Computation of Class Label

We analyzed the dataset and we encountered 73 patients who didn't have the respective class label (survival/death). Since the GRACE risk score is used for a

6 months prognosis, we used the admission date of a patient and death date to calculate if the death of the patient had happened within 6 months after hospital admission. For the 73 patients referenced above, there was no way to know if they had survived or died. Therefore, we discarded those 73 patients plus 42 more that had the information that did die but no death date was registered. After that, the patients that had a difference between the admission and death date superior to 6 months were considered as belonging to class 0 (surviving) for this study. In that way, below we clarify the meaning of the class in this problem (outcome of a patient).

- **Class=0 (Survived)**: 6 months after the hospital admission date, the patient was alive.
- **Class=1 (Deceased)**: 6 months after the hospital admission date, the patient was deceased.

After these operations, the class distribution of our dataset was the following: 86.06% of the patients survived and 13.94% died. Therefore, we have a case of an imbalanced dataset, which is common in cardiovascular risk death analysis.

Following the attribution of a class for each of the patients based on the date of admission and date of death, we evaluated the percentage of the missing values in each of the variables used in the GRACE risk score (represented in table 5.2). Therefore, we eliminated the patients that had missing values in those variables, finally arriving at 1392 patients with full GRACE variables and class information (approximately 90% of the original patients).

	Missing Values (%)
Age	0.00
Heart Rate	1.12
Systolic Blood Pressure	1.05
Creatinine	1.33
Killip Class	0.21
Troponin	0.00
STEMI	0.07
Cardiac Arrest	0.21

Table 5.2: Percentage of missing values for each of the variables used throughout our work.

5.2.3 Statistical Tests and Preliminary Analysis

A. Normality and Discriminating Power

Following the preprocessing of the dataset, we executed several statistical tests. The results are presented in table 5.3 and will be explained below.

	Chi-Square Test	Kolmogorov-Smirnov Test	Mann-Whitney U Test
Age		0.00	8.28^{-29}
Heart Rate		0.00	5.19^{-8}
Systolic Blood Pressure		0.00	7.24^{-6}
Creatinine		0.00	1.05^{-35}
Killip Class	2.24^{-41}		
Troponin	0.03		
STEMI	**1.00**		
Cardiac Arrest	1.86^{-22}		

Table 5.3: Results for the Chi-Square test (performed in the categorical variables), the Kolmogorov-Smirnov test, and Mann-Whitney U test (performed in the numerical variables). The p-values that aren't lower than 0.05 are marked in bold.

Kolmogorov-Sminorv Test

We compared the distribution of each numerical variable with the normal distribution. Since the p-values are all much smaller than 0.05 (as can be seen in table 5.3) we reject the null hypothesis that the numerical variables follow a normal distribution. In conclusion, we need to use non-parametric tests on the numerical variables.

Mann-Whitney Test and Boxplots of Numerical Variables

The distributions of two groups of patients (the individuals that survived and the individuals that died) were compared. Given that, all p-values are lower than 0.05 in the numerical variables, we can reject the null hypothesis that the groups follow the same distribution and therefore conclude that the numerical variables have discriminating power in the outcome of a patient.

The boxplots of the numerical variables can be seen in figure 5.3 where the median is represented in orange, the boxes represent the Q1 to Q3 quartile range and the dots represent the outliers. The whiskers that extend from the edges of the box intend to show the range of the data (Pandas Documentation, 2022). By analyzing the boxplots we can conclude, that in each variable, the survival group and the death group follow different distributions and have different medians. Furthermore, the boxplots display that the group of individuals that died is older, has higher

heart rate, lower systolic blood pressure, and higher creatinine. Those conclusions are consistent with the points given by the GRACE risk score to each one of these variables (table 2.1), since more points are attributed, the higher the age of the individual, the higher the heart rate, the higher the creatinine and the lower the systolic blood pressure.

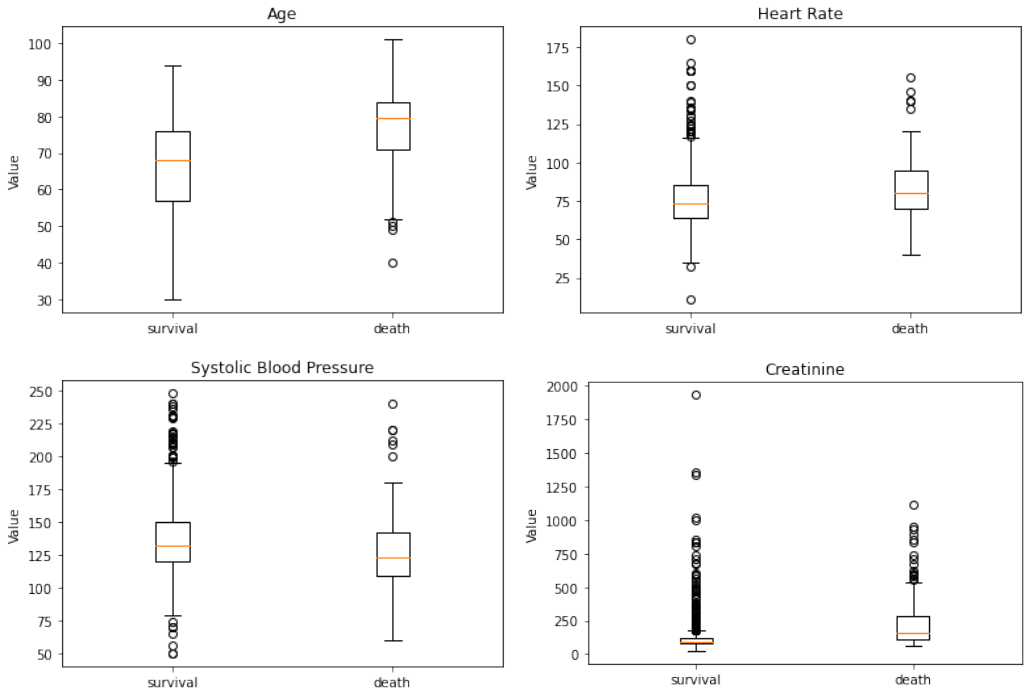

Figure 5.3: Boxplots of the numerical variables used in our work: age, heart rate, systolic blood pressure, and creatinine.

Means of Numerical Variables

The means for the groups of patients (survival/death) are represented in table 5.4 for each one of the numerical variables. It is important to note, that the same conclusions were drawn before based on the median (figure 5.3), since the group of patients that died is older, has lower systolic blood pressure, and has higher heart rate and higher creatinine values.

	Mean (Surviving Class)	**Mean (Death Class)**
Age (years)	66	77
Heart Rate (bpm)	75.65	82.96
Systolic Blood Pressure (mmHg)	135.42	126.16
Creatinine (μmol/L)	121.48	233.91

Table 5.4: Mean values for the numerical variables used in our work separated by class.

Chi-Squared Test and Histogram of the Categorical Variables

For the categorical variables, we performed the Chi-Squared test to evaluate the relationship between each one of the categorical variables and the outcome of the patients. In the case of the Killip class, the troponin, and the cardiac arrest variables, the p-values are lower than 0.05, so we can reject the null hypothesis that the observed frequencies for the variables match the frequencies that we would get by chance. This means that we can reject the hypothesis that the Killip class, the troponin, and the cardiac arrest variables are independent of the outcome. It is important to note that the p-value for the troponin variable is much greater than the p-value for the Killip and cardiac arrest variables, so the evidence against the null hypothesis is weaker. In the case of STEMI variable, we can't reject the hypothesis that this variable and the outcome have no association, since the p-value is greater than 0.05.

These results are in accordance with the information observed in figure 5.4, since in the Killip class, for the patients belonging to class 4, more individuals die than the ones that survive, despite overall our dataset containing more individuals that survive. In the same figure, we can reach the same conclusion for the cardiac arrest variable, since considering the patients in cardiac arrest, more individuals die than the ones who survive. In table 5.5 we can also see that the mortality rate rises as the Killip class goes from 1 to 4 and that the mortality rate in patients in cardiac arrest is much higher than the mortality rate in patients that are not in cardiac arrest. Those results are analogous to the GRACE risk score since higher points are attributed if the patient belongs to a higher Killip class. In the case of cardiac arrest, 43 points are attributed if a patient is in that condition (cardiac arrest=1) and none if the patient isn't.

On the other side, in table 5.5 it is possible to see that there is not a big difference between the mortality rates for people with elevated troponin and the mortality rate for people without elevated troponin levels. Regarding the STEMI variable, the mortality rates between patients with an ST-elevation and patients without an anomaly detected in the ECG are similar. The situation with the STEMI variable, as it will be explained in further sections, may be related to the fact that when a patient enters the hospital and an ST-elevation is detected in the ECG, the individual is treated with extreme urgency and in many situations, it can save his life. Therefore, the STEMI diagnostic leads to clinicians performing interventions in the patients that modify their global risk of mortality, lowering it.

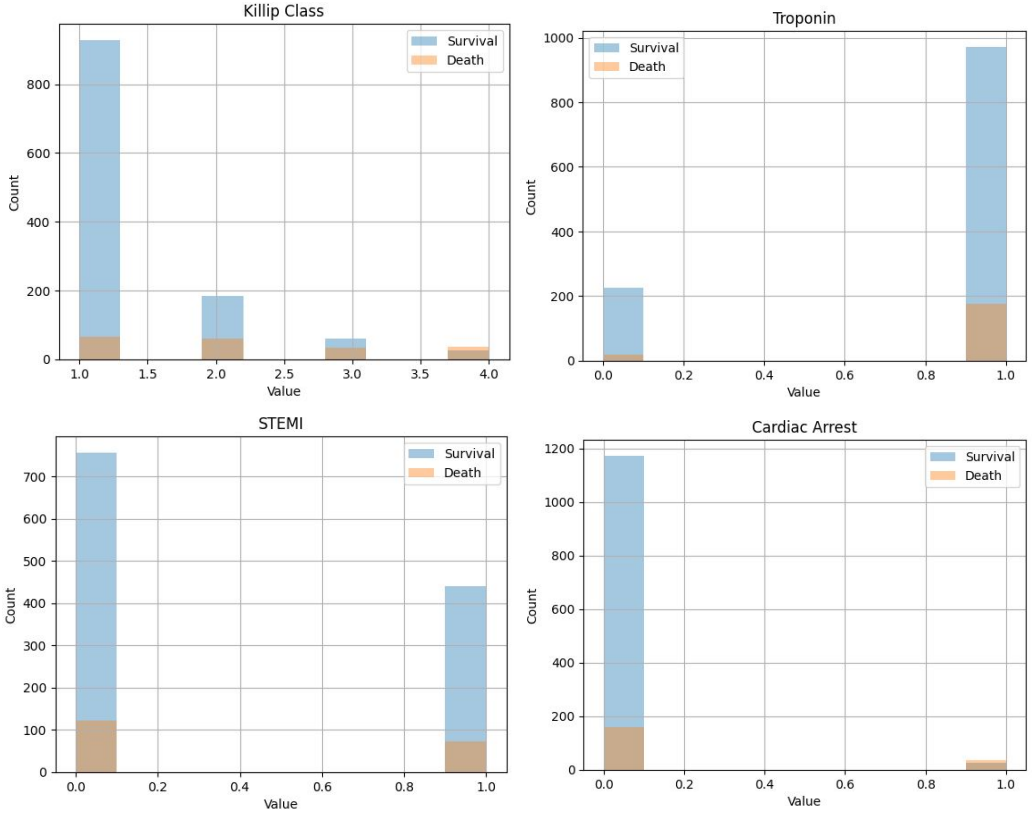

Figure 5.4: Distributions of the categorical variables used in our work: Killip class, troponin, STEMI and cardiac arrest.

