National College of Ireland

BSc (Honours) in Computing – Data Analytics

2021/2022

Sophia Hennouni

18102166

x18102166@student.ncirl.ie

Machine Learning Models to Predict in-Hospital Mortality of Heart Failure Patients in Intensive Care Units

Technical Report

Contents

Acronyms

AUC	Area Under the Curve
EDA	Exploratory Data Analysis
EHR	Electronic Health Record
GWTG-HF	Get With The Guidelines-Heart Failure
HF	Heart Failure
ICU	Intensive Care Unit
KDD	Knowledge Discovery in Databases
K-NN	K-Nearest Neighbour
MIMIC	Medical Information Mart for Intensive Care
PCA	Principal Component Analysis
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines
XGBOOST	Extreme Gradient Boosting

Executive Summary

Heart failure (HF) is a highly prevalent disorder worldwide but assessing the mortality risk of HF patients remains a challenge as the nature of the predictors is still poorly understood.

This project aimed to apply machine learning techniques to Electronic Health Records (EHRs) to develop a prediction model for in-hospital mortality of HF patients admitted to Intensive Care Unit (ICU). The main objective of the model was to determine whether a machine learning approach could improve the reliability of well-established scoring systems such as the *Get With The Guidelines-Heart Failure* (GWTG-HF).

The data used to train and validate the prediction model was from the Medical Information Mart for Intensive Care (MIMIC) III database. The analysis was carried out on a cohort of 1,177 ICU patients and the in-hospital mortality rate was 13.52%.

The features selection phase followed three approaches: correlation analysis, Principal Component Analysis (PCA) and a decision tree algorithm. The dataset was divided into a training set (70%) and a testing set (30%). Eight machine learning techniques were applied to the three sets of features: k-means clustering, k-nearest neighbour (K-NN), logistic regression, naïve Bayes, support vector machines (SVM), random forests, extreme gradient boosting ensembles (XGBoost) and repeated incremental pruning to produce error reduction (RIPPER). The area under the ROC curve (AUC), recall and F1 scores were computed for each set within each technique using stratified 10-fold cross validation, and the best performing set-model association was chosen.

The best performance was obtained with logistic regression on the features selected by the decision tree. The AUC score was 0.7625, with was marginally lower than the GWTG-HF scoring system (AUC = 0.7743). We can thus say the results developed in this research are largely comparable to existing gold standards for heart failure evaluation. Moreover, anion gap was consistently identified as a feature of importance, but it is not part of the GWTG-HF calculator. A potentially good attribute has indeed been identified.

Overall, machine learning models predicted mortality results very close to a well-established scoring system. Finally, with the newly identified variable (anion gap), existing models like GWTG-HF could also improve their prediction accuracy.

1. Introduction

1.1. Background

Heart Failure (HF) is a complex clinical syndrome in which the heart cannot pump enough blood into the body. It is caused by a structural or functional cardiac abnormality that prevents the ventricles to fill with or eject blood to the systemic circulation. This failure to meet the metabolic requirements is characterised by fatigue, shortness of breath (dyspnea) and signs of fluid retention such as pulmonary congestion (pulmonary edema) or ankle swelling (peripheral edema). HF is a highly prevalent disorder

affecting an estimated 26 million people worldwide, and it is associated with high morbidity and mortality rates (Malik *et al.*, 2022). As a life-threatening syndrome, HF is treated in priority with the advanced technology and care provided by Intensive Care Units (ICUs).

The European Society of Cardiology have pointed out the shortcomings of the existing prognosis markers of death for HF patients, which have a limited clinical applicability and remain imprecise (Ponikowski *et al.*, 2016). However, an accurate mortality prognosis is critical in identifying HF patients with an increased risk of poor outcome, for the ICU staff to provide them with more intensive treatment and closer follow-up. Examples of these existing prognosis tools are the Meta-Analysis Global Group in Chronic heart failure (MAGGIC), which assigns a risk score based on the mortality prediction at one and three years (Pocock *et al.*, 2013), and the American Heart Association's Get With The Guidelines-Heart Failure (GWTG-HF), which predicts in-hospital all-cause heart failure mortality (Peterson *et al.*, 2010). Both online risk calculators are shown in Appendix (sections 10.3. and 10.4.). Relying on traditional statistical methods, MAGGIC was created using multivariable piecewise Poisson regression (Pocock *et al.*, 2013), whereas GWTG-HF used multivariable logistic regression (Peterson *et al.*, 2010).

As healthcare institutions have become more data-driven, the availability of patients' digitalised health data, which are referred to as Electronic Health Records (EHRs), has increased the opportunities for clinical research (Sarwar et al., 2022). For instance, the number of hospitals equipped with digital systems went from 9.4% to 75.5% in seven years between 2008 and 2014 in the United States (Johnson *et al.*, 2016). In the meantime, the challenges of dealing with large amounts of complex and oftentimes sparse data can be addressed more effectively by machine learning methods, compared to the traditional analytics methods used for outcomes research, health economics, and epidemiology (Crown, 2015). In their review of 20 studies predicting the prognosis of hospitalised HF patients, Shin et al. have also found that machine learning methods were performing better than conventional statistical models (Shin *et al.*, 2021).

Therefore, the increased availability of EHRs and the exponential development of machine learning represent a major opportunity for data scientists to understand clinical data and use the discovered knowledge to provide decision-making support to healthcare professionals and improve patient care and outcome. The intent of this research was to challenge an existing system based on a traditional statistical analysis with a machine learning approach to demonstrate the potential of data science in EHR-based clinical research and in the general improvement of healthcare practice.

1.2. Aims

The goal of this project was to use EHRs to develop a machine learning model that could predict the in-hospital mortality of ICU patients admitted for heart failure. The main objective of the model was to determine whether a machine learning approach could improve the reliability of well-established scoring systems such as the GWTG-HF. GWTG-HF was the example of choice for this project, as all its parameters were also attributes from the dataset used to build the predictive model. Furthermore, the literature survey provided a performance score for the predictions made by the tool on the same data, enabling performance comparisons between the two models.

Besides, the literature survey highlighted some findings from previous research on the same data that provided a different set of features of similar size to the GWTG-HF scoring system. Both sets were therefore compared and contrasted with the important variables detected by the model developed as part of this project. If the machine learning-based model was to outperform the GWTG-HF scoring system, a new set of parameters could be proposed to increase the accuracy of the risk prediction, and therefore better support clinical decisions for an improved patient care and outcome.

1.3. Technology and Techniques

The datasets were imported and merged in R Studio. A few housekeeping steps were taken to get the data ready for exploration and save it in a comma-separated values file. Microsoft Power BI was used to perform the Exploratory Data Analysis (EDA). After importing the file, the data was manipulated using Power Query and the M language in the advanced editor to build a small relational schema that would allow the creation of visuals in Power BI Desktop. The necessary measures were created with Power BI's Data Analysis Expressions (DAX) for a full control over the calculation steps. The analysis and modelling were performed back in R Studio, and a variety of packages were used for the features analysis and to run and evaluate the performance of the algorithms. To a lesser extent, Microsoft Excel was used for quick and simple data manipulation tasks such as filtering, sorting or pivoting values to validate some of the commands executed in R.

In terms of data mining and machine learning techniques, three techniques were used for features selection and the models were built from eight algorithms. The features were selected through correlation analysis, principal component analysis and Classification And Regression Trees (CART). To build the models, k-means clustering was the only unsupervised machine learning approach tested, and the seven other techniques were supervised: k-nearest neighbour, logistic regression, naïve Bayes, support vector machines, random forests, extreme gradient boosting ensembles and repeated incremental pruning to produce error reduction. The statistical measures chosen to evaluate the models were correlation matrices indicators such as sensitivity and F1 scores, as well as the measure of the Area Under the Receiver Operating Characteristic (AUROC) curve and lastly the ANalysis Of VAriance (ANOVA) chi-squared test.

1.4. Structure

The next section of the report provides a detailed description of the data used for this project, along with a description of the main insights found during the EDA. Then, the methodology section will review two papers based on similar research, present the data science process model chosen approach, describe the pre-processing tasks carried out to prepare the dataset for modelling, and detail the modelling strategy that will be followed. The Analysis section is mainly focused on the features selection phase and leads to the Modelling section, which includes a full description of all the techniques and algorithms used in the project. The model selection and performance are presented in the Evaluation section, which then brings the report to its Conclusion and Discussion sections.

2. Data

2.1. Datasets

2.1.1. Heart Failures Data

The Medical Information Mart for Intensive Care (MIMIC) III database contains thousands of anonymised medical records of patients who were admitted to the Beth Israel Deaconess Medical Center Critical Care Unit (Boston MA, USA) between 2001 and 2012. Johnson et al. pointed out the uniqueness of this large and freely accessible critical care database (Johnson *et al.*, 2016), which offers multiple opportunities to carry out some research work on health matters that could never be explored before.

The MIMIC-III Heart Failures dataset has been made publicly available by a team of Chinese researchers from the Departments of Cardiology of the Qingdao University and the Fudan University of Shanghai (F. Li *et al.*, 2021). To gather these data, Li et al. identified patients aged 15 years old or more who were diagnosed a heart failure, which was a total of 13,389 records. From this subset, 162 non-ICU patients were excluded, as well as 7,179 patients with no record of Left Ventricular Ejection Fraction (LVEF) and 4,871 patients missing record of N-Terminal pro-Brain Natriuretic Peptide (NT-proBNP), which left them with a final dataset of 1,177 patient records.

The researchers then used Structured Query Language queries (PostgreSQL) to extract various demographic characteristics, vital signs, and laboratory variables, which were selected based on their domain knowledge, previous studies and availability in the database. Variables that contained multiple measurements were included as calculated means and the primary outcome was the vital status at hospital discharge time.

The MIMIC-III Heart Failures dataset, see Table 1, is the main data source of this project. All the attributes are listed in Appendix (section 10.5.) and the technical characteristics of the dataset are summarized in the table below.

URL	https://doi.org/10.5061/dryad.0p2ngf1zd
Format	Comma-separated values
Data structure	Structured
Size in MB	0.380
Number of instances	1,177
Number of attributes	51
Types of attributes	Categorical, Discrete, Continuous, Boolean

Table 1 Heart Failures Dataset Characteristics

2.1.2. Patients Data

The second dataset contains the Patients admission records and was found in a separate Kaggle project from Alexander Scarlat, “MIMIC3d aggregated data – ICU aggregated data as number of interactions between patient and hospital” (*MIMIC3d aggregated data*, 2019). This dataset is also a

subset of the MIMIC-III database, and it includes some additional information regarding the patients, such as age and ethnicity details. Its main characteristics are presented in the Table 2, shown below.

URL	https://www.kaggle.com/drscarlat/mimic3d
Format	Comma-separated values
Data structure	Structured
Size in MB	11.622
Number of instances	58,976
Number of attributes	28
Types of attributes	Categorical, Discrete, Continuous

Table 2 Patients Dataset Characteristics

The Heart Failures and Patients datasets could be combined in one single dataset using the Patient ID. Indeed, this attribute had distinct values in both files and every Patient ID from the Heart Failures dataset could be found in the Patients dataset, which confirmed that both tables were linked by this key field in the MIMIC-III relational database.

2.2. Exploratory Data Analysis

Once the raw data source files imported and merged in R Studio, both datasets were merged and three attributes from the Patient dataset were added to the main Heart Failure dataset features: gender, age and ethnicity. The gender was a duplication of a variable already included and could be dropped, but it did give useful information on the sex encoding. Age was not a duplication, but it was confirmed to be the age at heart failure, as opposed to the age at first admission to the hospital from the Patient data, which was dropped. Lastly, ethnicity was maintained as it was a new valid feature.

After the merging, the columns' headers were renamed to clarify medical acronyms, and an extra variable containing the outcome as a character value was also created to generate more descriptive visuals. Additionally, all observations were correctly labelled but one, so the record was removed completely from the set of records. The dataset was then exported to a .csv file which could be used to explore the data in Microsoft Power BI. A few formatting steps were included in the Power Query editor to prepare the datasets for visualisation, including the unpivoting of positive comorbidities which were kept in a separate table linked through the Patient ID to facilitate the creation of DAX measures. The link created a bi-directional one-to-many relationship from the Patients table to the Comorbidities table, as shown in the schema view of Power BI Desktop, see Figure 1.

A calculated column was added to group the Body Mass Index (BMI) values according to the international standard. Additionally, seven DAX measures were manually created to compute the number of patients and deaths, death ratio, number of male and female patients formatted for the card visuals, and finally the percentage of patients for both the main table and the comorbidities table due to the cross filter.

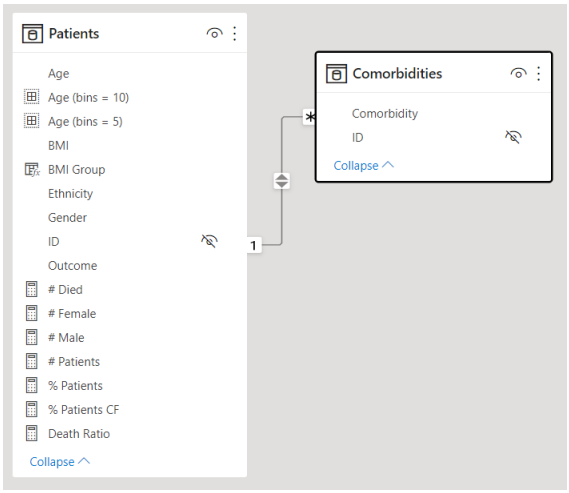


Figure 1 Comorbidities were pivoted in a new separate table linked to the Patients table by the Patient ID.

Once the data ready, the visuals were generated in a report to get some insights on the dataset, as shown in Figure 2 (top). The data contained 1,176 records after removing the unlabelled observation, amongst which 159 (13.52%) patients did not survive, see Figure 2 (bottom). This low number of death events matched findings from previous research (F. Li et al., 2021) and revealed an imbalanced class distribution.

Figure 2 EDA – Power BI Report Overview (top). Out of a cohort of 1,176 patients, 13.52% died in ICU (bottom).

Patients were aged between 19 and 99 years old when admitted to the ICU, and the histogram showed that the age distribution was left skewed, see Figure 3 (left). Indeed, the median age was 77 years old,

which was above the mean age of 74 years old and below the mode age of 89. As shown in the funnel in Figure 3 (middle), most patients were between 80 and 90 years old (40.4%), with a mortality ratio for this age bin of around 15%. The death ratio was very high before 40 years old (50%) and after 90 years old (40%) and appeared to be decreasing between 50 years old (14.9%) and 60 years old (5%), to then increase again steadily until 80 years old, to a 17.2% peak, see Figure 3 (right).

