

FACTORS ASSOCIATED WITH LIFESTYLE AND DIET
IN HEART FAILURE MORTALITY

LAU CHENG CHENG

MASTER OF MATHEMATICS

LEE KONG CHIAN FACULTY OF ENGINEERING AND
SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

AUGUST 2022

FACTORS ASSOCIATED WITH LIFESTYLE AND DIET IN HEART
FAILURE MORTALITY

By

LAU CHENG CHENG

A project report submitted to the Department of Mathematical and Actuarial
Sciences,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the Master of Mathematics
August 2022

ABSTRACT

FACTORS ASSOCIATED WITH LIFESTYLE AND DIET IN HEART FAILURE MORTALITY

Lau Cheng Cheng

With the expansion of COVID-19, the number of deaths from cardiac disease is rising. Due to advancements in healthcare infrastructure, the mortality rate from heart disease is trending downward in emerging nations. Age, serum creatinine, and serum sodium were significant according to the Chi-square test utilised in this investigation. The ANOVA test, in contrast, revealed important variables such as age, creatinine phosphokinase, ejection fraction, and platelets. Due to the low corrected R^2 value, analysis using linear regression is not promising. However, the analysis using logistic regression yields rather encouraging results, attaining an accuracy of 87.8 percent with the help of adjusted $MAPR^2 = 0.392$ and Hosmer-Lemeshow $p\text{-value} > 0.05$. This logistic model is built from the risk factor ejection fraction, serum creatinine, serum sodium, age and time to predict mortality for heart failure sufferers.

ACKNOWLEDGEMENT

I want to sincerely thank and appreciate Dr Goh Yann Ling and Dr Tan Wei Lun, who served as my supervisors, for their direction, patience, support, and advise while I finished my project.

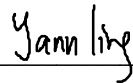
Lastly, I want to thank my family for being supportive of me throughout this project and for their patience with me.

I want to thank God for allowing me to overcome all the challenges before me.

APPROVAL SHEET

This project titled **“FACTORS ASSOCIATED WITH LIFESTYLE AND DIET IN HEART FAILURE MORTALITY”** was prepared by LAU CHENG CHENG and submitted as partial fulfilment of the requirements for the degree of Master of Mathematics at Universiti Tunku Abdul Rahman.

Approved by:



(Dr Goh Yann Ling)
Professor/Supervisor
Department of Mathematical and Actuarial Sciences
Faculty of LKC FES
Universiti Tunku Abdul Rahman

Date: 13/8/2022



(Dr Tan Wei Lun)
Professor/Co-supervisor
Department of Mathematical and Actuarial
Sciences Faculty of LKC FES
Universiti Tunku Abdul Rahman

Date: 13 August 2022

FACULTY OF ENGINEERING AND SCIENCE
UNIVERSITI TUNKU ABDUL RAHMAN

Date: 13 August 2022

SUBMISSION OF PROJECT REPORT

It is hereby certified that that **LAU CHENG CHENG** (ID No: **19UEM01834**) has completed this project entitled “FACTORS ASSOCIATED WITH LIFESTYLE AND DIET IN HEART FAILURE MORTALITY” under the supervision of Dr Goh Yann Ling (Supervisor) from the Department of Mathematical and Actuarial Sciences, Faculty of Engineering and Science, and Dr Tan Wei Lun (Co-Supervisor) from the Department of Mathematical and Actuarial Sciences, Faculty of Engineering and Science.

I understand that University will upload softcopy of my project in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

chenglau

(LAU CHENG CHENG)

DECLARATION

I thus declare that everything in the project, save for quotations and citations that have been properly acknowledged, is original work of mine. I further declare that it has not been submitted for any other degree at UTAR or at any other institution in the past or concurrently.

Name : LAU CHENG CHENG

Student ID : 19UEM01834

chenglau

Signed : _____

Date : 13 August 2022

TABLE OF CONTENTS

	Page
TITLE PAGE	i
ABSTRACT	ii
ACKNOWLEDGEMENT	iii
APPROVAL SHEET	iv
SUBMISSION SHEET	v
DECLARATION	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER	
1.0 INTRODUCTION	1
1.1 Background	1
1.2 Problem statement	2
1.3 Objectives	3
2.0 LITERATURE REVIEW	4
2.1 Heart failure mortality studies	4
2.2 Related Work	7
2.3 Logistic Regression model	9
3.0 Methodology	11
3.1 Data Description	11
3.2 Pre-Processing of Dataset	15
3.3 Analysis	15
4.0 RESULTS	17
4.1 Data Analysis	17
4.2 Inter-variable Analysis	25
5.0 CONCLUSIONS	47
REFERENCES	50
APPENDICES	61

LIST OF TABLES

Table		Page
3.1	Attribute description for the heat failure dataset from the UCI machine learning repository.	12
3.2	(a) The statistical outline of the numeric attributes. (b) The statistical outline of the binary attributes.	13
4.1	Frequency distribution of age group, gender, and categorical risk factor for DEATH_EVENT.	22
4.2	ANOVA test results of continuous risk factor for survived and dead.	24
4.3	chi-square test results for all variables for target of age group > 55.	24
4.4	chi-square test results for all variables for target of age group <= 55.	24
4.5	Regression test for the linear model.	26
4.6	Comparison model before (1) and after (2) excluding influential.	31
4.7	1 st Logistics Regression Model.	34
4.8	2 nd Logistics Regression Model.	36
4.9	3 rd Logistics Regression Model.	38
4.10	4 th Logistics Regression Model.	40
4.11	5 th Logistics Regression Model.	42
4.12	6 th Logistics Regression Model.	44

LIST OF FIGURES

Figures		Page
3.1	Visualisation of variables of the heart failure dataset.	14
4.1	Correlation matrix.	18
4.2	Visualisation of DEATH_EVENT with category variables.	19
4.3	Data visualisation DEATH_EVENT with sex and smoking	20
4.4	Visualisation of continuous variables of the heart failure dataset.	21
4.5	Residual plots for the model.	26
4.6	Scatter plot predictor value against log-odds DEATH_EVENT.	29
4.7	Cook's distance plot.	29
4.8	Residuals vs Leverage plot.	30
4.9	ROC curve for Model 1.	34
4.10	ROC curve for Model 2.	36
4.11	ROC curve for Model 3.	38
4.12	ROC curve for Model 4.	40
4.13	ROC curve for Model 5.	42
4.14	ROC curve for Model 6.	44
4.15	Summary evaluation values for each individual model.	46

CHAPTER 1

INTRODUCTION

1.1 Background

Heart attack, stroke, and heart failure are examples of cardiovascular diseases, which are conditions of the heart and blood vessels. Cardiovascular diseases (CVD) have the highest illness burden and are responsible for a total yearly loss of RM59.85 billion, according to the Ministry of Health Malaysia (2020). Premature deaths became the main cause of loss of productivity in Malaysia, accounting for 59.4 percent of it, versus diabetes accounting for 10 percent and cancer accounting for 30 percent.

Longer survival times for people with cardiovascular disorders are a result of better medical therapy for cardiac ailments. As the population ages and survival rates rise, heart failure occurs more frequently. Due to the creation and application of patient-specific life-prolonging medicines, there will likely be an increase in the cost of heart failure care. The total costs are anticipated to increase by 127 percent during the next 18 years, according to Mozaffarian et al. (2015). Because of the high cost of management, mortality, prevalence, and morbidity associated with heart failure, it has become a serious healthcare issue.

Heart failure happens when the heart cannot function efficiently as a pump supporting the blood flow through the body. Cough, wheezing, tiredness, worsen short breath, legs/abdomen swollen, and difficulty performing an active physical task are symptoms of heart failure.

Pillai and Ganapathi (2013) conclude that the leading cause of disease burden in South Asia by heart failure and is foreseen to rise. Managing the diet and lifestyle can stop the contribution of heart failure toward the economic burden.

1.2 Problem statement

The detection of heart failure depends entirely on symptoms and signs fraught with difficulties. Unpleasant lifestyles, such as a higher BMI, obesity, cholesterol, high salt meal, high sugar, smoking, liquor consumption, or drug abuse, play a vital role in developing health failure. High blood pressure, diabetes and hyperlipidaemia are the most common risk factors that worsen people with cardiovascular disease. Therefore, early detection and management are needed to prevent the disease from worsening and improve patient outcomes. The selected heart failure dataset has 12 variables, including these risk factors and can model the heart failure mortality caused by significant features.

Ahmad et al. (2017) analysed the dataset using Cox model and Kaplan Meier curves. Later Zahid et al. (2019) used this dataset to predict mortality based on gender. Chicco and Jurman (2020) further the dataset analysis by

applying machine learning classifiers and featured ranking to forecast the aliveness of heart failure patients. Das et al. (2021) evaluated and compared the accuracy of five different data mining algorithms using the same dataset.

1.3 Objectives

The high-risk cardiovascular group requires early detection and must go through health care to reduce their mortality risk and improve their daily lives. As a result, the goal of this study is to identify the risk factors that must be addressed to optimise survival among cardiac malfunction patients, enhance the detection of heart failure patients' mortality danger, and more effectively gauge the severity of a patient's condition.

CHAPTER 2

LITERATURE REVIEWS

2.1 Heart failure mortality studies

Lippi and Sanchis-Gomar (2020) reported that global heart failure is 64.34 million cases. Heidenreich et al. (2013) wrote that by 2030, heart failure in the United States will increase by 25 percent, from 2.42 percent in 2012 to 2.97 percent in 2030, and this brings the increase of heart failure by 46 percent. Lam (2015) summarised that in Southeast Asia, 9 million people suffer heart failure, with Malaysia and Singapore at 6.7 percent and 4.5 percent, respectively.

Although heart failure patient mortality has dropped, Bytyçi and Bajraktari (2015) noted that it is still unacceptably high. The mortality rates brought on by preserved ejection fraction are, nevertheless, lower than those brought on by reduced ejection fraction. According to Ponikowski et al. (2014), 17–45 percent of hospitalised patients with heart failure died within one year, and the majority passed away within five years. In 23 percent of COVID-19 heart failure patients, the mortality rate reached 52 percent, according to Zhou et al. (2020).

According to Dunlay et al. (2009) , risk factors known to be associated with heart failure range from lifestyle characteristics (smoking, physical

inactivity) to common medical conditions (hypertension, ischemic heart disease, atrial fibrillation, diabetes mellitus, obesity). Li et al. (2020) state that heart failure is the dominant mortality cause among older people as the charges affected by heart failure is estimated to be 1 percent for age above 50. Therefore, heart failure patients are categorized into groups aged < 55 years old and age \geq 55 years old.

Obesity has been identified by Savji et al. (2018) as a significant risk factor for heart failure with preserved ejection fraction (HFpEF). Since the risk of HFpEF increases by 34 percent for every standard deviation increase in body mass index, women are more likely than men to be obese globally (BMI). According to Rosano, Vitale and Seferovic (2017), diabetes mellitus patients have an incredibly high rate of acute and chronic heart failure, with 25 percent of patients experiencing chronic heart failure. The percentage demonstrates that people with diabetes have a greater risk of having heart malfunction.

According to Benjamin et al. (2018), there is a risk of 1.6 times more heart failure evolving for an individual with systolic blood pressure (SBP) > 160/90 mmHg than those with SBP >120/90 mmHg. Lloyd-Jones et al. (2002) stated that hypertension contributed 39 percent for men and 59 percent for women in developing heart failure.

Low haemoglobin continued to be a significant, independent indication of warded or death due from heart failure, according to Anand et al. (2004) investigation, which took into account a number of other factors. Anaemia has

been linked to higher mortality and morbidity rates in heart failure, according to Ezekowitz, McAlister and Armstrong (2003) study. According to Diana Rodriguez (2009), there is a direct link between anaemia and heart illness, with more than 48% of persons who are diagnosed with heart failure also having anaemia. The heart beats quickly and violently in response to low blood haemoglobin levels in order to meet the body's need for oxygen.

In the World Health Organization (2020) report, 18 percent of all deaths due to cardiovascular disease could be attributed to smoking. In Kamimura et al. (2018), smokers tend to be exposed to the risk of developing cardiovascular diseases.

Creatine phosphokinase (CPK), an enzyme found in muscles, has been suggested by Aujla RS and Patel (2022) to be a marker of cardiac damage. Creatine phosphokinase (CPK) levels should be between 10 and 120 micrograms per litre in a healthy individual. Heart attack and inflammation of the heart muscle can both be predicted by the abnormal levels.

The percentage of blood that leaves the heart with each beating is known as the ejection fraction, according to Healthwise Staff (2021). 55 percent or greater is a common range for the left ventricular ejection fraction. Reduced ejection fraction can be caused by heart disease, poor cardiac muscle, heart attack-related cardiac muscle damage, cardiac valve dysfunction, and chronic uncontrolled hypertension.

The usual number of platelets is 150.000 to 400.000. Mojadidi et al. (2016) reported that thrombocytopenia is high in heart failure reserve ejection fraction (HFrEF) patients. Therefore, platelet counts can be used to assess the patient with HFrEF. Mayo Clinic Staff (2021) stated that serum creatinine with the reading of 0.74 to 1.35 mg/dL is considered normal for men and 0.59 to 1.04 mg/dL for women.

According to Mahmood et al. (2019), salt, in tiny levels, is necessary for fluid balance as well as blood pressure, neurons, and muscle function. The sodium concentration in blood ranges from 135 to 145 milliequivalents/litre on average. Heart failure, however, could result from a value of less than 135 milliequivalents/litre.

2.2 Related Work

Cox regression and Kaplan-Meier curves were used in Ahmad et al. (2017) analysis of the dataset to model it utilising all the attributes and to support major risk variables that affect the patients' status. According to the study, factors such as old age, renal disease, hypertension, ejection fraction, and anaemia all have a role in heart failure patients' death.

The Cox's proportional hazards model in Zahid et al. (2019) is fitted to the variable that was chosen using lasso. To assess the model's goodness of fit, the likelihood ratio test was utilised, and the C-index was employed to gauge the effectiveness of the model. The results demonstrated that the risk factors

anaemia, smoking, and diabetes were harmful to female patients while platelets and ejection fraction were harmful to male patients.

When ten different survival prediction classifiers were used by Chicco and Jurman (2020) to predict the prognosis of patients, the results revealed that Random Forest had the highest accuracy (74 percent) of all the classifiers. The first two most significant factors (serum creatinine and ejection fraction) to fit into Random Forests were identified using four biostatistical methods. The accuracy of the three features—ejection fraction, serum creatinine, and time—selected for use in the logistic regression models was 83.3 percent, and it was 83.8 percent for all the features. The results show that serum creatinine and ejection fraction are sufficient to build a model that can predict a patient's prognosis for heart failure.