	Value	Percentage of Individuals (%)	Mortality Rate(%)
Killip Class	1	71.34	6.45
	2	17.53	24.59
	3	6.75	36.17
	4	4.38	59.02
Troponin	0	17.53	7.38
	1	82.47	15.33
STEMI	0	63.07	13.78
	1	36.93	14.20
Cardiac Arrest	0	95.62	11.87
	1	4.38	59.02

Table 5.5: Percentage of individuals that have each value in the categorical variables and mortality rate for each value.

B. Correlation

The correlation between the numerical and ordinal variables was evaluated through the Spearman correlation coefficient and the correlation between the binary variables through the Phi Coefficient. The correlation matrices are represented in figures 5.5 and 5.6.

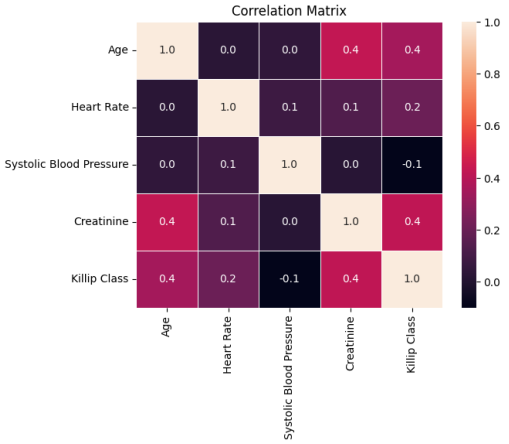

Figure 5.5: Correlation between the numerical and ordinal variables.

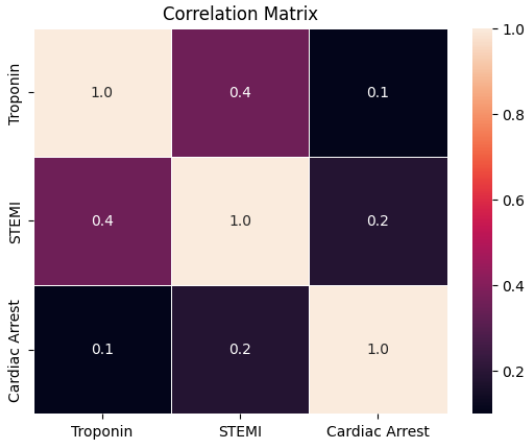

Figure 5.6: Correlation between the binary variables.

We can conclude that either between the numerical/ordinal variables or between the binary variables, no pairs of variables are highly correlated. Therefore, we have the assurance that our results will not be affected by correlation, except for the correlation between the Killip class and cardiac arrest (already discussed before in section 5.2.1).

5.3 Implementation and Evaluation of the Models

Posterior to analyzing the dataset, we started to implement our clinical reference (GRACE risk score) and several models, including logistic regression, naive Bayes, decision trees, and our proposed approach.

We verified that the results for the performance metrics were identical when fixing other conditions and varying only the method for data partition (holdout and stratified k-fold), however with the stratified k-fold we obtained lower standard deviations in the metrics, so we choose this method to divide our data. Therefore, our validation strategy consisted of a 10-times repeated 10-fold stratified cross-validation.

The performance of the models improved when oversampling was used, namely the sensitivity since the number of true positives increased. We implemented SMOTE-NC as our oversampling strategy and after some tests, we concluded that the best results were obtained when using 2 neighbors to interpolate the values for the added patients. Furthermore, the obtained class distribution (56% of patients belonging to class 0 and 44% belonging to class 1) has the best compromise between performance results and a lower number of new patients added. It is important to note, that by using oversampling, we slightly change the distribution of the variables. We decided to present in the next sections, the results obtained for the performance of the models using stratified k-fold to divide our data and using SMOTE-NC in the training portion of that data. As it was mentioned before, oversampling was not used in the GRACE risk score. In our approach, we also did not use oversampling as it will be explained further in section 5.3.5.

In this section we also present, the metrics used to evaluate quantitively the interpretability of the models, namely the stability measure, the confidence measure (95% CI on the geometric mean), and the descriptive accuracy measure (the Spearman correlation between the features' rank by importance attributed by the GRACE risk score and the features' rank by importance attributed by each one of the models). The stability measure reported is the mean and standard deviation of the stability in the 10 runs. The features' importance was obtained using the kernel SHAP method, by considering the absolute value of the Shapley values for each feature averaged for the entire test dataset instances. The features' importance was also averaged for the 10 different test partitions obtained with 1 run of stratified k-fold. We didn't average the features' importance for the 10 runs as we did with the performance and stability metrics due to the high computation time needed to perform that task. Furthermore, the SHAP force plot for a random test instance is analyzed for each one of the implemented models. The SHAP summary plot (figure 5.9) for a test partition of the dataset is presented for the GRACE risk score and the remaining plots are shown in appendix A, given that the conclusions regarding the features' effects are the same for all models.

5.3.1 GRACE

After applying the GRACE risk score to all patients in our dataset, we concluded that we have 375 (27%) patients belonging to the low-risk group, 406 (29%) patients belonging to the intermediate-risk group, and 611 (44%) patients in the high-risk group. Figure 5.7 represents the class distribution according to the risk stratification. We can conclude that the GRACE risk score performs a good

risk stratification since the mortality increases with the risk that was attributed to the patient (26.68 % mortality in the high-risk patients, followed by 5.42 % in the medium-risk patients and 2.4 % in the low-risk patients).

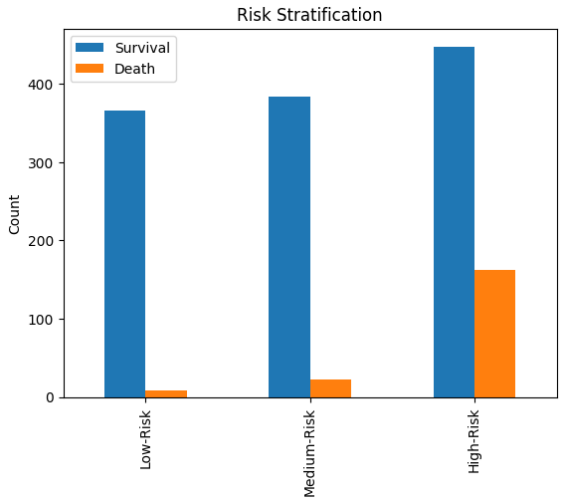

Figure 5.7: Class distribution according to risk stratification.

Transformation to a Binary Problem

After testing the 2 scenarios mentioned in section 4.2.1, we concluded that by grouping the intermediate-risk patients with the low-risk patients to form class 0 (survival) and considering high-risk patients the class 1 (death), we obtained better performance. Therefore, the results mentioned in the following sections were obtained considering that scenario.

A. Performance Evaluation

It is important to mention that with the GRACE risk score there aren't parameters to be learned during training like in Machine Learning models. Therefore, we report the performance metrics seen in table 5.6, only for the test partition of the dataset but there isn't any difference in the computation of the GRACE risk score in the train and test partitions except for their size. We chose to report the performance on the test partition, instead of applying the GRACE risk score to the whole dataset, to compare the GRACE with the remaining ML models on the same conditions.

	Geometric Mean (%)	**Sensitivity (%)**	**Specificity (%)**
GRACE Risk Score	72.11±0.07	84.02±0.07	62.1±0.00

Table 5.6: Performance metrics for the GRACE risk score.

B. Functionally-Grounded Evaluation of Interpretability

In table 5.7, the metrics for the quantitative evaluation of interpretability are displayed.

	Stability [-1,1]	**Geometric Mean Confidence Interval**	**Spearman Correlation [-1,1]**
GRACE Risk Score	0.506 ± 0.006	[68.2 % , 76.6%] (8.4%)	1

Table 5.7: Functionally-grounded evaluation of interpretability for the GRACE risk score.

Feature Importance

The features' rank by importance when applying the general GRACE risk score to our dataset is represented in table 5.8. We considered a Spearman correlation coefficient of 1 for the clinical reference since the features' rank for the other models was compared to the GRACE risk score features' rank.

Feature	**Rank - GRACE Risk Score**
Age	1
Heart Rate	6
Systolic Blood Pressure	4
Creatinine	5
Killip Class	3
Troponin	7
STEMI	2
Cardiac Arrest	8

Table 5.8: Ranks of the used features considering the features' importance returned by SHAP in our implementation of the GRACE risk score.

SHAP Force Plot

In figure 5.8, the force plot for an instance of the dataset is represented.

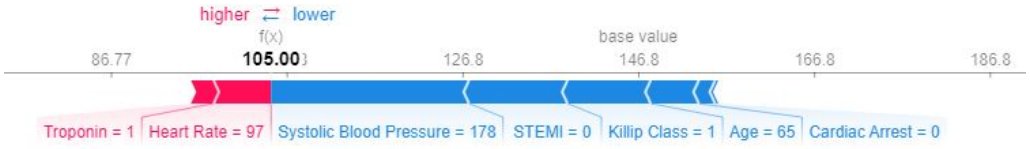

Figure 5.8: SHAP force plot for the GRACE risk score.

Several conclusions can be derived.

- **Base value:** The base value represents the average prediction for the whole dataset and for the GRACE that corresponds to a risk of approximately 147, that can be considered medium-high since a risk $\geq 141 (\geq 155)$ for NSTE-ACS (STEMI) patients is considered high. This result is consistent with the prior observations made about figure 5.7, since the majority of our patients fall under a high-risk diagnosis.
- **Predicted Risk:** The predicted risk for this patient is 105 and since the patient doesn't have a STEMI diagnosis, he/she can be considered a low-risk patient (score ≤ 108). According to the applied transformation to a binary problem, this patient would be considered as belonging to class 0 (survival).
- **Variable's Effects:** Note that only some variable's effects are represented, since the ones that aren't don't have a big positive/negative effect on the output (variables with low importance for this patient's output).
  - **Troponin:** The Shapley value for troponin is a force that pushes to increase the prediction for this patient (increase the risk). Since the value for troponin is one, this result is in accordance with the theoretical GRACE risk score since 15 points are attributed if the patient has high cardiac biomarkers (and none are attributed if the patient has not).
  - **Heart Rate:** The Shapley value for the heart rate variable is a force that pushes to increase the prediction for this patient. This result is equivalent to previous conclusions since the patient has a value of 97 bpm in this variable, higher than the average for the death class (82.96 bpm) represented in table 5.4. It is important to note that for this variable, higher values are associated with a higher probability of death, as can be seen in the boxplot of figure 5.3.
  - **Systolic Blood Pressure:** The Shapley value for the systolic blood pressure variable is a force that pushes to decrease the prediction for this patient. This result is in accordance with previous conclusions since the patient has a value of 178 mmHg bpm in this variable, higher than the average for the survival class (135.42 mmHg) represented in table 5.4. Note that for this variable, higher values are associated with the survival of the patient, as can be seen in the boxplot of figure 5.3.
  - **STEMI:** The Shapley value for the STEMI variable is a force that pushes to decrease the prediction for this patient. This result is analogous to the domain knowledge, since the patient has a value of 0 in this variable

and in the GRACE risk 30 points are attributed if the patient has an ST-elevation in the ECG (and none are attributed if the patient has not).

- **Killip Class:** The Shapley value for the Killip class variable is a force that pushes to decrease the prediction for this patient. This result is equal to previous conclusions, since the patient has a value of 1 in this variable and in table 5.5, we can see that the mortality rate in the patients that belong to Killip class 1 is very low compared to the mortality in the remaining Killip classes.
- **Age:** The Shapley value for the age variable is a force that pushes to decrease the prediction for this patient. This result is in accordance with previous conclusions since the patient has a value of 65 in this variable, very close to the average in the survival class (66 years) represented in table 5.4.
- **Cardiac Arrest:** The Shapley value for the cardiac arrest variable is a force that pushes to decrease the prediction for this patient. This result is in accordance with previous conclusions, since the patient has a value of 0 in this variable and in table 5.5, we can see that the mortality rate in the patients that are not in cardiac arrest is much lower than the mortality rate in the patients that are in this condition.