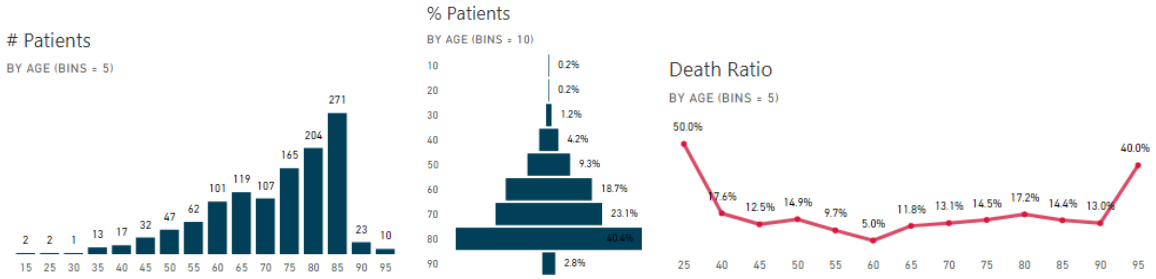


Figure 3 The age distribution was left skewed (left), 40.4% of the patients were between 80 and 90 years old (middle), with a mortality ratio for this age bin of around 15% (right).

As per Figure 4 (left), the cohort was comprised of 618 (52.6%) female patients and 558 (47.4%) male patients, with a mortality ratio of 12.8% and 14.3% respectively, suggesting that women had slightly higher chances to survive than men. Patients were in majority white (74.2%) and African American (12.4%), see Figure 4 (right). The pie chart in Figure 4 (middle left) showed that most patients in the cohort were either obese (41.5%) or overweight (30%). Interestingly, obese patients did not appear to have a higher mortality rate (8.8%), however five of the 24 underweighted patients did not survive (20.8%), as highlighted in Figure 4 (middle right).

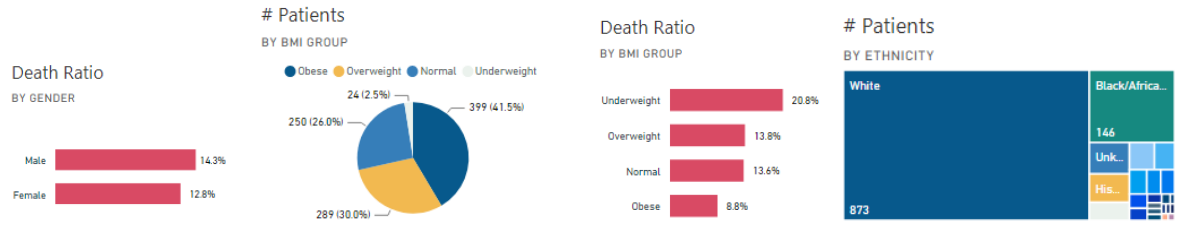


Figure 4 Women were slightly more represented in the cohort and their survival rate was 1.5 point higher than men's (left). Most HF patients admitted to the ICU were of white ethnicity (right) and either overweight or obese (middle left). Obese patients did not have a higher mortality rate (middle right).

With regards to comorbidities, hypertension was the most common condition with 844 patients affected (71.8%), but its mortality ratio was not the highest at 12%, see Figure 5. Although atrial fibrillation only affected 45.2% of the patients, the death ratio associated to it was the highest with 17.3%. Furthermore, ischaemic heart disease was only present in 101 patients (8.6%), but it was also the third highest mortality rate just after hypertension with 11.9% affected who did not survive.

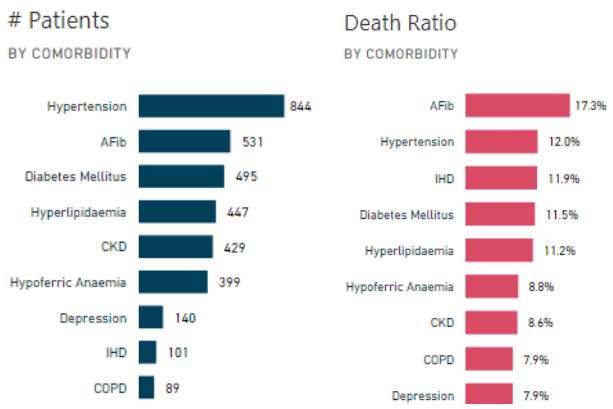


Figure 5 Although the most common comorbidity amongst patients was hypertension (844 patients – left), the deadliest one was atrial fibrillation (17.3% – right).

3. Methodology

3.1. Literature Review

Li et al. are the team of researchers who made the MIMIC-III Heart Failure dataset available under Public Domain Dedication licence (F. Li *et al.*, 2021). The purpose of their study was to better characterise the predictors of heart failure patients' in-hospital mortality. To achieve this, the team extracted a subset of records and attributes from the MIMIC-III database, which was then randomly split into training and testing sets in a 70-30 ratio. Then, the main mortality risk factors were identified using eXtreme Gradient Boosting (XGBoost) and Least Absolute Shrinkage and Selection Operator (LASSO) techniques. From these two sets of selected features, the researchers used Logistic Regression to build and validate the prediction models, which were finally compared based on C-index, calibration plot and decision curve analysis. Importantly for this project, they contrasted the performance of their models with the GWTG-HF scoring system. After bootstrapping validation, the final AUC scores for the XGBoost, LASSO regression, and GWTG-HF models were 0.8378, 0.8518 and 0.7743 respectively. Although the difference between the XGBoost model and the LASSO regression model was not statistically significant, the researchers chose XGBoost as it only included six variables against eight for the latter. The sensitivity score of the model was 53%, and the specificity was 95%. Finally, the team developed a nomogram that could be used to assess the in-hospital mortality risk score of heart failure patients based on the six model's variables: anion gap, lactate, calcium, blood urea nitrogen, chronic renal (or kidney) disease and diastolic blood pressure. The nomogram is shown in Appendix (section 10.6.).

Another team of researchers from Guangzhou in China worked on building a tool to stratify the mortality risk of ICU patients with heart failure (Luo *et al.*, 2022). Their study was also based on the MIMIC-III database as well as a second cohort of patients from the Telehealth Intensive Care Unit Collaborative Research Database (eICU-CRD), which was used as an external validation dataset. The eICU-CRD database contains the EHRs of 139,367 unique ICU patients and more than 200,000 ICU stays in 208 hospitals across the United States between 2014 and 2015. Luo et al. used PostgreSQL to extract the records of heart failure patients between 16 and 90 years old who were admitted to the ICU for the first time, without sepsis at admission and who stayed in the ICU for more than 24 hours. The researchers

excluded patients with no discharge information and records that contained physiologically impossible values. The attributes with more than 40% missing data and the records with more than 20% missing variables were dropped, and XGBoost was used to impute the remaining missing values. The final subset size was larger in this study than in Li et al.'s work, as 177 attributes and 5,676 records were extracted by Luo et al., against 51 attributes and 1,177 records previously. XGBoost was again used for the features selection and to build the model, along with two other models based on Elastic Net regression, GWTG-HF and Simplified Acute Physiology Score (SAPS) II, which is another ICU scoring systems for mortality risk. With a split ratio of 90-10, the model was trained on a larger set as well. Additionally, a 10-fold cross-validation was applied for model tuning. The calculated AUC on the internal testing set (10% of the data) were 0.667, 0.72, 0.817 and 0.831 for GWTG-HF, SAPS-II, logistic regression (Elastic Net) and XGBoost respectively. When tested on the external eICU-CRD validation set the XGBoost performance decreased slightly with an AUC of 0.809. Again, XGBoost was the best performing model overall, based on a selection of 24 features. To contrast with the previous study, the top six important variables present in Li et al.'s version of the MIMIC-III data that were detected by Luo et al.'s model were anion gap, urine output, blood urea nitrogen, age, calcium and respiratory rate. Glasgow Coma Scale (a clinical scale for measuring consciousness and estimate coma severity) pO2 maximum and glucose minimum were in the top six but are not included in this project's scope.

3.2. Data Science Process Model

Knowledge Discovery in Databases (KDD) is a Data Science process that aims to discover useful knowledge from data. The importance of KDD was demonstrated by Goodwin et al. in their research on the issues and opportunities of data mining for nursing knowledge (Goodwin *et al.*, 2003). In this paper, the team of researchers highlighted how clinical data could be used to discover knowledge in the field of healthcare by using the KDD process model. As illustrated in Figure 6, KDD is an iterative process that consists of five steps: Selection, Pre-Processing, Transformation, Data Mining and Interpretation/Evaluation. Unlike the CRoss Industry Standard Process for Data Mining (CRISP-DM), KDD does not incorporate Business Understanding or Deployment phases, which makes it more suited to a context of pure research (Quantum, 2019) and a perfect fit for this project.

The Selection stage aims to identify key attributes and observations from the main data source. This step is considered by Goodwin et al. as one of the main challenges in medical research as it heavily relies on domain knowledge. Indeed, the data used in this project was extracted from the MIMIC-III database by a team of medical professionals from the Cardiology departments of various Chinese hospitals. Once the target data is identified, the Pre-Processing phase can start to deal with missing and noisy data. The pre-processed data should be stripped of any inconsistencies, errors, or duplications. With regards to the use of Electronic Health Records (EHR), the anonymity of each record also has to be preserved. In the Transformation stage, any relevant action to reduce the dimensionality of the dataset should be taken, with techniques such as binning, or by creating new derived attributes, for instance.

At the core of the KDD process, the Data Mining phase is where the deep dive into the data can happen. In his Computer Science 831 notes, Hamilton introduces the choice of a data mining task (classification,

regression, clustering etc.) as an extra-step that is a pre-requisite to the selection of the appropriate(s) algorithm(s) (Hamilton, 2018). Finally, the last stage of KDD encompasses the Interpretation and Evaluation of the mined patterns, which eventually leads to the consolidation of the discovered knowledge.

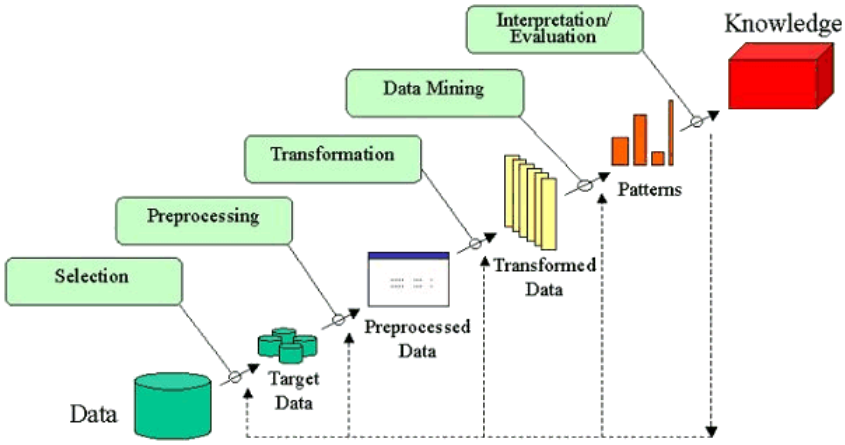


Figure 6 KDD is an iterative process that consists of five steps that aim to transform Data into Knowledge.

The details of the application of each KDD phase are presented in this report along with additional sections. The sections that are relevant to each phase are highlighted in the Table 3, shown below.

KDD Phase	Report Section
Selection	2.1. Datasets
Pre-Processing	3.3. Data Pre-Processing
Transformation	4.2. Features Selection
Data Mining	5. Modelling
Evaluation/Interpretation	6. Evaluation

Table 3 Each phase of the KDD process model is detailed in various sections of this report.

3.3. Data Pre-Processing

The EDA in Power BI was performed with minimal data cleaning on the raw data. After both datasets were merged, the columns were renamed, and two attributes were added to explicitly describe outcome and gender for a better readability of the visuals. The only significant step that was taken prior to the EDA was the dismissal of one unlabelled record, which brought the total number of observations from 1,177 to 1,176. To prepare the data for an analysis on correlation, a new column was created with a numeric imputation of ethnicities.

With regards to missing data, the analysis carried with the “naniar” package in R showed that the dataset was 97% complete. Out of 51 attributes in total, nineteen were missing values, including eleven attributes missing less than 3%, three attributes missing between 12% and 15% and five attributes missing between 18% and 25% of the data. In terms of records, 36.4% (n = 428) of the data was found to be completely observed. An upset plot was generated to visualise the data’s patterns of missingness, as shown in Figure 7. The plot revealed that a number of records were missing multiple values, with eleven features missing out of 51 (22%) in some cases. After considering discarding attributes and

observations below a completeness threshold, a more conservative approach was taken, and no features or records were dropped on the ground of missing data. This decision was made in concordance with a previous study where the authors advised against removing patients with incomplete records, as it can introduce bias in the analysis (F. Li *et al.*, 2021).

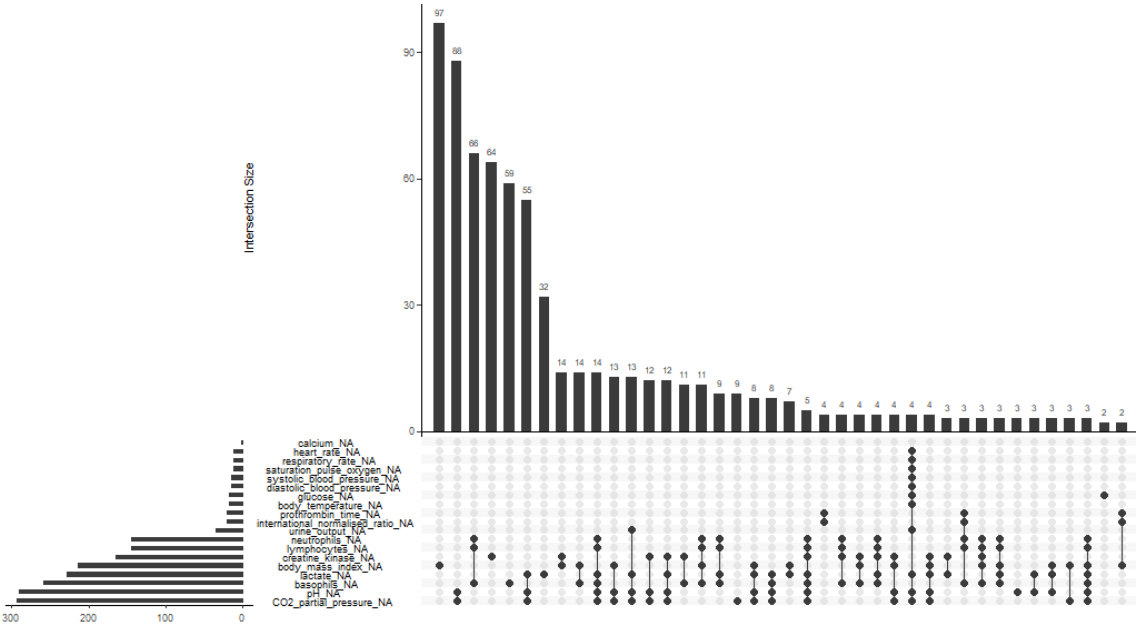


Figure 7 The upset plot generated in R ("naniar" package) showed that some records were missing a high number of attribute values.