Le et al. (2020) used decision trees and multilayer perceptron neural networks as their methodologies (MLP). Prior to fitting the outliers into the chosen models, the authors first eliminated the outliers from the dataset using the inter-quartile range. The accuracy of the decision tree is 86.57 percent, and the best model in comparison to other studies, the MLP, produces an accuracy of 88 percent.

Eletter et al. (2020) used support vector machines, generalized linear models, deep learning, random forest, and a naïve base to build classifiers. The accuracy of 87.78 percent obtained from generalized linear models and support vectors was the highest.

The dataset was examined using Naive Bayes Tree, Naive Bayes Classification, Bayes Network, Classification Regression, and LiBLinear by Das et al. (2021). The outcome demonstrates that each of the five machine learning methods is predictive with a respectable level of accuracy, but the Bayes network has the best accuracy with a rate of 79.28 percent.

Wang (2021) used eighteen different machine learning techniques to compare their performance on the dataset. The z-score with SMOTE is more accurate in predicting heart failure when compared to the z-score and min-max accuracy without SMOTE.

Zaman et al. (2021) want to enhance the technique for forecasting the survival of heart failure patients using the same dataset. To remedy the imbalance in the target class, SMOTE was used. K-Means, Fuzzy C-Means clustering, Random Forest, XGBoost, and Decision Tree are just a few of the methods that the authors used. They outperformed supervised learning, with accuracy rates of 62.24 percent for K-Means and 52.45 percent for Fuzzy C-Means, respectively.

2.3 Logistic Regression model

If the dependent variable is dichotomous, logistic regression is the most suitable statistical technique to predict the outcome (Fernandes et al., 2020). Zangmo and Tiensuwan (2018) stated that the logistic regression model identified significant factors in the patient's survival. The best-fitted model is

obtained by deviance analysis. Sakinc and Ugurlu (2013) stated that logistic regression could explain and check the hypotheses of binary, discrete or continuous variables. According to Hosmer, Lemeshow and Sturdivant (2013), if the outcome is two levels (0 and 1), the conditional mean is between zero and one. The equation gives the logit of the univariable logistic regression model,

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x$$

The equation gives the logit of the multivariable logistic regression model,

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

CHAPTER 3

METHODOLOGY

3.1 Data Description

The clinical heart failure record dataset was used in the investigation (UCI, 2020). The selection of this dataset was made possible by the fact that it combines categorical and continuous variables and includes the majority of the risk factors linked to cardiovascular disease.

The dataset consists of 299 heart failure patients' medical records, which are organised into 13 features in columns and rows. Table 3.1 displays a description of the features. There are 12 independent clinical features and 1 target variable, which is classified as either 0 (meaning patients survived) or 1 (meaning patients died). The information is gathered for a group of sufferers caring left ventricular systolic dysfunction in NYHA classes III and IV who are between the ages of 40 and 95 during their follow-up period, which lasts between 4 and 285 days.

Table 3.1: Attribute description for the heart failure dataset from the UCI machine learning repository.

Attribute	Description	Type of Attribute	Attribute Value Range
age	the patient's age	Numerical	[40, 95]
anaemia	low haemoglobin concentration	Binary	0 = non-anaemia 1 = anaemia
high_blood_pressure (BP)	if the patient has high blood pressure	Binary	0 = non-HB 1 = HB
creatinine_phosphokinase (CPK)	level of the CPK enzyme in the blood	Numerical	[23, 7861]
diabetes	if the patient has diabetes	Binary	0 = non-diabetes 1 = diabetes
ejection fraction (EF)	percentage of blood leaving the heart at each contraction	Numerical	[14, 80]
sex	gender	Binary	0 = Female 1 = Male
platelets	thrombocytes count in the blood	Numerical	[25.01, 850.00]
serum_creatinine	serum creatinine concentration in the blood	Numerical	[0.50, 9.40]
serum_sodium	serum sodium concentration in the blood	Numerical	[114, 148]
smoking	smoker or non-smoker	Binary	0 = non-smoker 1 = smoker
time	number of times following-up examination	Numerical	[4, 285]
target	prediction attribute	Binary	0 = survived 1 = death

The statistical characteristics of the numerical data, including the lowest, maximum, mean, standard deviation, and missing values, are reported in Table 3.2(a). The statistical details of the binary attributes, including label, count, proportion, and missing values, are provided in Table 3.2(b). Labels 0 (patients who lived) and 1 (patients who died) of the target class, which together accounted for 68 percent and 32 percent of the dataset, each had 203 occurrences and 96 instances, respectively. In the binary and numeric properties of the Heart Failure dataset, there are no missing values to be detected.

Table 3.2: (a) The statistical outline of the numeric attributes. (b) The statistical outline of the binary attributes.

(a)					
Attribute	Min.	Max.	Mean	StdDev	Missing
age	40	90	60.83	11.89	0
CPK	23	7861	581.84	970.29	0
ejection fraction	14	80	38.08	11.83	0
platelets	14	80	30.08	11.83	0
serum creatinine	0.5	9.4	1.39	1.03	0
serum sodium	113	148	136.63	4.41	0
time	4	285	130.26	77.61	0
(b)					
Attribute	Label	Count	Proportion	Missing	
anaemia	0	170	57%	0	
	1	129	43%		
high blood pressure	0	194	65%	0	
	1	105	35%		
diabetes	0	174	58%	0	
	1	125	42%		
sex	0	105	35%	0	
	1	194	65%		
smoking	0	203	68%	0	
	1	96	32%		
target	0	203	68%	0	
	1	96	32%		

Age, anaemia, creatinine phosphokinase (CPK), diabetes, ejection fraction (EF), blood pressure (BP), platelets, serum creatinine, serum sodium, gender, and smoking were the risk factors linked to lifestyle and diet that were recorded. They were seen as potential independent variables that might be used to account for heart failure-related death. Age, CPK, EF, platelets, serum creatinine, serum sodium, and time are all quantifiable data; anaemia, BP, diabetes, gender, and smoking were considered to be qualitative data. Figure 3.1 displays a visualisation of the Heart failure dataset's 13 variables.

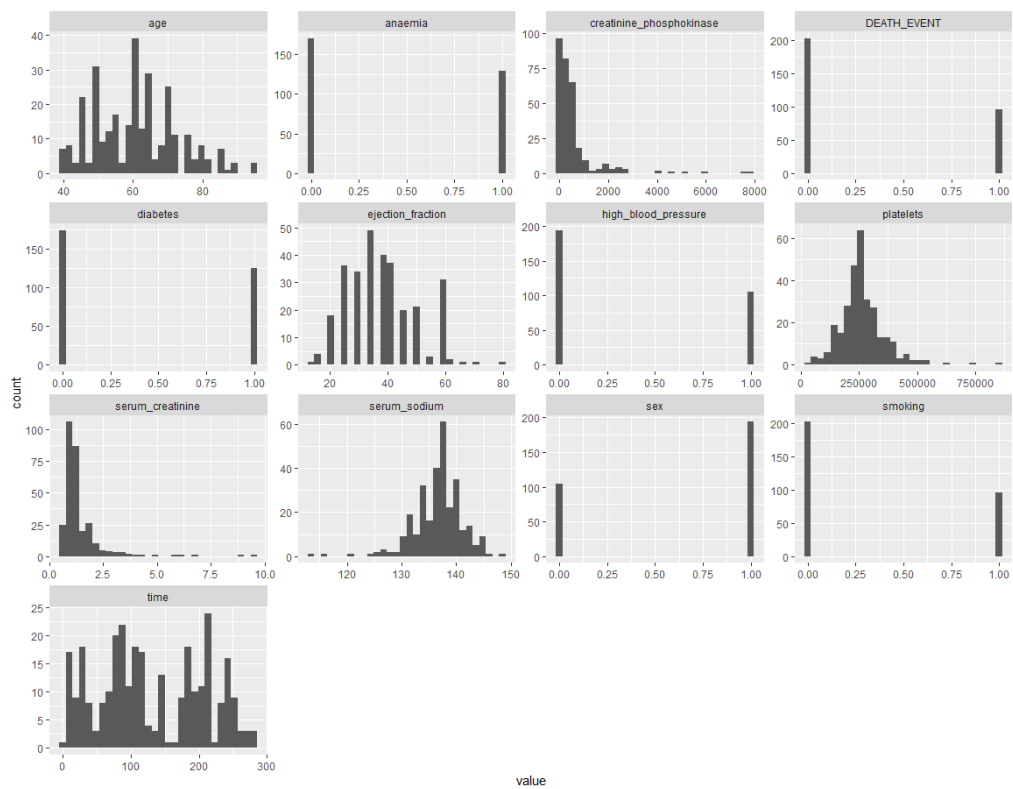


Figure 3.1: Visualisation of variables of the heart failure dataset.

3.2 Pre-Processing of Dataset

The time variable was excluded from the data set used for prediction since this information will not be available at the time of prediction. Generally, the normal serum creatinine levels for adult men are 0.74 to 1.35 mg/dL and adult women is 0.59 to 1.04 mg/dL. By using this standard, we can group the data of serum creatinine into two categories normal and non-normal. The normal value of platelet count is 150000-450000, and any value outside this range is considered abnormal. The platelet data could be grouped into "0" for normal and "1" for abnormal. The values of ejection fraction were divided into normal 41% - 75% and abnormal (<41% or >75%). A normal blood sodium level is between 135 and 145 mEq/L and value outside these range is abnormal. Creatinine phosphokinase was divided into category normal (10 – 120 mcg/l) and abnormal (<10 or >120 mcg/L). The heart failure patients were categorized into group age < 55 years old and age \geq 55 years old.

3.3 Analysis

In this study, the Chi-square test, a famous categorical statistical test, is used to determine the factors closely associated with the DEATH_EVENT. A p-value < 0.05 indicates insufficient evidence to reject the null hypothesis and inadequate evidence to suggest that the factor is independent. The dataset was first analysis by chi-square as the continuous variable had been categorise.

Then the dataset will also be analysis by ANOVA test for continuous features (age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, time). This is to clarify the significant of the continuous variable. Pearson correlation coefficients will be calculated to check for collinearity between the univariate prognostic indicators.

The dataset is further analysed by constructing linear and logistic regression models to explore relationships among variables. The linear models form by using age, creatinine_phosphokinase, ejection_fraction, serum_creatinine, serum_sodium and platelets as dependent variables. The logistic regression with DEATH_EVENT as the model outcome was generated and fit the combined data. All variables do not need to be normalised into homoscedasticity to execute logistic regression.

Due to the data distribution for the target class, DEATH_EVENT, which is not balanced, we must handle the data imbalance problem. Upsampling will be applied to the dataset to overcome the data imbalance problem.

Variance inflation factors (VIF) will be calculated for both linear and logistic if the model has more than one independent variable. (Hair et al., 2010) indicated that the multicollinearity issue does not exist when the variance inflation factors are less than five, and the model functions correctly.

CHAPTER 4

RESULTS

4.1 Data Analysis

In this study, 299 patients with heart failure were involved, and during the visiting periods, 203 patients died and 96 survived. Resampling at the latter phases of modelling is necessary since the results demonstrate that the levels are unbalanced because there are twice as many patients who survived. 68 percent more men than women have heart failure, according to statistics (35 percent).

According to the correlation matrix (Figure 4.1), age, serum creatinine, ejection fraction, and serum sodium are all strongly connected with DEATH EVENT. Except when it comes to sex and smoking, there is no significant link between attributes.

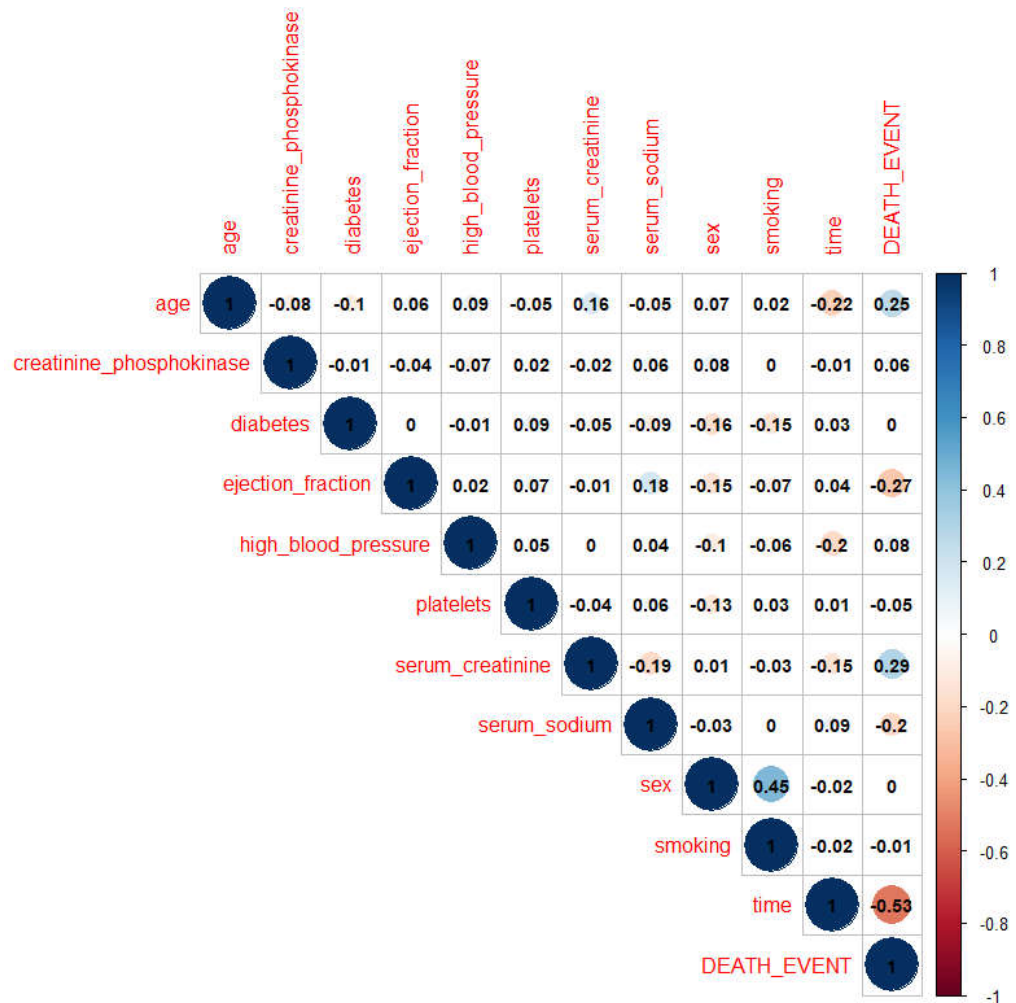


Figure 4.1: Correlation matrix.