SHAP Summary Plot

Figure 5.9 displays the summary plot outlining features' importance and features' effects. Regarding features' importance, they were already reported in table 5.8. Analyzing the features' effects, we can conclude that higher values on age, Killip class, creatinine, and heart rate, push to increase the prediction (increase the patient's risk). However, for the systolic blood pressure variable, lower values push to increase the prediction. For the binary variables, a value of 1 pushes to increase the prediction, and a value of 0 pushes to decrease the prediction. This conclusions are in accordance with the previous ones reported in this work, considering both the points attributed in the GRACE risk score (table 2.1), means and boxplots of the numerical variables (table 5.4 and figure 5.3, respectively) and the mortality rate in each of the values taken by the categorical variables (table 5.5). The SHAP summary plots for the implemented ML models are represented in appendix A.

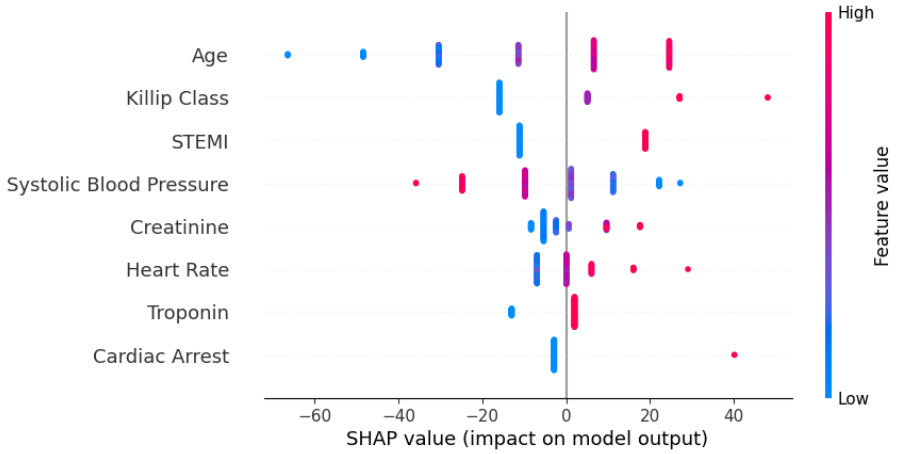

Figure 5.9: SHAP summary plot for the GRACE risk score.

5.3.2 Logistic Regression

Regarding the logistic regression model, different procedures had to be done in order to reach our main goal: obtain the logistic regression coefficients to calculate the odds ratio and interpret them for each variable (part B.).

Standardization

Primarily, it was needed to standardize the numerical variables (age, heart rate, systolic blood pressure, and creatinine) to zero mean and unit variance to be able to compare the magnitude of the coefficients of the numerical variables that would be returned by the model. The variable corresponding to the Killip class (containing values of 1 to 4) of a patient was also standardized because despite being a categorical variable, it is ordinal since a higher value of the Killip class represents a worse evaluation of the patient's condition. The categorical binary variables weren't standardized since their values are only 0 and 1, so their coefficients can already be compared with each other.

Correlated and Non-Significant Variables

Our clinical partner suggested that the correlation between some variables could be influencing our results. After investigation, we concluded that, since the cardiac arrest variable was calculated based on the Killip class, that was affecting the results of the logistic regression model regarding the significance level of the returned coefficients. We were able to obtain a statistically significant coefficient for the cardiac arrest variable, only when we left out of the model the Killip class. Therefore, we decided to not use the Killip class variable in our logistic regression model.

The STEMI variable had a coefficient that was not statistically significant. We tried to run the logistic regression model only with the age and STEMI variables and even so, the STEMI coefficient still had a p-value higher than 0.05. It is important to note, that this result is in agreement with the information on our dataset since the mortality rates in people with ST-elevation and people without it are very close in magnitude (table 5.5). Furthermore, the STEMI variable doesn't have an association with the outcome, as we concluded from the Chi-Square test represented in table 5.3. The explanation for these findings, as stated before in part A. of section 5.2.3 relies on the fact that when a patient enters the hospital with symptoms of myocardial infarction and an ST-elevation is detected in the ECG, the patient is treated immediately and with much more urgency than for example, someone with a Non-ST-Elevation Myocardial Infarction. Therefore, besides patients with STEMI having a higher risk of death, since they are treated immediately, some end up not dying, which explains the data collected and therefore the results we obtained with the STEMI coefficient. Following these conclusions, we decided to implement the model without the Killip class and STEMI variables.

A. Performance Evaluation

The performance evaluation of logistic regression is represented in table 5.9.

	Geometric Mean(%)	**Sensitivity(%)**	**Specificity(%)**
Train	76.95±0.26	74.06±0.48	79.97±0.15
Test	73.57±0.41	68.15±0.41	79.96±0.39

Table 5.9: Performance metrics for the logistic regression model.

B. Interpretability of Logistic Regression

Logistic Regression Coefficients

We extracted the coefficients returned by the model for each of the considered variables. In addition to the coefficients, we obtained the standard error, the p-values, and the 95% confidence interval for each one of the coefficients. All of these values are represented in table 5.10.

		LLR p-value	1.37^{-149}		
		Coefficients	**Standard Error**	**P-value**	**95% Confidence Interval (Coefficients)**
Numerical	**Age**	1.07	0.07	0.00	[0.93, 1.21]
	Heart Rate	0.46	0.06	0.00	[0.34,0.57]
	Systolic Blood Pressure	-0.27	0.06	0.00	[-0.39,-0.15]
	Creatinine	0.71	0.07	0.00	[0.56,0.85]
Categorical	**Troponin**	0.91	0.19	0.00	[0.54,1.29]
	Cardiac Arrest	1.10	0.27	0.00	[0.58,1.62]

Table 5.10: Coefficients values and their 95% Confidence Interval returned by the logistic regression model. The LLR p-value, the standard errors and p-values of the coefficients are also represented in the table. The represented values are the averaged values on the 10 runs performed.

The interpretation of the different values presented in the table is clarified below.

- **LLR Test P-Value**: Since $p - value < 0.05$, we reject the null hypothesis and conclude that this model performs better than a restricted model (model using only some of the variables that were used in this model).
- **P-values of the Coefficients**: Since $p - values < 0.05$, the coefficients are statistically significant and we can reject the null hypothesis that they are equal to zero.
- **95% CI of the Coefficients:** Since the intervals don't contain the zero value, the coefficients are statistically significant.
- **Standard Errors:** Since the standard errors are smaller than half of the magnitude of the coefficients, we conclude, once again, that the coefficients are statistically significant.

Regarding the magnitude of the coefficients, the biggest numerical predictor is the age of the patient, and the biggest categorical binary predictor the occurrence of cardiac arrest. In the GRACE, these two variables are also the ones that the scoring system attributes more points, precisely 43 points if a patient is in cardiac arrest and the mean of points for age is 45.5. Furthermore, the age variable can count for over 90 points if the person is older than 80.

Odds Ratio

Subsequently to obtaining the coefficients, we calculated their exponential to obtain the odds ratio that can be seen in table 5.11. Furthermore, the 95% confidence interval for the odds ratio was obtained by calculating the exponential of the confidence interval of the coefficients. The means and standard deviations of the numerical variables are also represented since we standardized them. As stated before, if the odds ratio is bigger (smaller) than 1, there is an increase (decrease) in the odds of dying for every standard deviation increase in the numerical variables.

For the binary variables, if the odds ratio is bigger than 1, there is an increase in the odds ratio for the people with the characteristic present compared to the people without it.

		Odds Ratio	95% Confidence Interval	Standard Deviation	Mean
Numerical	Age	2.92	[2.53,3.35]	13 years	71 years
	Heart Rate	1.58	[1.40,1.77]	18.26 bpm	78.76 bpm
	Systolic Blood Pressure	0.76	[0.68,0.86]	27.14 mmHg	131.17 mmHg
	Creatinine	2.03	[1.75,2.34]	164.84 μmol/L	172.38 μmol/L
Categorical	Troponin	2.48	[1.72,3.63]		
	Cardiac Arrest	3.00	[1.79,5.05]		

Table 5.11: Values of the odds ratio and their 95% Confidence Interval. For the numerical variables, it is also represented the mean and standard deviation values averaged on the 10 runs performed.

Analyzing the results of table 5.11, we can reach the conclusions shown below for each variable.

- **Age**:
  - Holding other variables fixed, there is an increase in the odds by a factor of 2.92 for every standard deviation increase in age (13 years).
  - These results are obviously in accordance with the GRACE risk score since it attributes more points to an individual if he or she is older.
- **Heart Rate:**
  - Holding other variables fixed, there is an increase in the odds of dying versus surviving by a factor of 1.58 for every standard deviation increase in the heart rate, approximately 18.26 bpm.
  - These results are obviously analogous to the GRACE risk score since it attributes more points to patients with higher heart rates.
- **Systolic Blood Pressure:**
  - Holding other variables fixed, there is a decrease in the odds of dying versus surviving by a factor of 0.76 for every standard deviation increase in the systolic blood pressure (27.14 mmHg).
  - The results are also in accordance with our clinical reference since if the levels of systolic blood pressure are higher, fewer points are attributed.
- **Creatinine:**
  - Holding other variables fixed, there is a increase in the odds of dying versus surviving by a factor of 2.03 for every standard deviation increase in the levels of creatinine (164.84 μmol/L).

  - The results are also equal to our clinical reference since if the levels of creatinine are higher, more points are attributed.
- **Troponin:**
  - Holding all the other features at fixed values, the odds of dying versus surviving increase by a factor of 2.48 for people with troponin levels elevated, compared to people without troponin levels elevated.
  - These results are analogous to our clinical reference since 15 points are attributed if an individual has elevated cardiac markers (and none if it has not elevated cardiac markers).
- **Cardiac Arrest:**
  - Holding all the other features at fixed values, the odds of dying versus surviving increase by a factor of 3.00 for people in cardiac arrest, compared to people not in cardiac arrest.
  - These results are analogous to our clinical reference since if a person is in cardiac arrest, 43 points are attributed to that variable (and none are if the person isn't in cardiac arrest).

C. Functionally-Grounded Evaluation of Interpretability

The functionally-grounded evaluation of interpretability is represented in table 5.12.

	Stability [-1,1]	**Geometric Mean Confidence Interval**	**Spearman Correlation [-1,1]**
Logistic Regression	0.634 ± 0.015	[67.3 % , 79.6%] (12.3%)	0.66

Table 5.12: Functionally-grounded evaluation of interpretability for the logistic regression model.

Feature Importance

The features' rank by importance considering the Shapley values returned by SHAP is represented in table 5.13, for the GRACE risk score and for logistic regression. The Spearman correlation between the 2 ranks was calculated and we obtained a value of 0.66 (represented in table 5.12).

Feature	Rank - Logistic Regression	Rank - GRACE Risk Score
Age	1	1
Heart Rate	3	4
Systolic Blood Pressure	5	2
Creatinine	2	3
Troponin	4	5
Cardiac Arrest	6	6

Table 5.13: Ranks of the used features considering the features' importance returned by SHAP for the logistic regression model and for the GRACE risk score.

We can compare the features' rank by importance with the magnitude of the logistic regression coefficients that were analyzed before (part B.). If we look at the features' rank regarding the numerical features, the order of importance is the same as the descending order of magnitude of the coefficients. Age is the biggest predictor, followed by creatinine, heart rate, and lastly systolic blood pressure.

SHAP Force Plot

In figure 5.10, the force plot for an instance of the dataset is represented.