Additionally, the “naniar” package enables to plot the percentage missing data by attribute, broken down by a factor. This functionality is especially relevant in the context of a classification task, as it allows to detect which attribute could potentially introduce bias in the prediction model due to a high proportion of imputed data. Applied to the HF dataset, the plot revealed that BMI and basophils had a high percentage (more than 20%) of missing data in the minority class (n = 159), as well as creatinine kinase to a lesser extent (around 15%). As shown in Figure 8, these features could be quickly identified visually, as the cells with higher percentages were filled with bright and contrasting colours (green and yellow). No further action was taken based on this visualisation, but it was kept as a reference to validate the features selection phase of the project.

Multivariate Imputation by Chained Equations (MICE) has been proven to be effective in the imputation of EHR data with low error (Cesare and Were, 2022) and its usage for laboratory variables showed promising results in several studies (J. Li *et al.*, 2021). This advanced procedure performs an iterative series of predictive models based on similar observed records until a convergence in the predictions is reached. The algorithm provided a log of one issue that was due to the expected collinearity between the outcome and the outcome description columns, so the data could be successfully completed.

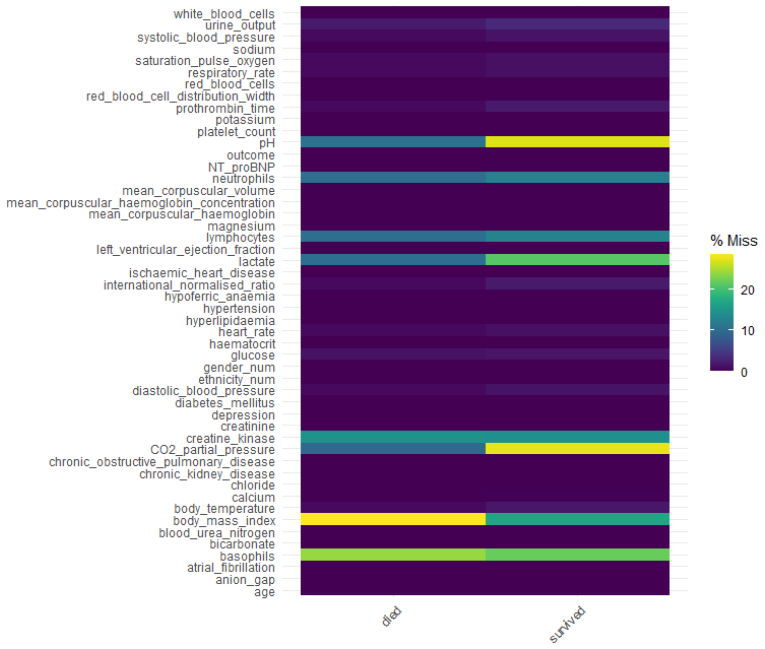


Figure 8 The plot of missing data broken down by factor highlighted three variables with a high percentage of missing values in the minority class: BMI, basophils and creatine kinase.

3.4. Modelling Strategy

Prior to building the model, a set of relevant features had to be selected from the 49 attributes of the MIMIC-III dataset. The three approaches taken – correlation test, principal component analysis and decision tree – resulted in three different sets of variables.

The goal of this predictive model is to categorise new observations based on known observations associated to a label. In machine learning, this is referred to as a supervised classification problem. Since the labels indicate the outcome of either death or survival, this is a binary classification scenario. Seven supervised machine learning algorithms were considered for this classification problem: K-Nearest Neighbour (K-NN), Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), Random Forest, eXtreme Gradient Boosting (XGBoost) and Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Prior to these techniques, the unsupervised learning approach K-Means Clustering was used to rule out any obvious segmentation of the data between the two classes.

Each model was trained on each of the three selected sets of features (correlation-based, PCA-based and tree-based), which were subsets of a main training set containing 70% (n = 824) of the total number of observations. The performance was then tested on the remaining 30% (n = 352) of unlabelled data using 10-fold cross-validation. The 70-30 split ratio was chosen to align with previous studies (Peterson *et al.*, 2010; F. Li *et al.*, 2021), as well as the number of folds for cross-validation (F. Li *et al.*, 2021; Luo *et al.*, 2022). The creation of a base training set containing all variables was to ensure that all models were tested on the same data, for a fair performance evaluation. The data partitioning was stratified for both the initial train-test split and the cross-validation folds to get a proportion of the minority class (death event) that reflected the actual proportion of the original dataset (13.52%). The stratification of the training and testing sets was obtained with the 'caret' R package.

The metrics used to compare the performance obtained by each set of variables within and between models were the recall and F1 scores. Indeed, Sarwar et al. have pointed out the predominance of class imbalance, where the class of interest is severely under-represented, when dealing with EHR data (Sarwar *et al.*, 2022). With a representation of 13.52% of the target class in the MIMIC-III dataset, the heart failure dataset is clearly imbalanced. Due to the low number of observations available to the models for training, it is much more challenging to detect true positives, which is the sensitivity or recall score, than true negatives. Therefore, the recall score and AUC were prioritised for the performance comparison. The F1-score, which is a function of precision and recall, was also computed to maintain a healthy balance between the true prediction rate of both positive and negative outcomes.

The model with the highest recall and F1 scores would then be explored further in the evaluation stage of the KDD process.

4. Analysis

4.1. Correlation Analysis

A matrix was generated to explore the potential correlations between some of the features and the target vector of outcomes. Out of 49 numeric attributes included in the analysis, five had a correlation coefficient greater than 0.20: anion gap (0.23), bicarbonate (-0.22), lactate (0.22), white blood cells (0.21), and blood urea nitrogen (0.20). These coefficients showed that patients with higher anion gap, white blood cells, lactate and blood urea nitrogen are more likely to die whereas patients with higher levels of bicarbonate have more chances to survive. This is also represented in Figure 9, as bicarbonate is the only boxplot where the box of higher values is filled with green, which is the colour associated to the survival outcome.

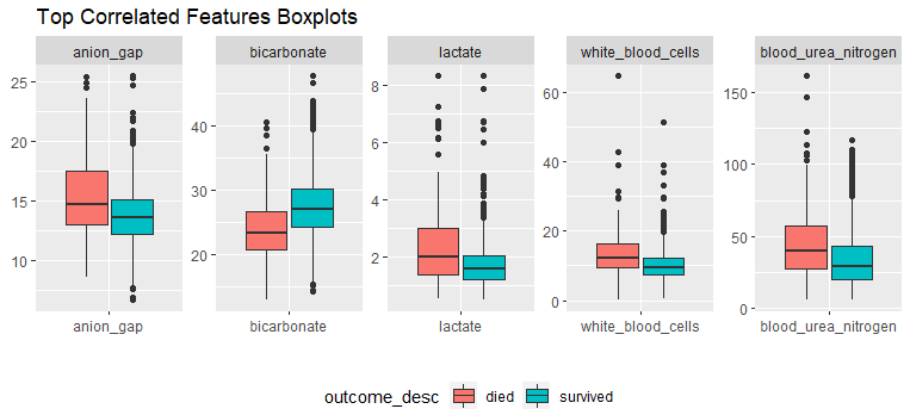


Figure 9 Patients with higher anion gap, white blood cells, lactate and blood urea nitrogen are more likely to die whereas patients with higher bicarbonate are more likely to survive as shown in the correlated features boxplots

Further research was carried out to get an understanding of these five indicators, based on two main sources: MSD manuals from Merck for consumers (*MSD Manual Consumer Version*, 2022) and for professionals (*MSD Manual Professional Edition*, 2022), as well as Medline Plus from the American National Institute of Health (NIH) (*MedlinePlus - Health Information from the National Library of Medicine*, 2022). Table 4 shows the main characteristics of these attributes.

Attribute	Type	Coefficient	Description
Anion gap	Laboratory variable	0.23	Human blood needs to remain electrically neutral, which means that the total cations (positively charged particles) equal the total anions (negatively charged particles). However, some ions are difficult to measure and only a few of them are normally obtained. Generally, the dominant measure of cation will be the sodium concentration, whereas the main measure of anions will be the sum of chloride and bicarbonate concentrations. Thus, the anion gap is the difference between the measured cations and the measured anions and represents the unmeasured anions that balance the positive charge produced by the cations.
Bicarbonate	Laboratory variable	-0.22	Bicarbonate is used to measure the metabolic acidosis. Metabolic acidosis occurs when a decrease in bicarbonate concentration lowers the blood pH below 7.35. Therefore, bicarbonate increases the concentration of bases and thus the blood pH, inversely to acids which decrease the pH. A normal blood pH is between 7.35 and 7.45. Metabolic acidosis falls into two categories: - high anion gap acidosis which can be caused by ketoacidosis, lactic acidosis, chronic kidney disease, or certain toxic ingestions; - normal anion gap acidosis which can be caused by gastrointestinal or renal bicarbonate loss.
Lactate	Laboratory variable	0.22	Lactic acid, or lactate, is a substance produced by muscle tissues and red blood cells that carries oxygen. Its level rise when the oxygen level decreases, in the case of a heart failure for instance. Lactic acid is used to diagnose lactic acidosis, which is the most common cause of metabolic acidosis (low pH caused by an excessive loss of bicarbonate) in hospitalised patients. It is characterised by a high anion gap due to an increased blood lactate, either due to a lactate overproduction or by a decreased lactate metabolism, or both.
White blood cells	Laboratory variable	0.21	Leucocytes, or leukocytes, are more commonly referred to as white blood cells. They are responsible for defending the human body against infections and they are categorised into five main types: neutrophils, lymphocytes, monocytes, eosinophils, and basophils. A low number of leucocytes is called leukopenia whilst a number higher than normal is a leucocytosis which can be an indication of an underlying disorder such as a leukaemia.
Blood urea nitrogen	Laboratory variable	0.20	Blood urea nitrogen (BUN) is an indicator of how well the kidneys are performing. The kidneys' primary function is to remove waste from the body. In case of a kidney disease, the increased amount of waste in the blood can lead to serious health issues such as high blood pressure, anaemia and heart disease. Like creatinine, urea

	nitrogen is a type of waste produced by the human body and normally removed by the kidneys. As shown in the correlation matrix, urea nitrogen and creatinine are both strongly correlated to the renal failure (or chronic kidney disease) comorbidity. Interestingly, these are also both correlated to the anion gap, which makes sense as the high anion gap acidosis can be caused by chronic kidney disease.

Table 4 Description of the five most correlated features to the outcome

With regards to the anion gap's positive correlation with blood urea nitrogen (0.52) and negative correlation with bicarbonate (-0.59), the relationship between these variables was highlighted in the above table. Indeed, a high level of blood urea nitrogen is oftentimes caused by a chronic kidney disease, which can itself increase the anion gap. Similarly, since the bicarbonate represents the measured anions and the anion gap is the difference between measured cations and measured anions, increasing the bicarbonate automatically reduces the anion gap.

The analysis on the top correlated features highlighted a possible causality relationship between the anion gap and bicarbonate features, which are both strongly correlated to the outcome, see Figure 10. A medical expert would be a key resource here to assess the relevance of having both features included in the prediction model. However, in the absence of such resource and as -0.59 is only a moderate correlation according to Hinkle et al.'s rule of thumb for interpreting coefficients (Hinkle et al., 2003, cited in Mukaka, 2012), both anion gap and bicarbonate were kept for the purpose of this project.

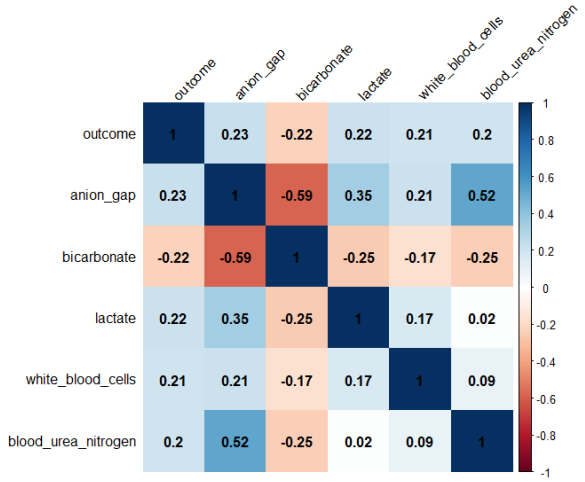


Figure 10 Anion gap is positively correlated to blood urea nitrogen but negatively correlated to bicarbonate as shown in the top correlated features correlation matrix.

4.2. Features Selection

4.2.1. Correlation Test

Following the correlation analysis carried out in the previous section, a strong correlation was found between the outcome and five variables: anion gap, bicarbonate, lactate, white blood cells and blood urea nitrogen. However, the correlation coefficient only shows the magnitude of the correlation, but it

does not indicate whether the correlation is significantly different from zero in the population. The Null and Alternate hypotheses to test this assumption are as follows:

H_0: $\rho = 0$ (there is no linear relationship between the two variables)

H_1: $\rho \neq 0$ (there is a linear relationship between the two variables)

As seen in Figure 11 (top), correlation coefficient and correlation test were combined into a single table through the use of the "correlation" package, which revealed that six variables had a correlation coefficient significantly different from zero in the population: anion gap, bicarbonate, lactate, white blood cells, systolic blood pressure and urine output. Despite having a correlation coefficient of 0.20, blood urea nitrogen had a p-value greater than 0.05 that did not allow to reject the null hypothesis. On the other hand, the test highlighted two further variables that were not in the top five coefficients but appeared to be significantly correlated to the outcome: urine output (p-value = 0.021) and systolic blood pressure (p-value = 0.037). The final selection of correlated features, along with their correlation coefficients and p-values are presented in Figure 11 (bottom).

```
# Correlation Matrix
# Parameter1 |                              Parameter2 |        r |         95% CI | t(47) |      p
# ----------------------------------------------------------------------------------------------
# outcome    |                                 lactate |     0.51 | [ 0.27,  0.69] |  4.07 | 0.009**
# outcome    |                             bicarbonate |    -0.49 | [-0.68, -0.24] | -3.89 | 0.015*
# outcome    |                            urine_output |    -0.48 | [-0.67, -0.23] | -3.77 | 0.021*
# outcome    |                               anion_gap |     0.47 | [ 0.23,  0.66] |  3.73 | 0.024*
# outcome    |                         white_blood_cells |     0.46 | [ 0.21,  0.66] |  3.61 | 0.033*
# outcome    |                    systolic_blood_pressure |    -0.46 | [-0.65, -0.20] | -3.56 | 0.037*
# outcome    |                        blood_urea_nitrogen |     0.40 | [ 0.13,  0.61] |  2.99 | 0.187
```

Attribute	Correlation Coefficient	Correlation Test p-value
Anion gap	0.23	p = 0.024
Bicarbonate	-0.22	p = 0.015
Lactate	0.22	p = 0.009
White blood cells	0.21	p = 0.033
Urine output	-0.18	p = 0.021
Systolic blood pressure	-0.13	p = 0.037

Figure 11 The correlation test showed that blood urea nitrogen was not statistically significant, but urine output and systolic blood pressure were two significantly correlated features (top). Blood urea nitrogen was therefore removed from the final selection, which also included urine output and systolic blood pressure (bottom).