The presence of anaemia will result in a higher mortality rate for patients with heart failure than for patients who survive with anaemia, according to the percent stacked bar chart in Figure 4.2(a). As predicted by the correlation matrix, the difference is relatively negligible. According to Figure 4.2(b), there is no difference in the number of fatalities between people with diabetes and those without the disease. Figure 4.2(c) demonstrates that, compared to patients who survive, individuals who die appear to have high blood pressure more frequently. As predicted by the correlation matrix, the difference is relatively negligible.

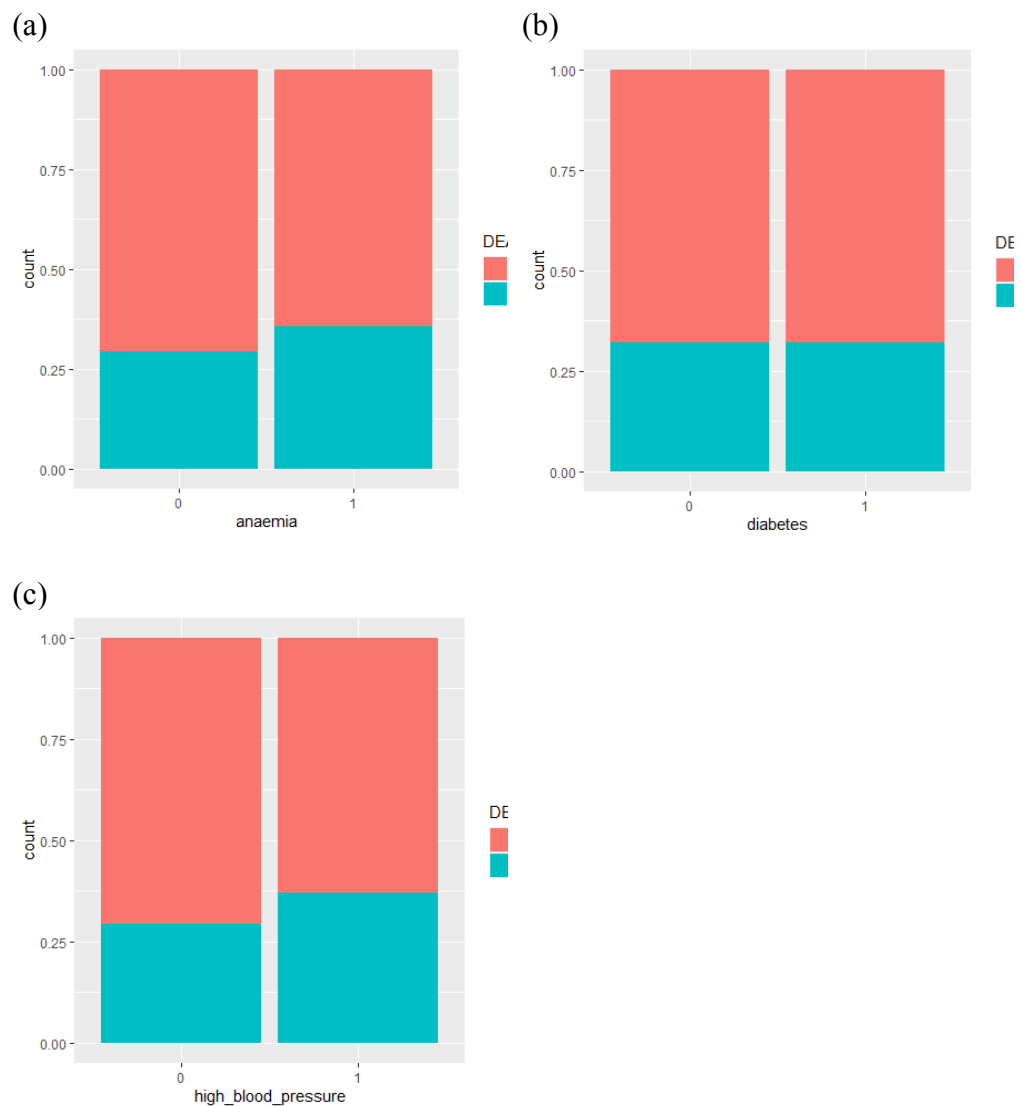


Figure 4.2: Visualisation of DEATH_EVENT with category variables.

Figures 4.3(a) and (b) demonstrate that there is little association between sex and smoking and patient death. The proportion of males and females pass away because of heart failure are equally the same, but there were more dead cases in smoking patients than non-smoking patients. Figure 4.3(c) shows a strong link between smoking and having sex.

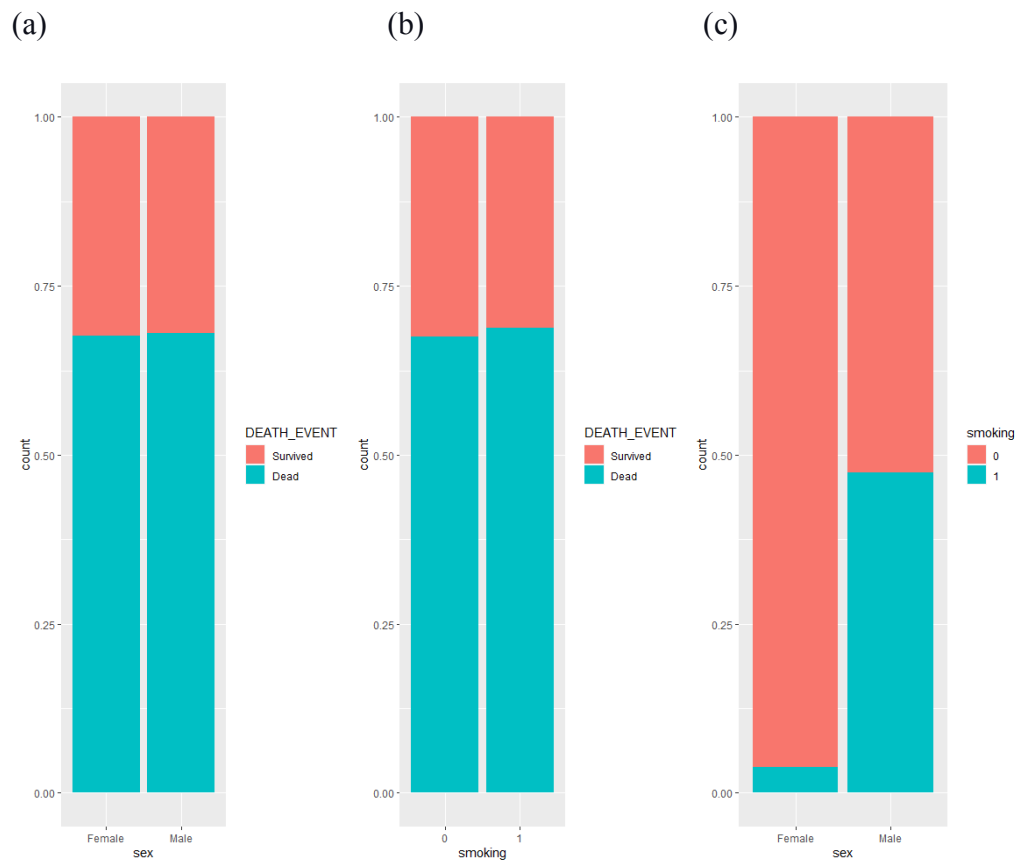


Figure 4.3: Data visualisation DEATH_EVENT with sex and smoking.

A higher mortality risk from heart failure is predicted for older patients, according to Figure 4.4. Sufferers with heart malfunction frequently have serum salt, platelets, and creatinine levels that are within the normal range in Figure 4.4. Serum creatinine is present in greater amounts in non-survivors than in survivors. For both groups, serum salt and platelets are within acceptable limits. Ejection fraction and CPK are found to be abnormal in the majority of heart failing patients. Although it is abnormally low (about 40%) among survivors, the level of ejection fraction is even lower (around 32%) in non-survivors.

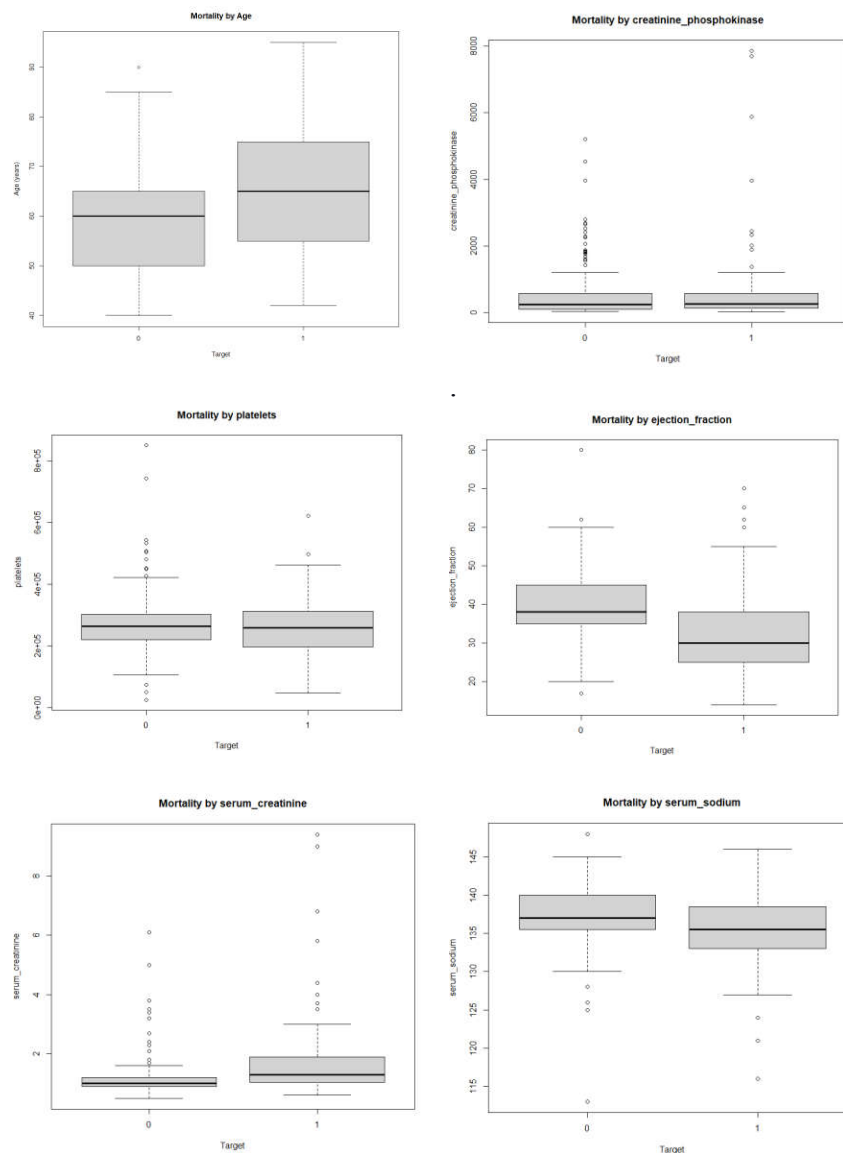


Figure 4.4: Visualisation of continuous variables of the heart failure dataset.

Table 4.1: Frequency distribution of age group, gender, and categorical risk factor for DEATH_EVENT.

	0 : Survived patients	1 : Dead patients	Row Total	chi- square	p-value
<= 55	87	25	112	7.1654	0.0074
> 55	116	71	187		
female : 0	71	34	105	0.0000	1.0000
male: 1	132	62	194		
no anaemia	120	50	170	1.0422	0.3073
anaemia	83	46	129		
no HB	137	57	194	1.5435	0.2141
HB	66	39	105		
no diabetes	118	56	174	0.0000	1.0000
diabetes	85	40	125		
no smoking	137	66	203	0.0073	0.9318
smoking	66	30	96		
CPK abnormal	145	77	222	2.1885	0.1390
CPK Normal	58	18	76		
serum creatinine abnormal	69	55	124	13.6350	0.0002
serum creatinine normal	134	41	175		
EF abnormal	143	77	220	2.7143	0.09945
EF normal	60	19	79		
platelets abnormal	25	16	41	0.7077	0.4002
platelets normal	178	80	258		
serum sodium abnormal	58	52	110	17.2770	3.231e- 05
serum sodium normal	145	44	189		
Total	203	96	299		

A chi-square test and a one-way ANOVA test are used as the first steps in the dataset analysis. In Table 4.1, the p-value for the chi-square test is shown, and in Table 4.2, the p-value for the one-way ANOVA test is shown. To ascertain whether there was a meaningful correlation between the target and risk factors, the test was run for each variable. Age, ejection fraction, creatinine

phosphokinase, and platelets underwent an ANOVA test with a p-value of less than 0.05.

The chi-square test for independence of DEATH_EVENT and age group yields a test statistic $\chi^2 = 7.1654$, and a p-value of 0.0074. The null hypothesis that there is no correlation within DEATH_EVENT and age group is rejected because $p < \alpha$ ($0.0074 < 0.05$). The evidence is sufficient to suggest that there is indeed an association between DEATH_EVENT and the patient's age group. Based on the data, it is likely that the age group > 55 is significant with an increase of DEATH_EVENT.

The second chi-square test for independence of DEATH_EVENT and serum_creatinine yields a test statistic $\chi^2 = 13.6350$, and a p-value of 0.0002. Since $p < \alpha$ ($0.0002 < 0.05$), the null hypothesis that there is no relationship between DEATH_EVENT and serum_creatinine is rejected. The evidence is sufficient to suggest that there is indeed an association between DEATH_EVENT and whether a patient has serum_creatinine abnormality. Based on the data, it is likely that having abnormal serum_creatinine is associated with an increased risk of DEATH_EVENT.

The third chi-square test for independence of DEATH_EVENT and serum_sodium yields a test statistic $\chi^2 = 17.2770$, which has a chi-square distribution with 1 degree of freedom under the null hypothesis. This test resulted in a p-value of 3.231×10^{-5} . Since $p < \alpha$ ($3.231 \times 10^{-5} < 0.05$), the null hypothesis that there is no association between DEATH_EVENT and serum_sodium is

rejected. The evidence is sufficient to suggest that there is indeed an association between DEATH_EVENT and whether a patient has serum_sodium abnormality. The information indicates that there is a strong likelihood that having abnormal serum sodium is linked to a higher risk of DEATH EVENT.