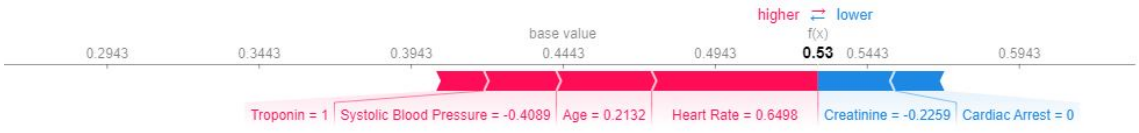

Figure 5.10: SHAP force plot for logistic regression.

Several conclusions can be derived.

- **Base value:** The base value represents the average prediction for the whole dataset and for logistic regression that is the probability of death of approximately 0.5.
- **Predicted Probability:** The predicted probability of belonging to class 1 for this patient is 0.53. Since it is higher than 0.5, we consider this patient to belong to class 1 (death).
- **Variable's Effects:** The numerical features values are standardized since we transformed them to zero mean and unit variance. The conclusions for the variable's effects are the same as the ones drawn for the force plot of an instance considering the GRACE risk score. Let's clarify the effect of the creatinine variable, which was not analyzed before.

– **Creatinine:** The Shapley value for creatinine is a force that pushes to decrease the prediction for this patient (decrease the probability of death). The value for this feature is -0.2259, and before the standardization, this value corresponded to 132.10 μmol/L. Therefore, this result is identical to previous conclusions, since it is close to the mean value of creatinine for the survival class (121.48 μmol/L), represented in table 5.4.

5.3.3 Naive Bayes

We used the categorical naive Bayes algorithm in Python to implement the naive Bayes model in our work.

Discretization of the numerical variables

In order to transform the numerical variables into categorical ones, for the variables that can be associated with medical conditions (heart rate, systolic blood pressure, and creatinine), we used the range of values that are affiliated with the conditions (section 2.1.2). For the age variable, we used the categories that the GRACE risk score applies. This information is represented in table 5.14. For the categorical ordinal and binary variables, the original categories were used.

Variable	Categories	Values		Medical Condition
Age (years)	0	$\leq$ 40		
	1	[40,50[		
	2	[50,60[		
	3	[60,70[		
	4	[70,80[		
	5	$\geq$80		
Heart Rate (bpm)	0	<60		Bradycardia
	1	[60,100]		Healthy Adult
	2	>100		Tachycardia
Systolic Blood Pressure (mmHg)	0	<120		Normal Blood Pressure
	1	[120,130[		Elevated Blood Pressure
	2	[130,140[		Hypertension Stage 1
	3	[140,180]		Hypertension Stage 2
	4	>180		Hypertensive Crisis
Creatinine (μmol/L)		**Men**	**Women**	
	0	<61.9	<53.0	Low Creatinine Values
	1	[61.9,114.9]	[53.0,97.2]	Healthy Adult
	2	>114.9	>97.2	High Creatinine Values

Table 5.14: Categories for the numerical variables used in the implementation of the naive Bayes model.

A. Performance Evaluation

The results of the performance evaluation of naive Bayes are represented in table 5.15.

	Geometric Mean(%)	**Sensitivity(%)**	**Specificity(%)**
Train	75.81±0.20	73.57±0.43	78.13±0.24
Test	74.39±0.33	71.57±0.70	77.94±0.25

Table 5.15: Performance metrics for the naive Bayes model.

B. Interpretability of Naive Bayes

The naive Bayes model returns the probability of each category in each feature given class 0 and class 1. The returned probabilities are represented in tables 5.16 to 5.23 and the values represented are the averaged on the 10 runs that were performed. The conclusions for each feature are also mentioned.

Age

Category / **Class**	**0**	**1**	**2**	**3**	**4**	**5**
Survival (0)	0.02	0.11	0.19	0.23	0.27	0.18
Death (1)	0.00	0.01	0.05	0.11	0.35	0.48

Table 5.16: Probabilities for each category in the age variable given class 0 or 1.

- For categories 0, 1, 2, and 3 (age<70 years), the probability for each category given class 0 is higher than the probability for each category given class 1. However, in categories 4 and 5 (age≥70 years), the probability for each category given class 1 is higher. Furthermore, the probability for each category given class death becomes higher as we go from category 0 to 5 (as the age of the patient increases).
- The results are in accordance with the GRACE risk score since the older the patient, the more points are attributed.

Heart Rate

Category / **Class**	**0**	**1**	**2**
Survival (0)	0.14	0.78	0.08
Death (1)	0.07	0.79	0.14

Table 5.17: Probabilities for each category in the heart rate variable given class 0 or 1.

- For category 1 (normal values of heart rate), the probability given class 0 and given class 1 is almost equal. However, for category 0 (low values of heart rate), the probability given class survival is higher than the probability given class death. On the other side, for category 2 (high values of heart rate) the probability given class death is higher.
- The results are in accordance with the GRACE risk score since more points are attributed to higher values of heart rate, therefore in those situations, the risk of mortality of the patient is higher.

Systolic Blood Pressure

Class \ Category	**0**	**1**	**2**	**3**	**4**
Survival (0)	0.25	0.20	0.14	0.35	0.06
Death (1)	0.39	0.20	0.13	0.25	0.03

Table 5.18: Probabilities for each category in the systolic blood pressure variable given class 0 or 1.

- For category 0 (lower values of systolic blood pressure), the probability given class death is higher than the probability given class survival. On the other side, for categories 3 and 4 (higher values of systolic blood pressure) the probability given class survival is higher.
- The results are analogous to the GRACE risk score since more points are attributed to lower values of systolic blood pressure.

Creatinine

Class \ Category	**0**	**1**	**2**
Survival (0)	0.03	0.66	0.32
Death (1)	0.00	0.24	0.75

Table 5.19: Probabilities for each category in the creatinine variable given class 0 or 1.

- The probability given class 1 is higher as we go from category 0 (lower creatinine values) to category 2 (higher creatinine values).
- The results are in accordance with the GRACE risk score since more points are attributed to higher values of creatinine.

Killip Class

Class \ Killip Class	1	2	2	4
Survival (0)	0.77	0.15	0.05	0.02
Death (1)	0.35	0.32	0.17	0.16

Table 5.20: Probabilities for each category in the Killip class variable given class 0 or 1.

- Considering the Killip class 1, the probability given class survival is higher than the probability given class death. However, for the Killip classes 2, 3, and 4, the probability given class death is higher.
- The results are in accordance with the GRACE risk score since more points are attributed to higher Killip classes, therefore in those situations, the risk of mortality of the patient is higher.

Troponin

Class \ Feature Value	0	1
Survival (0)	0.19	0.81
Death (1)	0.03	0.97

Table 5.21: Probabilities for each category in the troponin variable given class 0 or 1.

- Considering the feature value 0 for the troponin feature (not having elevated troponin), the probability given class survival is higher than the probability given class death. On the other side, considering the value 1 (having elevated troponin), the probability given class death is higher than the probability given class survival.
- The results are analogously the GRACE since the risk score only attributes points if the patient has elevated cardiac biomarkers.

STEMI

Class \ Feature Value	0	1
Survival (0)	0.63	0.37
Death (1)	0.69	0.31

Table 5.22: Probabilities for each category in the STEMI variable given class 0 or 1.

- In each of the feature values (0 and 1) the probabilities given class survival and death are almost equal.
- The results are in accordance with previous conclusions since the STEMI variable and the outcome of our problem do not have an association as proven by the Chi-Squared test represented in table 5.3 and the almost equality of the mortality rate in each of the feature values represented in table 5.5.

Cardiac Arrest

Class \ Feature Value	**0**	**1**
Survival (0)	0.98	0.02
Death (1)	0.88	0.12

Table 5.23: Probabilities for each category in the cardiac arrest variable given class 0 or 1.

- Considering the value 0 for the cardiac arrest feature, the probability given class survival is higher than the probability given class death. However, considering the value of 1 for cardiac arrest, the probability given class death is higher.
- The results are in accordance with the GRACE since the risk score only attributes points if the patient is in cardiac arrest.

C. Functionally-Grounded Evaluation of Interpretability

The functionally-grounded evaluation of interpretability for naive Bayes is represented in table 5.24.

	Stability [-1,1]	**Geometric Mean Confidence Interval**	**Spearman Correlation [-1,1]**
Naive Bayes	0.606 ± 0.008	[67.9 % , 79.2%] (11.3%)	0.26

Table 5.24: Functionally-grounded evaluation of interpretability for the naive Bayes model.

Feature Importance

The features' rank by importance considering the Shapley values returned by SHAP is represented in table 5.13, for the GRACE risk score and for the naive Bayes model. The Spearman correlation between the 2 ranks was calculated and we obtained a value of 0.26 (represented in table 5.24).

Feature	Rank - Naive Bayes	Rank - GRACE Risk Score
Age	2	1
Heart Rate	7	6
Systolic Blood Pressure	4	4
Creatinine	1	5
Killip Class	3	3
Troponin	5	7
STEMI	8	2
Cardiac Arrest	6	8

Table 5.25: Ranks of the used features considering the features' importance returned by SHAP for the naive Bayes model and for the GRACE risk score.

SHAP Force Plot

In figure 5.11, the force plot for an instance of the dataset is represented.

Figure 5.11: SHAP force plot for naive Bayes.

Several conclusions can be derived.

- **Base value:** The base value represents the average prediction for the whole dataset and for naive Bayes that is the probability of death of approximately 0.5.
- **Predicted Probability:** The predicted probability of belonging to class 1 for this patient is 0.81. Since it is higher than 0.5, we consider this patient to belong to class 1 (death).
- **Variable's Effects:** The numerical features were transformed into categorical ones, therefore all the feature values represented in the plot are categories. The transformation from numerical to categorical is represented in table 5.14. The conclusions for the variable's effects are the same as the ones drawn before.

5.3.4 Decision Trees

Parameters of the Decision Tree

In order to keep the decision tree model interpretable, we controlled some parameters, namely the maximum depth of the tree, which was fixed at 3. Another

parameter that was specified was the minimum number of samples required to be at a leaf node (5). Therefore, a split point at any depth was only considered if it left at least 5 training samples in each one of the branches (Scikit-learn, n.d.-b).

A. Performance Evaluation

The results regarding the performance metrics for the decision tree model are displayed in table 5.26.

	Geometric Mean(%)	**Sensitivity(%)**	**Specificity(%)**
Train	78.13±0.28	80.68±2.15	76.00±2.24
Test	71.98±1.26	70.81±1.60	73.96±2.15

Table 5.26: Performance metrics for the decision tree model.

B. Interpretability of Decision Trees

Our decision tree algorithm used different decision trees for different train partitions and runs performed. One of those decision trees can be visualized in figure 5.12. The decision rules for that particular decision tree are also analysed (equations 5.1 to 5.8). Furthermore, the features' importance for that particular iteration of the algorithm are presented in table 5.27.

Visualization of the Tree

The decision tree for a randomly selected train partition (9 of the 10 folds obtained with stratified k-fold) is represented in figure 5.12. In each subset, the Gini index, the number of samples in that subset, and the class for that subset are represented. The lower the Gini index the more pure the subset is, with a bigger difference between the number of instances that belong to each class (mentioned in value, the first number indicates the number of samples who belong to class 0 and the second the number of samples who belong to class 1). The class of each subset is chosen with majority voting between the samples of each subset (Galarnyk, 2019).