From this combined analysis of correlation coefficients and correlation tests, a first set of six features of relevance was identified: anion gap, bicarbonate, lactate, white blood cells, urine output and systolic blood pressure.

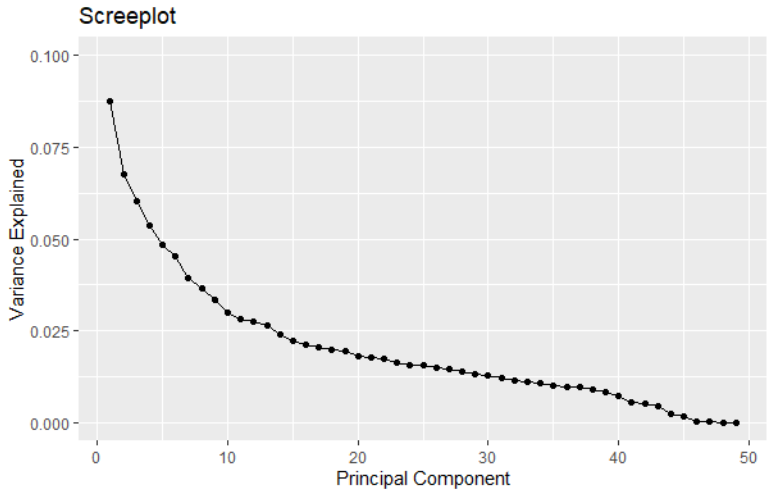
4.2.2. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure that computes components that maximise variance. Each component is ranked from the highest to the lowest information content, thus the first component can explain most of the variance in the data. The generated components can be used to reduce the dimensionality of large datasets and feed machine learning algorithms as new features, but this method would not provide the transparency that is needed to explain the model's

output. As the interpretation of the results in terms of individual variables is crucial when it comes to healthcare, a different use of PCA was made in this research.

In order to select the most relevant features that would give the best model performance, PCA was run on the scaled dataset and the first component was used as a base to determine which attributes could explain most of the data's variance.

The analysis revealed that this component could only explain 8.77% of the data, whilst the scree plot showed that 45 components out of 49 accounted for less than 5% of the variance only, see Figure 12 (left). Based on the variable contribution table of the pca object shown in Figure 12 (right), the five most contributing features of the first component were bicarbonate (10.47%), blood urea nitrogen (9.91%), anion gap (9.33%), creatinine (7.27%) and NT pro-BNP (6.86%).



	Dim.1
bicarbonate	10.468510812
blood_urea_nitrogen	9.909680345
anion_gap	9.327908246
creatinine	7.270024222
NT_proBNP	6.860479527
urine_output	4.881434634
CO2_partial_pressure	4.668504695
haematocrit	4.585188986
red_blood_cells	4.500198774
chronic_kidney_disease	3.983993926
body_mass_index	2.795757554
diastolic_blood_pressure	2.676045408

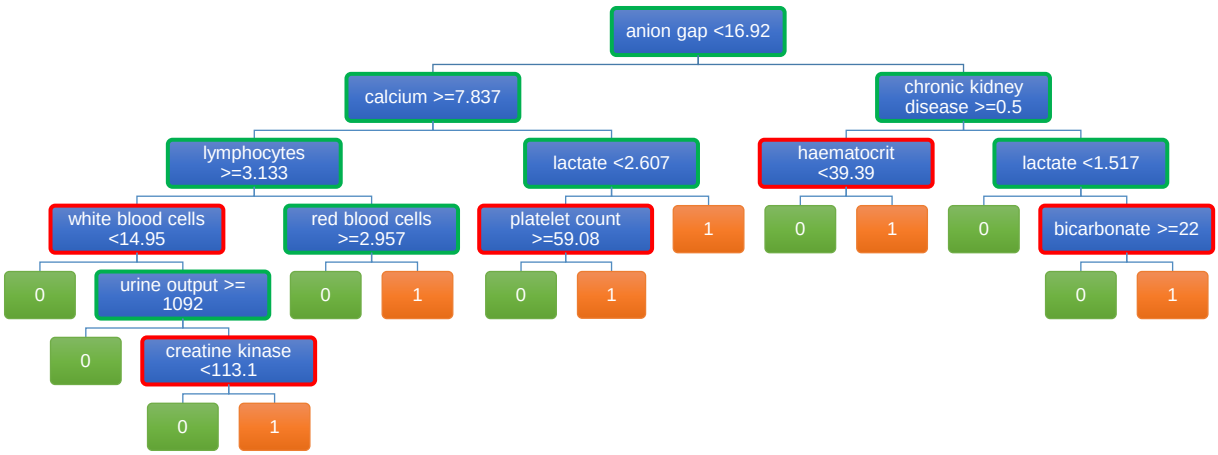
Figure 12 The scree plot displayed the little variance explained by most components (left) and five main features contributed to the first component that explained 8.74% of the data (right).

From the PCA, a second set of five features of relevance was identified: bicarbonate, blood urea nitrogen, anion gap, creatinine and NT pro-BNP.

4.2.3. Decision Tree

The last technique that was used to select a set of relevant features was to build a decision tree and explore the variables used at each node. No training set or test set were used here, as the goal was to determine which attributes the algorithm was choosing to create a logical path to the outcome. The visual output of the classifier was a 5-level deep decision tree, which is presented in Figure 13 (top).

As seen in Figure 13 (bottom), the 'variable.importance' attribute of the rpart object showed that anion gap (21.55) and lactate (17.30) were by far the most important features, followed by chronic kidney disease (12.33), lymphocytes (10.78), red blood cells (9.43), creatinine (9.15), urine output (8.91) and calcium (8.18). Although not all of these attributes were represented in the visual output of the tree generated in R, the features were selected based on this ranking and the remaining ones were not further considered.




```
> classifier_dt$variable.importance
                anion_gap                   lactate
               21.5504987                17.2999089
  chronic_kidney_disease                lymphocytes
               12.3306667                10.7798304
           red_blood_cells                creatinine
                9.4310160                 9.1516020
              urine_output                   calcium
                8.9071013                 8.1773327
              haematocrit                bicarbonate
                7.4631538                 5.9037993
        white_blood_cells           creatine_kinase
                5.7173148                 5.4034632
         prothrombin_time                 basophils
                4.5240822                 4.4071917
           platelet_count                   glucose
                4.1617457                 4.0152346
```

Figure 13 The visual output of the decision tree classifier (top) was 5-level deep and included features of higher importance (green border) along features of lower importance that were not selected (red border) based on the 'variable.importance' attribute of the tree (bottom).

From the decision tree analysis, a third set of eight features of relevance was identified: anion gap, lactate, chronic kidney disease, lymphocytes, red blood cells, creatinine, urine output and calcium.

4.2.4. Selection Summary

The summary table of the thirteen features selected highlights the importance of the anion gap laboratory measure, which was identified as a key feature by all the three approaches, see Table 5. Bicarbonate, lactate, urine output and creatinine were deemed as important in two approaches and the remaining eight attributes were specific to one selection method. Lastly, the features that could have introduced some bias in the study due to a high proportion of missing data in the minority class – BMI, basophils and creatine kinase – were not identified by any of the three selection methods.

Feature	COR Set	PCA Set	DT Set
Anion gap	✓	✓	✓
Bicarbonate	✓	✓	
Lactate	✓		✓
Urine output	✓		✓
Creatinine		✓	✓
White blood cells	✓		
Systolic blood pressure	✓		
Blood urea nitrogen		✓	

NT pro-BNP		✓	
Chronic kidney disease			✓
Lymphocytes			✓
Red blood cells			✓
Calcium			✓

Table 5 Anion gap was identified as important by the three feature selection approaches.

5. Modelling

5.1. K-Means Clustering (unsupervised)

Although it is an unsupervised machine learning technique, clustering can be useful for patient segmentation. In the context of this project, the hope was that a segmentation in two (k = 2) would reveal a cluster that would point to our two groups of patients who did and did not survive. The k-means algorithm found one first small clustered of 223 patients with a death ratio of 22% and a second larger cluster of 953 patients and a death ratio of 11.5%.

If we were to use the first cluster as the cluster of patients who did not survive, as the death ratio was higher, and the second cluster as the cluster of patients who did survive, the accuracy of the result would be 75.85% which is very low. Therefore, this first technique was not further considered.

5.2. K-Nearest Neighbour

The K-Nearest Neighbour (K-NN) algorithm classifies new data points based on the class of the nearest existing data points. The number of neighbour data points to consider for the classification is defined in advance by k, which is commonly set to 5. Various techniques can be used to determine the nearest neighbours, such as the Euclidian distance or the Manhattan distance. The new data point is allocated the majority class amongst the k nearest neighbours.

In the context of binary classification, k should be an odd number to facilitate the identification of the majority class by the algorithm. Additionally, using a small number of neighbours was the most intuitive approach to deal with the imbalanced data and account for the under-representation of positive outcomes. As setting k to 1 with a known training set would lead to an overfitting model, k was initialised to 3.

As seen in Table 6, all three sets of features performed very similar with K-NN, but the correlation set had slightly higher scores with 0.09 for recall and 0.19 for F1.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.09	0.07	0.07
F1 (10-fold CV)	0.19	0.17	0.17

Table 6 K-NN provided a slightly better performance with the set of correlated features (k = 3).

5.3. Logistic Regression

Logistic regression is a classification algorithm that predicts the probability of an outcome to occur, as opposed to predicting the outcome itself. The logistic regression function forms a "S" curve that represents the best fitting line for the vector of calculated probabilities and goes from 0 to 1, where:

- 0 is P(positive outcome) = 0, the likelihood of a positive outcome is 0%;
- 1 is P(positive outcome) = 1, the likelihood of a positive outcome is 100%;

The model also involves the definition of a threshold to classify the binary outcome. For instance, if the threshold is set to 0.5, 0 is the negative outcome, 1 is the positive outcome and the calculated probability is P(positive outcome) = 0.75, the model will predict 1.

Due to the imbalanced nature of the data, it was critical to optimise the value of the threshold so that it would not be too high, which would lower the true-positive rate, but not too low either to avoid a false-positive rate increase. Several threshold values were considered for this model, and 0.3 gave the best F1-score. With this threshold, the tree-based set was the best performing set of features with a recall score of 0.43 and a F1-score of 0.47, as described in Table 7.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.39	0.29	0.43
F1 (10-fold CV)	0.43	0.34	0.47

Table 7 The decision tree features provided the best performance with the Logistic Regression algorithm (threshold = 0.3).

5.4. Naïve Bayes

The Naïve Bayes modelling technique is a probabilistic classification approach based on Bayes' theorem. To classify a new data point, the formula compares the probability of getting a negative outcome given its set of features, versus the probability of getting a positive outcome given this same set of features. The new data point is assigned the outcome with the highest probability calculated from Bayes' formula. This machine learning technique is called "naïve" because the classifier "naively" relies on the assumption that all variables are independent, although most of times is not the case. In other words, the term "naïve" refers to the fact that Bayes' independence assumption is disregarded by the model.

This technique provided the best results when applied to the correlated features set, with a F1-score of 0.38 and a recall score of 0.34. As shown in Table 8, although the F1-score obtained with the decision tree set was very close (0.37), the recall score of 0.29 was clearly lower.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.34	0.30	0.29
F1 (10-fold CV)	0.38	0.32	0.37

Table 8 Naïve Bayes provided a better performance with the set of correlated features.

5.5. Support Vector Machines

The goal of the Support Vector Machines (SVM) algorithm is to define a boundary within the set of existing data points that separates the two classes and can then be used to classify new data points. Although it is a simple line in a two-dimensional area, the decision boundary becomes a hyperplane in a multidimensional space. The algorithm calculates the hyperplane by maximising the distance between the nearest data points (multidimensional vectors) of each outcome. Thus, these data points are "support vectors" to the creation of the maximum margin hyperplane, which is the classifier. The originality of this machine learning approach is its reliance on data points that are by nature the most ambiguous in the dataset.

When the data cannot be linearly separated, the SVM algorithm can determine the optimal landmarks that can be used by a variety of kernels to fit the hyperplane to more complex multidimensional shapes, such as gaussian (radial kernel), hyperbolic (sigmoid kernel) or polynomial. For this project, all linear, sigmoid, radial and polynomial kernels were tested, and the best results were obtained from the polynomial kernel with a number of degrees set to five.

The best performance was obtained with the set of correlated features, with a recall score of 0.21 and a F1-score of 0.37, see Table 9. The decision tree set had a very similar recall score (0.22), but the F1-score of 0.34 was below the performance of the correlated features set.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.21	0.15	0.22
F1 (10-fold CV)	0.37	0.29	0.34

Table 9 The correlated features set provided the best performance with SVM (polynomial kernel).

5.6. Random Forest

Random Forest is an ensemble learning technique, meaning that it combines the predictions made by several algorithms in a single enhanced model. Specifically, it leverages the output produced by multiple decision trees. To do so, the random forest algorithm selects a random subset of the training set, builds a decision tree based on these data points and repeat the process to get a set number of trees 'ntree'. For each new data point, the class predicted by the majority of 'ntree' trees is assigned to the data point.

The algorithm was tested with various 'ntree' parameters, and it did not show any improvement beyond 60 trees. Table 10 shows that although all sets performed similarly, a slightly better performance was obtained from the correlated features set, which gave a recall score of 0.21 against 0.19 for the two other sets.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.21	0.19	0.19
F1 (10-fold CV)	0.32	0.31	0.32

Table 10 The set of correlated features provided slightly better results with the random forest algorithm.

5.7. eXtreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble machine learning algorithm that can be used for classification problems. This technique consists in building a XGBoost tree based on multiple trees that predict residuals from various splitting thresholds. The algorithm computes the information gain of each threshold based on clusters of similar residuals as well as previous predictions. The tree maximising gain (and therefore reducing residuals) is then used to make a new prediction, which will be the starting point to a new tree that is fit to the new residuals. The process is repeated until the residuals cannot be further reduced, or if the maximum number of trees is reached.

Experimenting with more than 60 number of iterations('nrounds') or more than 5 levels deep ('max.depth') did not improve the performance further. Although all sets performed equivalently with this model, see Table 11, the decision tree set provided slightly better results with a F1-score of 0.36 and a recall score of 0.34.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.32	0.32	0.34
F1 (10-fold CV)	0.35	0.33	0.36

Table 11 The decision tree set provided a slightly better performance with XGBoost.