Table 4.2: ANOVA test results of continuous risk factor for survived and dead.

Variable	Mean	St. Dev.	F-value	df	p-value
age	60.83	11.89	20.44	1	0.0000
serum creatinine	1.39	1.03	1.173	1	0.2800
ejection fraction	38.08	11.83	23.09	1	0.0000
serum sodium	136.60	4.41	0.719	1	0.3970
creatinine phosphokinase	581.80	970.29	28.16	1	0.0000
platelets	263358.00	97804.24	11.77	1	0.0007

Further analysis of each category variables relatives to target by age group using chi-square as in Table 4.3 and Table 4.4 show that serum creatinine and serum sodium are significant in age group > 55. Unfortunately, none of the variables is significant for the age group < 55.

Table 4.3: chi-square test results for all variables for target of age group > 55

variables	sex	anaemia	CPK	diabetes	EF	HBP
p-value	0.7996	0.6886	0.3805	1.0000	0.2257	1.0000
variables	platelets	creatinine	sodium	smoking		
p-value	0.7136	0.0006	0.0011	0.4521		

Table 4.4: chi-square test results for all variables for target of age group <= 55

variables	sex	anaemia	CPK	diabetes	EF	HBP
p-value	0.5254	0.7051	0.1141	1.0000	0.2157	0.0536
variables	platelets	creatinine	sodium	smoking		
p-value	0.6717	0.3192	0.0536	0.1049		

4.2 Inter-variable Analysis

An effective model for this dataset would use age, serum creatinine, serum sodium, ejection fraction, creatinine phosphokinase, and platelets as the output responses and every other factor as an independent variable, with the exception of time and DEATH EVENT. This is evident from the initial data analysis. DEATH EVENT was "dropped" in this part since it is a binary variable that performs better as an output on its own. Time was left out since it is mostly an after-the-fact measure rather than a variable used to describe something.

By applying various dependent and independent factors, we formed eighteen linear models. As this study has a two hundred and ninety-nine sample size; therefore, there should be less than 20 independent variables in each model. In a model with less than two independent factors, we used a stepwise selection process to determine which subgroup of independent factors performed the prediction model excellent. Log-transformations and box cox-transformations were also applied to improve the model.

The outcome for the remaining linear models is placed at Appendix C, table C.1 – C.18. The best model generated was the stepwise-selected box cox-transformation model with an adjusted-R² of 13.4 percent.

Table 4.5: Regression test for the linear model below:

$$(\text{serum_creatinine})^{-0.99} \sim \text{age} + \text{ejection_fraction} + \text{high_blood_pressure} + \text{serum_sodium}$$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Error	p-value	Estimation	Std. Error	p-value
serum_creatinine						
intercept	-0.9304	0.5496	0.0915	-0.9502	0.5387	0.0788
sex	-0.0179	0.0413	0.6657			
anaemia	0.0100	0.0355	0.7787			
high_blood_pressure	0.0623	0.0362	0.0864	0.0601	0.0356	0.0924
smoking	0.0492	0.0413	0.2349			
diabetes	0.0010	0.0356	0.9785			
age	-0.0069	0.0015	0.0000	-0.0069	0.0014	0.0000
creatinine_phospokinase	0.0000	0.0000	0.4661			
platelets	0.0000	0.0000	0.9375			
ejection_fraction	0.0030	0.0015	0.0418	0.0030	0.0015	0.0421
serum_sodium	0.0152	0.0040	0.0002	0.0156	0.0039	0.0001
	Ajd R ² = 0.1215, p-value < 0.05			Ajd R ² = 0.1337, p-value < 0.05		

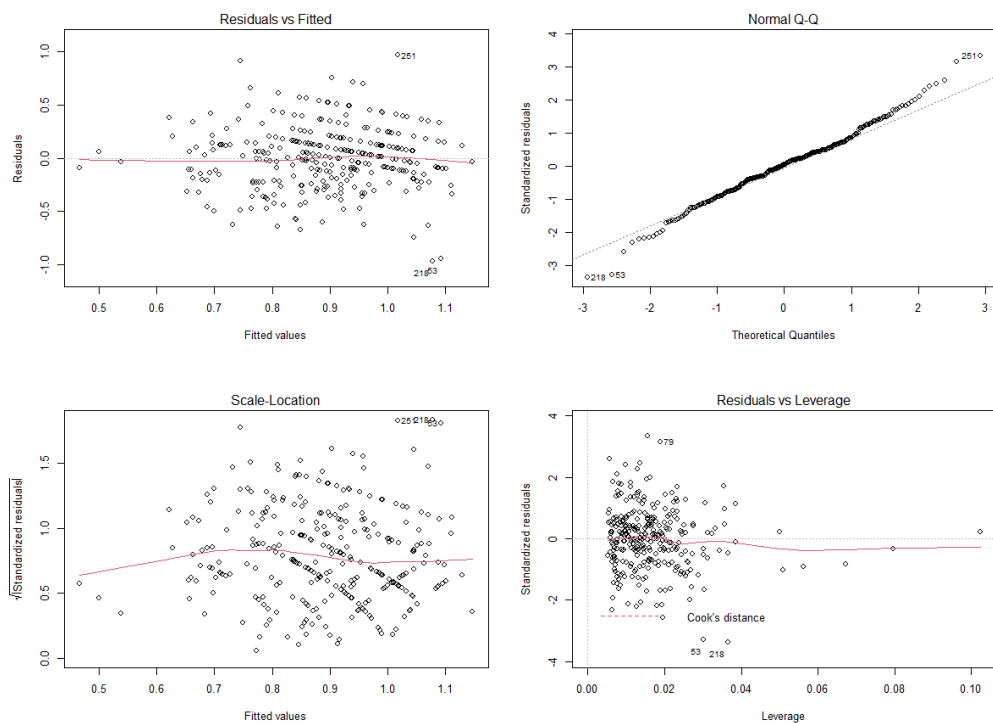


Figure 4.5 Residual plots for the model:
 $(\text{serum_creatinine})^{-0.99} \sim \text{age} + \text{ejection_fraction} + \text{high_blood_pressure} + \text{serum_sodium}$

The selected model is presented in Table 4.5 and has an adjusted-R² of 13.4 percent, which is the highest of any model used in this investigation. The log-likelihood function is maximised at a value of $\lambda = -0.99$, according to the Box-Cox transformation of the serum creatinine data. The low R² value demonstrates the poor performance of the age + ejection fraction + hypertension + serum sodium model. The linear regression appears to have acceptable validity, according on the residual plots. A nearly uniform distribution of residuals may be seen in the residuals vs. fitted plot. The usual Q-Q image reveals several small outliers. The residual data are said to be normally distributed if the Shapiro-Wilk normality test results are $p > 0.05$, which accepts the null hypothesis. The scale location plot can be used as a direct indicator of unequal variance. In the residuals vs. leverage graphic, there are no observations that are extraordinarily heavily weighted.

The majority of linear models have low adjusted-R², clear assumption violations in the residual plots, or both. The logistics regression model, which uses the DEATH EVENT as the dependent variable, was developed to continue the inquiry. It is a good idea to examine the basic assumptions for logistic regression prior to adapting a model to a dataset. These assumptions include a binary target, independent rather than paired data, predictor variables that do not substantially correlate with one another, predictor continuous variables that are linearly associated to the log probabilities of the target, and the absence of extreme outliers.

In this dataset, the target is a binary variable that is labelled as either 0 for survival or 1 for death (see Figure 3.1). The observations in the data collection, which each represent a different data point, demonstrate the independence of the data. As a result, the assumptions that the data are not matched and that the aim is to take two possible outcomes are met. To investigate the next assumption, that there is no severe multicollinearity among the explanatory variables, and this can be done in R by using `cor.test()`. The correlation would be excessively high if the value were higher than absolute 0.7. Figure 4.1 presents the results, which demonstrate that the third criterion is not falsified and that all numbers are below 0.7. The relationship between the predictor variables is not statistically significant. The notion is supported by the fact that the VIF reading was likewise below 5, indicating the multicollinearity is not a concern. According to a visual inspection of the scatter plot in Figure 4.6 between each predictor and the logit values, the continuous variables and the log-odds dependent variable seem to have a fairly linear connection.

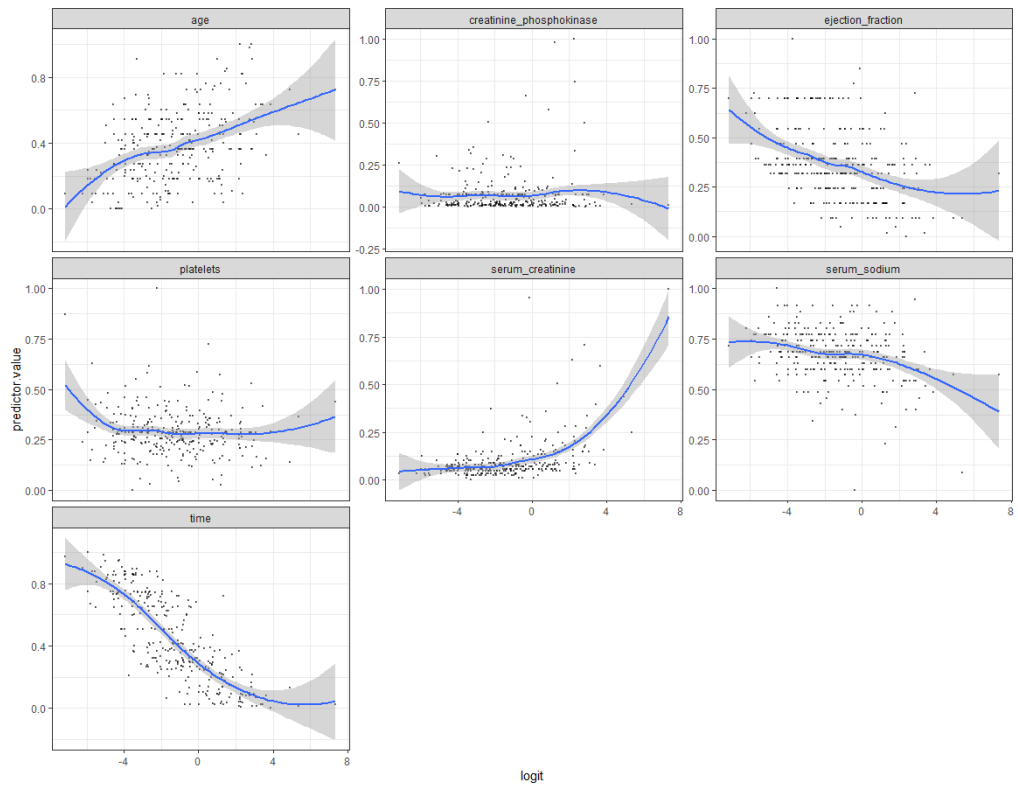


Figure 4.6 Scatter plot predictor value against log-odds DEATH_EVENT.

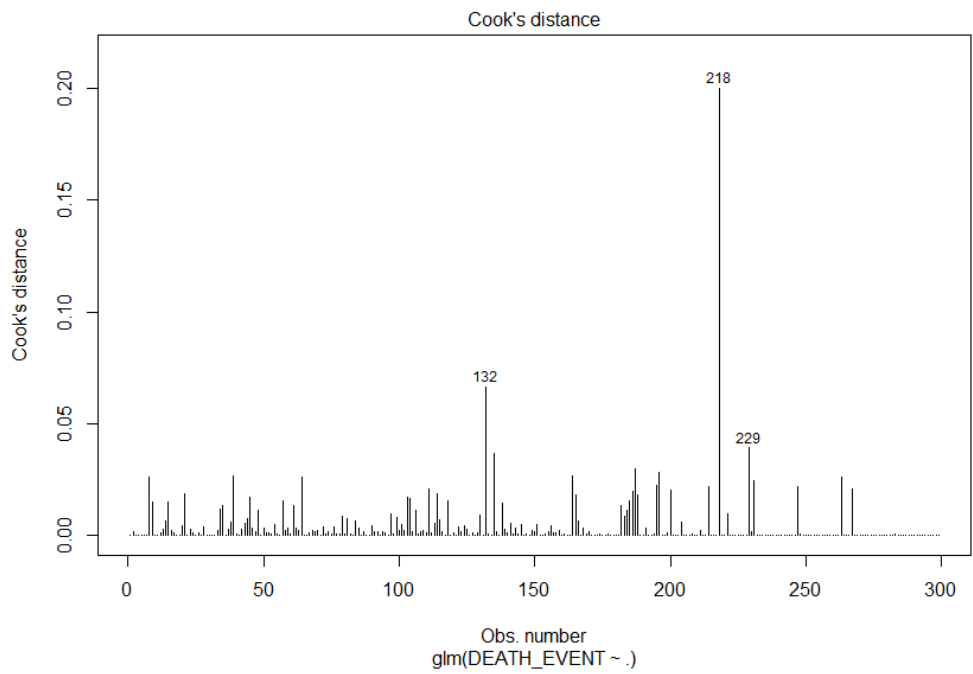


Figure 4.7 Cook's distance plot.

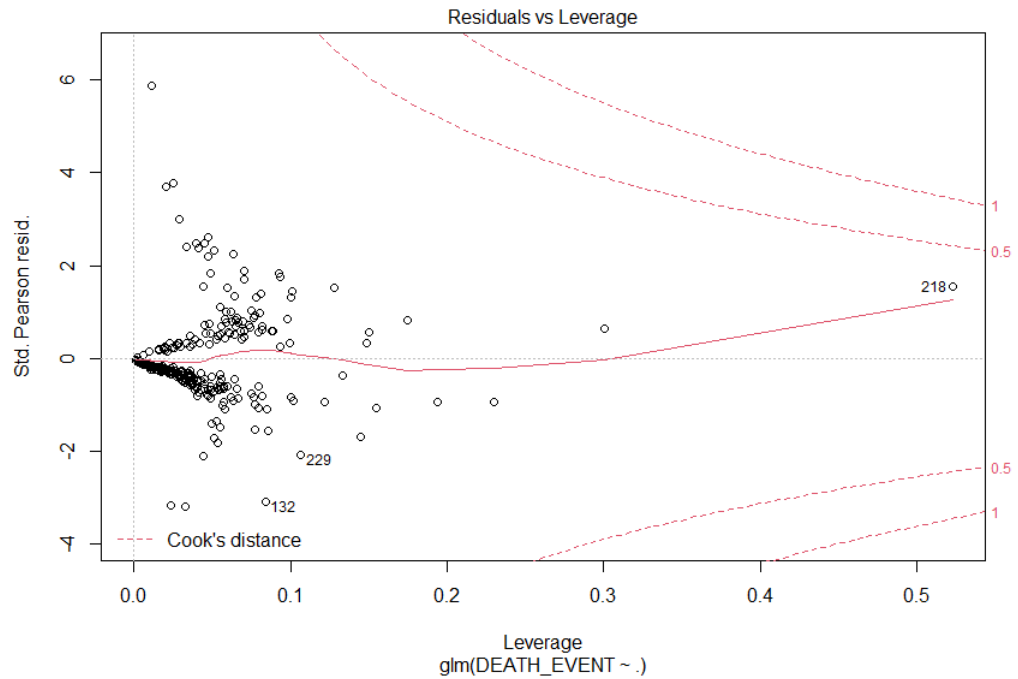


Figure 4.8 Residuals vs Leverage plot.