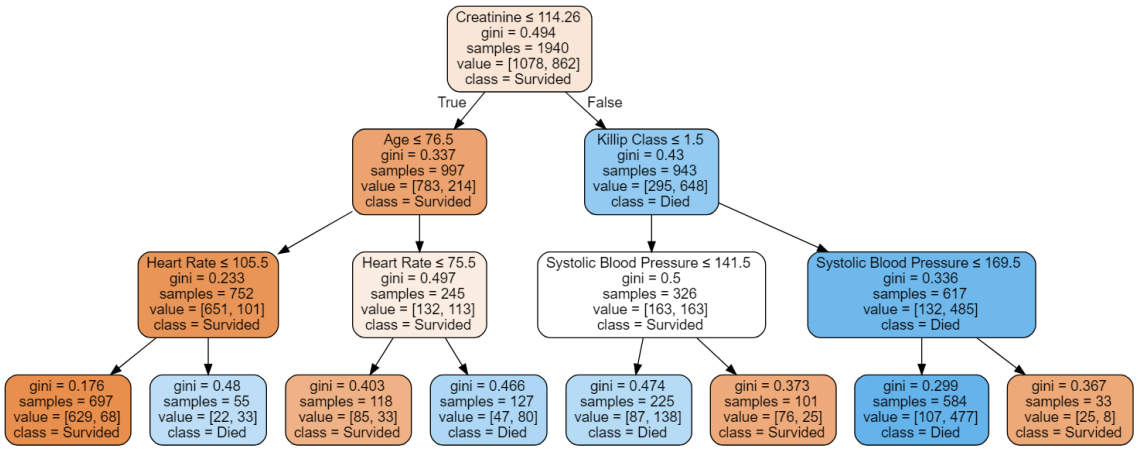

Figure 5.12: Decision tree for a randomly selected train partition of our dataset.

Decision Rules

The decision tree represented in figure 5.12 can be converted to a set of decision rules:

IF (Creatinine $\leq$ 114.26 μmol/L) AND (Age$\leq$76.5 years) AND (Heart Rate $\leq$ 105.5 bpm) THEN class=Survived (5.1)

IF (Creatinine$\leq$114.26 μmol/L) AND (Age$\leq$76.5 years) AND (Heart Rate >105.5 bpm) THEN class=Died (5.2)

IF (Creatinine$\leq$114.26 μmol/L) AND (Age>76.5 years) AND (Heart Rate $\leq$ 75.5 bpm) THEN class=Survived (5.3)

IF (Creatinine$\leq$114.26 μmol/L) AND (Age>76.5 years) AND (Heart Rate >75.5 bpm) THEN class=Died (5.4)

IF (Creatinine>114.26 μmol/L) AND (Killip $\leq$1.5) AND (Systolic Pressure$\leq$141.5 mmHg) THEN class=Died (5.5)

IF (Creatinine>114.26 μmol/L) AND (Killip $\leq$1.5) AND (Systolic Pressure>141.5 mmHg) THEN class=Survived (5.6)

IF (Creatinine>114.26 μmol/L) AND (Killip >1.5)
AND (Systolic Pressure$\leq$169.5 mmHg) THEN class=Died (5.7)

IF (Creatinine>114.26 μmol/L) AND (Killip >1.5)
AND (Systolic Pressure>169.5 mmHg) THEN class=Survived (5.8)

According to the GRACE risk score, more points are attributed to higher values of creatinine, therefore meaning a higher risk of mortality for the patient. We can verify that in some rules, the condition of having higher (lower) values of creatinine corresponds to the death (survival) class in the final subset, namely for rules 5.1, 5.3, 5.5, and 5.7. However, the same isn't verified in rules 5.2, 5.4, 5.6, 5.8. Furthermore, according to the GRACE risk score, more points are attributed to older patients. We can verify that in some rules, the condition of being older (younger) corresponds to the death (survival) class in the final subset, namely for rules 5.1 and 5.4. However, the same isn't verified in rules 5.2 and 5.3. The GRACE risk score assigns more points to patients with higher heart rates. We can verify that, the condition of having higher (lower) values of heart rate corresponds to the death (survival) class in all the final subsets of the rules that the heart rate variable was used, namely for rules 5.1, 5.2, 5.3 and 5.4. In the GRACE risk score, more points are assigned to patients that belong to a higher Killip class. We can verify that the condition of having a higher (lower) Killip class corresponds to the death (survival) class in some rules, namely for rules 5.6 and 5.7. However, the same isn't verified in rules 5.5 and 5.8. Finally, the GRACE risk score sets more points for patients with lower values of systolic blood pressure. We verify that the condition of having lower (higher) values in that variable corresponds to the death (survival) class in all of the rules that the variable is present, namely, 5.5, 5.6, 5.7, 5.8. In the rules 5.1 and 5.7 all of the conditions follow the rules of the GRACE risk score. It is interesting to verify that those final subsets are the ones with a lower Gini index.

Feature Importance

Feature	Importance
Age	**0.11**
Heart Rate	**0.11**
Systolic Blood Pressure	**0.11**
Creatinine	**0.58**
Killip Class	**0.09**
Troponin	0.00
STEMI	0.00
Cardiac Arrest	0.00

Table 5.27: Features' importance returned for the decision tree represented in figure 5.12. The values different from 0 are marked in bold.

Analyzing the features' importance represented in table 5.27, we can conclude that the troponin, STEMI, and cardiac arrest variables seem to have null importance on the model, as can be seen in figure 5.12, since they aren't included in the decision tree.

C. Functionally-Grounded Evaluation of Interpretability

The results of the quantitative evaluation of interpretability for the decision tree model are represented in table 5.28.

	Stability [-1,1]	**Geometric Mean Confidence Interval**	**Spearman Correlation [-1,1]**
Decision Tree	0.648 ± 0.029	[59.4 % , 77.1%] (17.7%)	0.66

Table 5.28: Functionally-grounded evaluation of interpretability for the decision tree model.

Feature Importance

The features' rank by importance considering the Shapley values returned by SHAP is represented in table 5.29, for the GRACE risk score and the decision tree model. The Spearman correlation between the 2 ranks was calculated and we obtained a value of 0.66 (represented in table 5.28).

Feature	Rank - Decision Tree	Rank - GRACE Risk Score
Age	2	1
Heart Rate	5	5
Systolic Blood Pressure	4	3
Creatinine	1	4
Killip Class	3	2
Troponin	6	6

Table 5.29: Ranks of the used features considering the features' importance returned by SHAP for the decision tree model and for the GRACE risk score.

By analyzing the Shapley values for the STEMI and cardiac arrest features, we concluded that they have null importance in the decision tree model as they did for the specific decision tree represented in figure 5.12. Therefore, despite being used in the decision tree model, we didn't consider them for the importance rank. The troponin feature has very small importance and that result is understandable since, for the decision tree represented in figure 5.12, it didn't have any importance. The Killip class has considerable higher importance here than it did for the decision tree of figure 5.12 (obtained for a random train partition), which is understandable given that in the ranks of the features we are considering the average of the features' importance for the 10 test partitions considering 1 run using stratified k-fold.

SHAP Force Plot

In figure 5.13, the force plot for an instance of the dataset is represented.

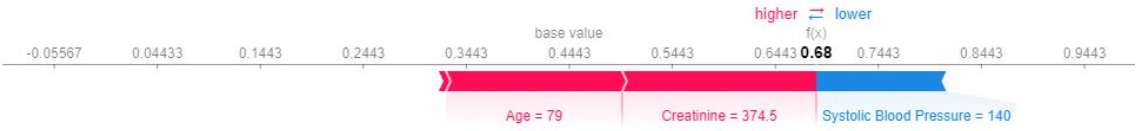

Figure 5.13: SHAP force plot for decision tree.

Several conclusions can be derived.

- **Base value:** The base value represents the average prediction for the whole dataset and for the decision tree model that is the probability of death of approximately 0.5.
- **Predicted Probability:** The predicted probability of belonging to class 1 for this patient is 0.68. Since it is higher than 0.5, we consider this patient to belong to class 1 (death).

- **Variable's Effects:** The conclusions for the variable's effects are the same as the ones drawn before.

5.3.5 Our Approach

In our proposed approach, we used several decision rules (mentioned and interpreted in part B.) by resorting to "virtual patients". Some considerations on the implementation of the clusters are presented, given that some variables had to be excluded. Furthermore, we used a personalization mechanism to give the probability that each rule is correct for each patient. An example for a patient is given in part B. We computed the mortality risk for each patient and transform it to obtain a binary outcome. Some considerations on that process are presented along with the results of the computation of the final step in our approach, the reliability measure.

Exclusion of the Troponin and STEMI Variables

The centers of the clusters obtained for the troponin and STEMI variables are represented in figure 5.14 and 5.15. As it can be seen, the centers of the clusters for the survival class correspond to the value 1 for troponin and STEMI and the centers of the clusters for the death class correspond to the value 0 for both variables. This would mean that the patients that survive have high cardiac biomarkers and a STEMI diagnosis. Since these conclusions are obviously in contradiction with the domain knowledge and with the points applied in the GRACE risk score, we decided to not use the STEMI and troponin variables. However, these results are consistent with prior conclusions obtained with our dataset, since from the results of table 5.3, we derived that the STEMI variable doesn't have an association with the outcome. Furthermore, in the logistic regression model, we also didn't consider this variable since it didn't have a statistically significant coefficient. In the naive Bayes model, the probabilities for each value of the STEMI variable (0 and 1) are close given either class (survival and death). Considering the troponin variable, as analyzed in table 5.5, there isn't a big difference between the mortality rates in each of the values of the variable.

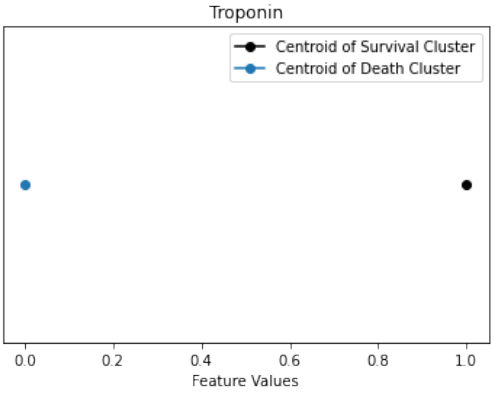

Figure 5.14: Means of the clusters for the troponin variable representing both patient's classes.

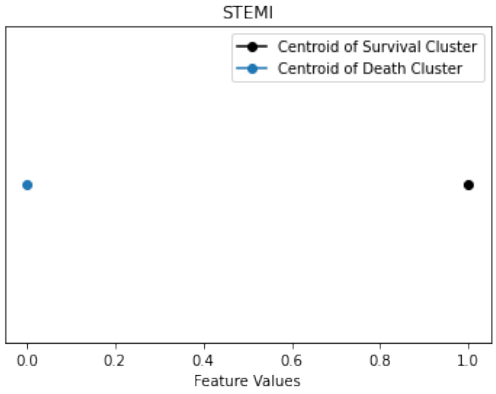

Figure 5.15: Means of the clusters for the STEMI variable representing both patient's classes.

Binarization of the mortality risk

To transform the mortality risk obtained for the patients into a class, we applied equation 4.6 with the threshold $P = 0.28$. This threshold was the one that allowed us to obtain better performance.

Reliability measure

We computed the reliability measure for each patient's outcome. Averaging the measure for the dataset we obtained a reliability of 57%. This value can be considered satisfactory. We decided to not use oversampling in our proposed approach given that when it was used, we obtained much lower values for the reliability measure (around 20%).

A. Performance Evaluation

The results concerning the performance evaluation for our proposed approach are displayed in table 5.30.

	Geometric Mean(%)	**Sensitivity(%)**	**Specificity(%)**
Train	75.57±0.04	74.34±0.09	76.82±0.03
Test	74.72±0.44	73.36±0.87	76.61±0.15

Table 5.30: Performance metrics for our proposed approach.