5.8. Repeated Incremental Pruning to Produce Error Reduction

The last technique is called RIPPER, which stands for 'Repeated Incremental Pruning to Produce Error Reduction'. It is a rule learning algorithm, meaning that new data points are classified according to the output of logical if-else statements, see Figure 14 for an example. RIPPER relies on pruning to improve the performance and accuracy of these logical rules based on a 'divide and conquer' approach. Similarly to decision trees, split thresholds are evaluated by computing information gain, and the branches that do not maximise this gain are pruned.

```
JRIP rules:
===========

(bicarbonate <= 24.909091) and (anion_gap >= 16.571429) and (white_blood_cells >= 13.4) and (glucose >= 103.2) => outcome=1 (19.0/1.0)
(urine_output <= 1510) and (bicarbonate <= 23.454545) and (platelet_count <= 146) => outcome=1 (34.0/13.0)
(urine_output <= 1237) and (white_blood_cells >= 11.777778) and (saturation_pulse_oxygen <= 94.366667) => outcome=1 (19.0/6.0)
 => outcome=0 (768.0/61.0)

Number of Rules : 4
```

Figure 14 The RIPPER rules obtained on all variables corroborated the features selection process, as most variables were already included in the selected sets.

As seen in Table 12, all sets performed equivalently with this technique, but the PCA set showed slightly better results with a F1-score of 0.32 against 0.29 for the two other sets.

Metric	COR Set	PCA Set	DT Set
Recall (10-fold CV)	0.21	0.20	0.21
F1 (10-fold CV)	0.29	0.32	0.29

Table 12 RIPPER provided a slightly better performance with the PCA-based set.

6. Evaluation

6.1. Model Selection

A comparative analysis of the scores obtained by the best performing set of each machine learning approach revealed that the best model was the Logistic Regression applied on the Decision Tree features set. The Logistic Regression model had a F1-score of 0.47 and a recall score of 0.43, and was ahead of Naïve Bayes and XGBoost, second and third best performing models respectively. With a F1-score of 0.37, SVM performed similarly to Naïve Bayes (0.38) and XGBoost (0.36), but the recall score of 0.21 was clearly under their shared score of 0.34. The same recall score was obtained with Random Forest and RIPPER, which also shared a lower F1-score of 0.32 compared to Naïve Bayes, XGBoost and SVM. As shown in Table 13, the lowest performance for both scores was given by the K-NN algorithm.

Model	Features Set	Recall	F1 Score
K-NN	Correlation Analysis	0.09	0.19
Logistic Regression	Decision Tree	0.43	0.47
Naïve Bayes	Correlation Analysis	0.34	0.38
SVM	Correlation Analysis	0.21	0.37
Random Forest	Correlation Analysis	0.21	0.32
XGBoost	Decision Tree	0.34	0.36
RIPPER	PCA	0.20	0.32

Table 13 Logistic Regression used in conjunction with the features selected from the decision tree analysis was the best performing model overall.

6.2. Area Under the Receiver Operating Characteristic Curve Analysis

Receiver Operating Characteristic (ROC) graphs provide a way to visualise the impact of the decision threshold applied to classify the probability scores on both True Positive (Recall/Sensitivity) and True Negative (Specificity) rates. A threshold of 0 results in a sensitivity score of 100% as we assume that all the outcomes are positive and thus correctly predict all the true positives, but since no negative outcome is predicted the specificity is 0%. Inversely, a threshold of 1 would predict all the outcomes as negative, resulting in the correct prediction of all true negatives with a specificity score of 100%, but the sensitivity score would be 0%. Sensitivity is on the y-axis and specificity is on the x-axis, therefore the threshold of 0 is represented at the top right of the graph and the threshold of 1 is at the bottom left. Any point on the dotted line means that the proportion of true positives equals the proportion of true negatives.

The Area Under the Curve (AUC) is the metric that is used to compare different ROC curves by computing the two-dimensional area underneath each curve. An AUC score of 100% indicates that all true positives and all true negatives were identified (Sensitivity = Specificity = 100%).

The goal of this analysis was to confirm the better performance of the logistic regression model against the XGBoost model. Indeed, XGBoost had shown very high scores in the initial testing of the modelling phase, before the change of threshold in the logistic regression model. As there was some variability in

the Recall and F1 scores obtained with the 10-fold cross-validation evaluation, the AUC was used to compare both ROC curves. The better performance of the logistic regression model was confirmed with an AUC score of 76.2%, but the XGBoost was very close with 76%, as shown in Figure 15.

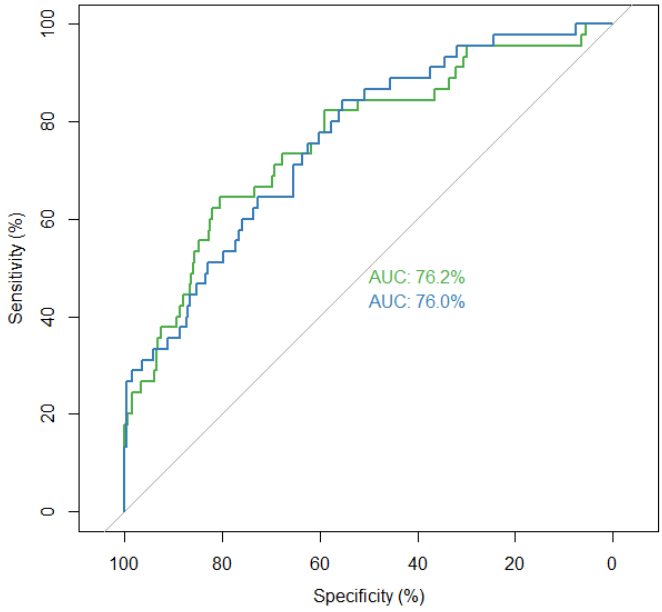


Figure 15 The ROC curves showed that despite both models performing better on the decision tree set, logistic regression had a slightly higher AUC score (76.2%, green curve) compared to XGBoost (76.0%, blue curve).

6.3. Features Importance

The summary statistics of the Logistic Regression model shown in Figure 17 (left) highlighted three features that did not have a significant impact on the model. Red blood cells, creatinine and lactate had p-values of 0.92, 0.37 and 0.14, all above the significance threshold of 0.05.

To test whether the performance of the model could be significantly improved by dropping some of these less significant features, an ANOVA chi-squared test was run to compare four classifiers organised in increasing number of attributes. The first classifier was fit without any of the three features, which were then added back in decreasing order of significance. Therefore, lactate was added back first since it had the smallest p-value (0.14), then creatinine (0.37) was added to the third model and finally the fourth classifier was back to the original set, including the least significant red blood cells attribute (0.92).

The ANOVA test showed that the addition of lactate had a small but noticeable impact on the reduction of the residual sum of squares (p-value = 0.09), however including creatinine or red blood cells did not improve further the reduction of residuals at the alpha value of 0.05. As presented in Figure 16, the second line of the table corresponding to the addition of lactate was followed by the ‘dot’ significance code (0.1), but the two following rows had no significance code at all.

```
Analysis of Deviance Table

Model 1: outcome ~ anion_gap + chronic_kidney_disease + lymphocytes +
    urine_output + calcium
Model 2: outcome ~ anion_gap + lactate + chronic_kidney_disease + lymphocytes +
    urine_output + calcium
Model 3: outcome ~ anion_gap + lactate + chronic_kidney_disease + lymphocytes +
    creatinine + urine_output + calcium
Model 4: outcome ~ anion_gap + lactate + chronic_kidney_disease + lymphocytes +
    red_blood_cells + creatinine + urine_output + calcium
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       818     542.09
2       817     539.23  1  2.86540   0.0905 .
3       816     538.39  1  0.83880   0.3597
4       815     538.38  1  0.00907   0.9241
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16 The analysis of variance (chi-squared) confirmed the importance of the lactate attribute in reducing the residuals.

Therefore, red blood cells and creatinine could be dropped, leaving the six following features in decreasing order of effect size: chronic kidney disease, calcium, anion gap, lactate, lymphocytes and urine output. Based on the logistic regression model output shown in Figure 17 (right), log odds of mortality = 1.7980132 + 0.2814691 x anion gap + 0.1758012 x lactate - 1.1362446 x chronic kidney disease - 0.0340972 x lymphocytes - 0.0002843 x urine output - 0.8172976 x calcium.

```
Call:
glm(formula = outcome ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9446  -0.5261  -0.3714  -0.2332   2.8223

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.6835164  1.7526912   0.961 0.336787
anion_gap               0.3122126  0.0578966   5.393 6.94e-08 ***
lactate                 0.1546312  0.1056885   1.463 0.143444
chronic_kidney_disease -1.0377625  0.2965035  -3.500 0.000465 ***
lymphocytes            -0.0333828  0.0156929  -2.127 0.033399 *
red_blood_cells        -0.0182572  0.1918992  -0.095 0.924204
creatinine             -0.1114387  0.1243057  -0.896 0.369992
urine_output           -0.0002972  0.0001118  -2.659 0.007844 **
calcium                -0.8234978  0.2134466  -3.858 0.000114 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 662.42  on 823  degrees of freedom
Residual deviance: 538.38  on 815  degrees of freedom
AIC: 556.38
```

```
Call:
glm(formula = outcome ~ ., family = binomial, data = training_set_anova)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9160  -0.5209  -0.3733  -0.2339   2.8330

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.7980132  1.7321152   1.038  0.29925
anion_gap               0.2814691  0.0466233   6.037 1.57e-09 ***
lactate                 0.1758012  0.1022509   1.719  0.08556 .
chronic_kidney_disease -1.1362446  0.2783061  -4.083 4.45e-05 ***
lymphocytes            -0.0340972  0.0155197  -2.197  0.02802 *
urine_output           -0.0002843  0.0001097  -2.592  0.00953 **
calcium                -0.8172976  0.2055595  -3.976 7.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 662.42  on 823  degrees of freedom
Residual deviance: 539.23  on 817  degrees of freedom
AIC: 553.23