A basic rule according to Bruce et al. (2020), an observation with Cook's distance more than $\frac{4}{n-p-1}$ where n represent number of observations and p represent number of predictor variables, can be regarded as an extreme outlier. The Cook's distance metric can be visualised to look at the most extreme outliers in the data. The plot, Figure 4.7, indicated the top three most extreme outliers' observations as being #132, #218, and #229. The Residuals vs. Leverage plot can assist us in identifying any influential insights. All the points in Figure 4.8 are inside the Cook's distance line, indicating the absence of the extreme influential points.

To investigate the relationship between "removing outliers" and survey respondent self-reported attributes, a straightforward logistic regression model was developed. The results from the two models do not differ by very

much, as shown by Table 4.6's summary. Because none of the three top extreme observations exceeded the 0.5 Cook's Distance limit, they are tolerable. Thus, the data set will continue to contain all 299 data points.

Table 4.6 Comparison model before (1) and after (2) excluding influential.

	(1)	(2)
(Intercept)	10.185 (5.657)	11.695 * (5.895)
age	0.047 ** (0.016)	0.049 ** (0.016)
anaemia1	-0.007 (0.360)	0.003 (0.374)
creatinine_phosphokinase	0.000 (0.000)	0.000 (0.000)
diabetes1	0.145 (0.351)	0.205 (0.360)
ejection_fraction	-0.077 *** (0.016)	-0.083 *** (0.018)
high_blood_pressure1	-0.103 (0.359)	-0.211 (0.375)
platelets	-0.000 (0.000)	-0.000 (0.000)
serum_creatinine	0.666 *** (0.181)	0.874 ** (0.293)
serum_sodium	-0.067 (0.040)	-0.079 (0.041)
sex1	-0.534 (0.414)	-0.539 (0.427)
smoking1	-0.013 (0.413)	-0.064 (0.423)
time	-0.021 *** (0.003)	-0.021 *** (0.003)
N	299	296
logLik	-109.777	-104.654
AIC	245.554	235.307

*** p < 0.001; ** p < 0.01; * p < 0.05.

The investigation showed that the chosen dataset satisfied the logistic regression's presumptions, hence the next section will go into detail about the models' discoveries. The dependent variable DEATH EVENT is used to construct six logistic models. These models will aid in determining the impact that factors such as time, sex, smoking, age, anaemia, diabetes, ejection fraction,

high blood pressure, platelets, serum creatinine, and serum sodium may have on an individual's likelihood of passing away.

To decrease the number of independent variables in the model for the study using logistic regression models, stepwise regression with stepAIC was utilised. For this model, the McFadden's Adjusted Pseudo R² (MAPR²) was utilised as the indicator of fit. Values closer to zero signify that the model has no predictive ability. The metric spans from 0 to slightly under 1. A McFadden's Pseudo R² score between 0.2 and 0.4 is considered to be good, according to Lee (2013). The formula for MAPR² is as follows:

$$R_{adj}^2 = 1 - \frac{\ln(L(M_{full})) - k}{\ln(L(M_{intercept}))}$$

There were four logistic regression models that had a MAPR² greater than 0.2, and two more that were below 0.2. The results of the analysis for those six models are presented in Tables 4.6 through 4.11. Appendices E through J contain additional details for each logistic regression model.

Using a Hosmer-Lemeshow goodness-of-fit test, the logistic model's goodness-of-fit was determined. A p-value of less than 0.05 was achieved, which supports the model's unsatisfactory fit. The Hosmer-Lemeshow test indicated that all of the logistic models in this study that did not use upsampling were perfectly fitted, with the exception of the model containing binary variables. No logistic variable or model had a VIF greater than 5 among all the logistic variables and models. The final models chosen for this study thus show that multicollinearity was not a concern. Use of a more sophisticated model is

advised, as it improves the model's accuracy, according to a likelihood ratio test with a p-value > 0.05 .

The results are validated using the accuracy, which is utilized to summarise the data in the confusion matrix. Values that are closer to one support the model's good data fit. The ratio of the true positive and true negative versus the total number is used to compute accuracy and provides a percentage of the fitted model's accuracy. It is denoted as

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

Area under the ROC curve (AUC), a metric used to assess the model's performance, spans from 0 to 1 above the threshold of $c = 0.5$. Being between 0.8 and 0.9 is a great range for the AUC.

Table 4.7: 1st Logistics Regression Model.

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p-value	Estimation	Std. Err	p-value
DEATH_EVENT						
intercept	4.2103	1.6239	p < 0.05	3.6035	1.3482	p < 0.05
age	2.4170	1.0549	p < 0.05	2.3992	1.0071	p < 0.05
anaemia	-0.6064	0.4772	0.2038			
creatinine_phospokinase	1.5372	1.6718	0.3578			
diabetes	0.0408	0.4365	0.9255			
ejection_fraction	-6.1920	1.5199	p < 0.05	-6.2245	1.4874	p < 0.05
high_blood_pressure	-0.1644	0.4545	0.7175			
platelets	-0.6219	2.0366	0.7601			
serum_creatinine	9.5429	2.5351	p < 0.05	9.4585	2.5502	p < 0.05
serum_sodium	-2.9710	1.6473	0.0713	-3.0077	1.5964	0.0596
sex	-0.2603	0.5177	0.6152			
smoking	-0.2067	0.5388	0.7013			
time	-5.6094	1.0329	p < 0.05	-5.3513	0.9557	p < 0.05
	McFadden R ² = 0.353, df = 13, p-value(hoslem.test) = 0.642			McFadden R ² = 0.392 df = 6, p-value(hoslem.test) = 0.434		

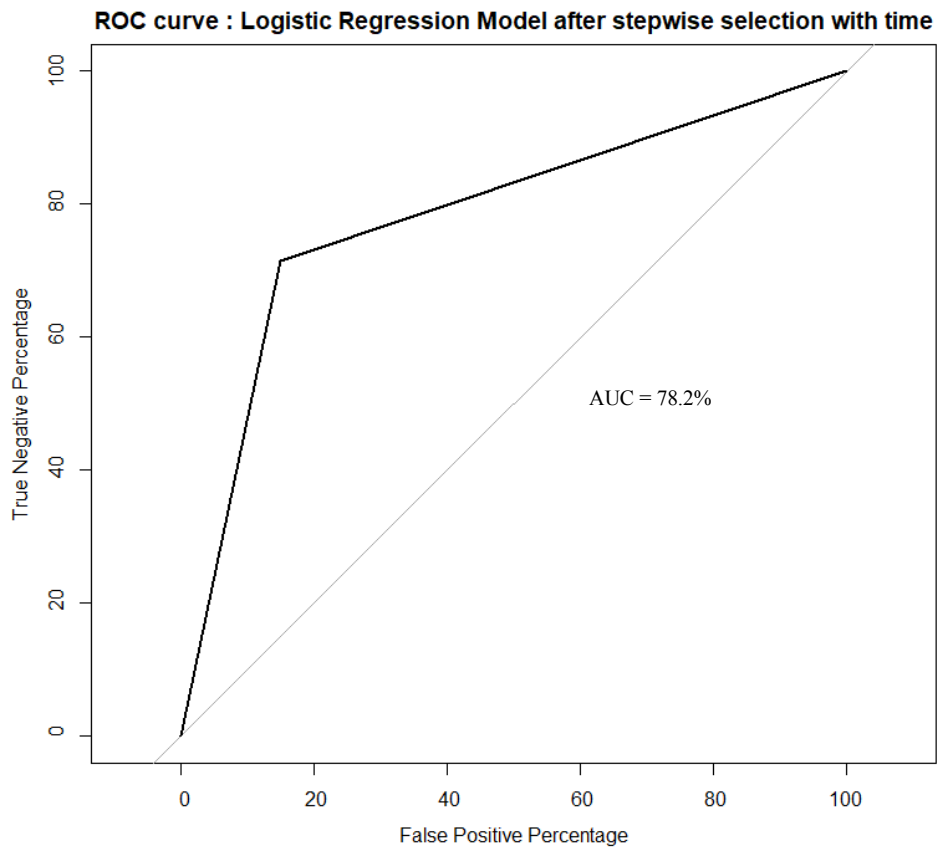


Figure 4.9: ROC curve for Model 1.

All predictor variables were used to generate the initial model, and the final model is justified using the stepwise process. Table 4.7 displays the estimates for the first model and codes them as

$$\text{DEATH_EVENT} \sim \text{age} + \text{ejection_fraction} + \text{serum_creatinine} \\ + \text{serum_sodium} + \text{time}$$

The McFadden's R^2 value of 0.392 is fairly high, indicating that the model has a high degree of predictive power and a very good fit to the data. The Hosmer-Lemeshow test with a p-value > 0.05 indicates that the model is well fitted. The accuracy produced by the model is 80.7% (see Appendix E), and the AUC is 78.2%. (see Figure 4.9).

Table 4.8: 2nd Logistics Regression Model.

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p-value	Estimation	Std. Err	p-value
DEATH_EVENT						
intercept	17.3069	6.1269	p < 0.05	17.5590	5.9669	p < 0.05
age	0.0565	0.0167	p < 0.05	0.0510	0.0160	p < 0.05
anaemia	-0.0723	0.3871	0.8518			
creatinine_phospokinase	0.0002	0.0002	0.3899			
diabetes	0.3798	0.3787	0.3159			
ejection_fraction	-0.1092	0.0207	p < 0.05	-0.1086	0.0202	p < 0.05
high_blood_pressure	0.2177	0.3849	0.5717			
platelets	0.0000	0.0000	0.6328			
serum_creatinine	1.1909	0.2792	p < 0.05	1.1382	0.2634	p < 0.05
serum_sodium	-0.1241	0.0445	p < 0.05	-0.1180	0.0426	p < 0.05
sex	0.0578	0.4441	0.8964			
smoking	-0.1353	0.4614	0.7693			
time	-0.0195	0.0030	p < 0.05	-0.0203	0.0029	p < 0.05
	McFadden R ² = 0.424 df = 13 p-value(hoslem.test) = 0.526			McFadden R ² = 0.451 df = 6 p-value(hoslem.test) = 0.000		

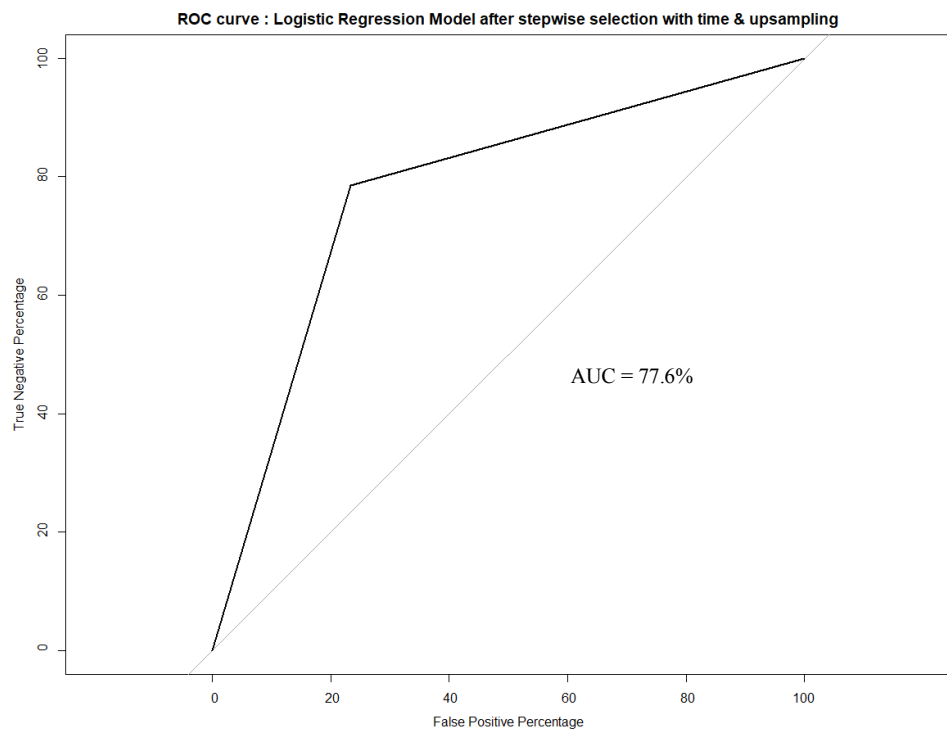


Figure 4.10: ROC curve for Model 2.

All predictor variables were used to generate the second model, and upsampling was used to address the imbalance target. The stepwise estimates for the final model are shown in Table 4.8 and are coded as

$$\text{DEATH_EVENT} \sim \text{age} + \text{ejection_fraction} + \text{serum_creatinine} \\ + \text{serum_sodium} + \text{time}$$

The model has a very high degree of prediction capacity and a very strong fit to the data, as indicated by McFadden's R^2 values of 0.451. With a p-value of less than 0.05, the Hosmer-Lemeshow test demonstrates that the model is not well fitted. The accuracy of the model is 77.3% (see Appendix F), while the AUC is 77.6%. (see Figure 4.10).

Table 4.9: 3rd Logistics Regression Model.