B. Intepretability of Our Approach

Original Set of Rules

In our approach, we used several decision rules, some of them were the ones used in the work of (Valente et al., 2022), namely rules 5.9 to 5.12. Furthermore, for the remaining variables, we obtained the decision rules ourselves, by resorting to k-means clustering following the methodology mentioned in chapter 2. After some tests, we concluded that the threshold $L = 0.5$ (following the notation of rule 4.3) was the one that improved the performance the most. The rules we obtained were transformed to the notation of equation 4.1.

$$\text{IF (Age}\geq 68 \text{ years) THEN } \hat{t_i} = 1 \tag{5.9}$$

$$\text{IF (Heart Rate}\geq 84 \text{ bpm) THEN } \hat{t_i} = 1 \tag{5.10}$$

$$\text{IF (Systolic Blood Pressure}\leq 135 \text{ mmHg) THEN } \hat{t_i} = 1 \tag{5.11}$$

$$\text{IF (Killip Class}\geq 2) \text{ THEN } \hat{t_i} = 1 \tag{5.12}$$

$$\text{IF (Cardiac Arrest}\geq 0.5) \text{ THEN } \hat{t_i} = 1 \tag{5.13}$$

$$\text{IF (Creatinine}\geq 392.49 \ \mu\text{mol/L) THEN } \hat{t_i} = 1 \tag{5.14}$$

The rules follow the same rationale applied by the GRACE risk score since more points are attributed, meaning a higher risk of mortality, to older patients and in equation 5.9 for the patients with more than 68 years, the rule suggests an output of 1 (death). The same happens for the heart rate (5.10), Killip class (5.12) and creatinine (5.14) variables, with higher values being attributed an output of 1 and in the GRACE risk score, high values on those variables are also associated with a higher risk of mortality. Furthermore, for the systolic blood pressure (5.11), both in our rule and in the clinical reference, we verify the opposite scenario, with lower values being attributed more points. Finally, for the cardiac arrest variable, in our rule (5.13), a value$\geq$0.5 (meaning a value of 1 since it is a binary variable) gets an output of 1. This rule is in accordance with the GRACE risk score since, in this variable, it only attributes points to patients if they are in cardiac arrest.

Rules Acceptance for a Patient

In table 5.31, are represented the feature values of a specific patient and the rules outputs obtained by applying rules 5.9 to 5.14. The predicted rules acceptance returned by our ML model as part of our personalization mechanism are also represented. The mortality risk of the patient was computed through equations

4.4 and 4.5, assuming an output of -1 for the negative rules instead of 0. Since this patient has a mortality risk of 0.36 and we considered a threshold of 0.28, we consider that this patient belongs to class 1 (death). The reliability measure was computed through equation 4.7 using the mean acceptance of the positive and negative rules.

Patient			
Feature	**Feature Value**	**Rule Output**	**Rule Acceptance**
Age	87 years	1	0.08
Heart Rate	98 bpm	1	0.27
Systolic Blood Pressure	120 mmHg	1	0.30
Creatinine	140.8 μmol/L	0	0.69
Killip Class	1	0	0.87
Cardiac Arrest	0	0	0.77
Average Acceptance of Positive Rules	0.22	**Average Acceptance of Negative Rules**	0.78
Mortality Risk	0.36	**Reliability**	0.56

Table 5.31: Representation of feature values, rule outputs, and predicted rules acceptance for a specific patient. Furthermore, we present the average acceptance of positive and negative rules, the mortality risk, and the reliability measure.

C. Functionally-Grounded Evaluation of Interpretability

The results of the functionally-grounded evaluation of interpretability for our approach are represented in table 5.32.

	Stability [-1,1]	**Geometric Mean Confidence Interval**	**Spearman Correlation [-1,1]**
Our Approach	0.506 ± 0.009	[68.9 % , 80.1%] (11.2%)	0.83

Table 5.32: Functionally-grounded evaluation of interpretability for our proposed approach.

Feature Importance

The features' rank by importance considering the Shapley values returned by SHAP is represented in table 5.33, for the GRACE risk score and for our proposed approach. The Spearman correlation between the 2 ranks was calculated and we obtained a value of 0.83 (represented in table 5.32).

Feature	Rank - Our Approach	Rank - GRACE Risk Score
Age	1	1
Heart Rate	4	5
Systolic Blood Pressure	3	3
Creatinine	6	4
Killip Class	2	2
Cardiac Arrest	5	6

Table 5.33: Ranks of the used features considering the features' importance returned by SHAP for our approach and for the GRACE risk score.

SHAP force plot

In figure 5.16, the force plot for an instance of the dataset is represented.

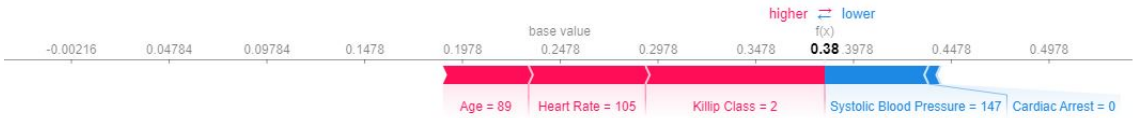

Figure 5.16: SHAP force plot for our proposed approach.

Several conclusions can be derived.

- **Base value:** The base value represents the average prediction for the whole dataset and for our approach that value is approximately the threshold considered to binarize the mortality risk (0.28). The average prediction is smaller in our approach (skewed to the survival side) than with the logistic regression, naive Bayes, and decision tree models since in this case we didn't use oversampling as we did in the remaining ML models. Therefore, we also use a different threshold than the standard 0.5 used by the other models.
- **Predicted Probability:** The mortality risk for this patient is 0.38. Since it is higher than 0.28, we consider this patient to belong to class 1 (death).
- **Variable's Effects:** The conclusions for the variable's effects are the same as the ones drawn before.

5.4 Results Overview

5.4.1 Performance Evaluation

Table 5.34 details the performance of the clinical reference and the ML models.

Model	Dataset Partition	Geometric Mean (%)	Sensitivity (%)	Specificity (%)
GRACE Risk Score	**Test**	72.11±0.07	84.02±0.07	62.10±0.00
Logistic Regression	**Train**	76.95±0.26	74.06±0.48	79.97±0.15
	Test	73.57±0.41	68.15±0.41	79.96±0.39
Naive Bayes	**Train**	75.81±0.20	73.57±0.44	78.13±0.24
	Test	74.39±0.33	71.57±0.70	77.94±0.25
Decision Tree	**Train**	78.13±0.28	80.68±2.15	76.00±2.24
	Test	71.98±1.26	70.81±1.60	73.96±2.15
Our Approach	**Train**	75.57±0.04	74.34±0.09	76.82±0.03
	Test	74.72±0.44	73.36±0.87	76.61±0.15

Table 5.34: Overview of the performance of the different models.

Regarding the geometric mean, both naive Bayes and our proposed approach have good results in that metric, in the train and test validations. However, we can consider that our approach has better general performance given that in the test validation, it has higher sensitivity (73.36±0.87) than the naive Bayes model (71.57±0.70) without a meaningful loss of specificity. We are interested in good results in the sensitivity metric since it allows us to understand the percentage of deceased patients that the model correctly predicted. In healthcare, we generally want to avoid FN more than FP, since we consider that is worse to don't have a diagnosis and be sick than to have a diagnosis and be healthy. Regarding the train validation, the naive Bayes model has very close sensitivity results (73.57± 0.44) to our approach (74.34±0.09). The difference in sensitivity is only verified for the test partition since, in the naive Bayes model, oversampling was used in the train partition, and in our approach, no oversampling was performed. Therefore, the naive Bayes model is trained with 55% of patients belonging to the survival class and 44% to the death class, and then, in the test partition, it only has about 14% of patients belonging to class death. The logistic regression and decision tree models have good results regarding the geometric mean metric in the train validation, 76.95±0.26, and 78.13±0.28, respectively. However, in the test validation, these values are smaller (73.57±0.41 and 71.98±1.26), due to the loss that the sensitivity metric suffers in

the test validation of these models, again due to oversampling being used in the train partition.

Regarding sensitivity, the GRACE risk score is the model that performs better (84.02±0.07), followed by our proposed approach (73.36±0.87). However, the GRACE risk score is also the model with poorer performance regarding specificity (62.10±0.00). This can be explained by the fact that with the clinical reference, most of the patients are identified as being high-risk patients (class death), however, some of them don't end up dying due to interventions by doctors. Therefore, the diagnosis of high risk lowers the mortality rate in those patients by them having access to early intervention.

Concerning specificity, the method that has the best results is the logistic regression model (79.96±0.39 in the test validation), with the naive Bayes and our proposed approach also obtaining good results, 77.94±0.25 and 76.61±0.15, respectively.

To conclude, for our dataset, the model that has the best performance is our approach, given that it is one of the models to have a higher geometric mean on the test validation. Furthermore, following the GRACE risk score it is the method that has higher sensitivity.

5.4.2 Interpretability: Functionally-Grounded Evaluation

Table 5.35 shows the overview of the quantitative evaluation of interpretability.

Model	Stability [-1,1]	Geometric Mean Confidence Interval	Spearman Correlation [-1,1]
GRACE Risk Score	0.506±0.006	[68.2 % , 76.6 %] (8.4%)	1
Logistic Regression	0.634±0.015	[67.3 % , 79.6 %] (12.3%)	0.66
Naive Bayes	0.606±0.008	[67.9 % , 79.2 %] (11.3%)	0.26
Decision Tree	0.648±0.029	[59.4 % , 77.1 %] (17.7%)	0.66
Our Approach	0.506±0.009	[68.9 % , 80.1 %] (11.2%)	0.83

Table 5.35: Overview of the functionally-grounded evaluation of the model's interpretability.

As discussed earlier, the stability measure must be evaluated with the geometric mean CI. The decision tree model is the one with greater stability (0.648 ± 0.029), however, it is also the one with the largest CI (17.7 %), so we can't be confident in its results. This conclusion conforms with the theoretical information regarding decision trees, which are mentioned to be unstable. Furthermore, the lower range of the geometric mean confidence interval is the lowest of all models (59.4%). That result is in accordance with the information displayed in table 5.34, since the decision tree was the model with the worst geometric mean performance on the test set. The GRACE risk score has the narrower confidence interval (8.4%) but it is also one of the models to have the worst stability (0.506 ± 0.006). This trade-off between the stability measure and the confidence interval on the geometric mean can be explained because when computing the stability measure for a single patient if the 3 neighbors used in the computation have the same label as the patient in question, the stability measure is almost 1. Therefore, the overall stability of the model (average of each patient's stability) also improves. If the patients close to the decision boundary of the classifier have a stability of 1, this would mean that the classifier has a poor ability to discriminate between classes (and therefore a worst geometric mean result). Having this in mind, the naive Bayes model is the ML model which allows for a better compromise between the stability and geometric mean CI measures, with a stability of 0.606± 0.008 and a relatively narrow CI (11.3%).

However, the naive Bayes model is the worst performing model regarding the comparison between the importance it gives to each feature with the importance the GRACE risk score gives (negligible correlation of 0.26). Regarding the similarity with the GRACE risk score features' rank by importance, our approach is the best model with a high correlation of 0.83. The age of the patient is the most important feature in both the clinical reference and our model. The features' rank considering the logistic regression and decision tree models has a moderate correlation of 0.66 with the features' rank considered by the clinical reference. However, it is important to consider, that besides obtaining the same value for the Spearman correlation, the logistic regression model has higher quality in the explanations provided than the decision tree model. Firstly, the logistic regression model considers the age variable to be the most important like the clinical reference does, while the decision tree considers the creatinine feature to be the most relevant. Furthermore, as discussed in part B. of section 5.3.4, some of the conditions in decision rules returned by the decision tree model don't follow the same rationale applied in the GRACE risk score.

In summary, our proposed approach is the model that offers the best compromise between the 3 proxies used to evaluate quantitatively the

interpretability of the models, having the best correlation with the GRACE risk score regarding the importance of the features. Furthermore, it has the same stability as the GRACE risk score (0.506) and a geometric mean CI with the same width as the confidence interval for the naive Bayes model.