Number of Fisher Scoring iterations: 5
```

Figure 17 The initial summary statistics of the Logistic Regression model (left) revealed two predictors of no significance – red blood cells and creatinine – which were removed in the final version of the model (right).

7. Conclusions

Out of eight machine learning techniques applied to the MIMIC-III heart failure data, the logistic regression model provided the best performance with an AUC score of 0.7625. This score was not far off the AUC of 0.7743 provided by the GWTG-HF scoring system (F. Li *et al.*, 2021), but it was way below the AUC of 0.8378 delivered by Li et al. with their XGBoost model. The authors also had a better sensitivity score of 0.53, against 0.43 for the model developed as part of this project. As the XGBoost modelling technique was also tested in this research but did not perform as well as the logistic regression, it is likely that the difference lied in the features selection process applied. The approach of Li et al. consisting in using XGBoost for both features selection and modelling might have enhanced even further the results.

Originally based on the set of eight most important features designated by the decision tree, the list of features list was reduced to only keep the attributes that were significantly contributing towards the model. The final set included chronic kidney disease, calcium, anion gap, lactate, lymphocytes and urine output. Four of these variables – anion gap, lactate, calcium and chronic kidney disease – were also selected by Li et al., but blood urea nitrogen and diastolic blood pressure were preferred to lymphocytes and urine output. Blood urea nitrogen was found in the first PCA component, but none of the three approaches taken in this study found diastolic blood pressure to be a feature of importance. Blood urea nitrogen is also a parameter used by the GWTG-HF scoring system, along with systolic blood pressure, sodium, age, heart rate, chronic obstructive pulmonary disease and ethnicity (optional). Although systolic blood pressure was identified in the correlation test, none of the other GWTG-HF variables were found to be important through the selection process. However, the absence of the anion gap amongst the variables used by the GWTG-HF system was surprising, considering the predominance of this feature across all the selection strategies implemented in both this project and the literature (F. Li *et al.*, 2021; Luo *et al.*, 2022).

The results of this study are very close to what seasoned research teams have produced, see for instance GWTG-HF system. Noteworthily, these promising results were obtained in only four months. This project is a proof of concept that machine learning offers endless opportunities to explore EHRs for clinical research, but it also highlights the intricacy of a field that keeps expanding with ever more techniques and algorithms. Of course, when dealing with human lives, the impact of any erroneous predictions could be disastrous.

8. Further Research

With additional time and resources on this project, the main goal would be to improve the prediction model to outperform the GWTG-HF system. Many improvements could be done to these models and methodology. First of all, the approaches taken by Li et al. for features selections based on XGBoost and LASSO (F. Li *et al.*, 2021) are just two out of the many feature selection strategies available. They showed that there are a variety of ways to identify predictors, and therefore more approaches could be tested.

Similarly, MICE allows for customised imputation methods and the imputation of missing data may also be a factor in the predictive performance of the model. Again, more approaches may be tested and contrasted for imputation. More generally, the functions used to create the classifiers came with a set of advanced parameters which should be explored to test and evaluate the impact of relevant options on the model's performance.

Lastly, other studies seem to suggest that oversampling techniques such as SMOTE can improve prognosis predictions on heart failure patients (Kim *et al.*, 2020; Ishaq *et al.*, 2021). Once the model's performance increased, it should be tested externally as per the validation process followed by Luo et al. in their similar research (Luo *et al.*, 2022), and the mortality time component could also be considered by extending the scope to a survival analysis.

9. References

Websites

Hamilton, H. (2018) *KDD Process/Overview*. Available at: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html (Accessed: 7 April 2022).

Malik, A. *et al.* (2022) 'Congestive Heart Failure', in *StatPearls*. Treasure Island (FL): StatPearls Publishing. Available at: http://www.ncbi.nlm.nih.gov/books/NBK430873/ (Accessed: 12 May 2022).

MedlinePlus - Health Information from the National Library of Medicine (2022). Available at: https://medlineplus.gov/ (Accessed: 17 December 2021).

MIMIC3d aggregated data (2019). Available at: https://kaggle.com/drscarlat/mimic3d (Accessed: 28 November 2021).

MSD Manual Consumer Version (2022) *MSD Manual Consumer Version*. Available at: https://www.msdmanuals.com/home (Accessed: 17 December 2021).

MSD Manual Professional Edition (2022) *MSD Manual Professional Edition*. Available at: https://www.msdmanuals.com/professional (Accessed: 17 December 2021).

Quantum (2019) *Data Science project management methodologies*, *Medium*. Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb (Accessed: 7 April 2022).

Saltz, J. (2020) 'CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects', *Data Science Process Alliance*, 30 November. Available at: https://www.datascience-pm.com/crisp-dm-still-most-popular/ (Accessed: 22 December 2021).

Journals

Cesare, N. and Were, L.P.O. (2022) 'A multi-step approach to managing missing data in time and patient variant electronic health records', *BMC Research Notes*, 15(1), p. 64. doi:10.1186/s13104-022-05911-w.

Crown, W.H. (2015) 'Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions', *Value in Health*, 18(2), pp. 137–140. doi:10.1016/j.jval.2014.12.005.

Goodwin, L. *et al.* (2003) 'Data mining issues and opportunities for building nursing knowledge', *Journal of Biomedical Informatics*, 36(4), pp. 379–388. doi:10.1016/j.jbi.2003.09.020.

Ishaq, A. *et al.* (2021) 'Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques', *IEEE Access*, 9, pp. 39707–39716. doi:10.1109/ACCESS.2021.3064084.

Johnson, A.E.W. *et al.* (2016) 'MIMIC-III, a freely accessible critical care database', *Scientific Data*, 3(1), p. 160035. doi:10.1038/sdata.2016.35.

Kim, Y.-T. *et al.* (2020) 'A Comparison of Oversampling Methods for Constructing a Prognostic Model in the Patient with Heart Failure', in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*. pp. 379–383. doi:10.1109/ICTC49870.2020.9289522.

Li, F. *et al.* (2021) 'Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database', *BMJ Open*, 11(7), p. e044779. doi:10.1136/bmjopen-2020-044779.

Li, J. *et al.* (2021) 'Imputation of missing values for electronic health record laboratory data', *npj Digital Medicine*, 4(1), pp. 1–14. doi:10.1038/s41746-021-00518-0.

Luo, C. *et al.* (2022) 'A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure', *Journal of Translational Medicine*, 20(1), p. 136. doi:10.1186/s12967-022-03340-8.

Mukaka, M. (2012) 'A guide to appropriate use of Correlation coefficient in medical research', *Malawi Medical Journal : The Journal of Medical Association of Malawi*, 24(3), pp. 69–71.

Peterson, P.N. *et al.* (2010) 'A Validated Risk Score for In-Hospital Mortality in Patients With Heart Failure From the American Heart Association Get With the Guidelines Program', *Circulation: Cardiovascular Quality and Outcomes*, 3(1), pp. 25–32. doi:10.1161/CIRCOUTCOMES.109.854877.

Pocock, S.J. *et al.* (2013) 'Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies', *European Heart Journal*, 34(19), pp. 1404–1413. doi:10.1093/eurheartj/ehs337.

Ponikowski, P. *et al.* (2016) '2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC', *European Heart Journal*, 37(27), pp. 2129–2200. doi:10.1093/eurheartj/ehw128.

Roger, V.L. *et al.* (2004) 'Trends in Heart Failure Incidence and Survival in a Community-Based Population', *JAMA*, 292(3), pp. 344–350. doi:10.1001/jama.292.3.344.

Sarwar, T. *et al.* (2022) 'The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges', *ACM Computing Surveys*, 55(2), pp. 1–40. doi:10.1145/3490234.

Shin, S. *et al.* (2021) 'Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality', *ESC Heart Failure*, 8(1), pp. 106–115. doi:10.1002/ehf2.13073.

10. Appendices

10.1. Project Proposal

Project Proposal

Objectives

This project aims to provide a statistical analysis of medical records from patients who suffered a heart failure. The first part of the project will consist in using statistical tools to get some insights into the data and identify significant features in heart failure mortality rates, and the second part will be a survival analysis to understand the temporality of death events. The finality of this project is to build a model that predicts the outcome of a heart failure as well as the expected survival time of the patient.

Therefore, this analysis of the medical records will allow me to answer the following questions:

- What are the significant health features that lead to an accurate prediction of the chances to survive a heart failure?
- What are the chances of survival of a patient after suffering a heart failure?

Aside from the technical interest of this Machine Learning project, the question of the digitalisation of patients' health records, which are referred to as Electronic Health Records (EHR), is a major opportunity for Data Analysts and Data Scientists to explore high-quality medical data. Hospitals have been moving towards the adoption of digitalised systems over the last few years, and in the United States, for instance, "the number of non-federal acute care hospitals with basic digital systems increased from 9.4 to 75.5% over the 7 year period between 2008 and 2014" (Johnson *et al.*, 2016). It is therefore crucial that Data Analysts and Data Scientists are not put off by a lack of field knowledge, as the data could reveal some trends or behaviours that Doctors may have missed due to some potential bias induced by their medical expertise. The opportunities presented by the increased availability of anonymized Electronic Health Records (EHR) will only grow overtime and understanding them could lead to medical discoveries that would eventually contribute to saving lives.

Background

The initial goal for this project was to build a predictive maintenance model based on machines failures. Using data to forecast the future is a fascinating aspect of data analytics that deemed interesting to explore further. I thought of working on predictive maintenances as it is a topic that is predominant in my current position.

However, the inherently sensitive nature of this data made it difficult to find some suitable datasets. My project supervisor was conscious of this roadblock, and he advised me to expand my research to a broader sense of the concepts I wanted to tackle. Indeed, predicting a failure could also apply to a car

breakdown based on some mechanical features and a number of kilometres on the roads, or even to an athlete who would be more likely to get injured after years of intense training.

With that in mind, I started to browse some new datasets about many various subjects, and this further deep dive into the world wide web led me to some very interesting datasets containing some anonymised medical records. After identifying a few interesting sources and as I found the idea of exploring some human data more appealing than looking into machine components, I finally settled on a project that would deal with heart data. Thus, I would not be exploring machine failures but heart failures, and I would not perform statistics based on sensors but on health measurements. This would also give me an opportunity to study the survival analysis theory, which I read about on my supervisor's recommendation.

Since I do not have any prior knowledge in relation with heart failures or health in general, I am also hoping to demonstrate how Data Analytics can shed some light on any type of information, regardless of the field complexity, as long as the techniques are thoroughly applied and the interpretations cautiously approached.

State of the Art

Two literature reviews have been conducted so far, but further research need to be done.