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p-value	Estimation	Std. Err	p-value
DEATH_EVENT						
intercept	0.2452	1.2645	0.8462	0.4792	1.1233	0.6696
age	3.1594	0.9277	p < 0.05	3.0675	0.8830	p < 0.05
anaemia	0.0320	0.3947	0.9354			
creatinine_phospokinase	2.3990	1.3354	0.0724	2.1873	1.2339	0.0763
diabetes	0.1340	0.3742	0.7202			
ejection_fraction	-5.7998	1.3670	p < 0.05	-5.6872	1.3460	p < 0.05
high_blood_pressure	0.3777	0.3874	0.3297			
platelets	0.2786	1.7479	0.8734			
serum_creatinine	9.3180	2.7456	p < 0.05	8.9578	2.5852	p < 0.05
serum_sodium	-2.4104	1.4992	0.1079	-2.4601	1.4964	0.1002
sex	-0.1614	0.4475	0.7184			
smoking	-0.1319	0.4631	0.7758			
	McFadden R ² = 0.197, df = 12, p-value(hoslem.test) = 0.317			McFadden R ² = 0.235 df = 6, p-value(hoslem.test) = 0.057		

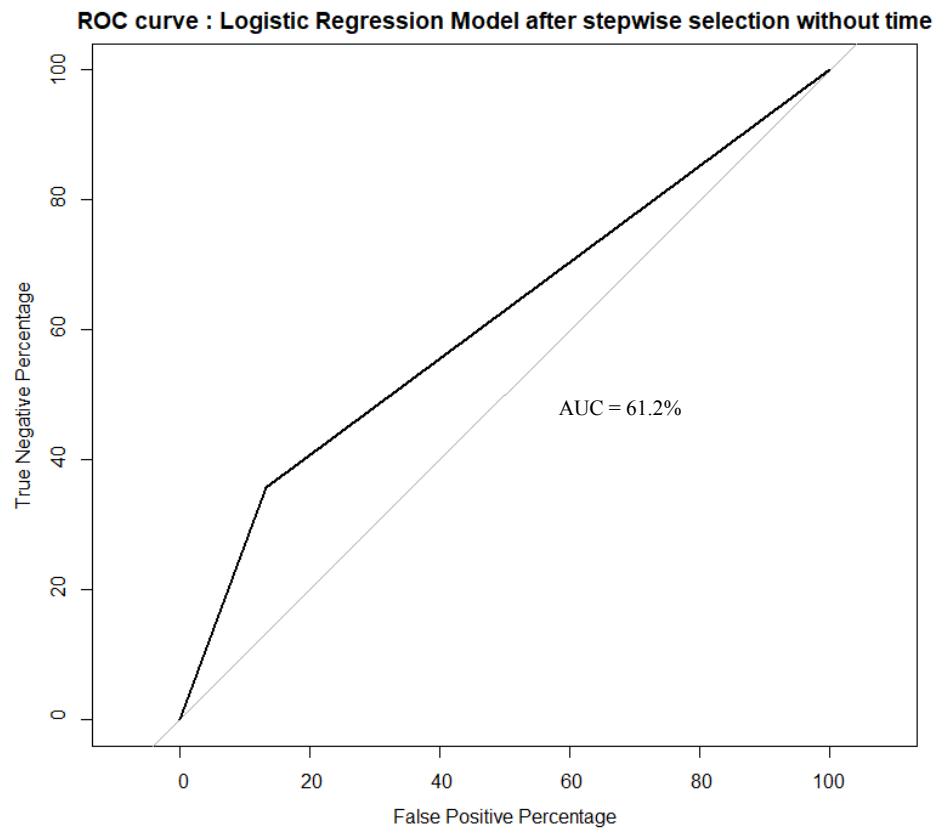


Figure 4.11: ROC curve for Model 3.

Apart from time, all predictor variables were used to build the third model. The stepwise estimates for the final model are shown in Table 4.9 and are coded as

$$\text{DEATH_EVENT} \sim \text{age} + \text{creatinine_phosphokinase} + \text{ejection_fraction} \\ + \text{serum_creatinine} + \text{serum_sodium}$$

The model has a high degree of predictive power and a very excellent fit to the data, as indicated by McFadden's R^2 values of 0.235. The model fits the data well, as shown by the Hosmer-Lemeshow test, which has a p-value above 0.05. Its accuracy, which is 70.5 percent (see Appendix G), and AUC, which is 61.2 percent (see Figure 4.11), are both subpar.

Table 4.10: 4th Logistics Regression Model.

Dependent variable	Without Stepwise			With Stepwise		
	Estimation	Std. Err	p-value	Estimation	Std. Err	p-value
DEATH_EVENT						
intercept	-0.0579	1.0236	0.9549	0.0664	0.9676	0.9453
age	3.3351	0.7674	p < 0.05	3.3392	0.7659	p < 0.05
anaemia	0.1980	0.3257	0.5437			
creatinine_phospokinase	2.1308	1.2543	0.0894	1.9441	1.1828	0.1003
diabetes	0.5581	0.3099	0.0717	0.5661	0.3076	0.0657
ejection_fraction	-5.1810	1.0147	p < 0.05	-5.1698	1.0155	p < 0.05
high_blood_pressure	0.5683	0.3184	0.0743	0.5706	0.3087	0.0646
platelets	1.8959	1.3091	0.1476	1.8655	1.3012	0.1517
serum_creatinine	10.7065	2.6789	p < 0.05	10.4108	2.6019	p < 0.05
serum_sodium	-2.9318	1.2545	p < 0.05	-2.9093	1.2499	p < 0.05
sex	0.0096	0.3717	0.9794			
smoking	0.0515	0.3869	0.8942			
	McFadden R ² = 0.247 df = 12 p-value(hoslem.test) = 0.164			McFadden R ² = 0.261 df = 9 p-value(hoslem.test) = 0.005		

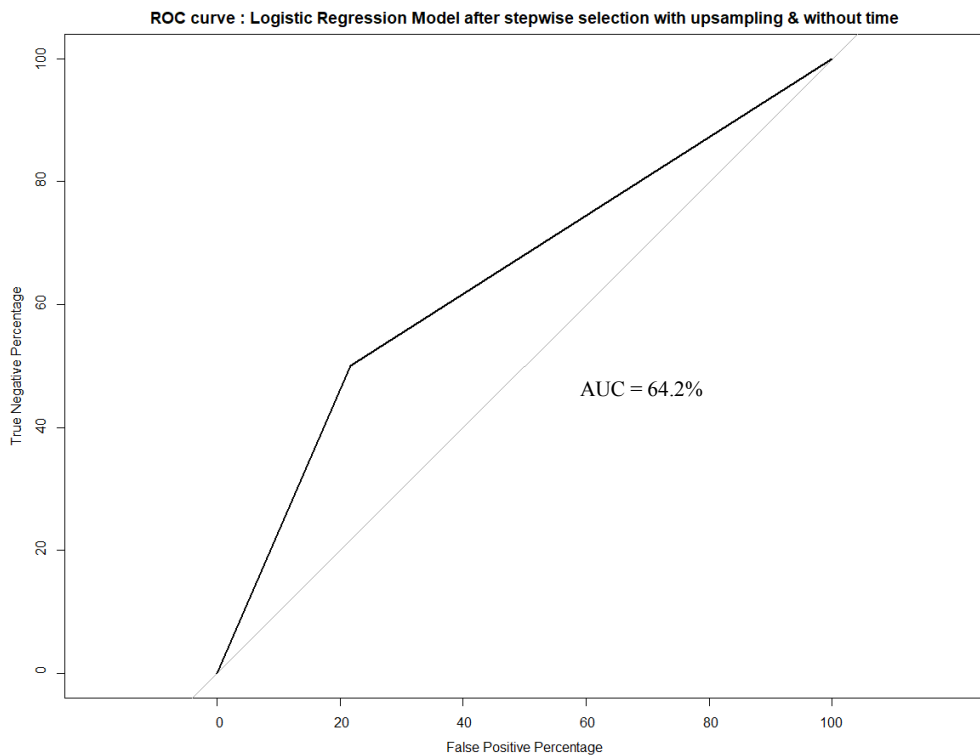


Figure 4.12: ROC curve for Model 4.

All predictor variables, apart from time and upsampling applied to the hand imbalance target, were used to generate the fourth model. The stepwise estimates are shown in Table 4.10 along with the final model's justification, which is coded as

```
DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes  
+ ejection_fraction + high_blood_pressure + platelets  
+ serum_creatinine + serum_sodium
```

The model has a high degree of predictive power and a very excellent fit to the data, as indicated by McFadden's R^2 values of 0.261. With a p-value under 0.05, the Hosmer-Lemeshow test reveals that the model is not adequately fitted. Its accuracy, which is 69.3 percent (see Appendix H), and AUC, which is 64.2 percent (see Figure 4.12), are poor.

Table 4.11: 5th Logistics Regression Model.

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p-value	Estimation	Std. Err	p-value
DEATH_EVENT						
intercept	-3.5496	0.7535	p < 0.05	-2.6841	0.4907	p < 0.05
age	0.6625	0.3695	0.0730	0.6601	0.3551	0.0631
anaemia	0.0551	0.3438	0.8727			
creatinine_phospokinase	0.5392	0.4328	0.2128			
diabetes	-0.1057	0.3442	0.7587			
ejection_fraction	0.9280	0.4262	p < 0.05	0.8790	0.4125	p < 0.05
high_blood_pressure	0.5740	0.3553	0.1062			
platelets	0.6246	0.4967	0.2086			
serum_creatinine	0.7034	0.3474	p < 0.05	0.6362	0.3297	0.0536
serum_sodium	1.1713	0.3446	p < 0.05	1.1890	0.3287	p < 0.05
sex	0.2304	0.4096	0.5737			
smoking	-0.1759	0.4067	0.6654			
	McFadden R ² = 0.056 df = 12, p-value(hoslem.test) = 0.367			McFadden R ² = 0.088 df = 5, p-value(hoslem.test) = 0.572		

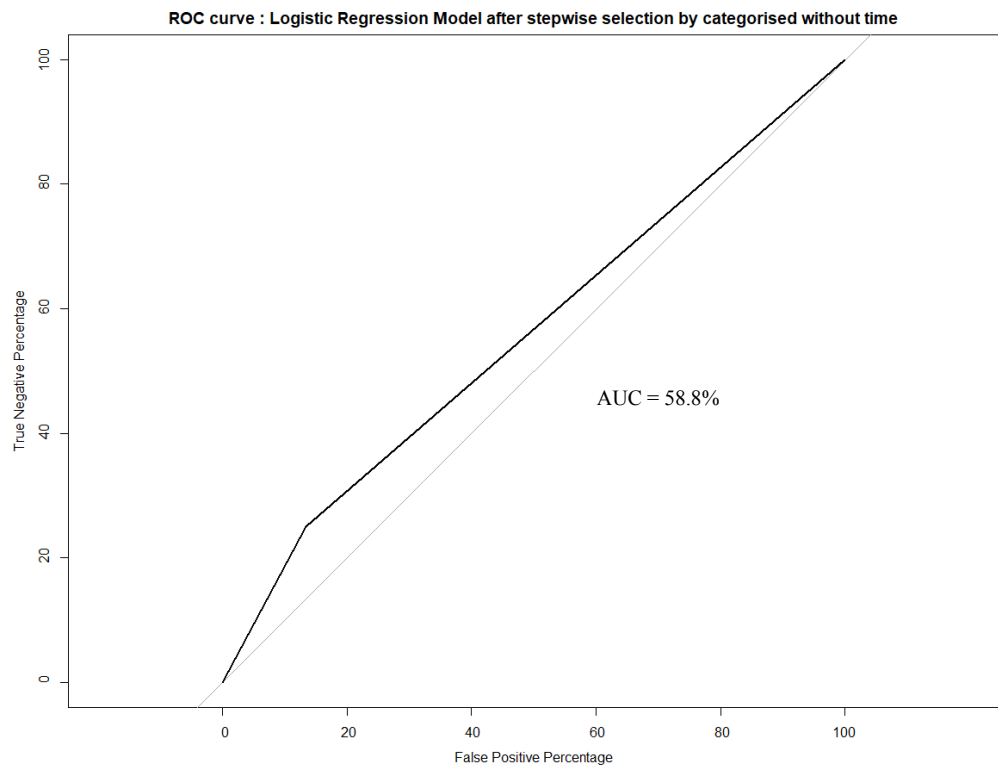


Figure 4.13: ROC curve for Model 5.

All predictor variables—apart from time—were used to build the fifth model. The two level "0" and "1" categories are used to classify each of the chosen variables. The stepwise estimates are displayed in Table 4.11 and the final model is justified as follows:

$$\text{DEATH_EVENT} \sim \text{serum_sodium} + \text{ejection_fraction} \\ + \text{serum_creatinine} + \text{age}$$

According to the results of the Hosmer-Lemeshow test, p-values greater than 0.05 imply a good fit. However, McFadden's R^2 values of 0.088 show that the model does not adequately fit the data. This model is unable to fit the data as well as Model 1 does, as shown by its accuracy of 67.0 percent (see Appendix I) and AUC of 58.8 percent (see Figure 4.13).

Table 4.12: 6th Logistics Regression Model.

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p-value	Estimation	Std. Err	p-value
DEATH_EVENT						
intercept	-3.2367	0.6238	p < 0.05	-2.6086	0.4406	p < 0.05
age	0.7890	0.3065	p < 0.05	0.7469	0.2963	p < 0.05
anaemia	0.0370	0.2816	0.8954			
creatinine_phospokinase	0.4465	0.3461	0.1970			
diabetes	0.2604	0.2768	0.3468			
ejection_fraction	0.9757	0.3401	p < 0.05	1.0538	0.3335	p < 0.05
high_blood_pressure	0.8391	0.2981	p < 0.05	0.8093	0.2910	p < 0.05
platelets	0.8383	0.3988	p < 0.05	0.9448	0.3894	p < 0.05
serum_creatinine	0.7249	0.2944	p < 0.05	0.6450	0.2811	p < 0.05
serum_sodium	1.1811	0.2841	p < 0.05	1.1984	0.2797	p < 0.05
sex	0.2708	0.3447	0.4322			
smoking	-0.1061	0.3348	0.7513			
	McFadden R ² = 0.109 df = 12, p-value(hoslem.test) = 0.129			McFadden R ² = 0.126 df = 7, p-value(hoslem.test) = 0.342		

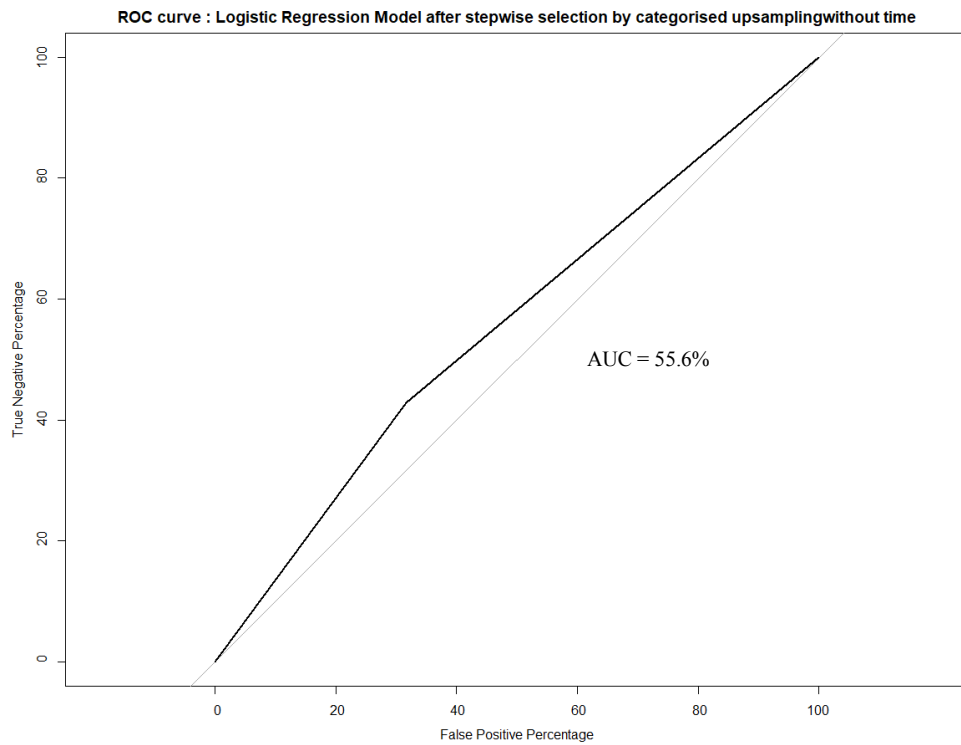


Figure 4.14: ROC curve for Model 6.