6

Conclusions

The risk stratification of ACS patients is done by using risk scores, for example, the GRACE score. Due to the limitations of risk scores and the recent research progress and good performance results of ML, our goal was to develop Machine Learning models to predict the 6-month mortality of ACS patients. Our implemented models need to be interpretable since, in critical domains like healthcare, the models need to provide explanations for their predictions in order to be accepted by physicians. We analyzed the explanations provided by each implemented model and concluded that for most models, they follow the rationale of the GRACE risk score. However, the decision tree model had some conditions in some rules that aren't aligned with the clinical knowledge.

In the literature, it is recurrently mentioned the necessity of measures to quantify interpretability. We started by analyzing the existent state-of-the-art techniques and concluded that interpretability can be quantitatively evaluated by using a set of proxies. One of those consists in the quality of explanations, that we propose to be evaluated considering the domain knowledge, e.g. our clinical reference (the GRACE risk score). We computed for each model, the features' importances using kernel SHAP and ranked the features by importance. That rank is compared with the rank attributed by the GRACE risk score using the Spearman correlation. We conclude that this metric can be used to evaluate models regarding the quality of explanations. Furthermore, it is important to highlight that, although the Spearman correlation can be used to have a simplified quantitative measure, the features' rank should always be checked and compared (since two models may have the same Spearman correlation but the most important feature in each model can be different). The measures of stability and confidence can also be used in future works, however, they should be analyzed together since there is a trade-off between them. Although we used the 95% geometric mean CI for the confidence measure, other confidence intervals (for example, the 99% CI) on other metrics (for example, accuracy) can be adopted.

Furthermore, we proposed a hybrid approach to solve besides interpretability, the personalization issue of ML models. First, we create decision rules for all the patients using the ACS risk factors and resorting to clustering. In a second step, we train an ML classifier to predict the probability that each rule is correct for each patient. Therefore, we mimic the actions a clinician would take when doing a prognostic. Our proposed approach is the model with the best performance with a geometric mean of 74.72%, a sensitivity of 73.36%, and a specificity of 76.61% in the test validation. Regarding the evaluation of interpretability, our method has the highest correlation with the features' rank attributed by the GRACE risk score (0.83). Besides that, it has the same stability as the clinical reference (0.506) and a relatively narrow CI (11.2 %). Furthermore, besides being an interpretable approach and addressing the personalization issue, it provides the reliability of the estimated mortality risk. In summary, the results suggest that our proposed approach has the potential to be used in a real-life clinical scenario given its performance (better than the currently used risk score) and interpretability (similar results to the currently used risk score). We consider that our method can contribute to assist the decision-making process of clinicians in a trustful manner.

In future work, developing a measure of simplicity would be interesting, as it is one of the properties mentioned in the literature as imperative for a model to be interpretable. Furthermore, our proposed approach can be improved, namely the reliability measure step. The proposed ML approach in this work and the interpretability measures can be further validated in other datasets with larger and more heterogeneous patient information since the data used in this work is reflective of a small Portuguese sample size. Finally, in order to analyze how it performs, our proposed methodology can be applied to other healthcare problems and even, other domains.

Bibliography

Abedin Babak. Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective // Internet Research. 3 2022. 32. 425–453.

Ahmad Muhammad Aurangzeb, Eckert Carly, Teredesai Ankur. Interpretable machine learning in healthcare // Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018. 559–560.

Alivecor . What is an ECG? 2022. Retrieved June 2022, from `https://www.alivecor.com/education/ecg.html`.

American Heart Association . Understanding blood pressure readings. n.d. Retrieved June 2022, from `https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings`.

Analytics Vidhya . Overcoming class imbalance problem using SMOTE. Oct 2020. Retrieved July 2022, from `https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/`.

Anderson Richard P., Jin Ruyun, Grunkemeier Gary L. Understanding logistic regression analysis in clinical reports: an introduction // The Annals of thoracic surgery. 3 2003. 75. 753–757.

Antman E. M., Cohen M., Bernink P. J.L.M., McCabe C. H., 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 // Journal of the American Medical Association. 8 2000. 284. 835–842.

Aras Arda Can. Explaining what learned models predict: In which cases can we trust machine learning models and when is caution required? Ankara, Turkey,

2020. Bilkent University.

Araújo Gonçalves Pedro de, Ferreira Jorge, Aguiar Carlos, Seabra-Gomes Ricardo. TIMI, PURSUIT, and GRACE risk scores: sustained prognostic value and interaction with revascularization in NSTE-ACS // European heart journal. 2005. 26, 9. 865–872.

Arya Vijay, Bellamy Rachel KE, Chen Pin-Yu, Dhurandhar Amit, Hind Michael, Hoffman Samuel C, Houde Stephanie, Liao Q Vera, Luss Ronny, Mojsilović Aleksandra, others . One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques // arXiv preprint arXiv:1909.03012. 2019.

Boersma Eric, Pieper Karen S., Steyerberg Ewout W., Wilcox Robert G., Chang Wei Ching, Lee Kerry L., Akkerhuis K. Martijn, Harrington Robert A., Deckers Jaap W., Armstrong Paul W., Lincoff A. Michael, Califf Robert M., Topol Eric J., Simoons Maarten L. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. The PURSUIT Investigators // Circulation. 6 2000. 101. 2557–2567.

British Heart Foundation . Cardiac arrest. 2021. Retrieved June 2022, from `https://www.bhf.org.uk/informationsupport/conditions/cardiac-arrest`.

Brownlee Jason. How to calculate bootstrap confidence intervals for machine learning results in python. Jun 2017. Retrieved July 2022, from `https://machinelearningmastery.com/calculate-bootstrap-confidence-intervals-machine-learning-results-python/`.

Brownlee Jason. Confidence intervals for machine learning. May 2018. Retrieved July 2022, from `https://machinelearningmastery.com/confidence-intervals-for-machine-learning/`.

Brownlee Jason. Tour of evaluation metrics for imbalanced classification. 2020. Retrieved July 2022, from `https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/`.

Burkart Nadia, Huber Marco F. A survey on the explainability of supervised machine learning // Journal of Artificial Intelligence Research. 1 2021. 70. 245–317.

CanadiEM . Sirens to scrubs: Acute coronary syndromes, part one - Beyond door-to-balloon. 2018. Retrieved June 2022, from `https://canadiem.org/acute-coronary-syndrome-beyond-door-to-balloon/`.

Carrington André, Fieguth Paul, Chen Helen. Measures of model interpretability for model selection // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018. 11015 LNCS. 329–349.

Carvalho Diogo V., Pereira Eduardo M., Cardoso Jaime S. Machine learning interpretability: A survey on methods and metrics // Electronics 2019, Vol. 8, Page 832. 7 2019. 8. 832.

Chawla Nitesh V., Bowyer Kevin W., Hall Lawrence O., Kegelmeyer W. Philip. SMOTE: Synthetic minority over-sampling technique // Journal of Artificial Intelligence Research. 2002. 16. 321–357.

Choueiry George. Interpret logistic regression coefficients – Quantifying health. n.d. Retrieved January 2022, from `https://quantifyinghealth.com/interpret-logistic-regression-coefficients/`.

Cleveland Clinic . What to know about your heart rate and pulse. n.d. Retrieved June 2022, from `https://my.clevelandclinic.org/health/diagnostics/17402-pulse--heart-rate`.

Delua Julianna. Supervised vs. unsupervised learning: what's the difference? Mar 2021. Retrieved July 2022, from `https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning`.

Domingos Pedro. A few useful things to know about machine learning // Commun. ACM. 10 2012. 55. 78–87.

Doshi-Velez Finale, Kim Been. Towards a rigorous science of interpretable machine learning // arXiv preprint arXiv:1702.08608. 2017.

Elbarouni Basem, Goodman Shaun G., Yan Raymond T., Welsh Robert C., Kornder Jan M., DeYoung J. Paul, Wong Graham C., Rose Barry, Grondin François R., Gallo Richard, Tan Mary, Casanova Amparo, Eagle Kim A., Yan Andrew T. Validation of the Global Registry of Acute Coronary Event (GRACE) risk score for in-hospital mortality in patients with acute coronary syndrome in Canada // American Heart Journal. 9 2009. 158. 392–399.

Folkman Tyler. How to add confidence intervals to any model. Nov 2019. Retrieved July 2022, from `https://towardsdatascience.com/how-to-add-confidence-intervals-to-any-model-7bbb9f80fd9c`.

Freitas Alex A., Wieser Daniela C., Apweiler Rolf. On the importance of comprehensible classification models for protein function prediction // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 1 2010. 7. 172–182.

Galarnyk Michael. Understanding decision trees for classification (python). Jul 2019. Retrieved July 2022, from `https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952`.

Gjesdal Grunde, Braun Oscar, Smith J. Gustav, Scherstén Fredrik, Tydén Patrik. Blood lactate is a predictor of short-term mortality in patients with myocardial infarction complicated by heart failure but without cardiogenic shock // BMC Cardiovascular Disorders. 1 2018. 18.

Gonçalves Ana Margarida Lopes. Regressão logística aplicada à pesquisa de preditores de morte. Coimbra, Portugal, 9 2013. University of Coimbra.

Granger Christopher B., Goldberg Robert J., Dabbous Omar, Pieper Karen S., Eagle Kim A., Cannon Christopher P., Werf Frans V. Van de, Avezum Álvaro, Goodman Shaun G., Flather Marcus D., Fox Keith A.A. Predictors of hospital mortality in the global registry of acute coronary events // Archives of internal medicine. 10 2003. 163. 2345–2353.

Han Jiawei, Kamber Micheline, Pei Jian. Data mining : Concepts and techniques. 2011. 3rd ed. Elsevier.

Hashmi Kashif A, Adnan Fahar, Ahmed Omer, Yaqeen Syed Rafay, Ali Javaria, Irfan Muhammad, Edhi Muhammad M, Hashmi Atif A. Risk assessment of patients after ST-segment elevation myocardial infarction by Killip Classification: an institutional experience // Cureus. 2020. 12, 12.

Hindle Bethan J., Childs Philip H. Warren Dylan Z. APS 240: Data analysis and statistics with R. 2021. Available at `https://dzchilds.github.io/stats-for-bio/`.

Hussain Hussain, Fadel Aya. Malignant hypertension without end-organ damage secondary to stressful condition in a Female // Cureus. Aug 2020. 12, 8.

Hwang Calvin, Levis Joel T. ECG diagnosis: ST-elevation myocardial infarction // The Permanente journal. 3 2014. 18. e133.

IBM Cloud Education . What is machine learning? Jul 2020. Retrieved July 2022, from `https://www.ibm.com/cloud/learn/machine-learning`.

Ille Tatjana, Milic Natasa. Encyclopedia of public health. 2008. 1st ed. Springer Dordrecht.

Imbalanced-learn . User guide. 2014. Retrieved July 2022, from `https://imbalanced-learn.org/dev/introduction.html`.

International Diabetes Federation . Cardiovascular disease. 2021. Retrieved June 2022, from `https://idf.org/our-activities/care-prevention/cardiovascular-disease.html`.

Italo José. KNN (K-Nearest Neighbors) #1. Jul 2018. Retrieved July 2022, from `https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e`.

Jansen Stefan. Hands-on machine learning for algorithmic trading : design and implement investment strategies based on smart algorithms that learn from data using Python. Birmingham, United Kingdom: Packt Publishing, 2018.

Lahav Owen, Mastronarde Nicholas, Schaar Mihaela van der. What is interpretable? using machine learning to design interpretable decision-support systems // arXiv preprint arXiv:1811.10799. 2018.

Liao T Warren. Clustering of time series data-a survey // Pattern Recognition. 2005. 38. 1857–1874.

Linardatos Pantelis, Papastefanopoulos Vasilis, Kotsiantis Sotiris. Explainable ai: A review of machine learning interpretability methods // Entropy. 2020. 23, 1. 18.

López Fernando. SHAP: Shapley Additive Explanations. Jul 2021. Retrieved July 2022, from at `https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3`.