The first review was based on a paper entitled "Trends in Heart Failure Incidence and Survival in a Community-Based Population" (Roger *et al.*, 2004). In this study published in 2004, the researchers were aiming to demonstrate that the survival after heart failure had "improved over time but that secular trends have diverged by sex". To carry out this task, the authors did not use the MIMIC-III database but a cohort study of 4,537 participants who lived in Olmsted County, Minnesota, and had taken part in the Rochester Epidemiology Project conducted in the county. This first example of a similar study was taking into consideration a time element that the MIMIC-III database does not provide, but it could be used as a reference paper for the survival analysis. Interestingly, the research points out how this major public health issue is misunderstood, mainly due to a lack of quality of the available data. The cohort was also mostly comprised of white Americans, but the attributes used were quite different. Further medical research would be necessary in order to understand shared variables, as they could be named differently but defining the same medical observation. The research methodology included a Poisson regression model and proportional hazards modelling, and all the analysis were carried out separately for men and women.

The second review was on "Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database" (F. Li *et al.*, 2021). This research was conducted in 2021 and used the exact same data from MIMIC-III. The objective of this study was also to develop a prediction model for in-hospital mortality of heart failure patients. Some advanced techniques were used on the data, such as extreme gradient boosting (XGBoost), least absolute shrinkage and selection operator (LASSO), multivariate logistic regression, C-index, calibration plot and decision curve analysis. After validation and pairwise comparisons, the

XGBoost model was deemed as the most reliable model to predict heart failure mortality. Although this research provides some important insights into the MIMIC-III data, the time element I have incorporated into my project is not part of its scope.

Data

This first part of this project on mortality prediction requires having access to medical records that include some medical measurements as well as a binary attribute for the life-or-death unambiguous event. After some extensive research, two suitable datasets have been identified which come from a single database: the Medical Information Mart for Intensive Care, also called MIMIC-III database. Both datasets are in the shape of Comma-Separated Value files provided by researchers, one contains the Heart Failures data and the other holds all of the patients' records. Once merged, these two sources will be the primary dataset that supporting the first part of the analysis, including some of the patients' demographics details, comorbidities, laboratory variables, and most importantly the died or survived outcome indicator. The characteristics of both datasets are summarized in the tables below.

URL	https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.0p2ngf1zd
Format	Comma-separated values
Data structure	Structured
Size in MB	0.380
Number of instances	1,177
Number of attributes	51
Types of attributes	Categorical, Discrete, Continuous, Boolean

Table 14 Heart Failures Dataset Characteristics

URL	https://www.kaggle.com/drscarlat/mimic3d
Format	Comma-separated values
Data structure	Structured
Size in MB	11.622
Number of instances	58,976
Number of attributes	28
Types of attributes	Categorical, Discrete, Continuous

Table 15 Patients Dataset Characteristics

For the second part of the project on survival analysis, an additional time indicator will be needed in order to analyse and predict the survival temporality. Two datasets from the UCI database have been found that fulfil this requirement. The Echocardiogram dataset comprises a 'survival' attribute that indicates the number of months each patient has survived, and the Heart Failure Clinical Records dataset includes a 'time' attribute for the same indicator. Both datasets' characteristics are summarised in the tables below.

URL	https://archive.ics.uci.edu/ml/datasets/echocardiogram
Format	Comma-separated values
Data structure	Structured
Size in kB	19
Number of instances	132
Number of attributes	13

Types of attributes	Discrete, Continuous, Boolean

Table 16 UCI - Echocardiogram Dataset Characteristics

URL	https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
Format	Comma-separated values
Data structure	Structured
Size in kB	12
Number of instances	299
Number of attributes	13
Types of attributes	Discrete, Continuous, Boolean

Table 17 UCI - Heart Failure Clinical Records Dataset Characteristics

Methodology & Analysis

The CRISP-DM methodology will be the reference for this project. According to the Data Science Process Alliance, it is the "most popular framework for executing data science projects" (Saltz, 2020), and I was keen to choose a methodology that was widely recognised in the data science community.

CRISP-DM stands for CRoss-Industry Standard Process for Data Mining and was created in late 1996 to develop a concerted approach on data mining. The CRISP-DM Consortium established a standard process model that is described in "CRISP-DM 1.0 – Step-by-step data mining guide" (Chapman *et al*., 2000) and that can be applied to a wide range of industries, tools and applications.

The CRISP-DM reference model consists of six phases that define the data mining project life cycle: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

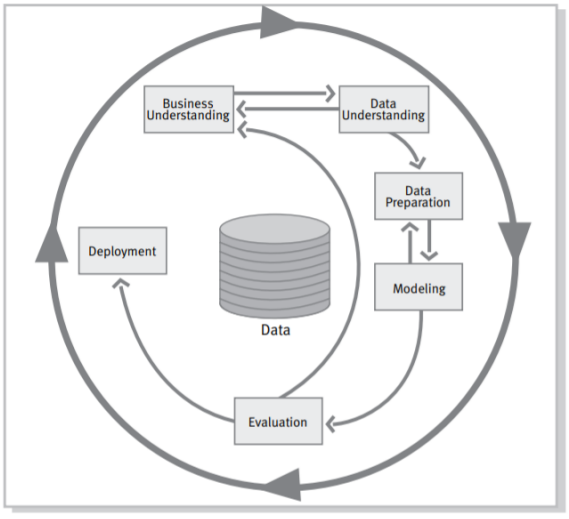


Figure 18 Phases of the CRISP-DM reference model

The CRISP-DM phases will be organised around the two project questions and corresponding datasets and iterated through in the order presented below. The deployment phase has yet to be clarified.

CRISP-DM Phase	Project Question	Report Section
Business Understanding	N/A	1. Introduction
Data Understanding	Mortality prediction (MIMIC-III data)	2. Data Understanding
Data Preparation	Mortality prediction (MIMIC-III data)	4. Analysis
Modelling	Mortality prediction (MIMIC-III data)	5. Modelling
Evaluation	Mortality prediction (MIMIC-III data)	6. Evaluation
Data Understanding	Survival analysis (UCI data)	2. Data Understanding
Data Preparation	Survival analysis (UCI data)	4. Analysis
Modelling	Survival analysis (UCI data)	5. Modelling
Evaluation	Survival analysis (UCI data)	6. Evaluation
Deployment	Mortality prediction (MIMIC-III data) and survival analysis (UCI data)	TBD

Table 18 Report Sections Mapping to the CRISP-DM Phases

Within each phase, project activities will be carried out following roughly the CRISP-DM generic tasks, but not all outputs will be provided as the reference documentation for this research consists in the current project report.

Technical Details

The primary technology I will be using throughout this project is R within the R Studio, which is the IDE of choice when programming in R. Various R packages will be used to clean, explore and pre-process the datasets. SPSS will also be used to get some statistical insights into the data and describe it with visuals, as well as Microsoft Excel, which remains a handy tool when it comes to generating quick and simple manipulation tasks such as filtering, sorting or pivoting data.

Once the data is ready for data mining and modelling, both R and Python will be compared in order to choose the most suitable tool. Data mining techniques and advanced statistics will be studied in the second semester, which should also help in defining the right technology to use in this central part of the project. After some preliminary research ahead of the classes, it would appear that scikit-learn is an efficient Python library for predictive data analytics.

In terms of data mining and machine learning techniques, Bayes classification and logistic regression will be used for the first part of the research, which consists in predicting a categorical value – life or death – and therefore belongs to the scope of supervised learning by binary classification algorithm. Further techniques might be applied following the actual data mining course that will be taught from January. Besides, the survival analysis should give me the opportunity to apply linear regression as well as more specific approaches, such as the Kaplan-Meier plot and the Cox proportional hazards regression techniques, which I will need to research in my own time in order to complete the second part of the analysis. Following a more in-depth literature review, hypothesis testing could also be used to contrast my own findings with existing research work.

With regards to visualisation and final output, Microsoft Power BI is a powerful tool which I would like to incorporate to this project in a way that has yet to be defined, depending on the outcomes of the

modelling phase. For instance, a Power BI report could be used to create interactive visuals that would allow the user to choose variables and observe their impact on the prediction.

Finally, Microsoft Word and PowerPoint will also be used to document the project and provide the required deliverables, and the project's activities will be planned in the free cloud-based project planning solution TeamGantt.

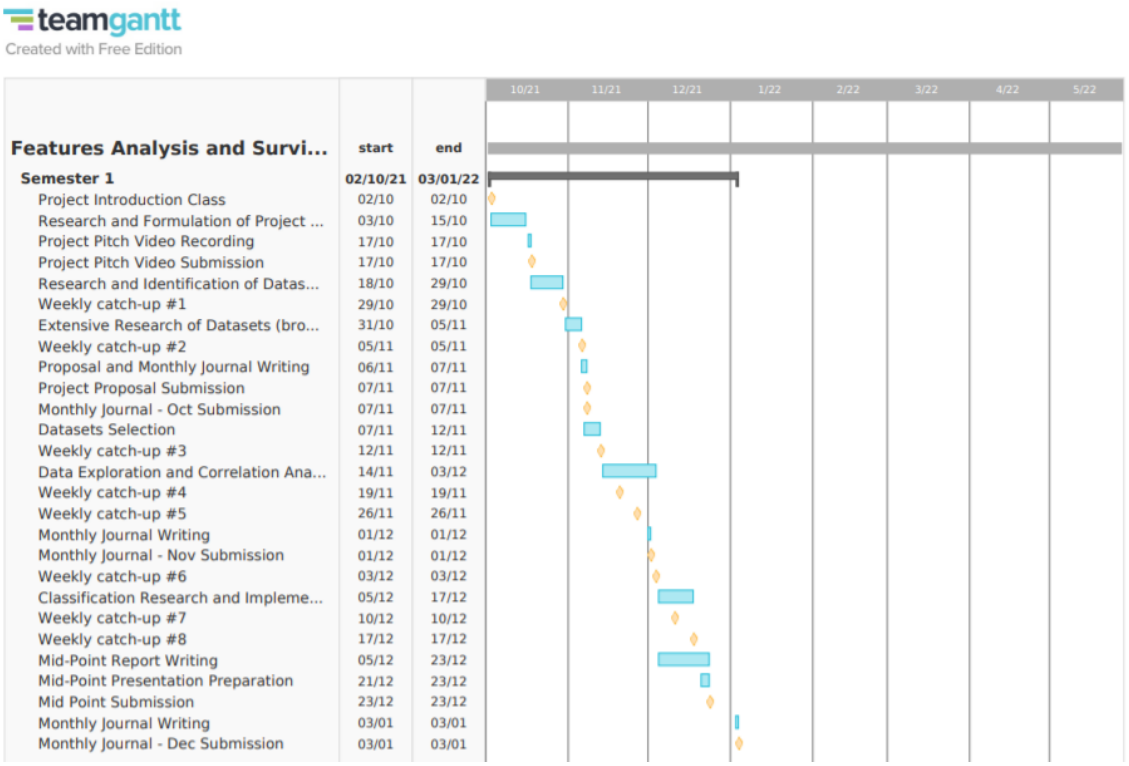
Project Plan

The following project plan has been created with TeamGantt, a free online planification tool that generates Gantt charts, amongst other features.

The project plan includes:

- Moodle submission dates
- meeting dates with project supervisor (22 occurrences)
- project activities in accordance with the CRISP-DM methodology

I have articulated it around the College's final year own schedule, which I found was the most sensible approach to make sure no submission milestone would be missed.



Features Analysis and Survi...	start	end
Semester 1	02/10/21	03/01/22
Project Introduction Class	02/10	02/10
Research and Formulation of Project ...	03/10	15/10
Project Pitch Video Recording	17/10	17/10
Project Pitch Video Submission	17/10	17/10
Research and Identification of Datas...	18/10	29/10
Weekly catch-up #1	29/10	29/10
Extensive Research of Datasets (bro...	31/10	05/11
Weekly catch-up #2	05/11	05/11
Proposal and Monthly Journal Writing	06/11	07/11
Project Proposal Submission	07/11	07/11
Monthly Journal - Oct Submission	07/11	07/11
Datasets Selection	07/11	12/11
Weekly catch-up #3	12/11	12/11
Data Exploration and Correlation Ana...	14/11	03/12
Weekly catch-up #4	19/11	19/11
Weekly catch-up #5	26/11	26/11
Monthly Journal Writing	01/12	01/12
Monthly Journal - Nov Submission	01/12	01/12
Weekly catch-up #6	03/12	03/12
Classification Research and Impleme...	05/12	17/12
Weekly catch-up #7	10/12	10/12
Weekly catch-up #8	17/12	17/12
Mid-Point Report Writing	05/12	23/12
Mid-Point Presentation Preparation	21/12	23/12
Mid Point Submission	23/12	23/12
Monthly Journal Writing	03/01	03/01
Monthly Journal - Dec Submission	03/01	03/01

Figure 19 Project Plan - Semester 1

Semester 2	**17/01/22**	**08/05/22**
Modeling Techniques Exploration	17/01	28/01
Weekly catch-up #9	28/01	28/01
Predictive Model Creation	30/01	18/02
Monthly Journal Writing	01/02	01/02
Monthly Journal - Jan Submission	01/02	01/02
Weekly catch-up #10	04/02	04/02
Weekly catch-up #11	11/02	11/02
Weekly catch-up #12	18/02	18/02
Model Assessment and Evaluation	20/02	04/03
Weekly catch-up #13	25/02	25/02
Monthly Journal Writing	01/03	01/03
Monthly Journal - Feb Submission	01/03	01/03
Weekly catch-up #14	04/03	04/03
Survival Analysis - EDA and Pre-Proc...	06/03	25/03
Weekly catch-up #15	11/03	11/03
Weekly catch-up #16	18/03	18/03
Weekly catch-up #17	25/03	25/03
Survival Analysis Research and Impl...	20/03	01/04
Weekly catch-up #18	01/04	01/04
Monthly Journal Writing	02/04	02/04
Monthly Journal - Mar Submission	02/04	02/04
Assessment and Evaluation	03/04	08/04
Weekly catch-up #19	08/04	08/04
Output Visualisation	10/04	22/04
Weekly catch-up #20	15/04	15/04
Weekly catch-up #21	22/04	22/04
Final Documentation Writing	22/04	08/05
Weekly catch-up #22	29/04	29/04
Monthly Journal Writing	01/05	01/05
Monthly Journal - Apr Submission	01/05	01/05
Final Submission	08/05	08/05
Project Showcase	**09/05/22**	**27/05/22**
Showcase Preparation	09/05	22/05
Project Showcase	23/05	27/05

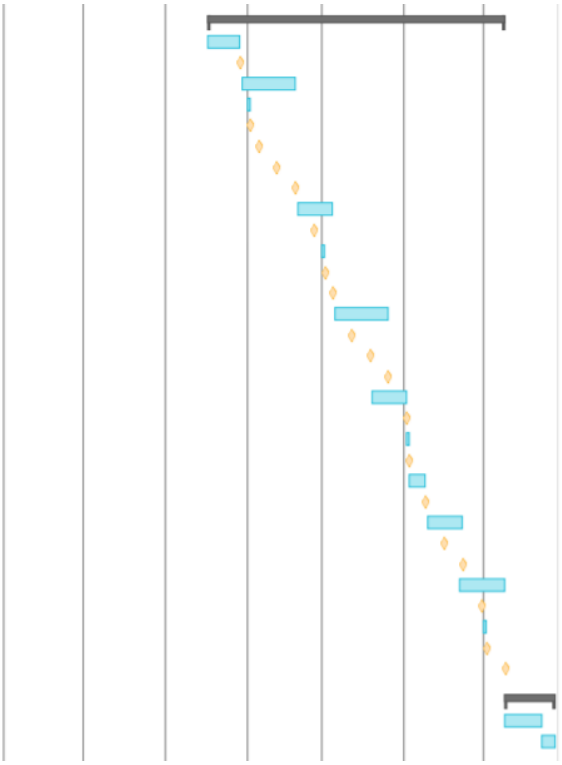


Figure 20 Project Plan - Semester 2

10.2. Reflective Journals

Month	October

What?

The recording of the project pitch video that was due on October 17th was the perfect opportunity to gather my initial thoughts about the project idea. Since I started a new role in the semiconductor industry in January 2021, I was eager to choose a subject matter that was closely related to my work.

The idea I presented in the project pitch video was to conduct an analysis of machines' failures in order to get some insights on their lifespan. Taking this analysis one step further, I could create a forecasting model that would predict the optimum maintenance time of the machines.

The challenges I had identified in the project pitch were with regards to finding some real-world data, which is by nature extremely sensitive, and also my lack of engineering knowledge for a project that announced itself as quite technical.

At the end of October, I was also contacted by my project supervisor Giovani Estrada, who has an extensive experience in the field of semiconductors, and we had our first meeting on October 29th. We agreed to schedule some catch-up meetings on a weekly basis.

So What?