Apart from time, all predictor variables were used to generate the sixth model. The two levels of "0" and "1" are used to categorise each of the chosen variables. Application of upsampling to the deal with the imbalance target. Table 4.12 displays the estimates for the sixth model and codes them as

```
DEATH_EVENT ~ agegpn + efractionn + high_blood_pressure + plat  
+ screatn + sodium
```

Although the results of the Hosmer-Lemeshow test suggest that p-values > 0.05 imply that it is a decent match, the value of McFadden's R^2 0.126 indicates that the models do not fit the data very well. Model 1 appears to perform better than the accuracy of 60.2% (see Appendix J) and AUC of 55.6% (see Figure 4.14).

The MAPR², Hosmer-Lemeshow, VIF, likelihood ratio test, accuracy, and AUC have all been used to evaluate each model to determine which one provides the best fit. Evidently, Figure 4.15 from the first model is the optimal model. The selected model consists of five predictor variables (IV) against one dependent variable (DV) and the fit is expressed as

$$\text{DEATH_EVENT} \sim \text{age} + \text{ejection_fraction} + \text{serum_creatinine} + \text{serum_sodium} + \text{time}$$

The logit equation with the estimations included is written as,

$$\begin{aligned} \log_e(\text{odd}[\text{event}]) \\ = 3.6035 + 2.3992x_1 - 6.2245x_2 + 9.4585x_3 - 3.0077x_4 \\ - 5.3513x_5 \end{aligned}$$

where $x_1 = \text{age}$, $x_2 = \text{ejection_fraction}$, $x_3 = \text{serum_creatinine}$, $x_4 = \text{serum_sodium}$, $x_5 = \text{time}$

Model	MAPR ²	Hosmer-Lemeshow goodness-of-fit test	VIF	Likelihood ratio test	ROC - AUC	Accuracy
1 st	0.392	p-value = 0.434	Below 5	p-value > 0.05	78.2%	80.7%
2 nd	0.451	p-value = 0.000	Below 5	p-value > 0.05	77.6%	77.3%
3 rd	0.235	p-value = 0.057	Below 5	p-value > 0.05	61.2%	70.5%
4 th	0.261	p-value = 0.005	Below 5	p-value > 0.05	64.2%	69.3%
5 th	0.088	p-value = 0.572	Below 5	p-value > 0.05	58.8%	67.0%
6 th	0.126	p-value = 0.342	Below 5	p-value > 0.05	55.6%	60.2%

Figure 4.15: Summary evaluation values for each individual model.

CHAPTER 5

CONCLUSIONS

The heart failure dataset comprises independent variables that can predict the dependent variable, according to the results of the chi-square test and the ANOVA test, therefore a model can be built using this data. A heart failure patient's likelihood of survival is maximised by using models that identify the variables that are most crucial to controlling the condition. Additionally, early detection of any potential risk factors for heart failure is possible with the aid of the models.

The successful creation of eighteen linear regression models shows that the models are not good models, as indicated by the adjusted-R² values that are less than 0.3. The best linear regression model generated by Table 4.5 yields an adjusted-R² value of 0.1337. The model includes the target variable “serum_creatinine” and predictor variables “age”, “ejection_fraction”, “high_blood_pressure” and “serum_sodium”. The logit equation with the inclusion of the estimate is expressed as

$$\hat{Y}^{-0.99} = -0.9502 - 0.0069x_1 + 0.0030x_2 + 0.0601x_3 + 0.0156x_4$$

Unfortunately, the adjusted-R² of this model, which was constructed, is below 0.3, making it weak.

Six logistic regression models with DEATH EVENT as the response variable demonstrate that models without upsampling provided a better fit to the data than models with upsampling. The adjusted MAPR² fell between the recommended limits of 0.2 and 0.4 for an acceptable model. The Hosmer-Lemeshow goodness-of-fit test of these models, which had p-values above 0.05, indicated that they were good models. A value of fewer than five is produced via variance inflation factors. Therefore, this might attest to the fact that multicollinearity is not a serious issue for the model. According to the results of the entire likelihood ratio test, the models should employ the more sophisticated model that includes the constant because it improves their accuracy.

The binary logistic model with the applied stepwise selection without upsampling model has the greatest percentage accuracy with a value of 80.7 percent, according to a comparison utilising the information from the confusion matrix. The models performed weaker with datasets balanced by up-sampling. The performance of the models by categorised the continuous variables into normal and abnormal is the weakest with the accuracy 67 percent and 60 percent.

Given the accuracy rate of 80.7 percent on the test data, the chosen logistic regression model includes the variables “age”, “ejection_fraction”, “serum_creatinine”, “serum_sodium” and “time”, has been found to be reliable. The logit equation with the inclusion of the estimate is expressed as

$$\begin{aligned} \text{Logit}(p(x)) = & 3.6035 + 2.3992x_1 - 6.2245x_2 + 9.4585x_3 - 3.0077x_4 \\ & - 5.3513x_5 \end{aligned}$$

It demonstrates that ejection fraction, serum creatinine, serum sodium, age, and time are the most significant and necessary factors to predict the mortality of heart failure patients when compared to using all features.

Using the same dataset, it would be intriguing to conduct additional analysis to compare the survival rates of people with heart failure in various age groups.

REFERENCES

Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M. and Raza, M.A., 2017. Survival analysis of heart failure patients: A case study. *PLoS ONE*, 12(7). <https://doi.org/10.1371/journal.pone.0181001>.

Anand, I., McMurray, J.J.V., Whitmore, J., Warren, M., Pham, A., McCamish, M.A. and Burton, P.B.J., 2004. Anemia and its relationship to clinical outcome in heart failure. *Circulation*, 110(2), pp.149–154. <https://doi.org/10.1161/01.CIR.0000134279.79571.73>.

Aujla RS and Patel, R., 2022. *Creatine Phosphokinase*. [online] StatPearls. Available at: <<https://www.ncbi.nlm.nih.gov/books/NBK546624/>> [Accessed 28 July 2022].

Benjamin, E.J., Virani, S.S., Callaway, C.W., Chamberlain, A.M., Chang, A.R., Cheng, S., Chiuve, S.E., Cushman, M., Dellinger, F.N., Deo, R., de Ferranti, S.D., Ferguson, J.F., Fornage, M., Gillespie, C., Isasi, C.R., Jiménez, M.C., Jordan, L.C., Judd, S.E., Lackland, D., Lichtman, J.H., Lisabeth, L., Liu, S., Longenecker, C.T., Lutsey, P.L., MacKey, J.S., Matchar, D.B., Matsushita, K., Mussolino, M.E., Nasir, K., O’Flaherty, M., Palaniappan, L.P., Pandey, A., Pandey, D.K., Reeves, M.J., Ritchey, M.D., Rodriguez, C.J., Roth, G.A., Rosamond, W.D., Sampson, U.K.A., Satou, G.M., Shah, S.H., Spartano, N.L., Tirschwell, D.L., Tsao, C.W., Voeks, J.H., Willey, J.Z., Wilkins, J.T., Wu, J.H.Y., Alger, H.M., Wong, S.S. and Muntner, P., 2018. Heart disease and stroke statistics - 2018 update: A report from the American Heart Association. *Circulation*, 137(12), pp.E67–E492. <https://doi.org/10.1161/CIR.0000000000000558>.

Bruce, P., Bruce, A., Gedeck, P. and Safari, an O.M.Company., 2020. *Practical Statistics for Data Scientists, 2nd Edition*. Second Edition ed. O’Reilly Media.

Bytyçi, I. and Bajraktari, G., 2015. *Mortality in heart failure patients. Anadolu Kardiyoloji Dergisi*, <https://doi.org/10.5152/akd.2014.5731>.

Chicco, D. and Jurman, G., 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1). <https://doi.org/10.1186/s12911-020-1023-5>.

Das, P., Jain, S., Sharma, C. and Shambhu, S., 2021. *Prediction of Heart Disease Mortality Rate Using Data Mining*. [online] Available at: <<http://ceur-ws.org/Vol-2823/Paper18.pdf>> [Accessed 30 July 2022].

Diana Rodriguez, 2009. *Anemia and Your Heart | Everyday Health*. [online] Available at: <<https://www.everydayhealth.com/heart-health/anemia.aspx>> [Accessed 28 July 2022].

Dunlay, S.M., Weston, S.A., Jacobsen, S.J. and Roger, V.L., 2009. Risk Factors for Heart Failure: A Population-Based Case-Control Study. *American Journal of Medicine*, 122(11), pp.1023–1028. <https://doi.org/10.1016/j.amjmed.2009.04.022>.

Eletter, S., Yasmin, T., Elrefae, G., Aliter, H. and Elrefae, A., 2020. Building an intelligent telemonitoring system for heart failure: The use of the internet of things, big data, and machine learning. In: *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*. [online] Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ACIT50332.2020.9300113>.

Ezekowitz, J.A., McAlister, F.A. and Armstrong, P.W., 2003. Anemia is common in heart failure and is associated with poor outcomes: Insights from a cohort of 12 065 patients with new-onset heart failure. *Circulation*, 107(2), pp.223–225. <https://doi.org/10.1161/01.CIR.0000052622.51963.FC>.

Fernandes, A.A.T., Filho, D.B.F., da Rocha, E.C. and da Silva Nascimento, W., 2020. Read this paper if you want to learn logistic regression. *Revista de Sociologia e Politica*, 28(74), pp.1/1-19/19. <https://doi.org/10.1590/1678-987320287406EN>.

Hair, J.F.Jr., Anderson, R.E., Babin, B.J. and Black, W.C., 2010. *Multivariate Data Analysis : A Global Perspective*. 7th ed ed. Upper Saddle River (N.J.): Pearson Education.

Healthwise Staff, 2021. *Heart Failure With Reduced Ejection Fraction (Systolic Heart Failure)*. [online] MyHealth.Alberta.ca. Available at: <<https://myhealth.alberta.ca/Health/Pages/conditions.aspx?hwid=tx4090abc>> [Accessed 28 July 2022].

Heidenreich, P.A., Albert, N.M., Allen, L.A., Bluemke, D.A., Butler, J., Fonarow, G.C., Ikonomidis, J.S., Khavjou, O., Konstam, M.A., Maddox, T.M., Nichol, G., Pham, M., Piña, I.L. and Trogdon, J.G., 2013. Forecasting the impact of heart failure in the united states a policy statement from the american heart association. *Circulation: Heart Failure*, 6(3), pp.606–619. <https://doi.org/10.1161/HHF.0b013e318291329a>.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied Logistic Regression: Third Edition. Applied Logistic Regression: Third Edition*. wiley. <https://doi.org/10.1002/9781118548387>.

Kamimura, D., Cain, L.R., Mentz, R.J., White, W.B., Blaha, M.J., Defilippis, A.P., Fox, E.R., Rodriguez, C.J., Keith, R.J., Benjamin, E.J., Butler, J., Bhatnagar, A., Robertson, R.M., Winniford, M.D., Correa, A. and Hall, M.E., 2018. Cigarette smoking and incident heart failure: Insights from the jackson heart study. *Circulation*, 137(24), pp.2572–2582. <https://doi.org/10.1161/CIRCULATIONAHA.117.031912>.

Lam, C.S.P., 2015. *Heart failure in Southeast Asia: facts and numbers. ESC Heart Failure*, <https://doi.org/10.1002/ehf2.12036>.

Le, M.T., Vo, M.T., Mai, L. and Dao, S.V.T., 2020. Predicting heart failure using deep neural network. *2020 International Conference on Advanced Technologies for Communications (ATC)*. *IEEE*, 2020. [online] Available at: <<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9255445>> [Accessed 31 July 2022].

Li, H., Hastings, M.H., Rhee, J., Trager, L.E., Roh, J.D. and Rosenzweig, A., 2020. *Targeting age-related pathways in heart failure*. *Circulation Research*, <https://doi.org/10.1161/CIRCRESAHA.119.315889>.

Lippi, G. and Sanchis-Gomar, F., 2020. Global epidemiology and future trends of heart failure. *AME Medical Journal*, 5, pp.15–15. <https://doi.org/10.21037/amj.2020.03.03>.

Lloyd-Jones, D.M., Larson, M.G., Leip, E.P., Beiser, A., D'Agostino, R.B., Kannel, W.B., Murabito, J.M., Vasan, R.S., Benjamin, E.J. and Levy, D., 2002. Lifetime risk for developing congestive heart failure: The Framingham Heart Study. *Circulation*, 106(24), pp.3068–3072. <https://doi.org/10.1161/01.CIR.0000039105.49749.6F>.

Mahmood, T., Raj, K., Ehtesham, M., Bhimani, J., Jabeen, S. and Tahir, A., 2019. Serum Sodium Profile of Congestive Heart Failure Patients and its Impact on Their Outcome at Discharge. *Cureus*. <https://doi.org/10.7759/cureus.5462>.

Mayo Clinic Staff, 2021. *Creatinine tests*. [online] Mayo Clinic. Available at: <<https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646>> [Accessed 28 July 2022].