Margot Vincent, Luta George. A new method to compare the interpretability of rule-based algorithms // AI. 2021. 2, 4. 621–635.

Mayo Clinic . Acute coronary syndrome. 2021a. Retrieved June 2022, from `https://www.mayoclinic.org/diseases-conditions/acute-coronary-syndrome/symptoms-causes/syc-20352136`.

Mayo Clinic . Arteriosclerosis / atherosclerosis. 2021b. Retrieved June 2022, from https://www.mayoclinic.org/diseases-conditions/arteriosclerosis-atherosclerosis/symptoms-causes/syc-20350569.

Mayo Clinic . Myocardial ischemia - Symptoms and causes. n.d. Retrieved June 2022, from https://www.mayoclinic.org/diseases-conditions/myocardial-ischemia/symptoms-causes/syc-20375417.

Medical News Today . Normal troponin levels: Healthy ranges and what high levels mean. n.d. Retrieved June 2022, from https://www.medicalnewstoday.com/articles/325415.

MedlinePlus . Troponin test. n.d. Retrieved June 2022, from https://medlineplus.gov/lab-tests/troponin-test/.

Meek S. ABC of clinical electrocardiography: Introduction. I—Leads, rate, rhythm, and cardiac axis // BMJ. Feb 2002. 324, 7334. 415–418.

Menard Scott. Standards for standardized logistic regression coefficients // Social Forces. 6 2011. 89. 1409–1428.

Molnar Christoph. Interpretable machine learning: A guide for making black box models explainable. 2022. 2nd ed. Independently published.

Molnar Christoph, Casalicchio Giuseppe, Bischl Bernd. Interpretable machine learning – A brief history, atate-of-the-art and challenges // Communications in Computer and Information Science. 2020. 1323. 417–431.

Mukaka M. M. A guide to appropriate use of correlation coefficient in medical research // Malawi Medical Journal. 2012. 24. 69.

Murdoch W. James, Singh Chandan, Kumbier Karl, Abbasi-Asl Reza, Yu Bin. Definitions, methods, and applications in interpretable machine learning // Proceedings of the National Academy of Sciences. 10 2019. 116. 22071–22080.

Müller Andreas C. Data splitting strategies — Applied machine learning in python. 2020. Retrieved June 2022, from https://amueller.github.io/aml/04-model-evaluation/1-data-splitting-strategies.html.

Nandeshwar Ashutosh R. Models for calculating confidence intervals for neural networks. Morgantown, Virginia, United States of America, 5 2006. West Virginia University.

Nguyen An-phi, Martínez María Rodríguez. On quantitative aspects of model interpretability // arXiv preprint arXiv:2007.07584. 2020.

Panch Trishan, Szolovits Peter, Atun Rifat. Artificial intelligence, machine learning and health systems // Journal of Global Health. 2018. 8.

Pandas Documentation . 2022. Retrieved July 2022, from `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html`.

Prawtama Kelvin. The tale of probability: Naive Bayes. Oct 2021. Retrieved August 2022, from `https://medium.com/@kdatainc/the-tale-of-probability-naive-bayes-da91d732da52`.

Qayyum Adnan, Qadir Junaid, Bilal Muhammad, Al-Fuqaha Ala. Secure and robust machine learning for healthcare: A survey // IEEE Reviews in Biomedical Engineering. 1 2020. 14. 156–180.

Raschka Sebastian. STAT 451: Introduction to machine learning lecture notes. Wisconsin, United States of America, 2020. University of Wisconsin–Madison.

República Portuguesa . AI Portugal 2030. 2022. Retrieved July 2022, from `https://www.portugal.gov.pt/pt/gc23`.

Sampson Michael, McGrath Anthony. Understanding the ECG. Part 1: Anatomy and physiology // British Journal of Cardiac Nursing. 11 2015. 10. 548–554.

Schmidt Philipp, Biessmann Felix. Quantifying interpretability and trust in machine learning systems // arXiv preprint arXiv:1901.08558. 2019.

SciPy . 2022a. Retrieved July 2022, from `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html`.

SciPy . 2022b. Retrieved July 2022, from `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html`.

SciPy . 2022c. Retrieved July 2022, from `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html`.

Scikit-learn . 3.1. Cross-validation: evaluating estimator performance. 2013. Retrieved July 2022, from `https://scikit-learn.org/stable/modules/cross_validation.html#`.

Scikit-learn . 2.3. Clustering. 2019. Retrieved July 2022, from `https://scikit-learn.org/stable/modules/clustering.html#k-means`.

Scikit-learn . 1.6. Nearest neighbors. 2022. Retrieved July 2022, from https://scikit-learn.org/stable/modules/neighbors.html#.

Scikit-learn . Naive Bayes. n.d.-a. Retrieved July 2022, from https://scikit-learn.org/stable/modules/naive_bayes.html.

Scikit-learn . Decision trees. n.d.-b. Retrieved July 2022, from https://scikit-learn.org/stable/modules/tree.html#tree.

Sheskin David J. Handbook of parametric and nonparametric statistical procedures. Boca Raton, Florida, United States of America: Chapman and Hall/CRC, 6 2020.

Smith Jennifer N., Negrelli Jenna M., Manek Megha B., Hawes Emily M., Viera Anthony J. Diagnosis and management of acute coronary syndrome: an evidence-based update // Journal of the American Board of Family Medicine : JABFM. 3 2015. 28. 283–293.

StatsTest.com . Chi-Square test of independence - StatsTest.com. Oct 2020a. Retrieved July 2022, from https://www.statstest.com/chi-square-test-of-independence/.

StatsTest.com . Phi coefficient - StatsTest.com. Apr 2020b. Retrieved July 2022, from https://www.statstest.com/phi-coefficient/.

StatsTest.com . Spearman's rho - StatsTest.com. Apr 2020c. Retrieved July 2022, from https://www.statstest.com/spearmans-rho/.

Sunasra Mohammed. Performance metrics for classification problems in machine learning. 2017. Retrieved July 2022, from https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.

UCSF Health . Creatinine blood test. n.d.a. Retrieved June 2022, from https://www.ucsfhealth.org/medical-tests/creatinine-blood-test.

UCSF Health . Troponin test. n.d.b. Retrieved June 2022, from https://www.ucsfhealth.org/medical-tests/troponin-test.

Valente Francisco, Henriques Jorge, Paredes Simão, Rocha Teresa, Carvalho Paulo de, Morais João. Improving the compromise between accuracy, interpretability and personalization of rule-based machine learning in medical problems // 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). 2021a. 2132–2135.

Valente Francisco, Henriques Jorge, Paredes Simão, Rocha Teresa, Carvalho Paulo de, Morais João. A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario // Artificial Intelligence in Medicine. 7 2021b. 117. 102113.

Valente Francisco, Paredes Simão, Henriques Jorge, Rocha Teresa, Carvalho Paulo de, Morais João. Interpretability, personalization and reliability of a machine learning based clinical decision support system // Data Mining and Knowledge Discovery. 2022. 36, 3. 1140–1173.

Viera Anthony J, Sheridan Stacey L. Global risk of coronary heart disease: assessment and application // American family physician. 2010. 82, 3. 265–274.

Waa Jasper Van Der, Diggelen Jurriaan Van, Neerincx Mark A, Raaijmakers Stephan. ICM: an intuitive model independent and accurate certainty measure fovan derne learning. // ICAART (2). 2018a. 314–321.

Waa Jasper Van Der, Diggelen Van Jurriaan, Neerincx Mark. The design and validation of an intuitive confidence measure // Workshop On Explainable Smart Systems (EXSS). 2. 2018b. 1.

Waa Jasper Van Der, Schoonderwoerd Tjeerd, Diggelen Jurriaan Van, Neerincx Mark. Interpretable confidence measures for decision support systems // International Journal of Human-Computer Studies. 2020. 144. 102493.

Weiss Gary M. Cost-Sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? // Proceedings of the 2007 International Conference on Data Mining. 2007.

Weng Stephen F., Reps Jenna, Kai Joe, Garibaldi Jonathan M., Qureshi Nadeem. Can machine-learning improve cardiovascular risk prediction using routine clinical data? // PloS one. 4 2017. 12.

World Health Organization . Cardiovascular diseases (CVDs). n.d. Retrieved June 2022, from `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`.

XLSTAT . What is a statistical test? 2022. Retrieved July 2022, from at `https://help.xlstat.com/6758-what-statistical-test`.

Zhang Jesse. Estimating confidence intervals on accuracy in classification in machine learning. Alaska, United States of America, 2019. University of Alaska Fairbanks.

Zhang Junfeng, Chen Wei, Gao Mingyi, Shen Gangxiang. K-means-clustering-based fiber nonlinearity equalization techniques for 64-QAM coherent optical communication system // Optics Express. 10 2017. 25. 27570.

Zhou Jianlong, Gandomi Amir H., Chen Fang, Holzinger Andreas. Evaluating the quality of machine learning explanations: A survey on methods and metrics // Electronics 2021, Vol. 10, Page 593. 3 2021. 10. 593.

Appendices

This page is intentionally left blank.

A

SHAP Summary Plots

This appendix details the SHAP summary plots (figures A.1 to A.4) for the implemented ML models. The features' effects are the same ones reported for the GRACE risk score in figure 5.9. The conclusions concerning the features' importance are the same as those discussed in section 5.3 for all models except for the decision tree model.

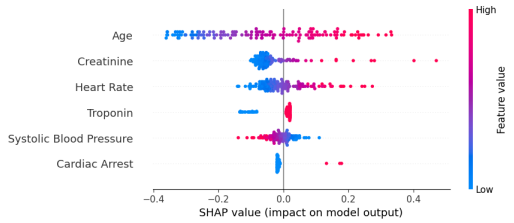

Figure A.1: SHAP summary plot for logistic regression.

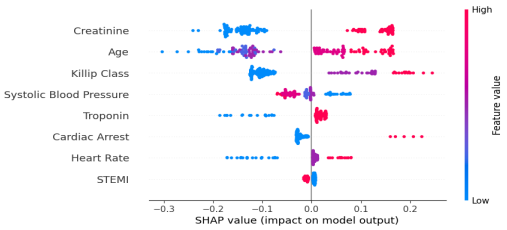

Figure A.2: SHAP summary plot for naive Bayes.

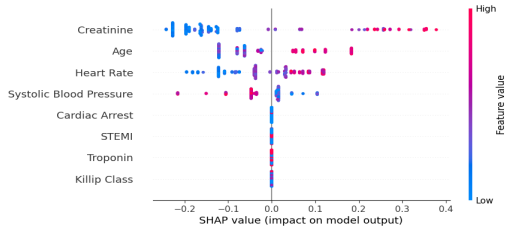

Figure A.3: SHAP summary plot for the decision tree model.

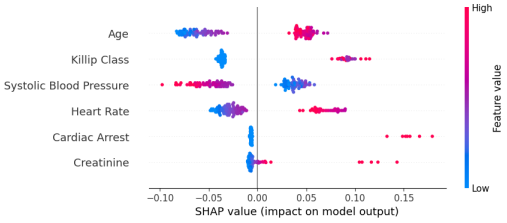

Figure A.4: SHAP summary plot for our proposed approach.

Regarding the decision tree model, the order of importance of the features isn't the same as the one verified in table 5.29, which can be explained by the fact that the SHAP summary plot was obtained by using a test partition and the values on table 5.29 were obtained by averaging the features' importance on the 10 test partitions obtained with a run of stratified k-fold. We can verify in the summary plot that the features STEMI, cardiac arrest, troponin, and Killip class have null or almost null importance. That result is in accordance with previous conclusions, namely when

analysing the features' importance for the decision tree of figure 5.12 (registered in table 5.27). Furthermore, we didn't consider the STEMI and cardiac arrest variables for the features' rank (registered in table 5.29) since they have null Shapley values.