During our first meeting, Giovani informed me that my project idea was validated under the condition that I had to clarify the project's key questions and find some relevant dataset that would allow me to answer these questions. From our discussion, the main practical benefits from the project output should be to minimise downtime, maximise throughput and predict maintenances.

He also shared some resources that will give me a glimpse into the concepts we will be touching on, and he warned me that there will be some mathematics involved. This could be a challenge, but I am hoping that put in the perspective of a work-related project that I chose myself, the concepts will make more sense and I can overcome it.

Another upcoming challenge is the writing of the project proposal that is due for November 7th, especially with regard to the last sections that concern the project's methodology, technical details and planning. I do not feel I have reached the stage of having these mapped out just yet, but I intend to read through past papers and talk to Giovani about it in our next meeting.

Month	November

What?

My supervisor and I have been meeting on a weekly basis over the month of November. The main highlight of this month was a shift in the project's topic, which was initially meant to concern predictive maintenances. But after expanding the concept itself of predicting failures, as suggested by my supervisor, I finally settled on doing my project on heart failures.

The idea of predicting an outcome would still be there, but I will be predicting the patients' life instead of machines' life. This change was made possible after I found a very interesting database, called MIMIC-III, that makes real medical records freely available to researchers. This subject will also help me to put the survival theory I read about on my supervisor's recommendation into practice, with the help of additional datasets containing some data on patients' survival time after a heart failure.

At the end of November, I had therefore chosen a topic that is using the same predictive data analysis techniques, but in an area that is new and a bit more appealing to me as it involves a human factor. I had also found interesting datasets that would help me answer the questions asked in my project.

So What?

The different take on predictive analysis was approved by my supervisor, and I could define two questions that my project will need to answer:

- What are the significant health features that lead to an accurate prediction of the chances to survive a heart failure?
- What are the chances of survival of a patient after suffering a heart failure?

The first question will lead to the creation of a predictive model and the second will set the ground for an application of the survival analysis techniques. These two main features are challenges in themselves, as I know nothing about these topics for now. During this month though, we mainly focused on exploring the datasets, and I could successfully merge them and create a correlation matrix to get an overview of the related variables.

The other challenge I have identified is with regards to the topic itself. Indeed, the basic exploration performed on the data revealed my complete lack of medical knowledge, and I will have to investigate these variables in order to provide a better analysis.

Month	December

What?

December was a critical milestone in the project plan as the mid-point implementation was due on the 23rd at noon. The expected deliverables for this submission were:

- Mid-Point Documentation (to include Proposal as an Appendix)
- Link to Video Presentation, Demonstration and Q&A
- PowerPoint Slides

So What?

The main challenge was therefore to finalise the deliverables on time.

I went through the entire documentation, filled the gaps and made some corrections based on my supervisor's feedback on the previously shared draft. This review was covering both the project proposal – which was to be included as an Appendix – and the first half of the final project report. In parallel, I was also reviewing the R code to improve some of the visuals, clean the comments and delete irrelevant lines. Additionally, the overall structure of the code was reorganised to ensure that the script would flow seamlessly during the presentation.

Once my notes and slides ready, I could finally record my presentation on Microsoft Teams. This step revealed itself extremely time-consuming as it took me numerous trials to fit into the time limit of the video, which was set to ten minutes. However, I did manage to submit all of the deliverables on time.

Now What?

After the mid-point submission, I am planning to focus on my preparation for the exams that will take place remotely at the beginning of January. After the exams, I will take some time off this project so I can get a fresh start on it after the second semester kicks in. At this point, I should review the project plan and continue the implementation of the machine learning model in R.

The weekly catchups with my supervisor will resume early February.

Month	January

Nothing to report for January as the two first weeks of the month were dedicated to the exams and the two remaining weeks were the mid-year (very much needed) break.

I am planning to resume the work on my project next month, starting where I left off with the logistic regression and the Bayes classification models.

Month	February

What?

The second semester being much lighter in terms of number of modules taught, and these modules being eventually relevant for the implementation of my project, I was able to make some good progress.

I could apply different Machine Learning (ML) techniques seen in the Data and Web Mining module to my dataset. One of the key points I understood during these first ML implementations was that visualisations in were not needed. After spending a lot of time trying to find some ways to plot the models' output, my supervisor clarified this point during one of our weekly meetings, which allowed me to better grasp the important concept of data dimensionality.

So What?

The ML techniques that were implemented this month were:

- K-Means clustering
- K-Nearest Neighbour
- Naïve Bayes
- Logistic Regression
- Decision Tree

For each of these techniques, a fair amount of online research was done in addition to the reading of the class material in order to understand the intuition behind them.

Now What?

The next step will consist in applying more techniques to my project and compare the various levels of accuracy, sensitivity and specificity obtained. Considering the medical nature of this work, the sensitivity score will be a crucial factor in the final model selection.

Month	March

What?

I managed to make some good progress on the Machine Learning (ML) techniques implementation this month. This was the continuation of the work started in February and which is at the core of my project. By applying these various algorithms to the data, I became more confident with the overall ML approach and started to play around the parameters, as advised by my supervisor.

So What?

The ML techniques that were implemented this month were:

- Random Forest
- Support Vector Machine
- XGBoost
- Principal Component Analysis

For each of these techniques, a fair amount of online research was done in addition to the reading of the class material in order to understand the intuition behind them.

Now What?

The Principal Component Analysis (PCA) was the biggest challenge this month as I was not sure how to use it correctly. From the discussion with my supervisor, I now understand that I cannot use the components as they are a 'black box' that would be relevant in the context of pattern study, however not insightful enough for data analysis. For this project, PCA should only be used to select features.

I also realised that the Decision Tree algorithm was performed on a subset of variables only, although it should have covered the entire dataset. This tree is useful for features selection as well.

The next month will be the last full month before the final submission of the project, and the following tasks will have to be completed:

- creation of two new subsets based on PCA and Decision Tree analysis
- application of algorithms to all three subsets
- performance analysis
- survival analysis
- report writing

Month	April

What?

The last few weeks were dedicated to adding one more modelling technique to the code (RIPPER), creating new subsets of variables based on the PCA and Decision Tree analysis and evaluating the performance of the models. A ten-fold cross-validation was also added to get a more accurate idea of the sensitivity score obtained with each technique.

So What?

A new function was built to evaluate each model's performance in only one command. The evaluation metrics calculated in the function aimed to print the following scores: Accuracy, Sensitivity, F1 and ROC. During the Data and Web Mining terminal assessment earlier this month, I found out from the literature review that the ROC score was the metric of choice when dealing with imbalanced datasets. Therefore, based on this finding that was confirmed by my supervisor, I decided to focus solely on this metric for the final comparative analysis of all features subsets and models used in my project. The ROC scores and graphs would be generated from the ten-fold cross validation code in order to increase the evaluation accuracy.

Now What?

After the latest weekly catch-up with my supervisor, the following improvements will have to be made to the project:

- do more research to put my work into perspective with existing assumptions
- clarify the results obtained and relate them to a more precise and impactful research question, and build a story from it
- review the report sections already submitted accordingly

With regards to the survival analysis, this would be a nice addition if there is extra time, but the priority is to improve the core analysis.

A Power BI visualisation of the MIMIC-III data would be a strong addition as it involves a different technology.

Finally, the report should not exceed 6,000 words (20-25 pages), excluding references and appendices.

10.3. Get With The Guidelines-Heart Failure Risk Score

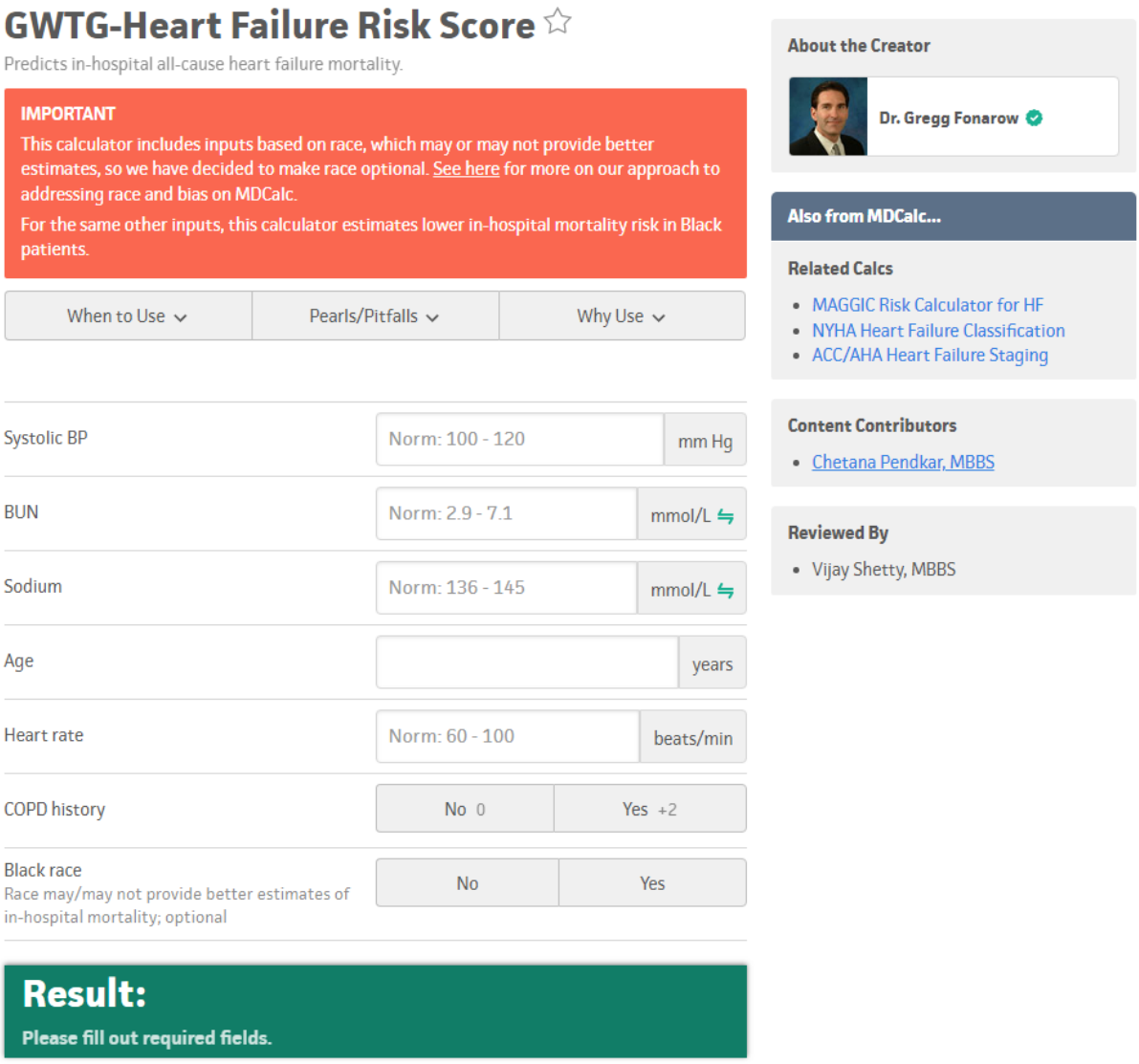


Figure 21 GWTG-HF Heart Failure risk calculator available at: https://www.mdcalc.com/gwtg-heart-failure-risk-score

10.4. Meta-Analysis Global Group in Chronic Heart Failure Risk Calculator

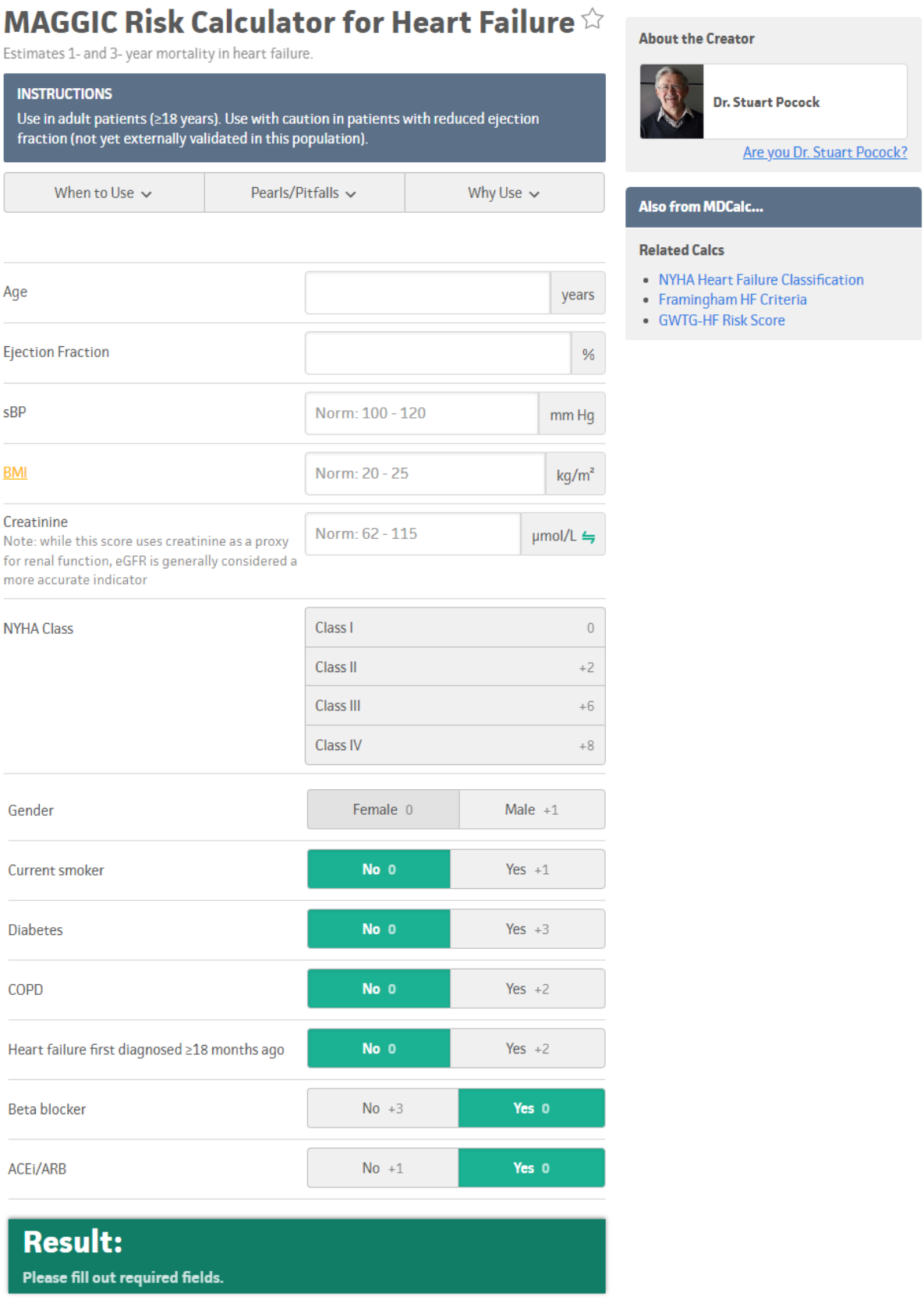


Figure 22 MAGGIC Heart Failure risk calculator available at: https://www.mdcalc.com/maggic-risk-calculator-heart-failure

10.5. MIMIC-III Heart Failures Dataset – Attributes Table

Attribute	Category	Description
outcome	target	survival or death
age	demographics	age at age of heart failure
BMI	demographics	body mass index
gendera	demographics	gender
Diastolic blood pressure	vital signs	diastolic blood pressure
heart rate	vital signs	heart rate
Respiratory rate	vital signs	respiratory rate
SP O2	vital signs	saturation pulse oxygen
Systolic blood pressure	vital signs	systolic blood pressure
temperature	vital signs	body temperature
Urine output	vital signs	urine output
atrialfibrillation	comorbidities	atrial fibrillation
CHD with no MI	comorbidities	ischaemic heart disease
COPD	comorbidities	chronic obstructive pulmonary disease
deficiencyanemias	comorbidities	hypoferric anaemia
depression	comorbidities	depression
diabetes	comorbidities	diabetes mellitus
Hyperlipemia	comorbidities	hyperlipidaemia
hypertensive	comorbidities	hypertension
Renal failure	comorbidities	chronic kidney disease
Anion gap	laboratory variables	anion gap
Basophils	laboratory variables	basophils
Bicarbonate	laboratory variables	bicarbonate
Blood calcium	laboratory variables	calcium
Blood potassium	laboratory variables	potassium
Blood sodium	laboratory variables	sodium
Chloride	laboratory variables	chloride
Creatine kinase	laboratory variables	creatine kinase
Creatinine	laboratory variables	creatinine
EF	laboratory variables	left ventricular ejection fraction
glucose	laboratory variables	glucose
hematocrit	laboratory variables	haematocrit
INR	laboratory variables	international normalised ratio
Lactic acid	laboratory variables	lactate
Leucocyte	laboratory variables	white blood cells
Lymphocyte	laboratory variables	lymphocytes
Magnesium ion	laboratory variables	magnesium
MCH	laboratory variables	mean corpuscular haemoglobin

MCHC	laboratory variables	mean corpuscular haemoglobin concentration
MCV	laboratory variables	mean corpuscular volume
Neutrophils	laboratory variables	neutrophils
NT-proBNP	laboratory variables	NT-proBNP
PCO2	laboratory variables	partial pressure of CO2 in arterial blood
PH	laboratory variables	hydrogen ion concentration
Platelets	laboratory variables	platelet count
PT	laboratory variables	prothrombin time
RBC	laboratory variables	red blood cells
RDW	laboratory variables	red blood cell distribution width
Urea nitrogen	laboratory variables	blood urea nitrogen

10.6. Risk of in-hospital mortality nomogram (Li et al.)

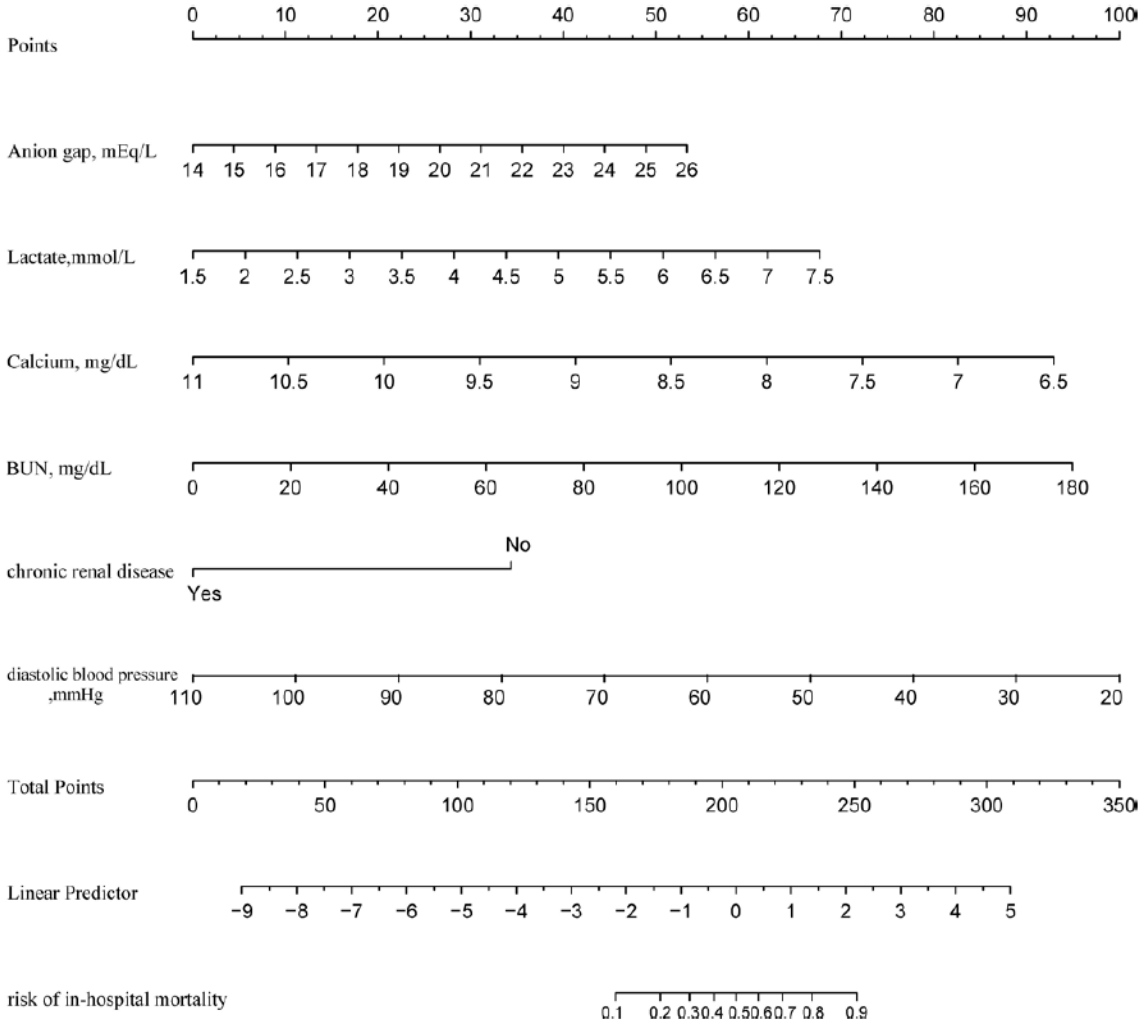


Figure 23 The nomogram developed by Li et al. included all six variables used in the XGBoost model (F. Li et al.*, 2021).*