Ministry of Health Malaysia, 2020. *THE IMPACT OF NONCOMMUNICABLE DISEASES AND THEIR RISK FACTORS ON MALAYSIA'S GROSS DOMESTIC PRODUCT*. [online] Available at: <https://www.moh.gov.my/index.php/database_stores/attach_download/554/64> [Accessed 27 July 2022].

Mojadidi, M.K., Galeas, J.N., Goodman-Meza, D., Eshtehardi, P., Msaouel, P., Kelesidis, I., Zaman, M.O., Winoker, J.S., Roberts, S.C., Christia, P. and Zolty, R., 2016. Thrombocytopaenia as a Prognostic Indicator in Heart Failure with Reduced Ejection Fraction. *Heart Lung and Circulation*, 25(6), pp.568–575. <https://doi.org/10.1016/j.hlc.2015.11.010>.

Mozaffarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M., Das, S.R., de Ferranti, S., Després, J.-P., Fullerton, H.J., Howard, V.J., Huffman, M.D., Isasi, C.R., Monik, ;, Jiménez, C., Judd, S.E., Kissela, B.M., Lichtman, J.H., Lisabeth, L.D., Liu, S., Mackey, R.H., Magid, D.J., Mcguire, D.K., Mohler Iii, E.R., Moy, C.S., Muntner, P., Mussolino, M.E., Nasir, K., Neumar, R.W., Nichol, G., Palaniappan, L., Pandey, D.K., Reeves, M.J., Rodriguez, C.J., Rosamond, W., Sorlie, P.D., Stein, J., Towfighi, A., Turan, T.N., Virani, S.S., Woo, D., Yeh, R.W. and Turner, M.B., 2015. *Heart Disease and Stroke Statistics-2016 Update A Report From the American Heart Association*. [online] *Circulation*, Available at: <<http://my.americanheart.org/statements>>.

Pillai, H.S. and Ganapathi, S., 2013. Heart Failure in South Asia. *Current Cardiology Reviews*, [online] 9(2), p.102. <https://doi.org/10.2174/1573403X11309020003>.

Ponikowski, P., Anker, S.D., AlHabib, K.F., Cowie, M.R., Force, T.L., Hu, S., Jaarsma, T., Krum, H., Rastogi, V., Rohde, L.E., Samal, U.C., Shimokawa, H., Budi Siswanto, B., Sliwa, K. and Filippatos, G., 2014. *Heart failure: preventing disease and death worldwide. ESC Heart Failure*, <https://doi.org/10.1002/ehf2.12005>.

Rosano, G.M., Vitale, C. and Seferovic, P., 2017. Heart Failure in Patients with Diabetes Mellitus. *Cardiac Failure Review*, [online] 03(01), p.52. <https://doi.org/10.15420/cfr.2016:20:2>.

Sakinc, I. and Ugurlu, E., 2013. *A Logistic Regression Analysis to Examine Factors Affecting Gender Diversity on the Boardroom: ISE Case **. [online] *International Journal of Business and Social Science*, Available at: <www.ijbssnet.com>.

Savji, N., Meijers, W.C., Bartz, T.M., Bhambhani, V., Cushman, M., Naylor, M., Kizer, J.R., Sarma, A., Blaha, M.J., Gansevoort, R.T., Gardin, J.M., Hillege, H.L., Ji, F., Kop, W.J., Lau, E.S., Lee, D.S., Sadreyev, R., van Gilst, W.H., Wang, T.J., Zanni, M. v., Vasani, R.S., Allen, N.B., Psaty, B.M., van der Harst, P., Levy, D., Larson, M., Shah, S.J., de Boer, R.A., Gottdiener, J.S. and Ho, J.E., 2018. The Association of Obesity and Cardiometabolic Traits With Incident HFpEF and HFrEF. *JACC: Heart Failure*, 6(8), pp.701–709. <https://doi.org/10.1016/j.jchf.2018.05.018>.

UCI, 2020. *UCI Machine Learning Repository: Heart failure clinical records Data Set*. [online] Available at: <<https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>> [Accessed 29 July 2022].

Wang, J., 2021. Heart failure prediction with machine learning: A comparative study. In: *Journal of Physics: Conference Series*. IOP Publishing Ltd. <https://doi.org/10.1088/1742-6596/2031/1/012068>.

World Health Organization, 2020. *TOBACCO & CORONARY HEART DISEASE*. [online] Available at: <[https://www.google.com/search?q=World+Health+Organization+\(2017\)+reported+that+10%25+of+all+deaths+due+to+cardiovascular+disease+could+be+attributed+to+smoking.&aq=chrome..69i57.4863j0j15&sourceid=chrome&ie=UTF-8#:~:text=https%3A//apps.who.int/iris/rest/bitstreams/1303433/retrieve](https://www.google.com/search?q=World+Health+Organization+(2017)+reported+that+10%25+of+all+deaths+due+to+cardiovascular+disease+could+be+attributed+to+smoking.&aq=chrome..69i57.4863j0j15&sourceid=chrome&ie=UTF-8#:~:text=https%3A//apps.who.int/iris/rest/bitstreams/1303433/retrieve)> [Accessed 28 July 2022].

Zahid, F.M., Ramzan, S., Faisal, S. and Hussain, I., 2019. Gender based survival prediction models for heart failure patients: A case study in Pakistan. *PLoS ONE*, 14(2). <https://doi.org/10.1371/journal.pone.0210602>.

Zaman, S.M.M., Qureshi, W.M., Raihan, Md.M.S., Monjur, O. and Shams, A. bin, 2021. Survival Prediction of Heart Failure Patients using Stacked Ensemble Machine Learning Algorithm. [online] <https://doi.org/10.48550/arxiv.2108.13367>.

Zangmo, C. and Tiensuwan, M., 2018. Application of logistic regression models to cancer patients: A case study of data from Jigme Dorji Wangchuck National Referral Hospital (JDWNRH) in Bhutan. In: *Journal of Physics: Conference Series*. Institute of Physics Publishing. <https://doi.org/10.1088/1742-6596/1039/1/012031>.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H. and Cao, B., 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229), pp.1054–1062. [https://doi.org/10.1016/S0140-6736\(20\)30566-3](https://doi.org/10.1016/S0140-6736(20)30566-3).

APPENDIX A

DATA DISTRIBUTION FOR CONTINUOUS VARIBALES

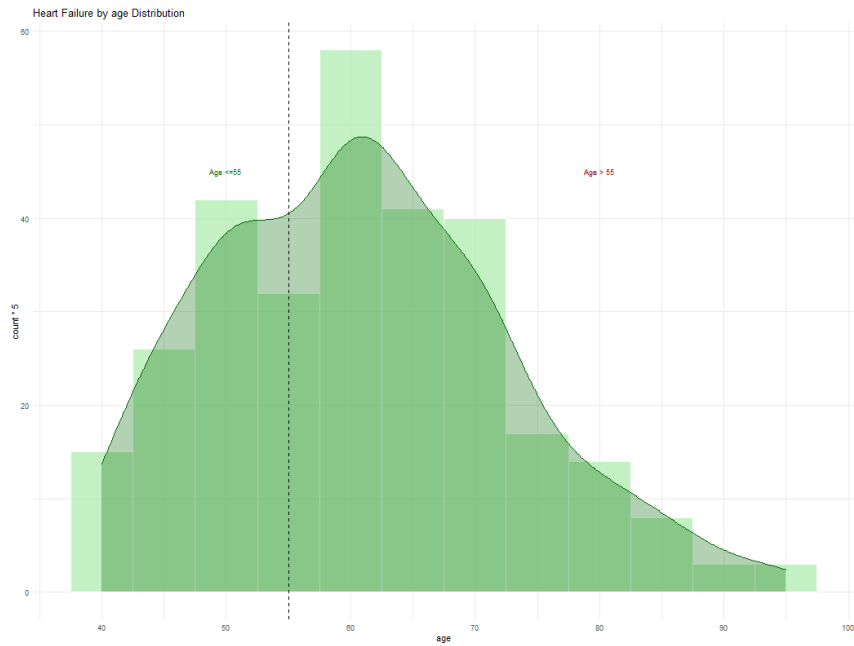


Figure A.1: Heart failure by age distribution

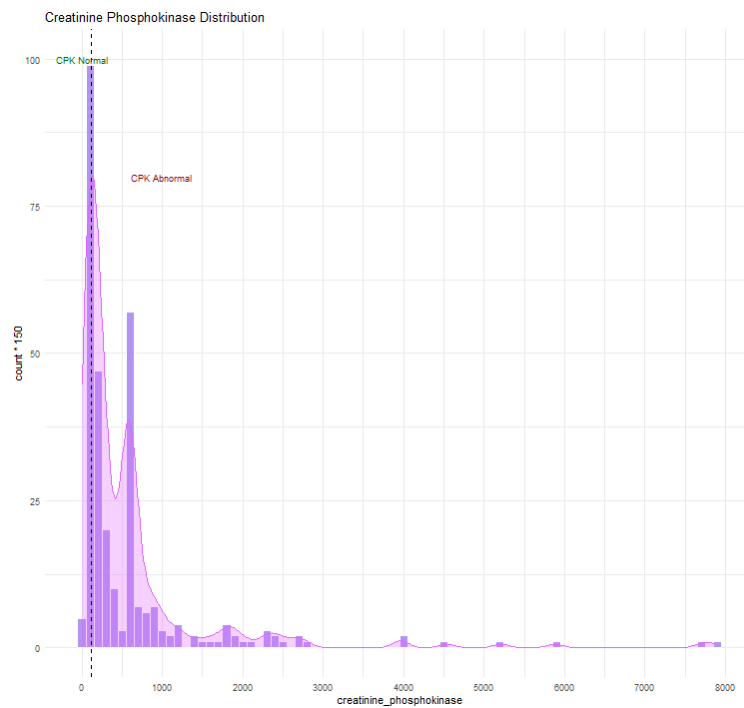


Figure A.2: Heart failure by creatinine_phosphokinase distribution

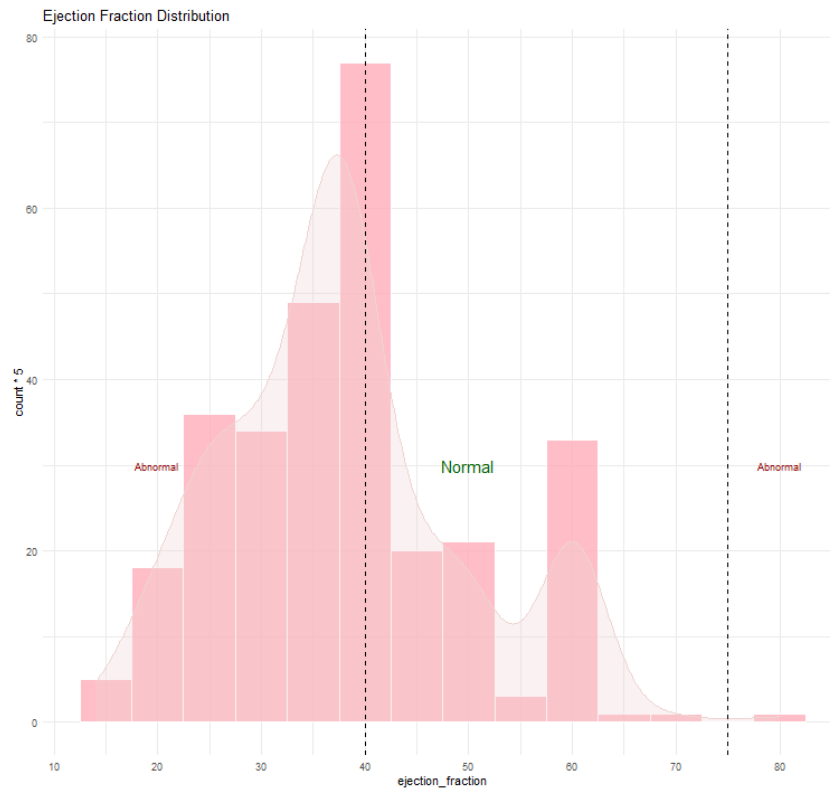


Figure A.3: Heart failure by ejection_fraction distribution

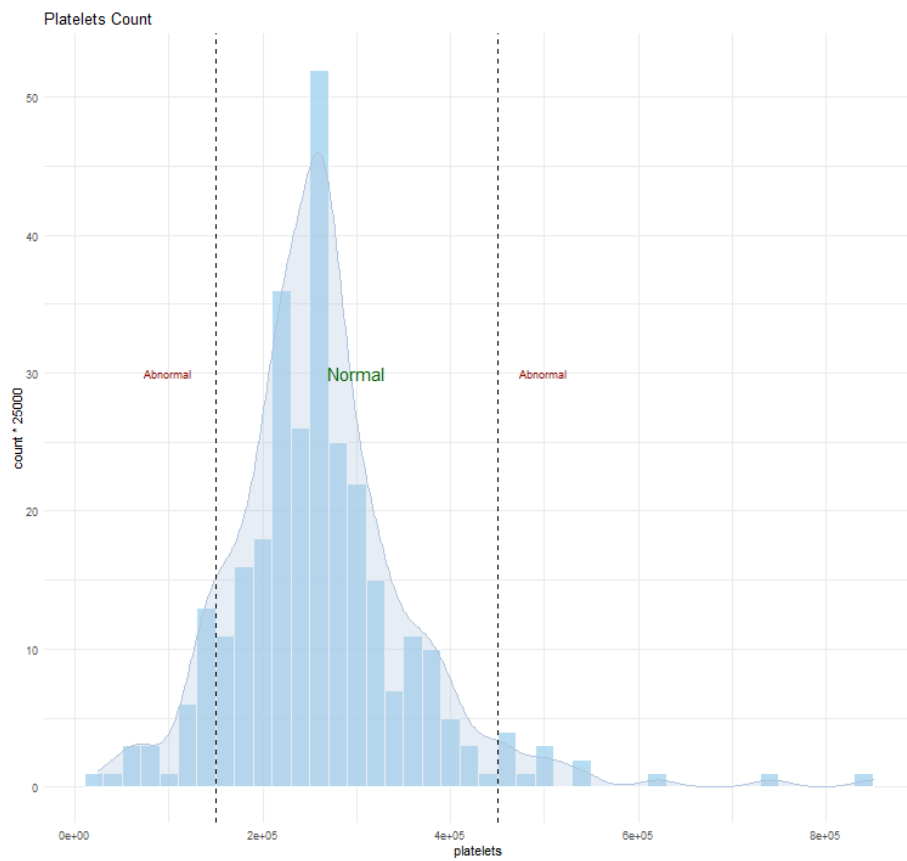


Figure A.4: Heart failure by platelets distribution

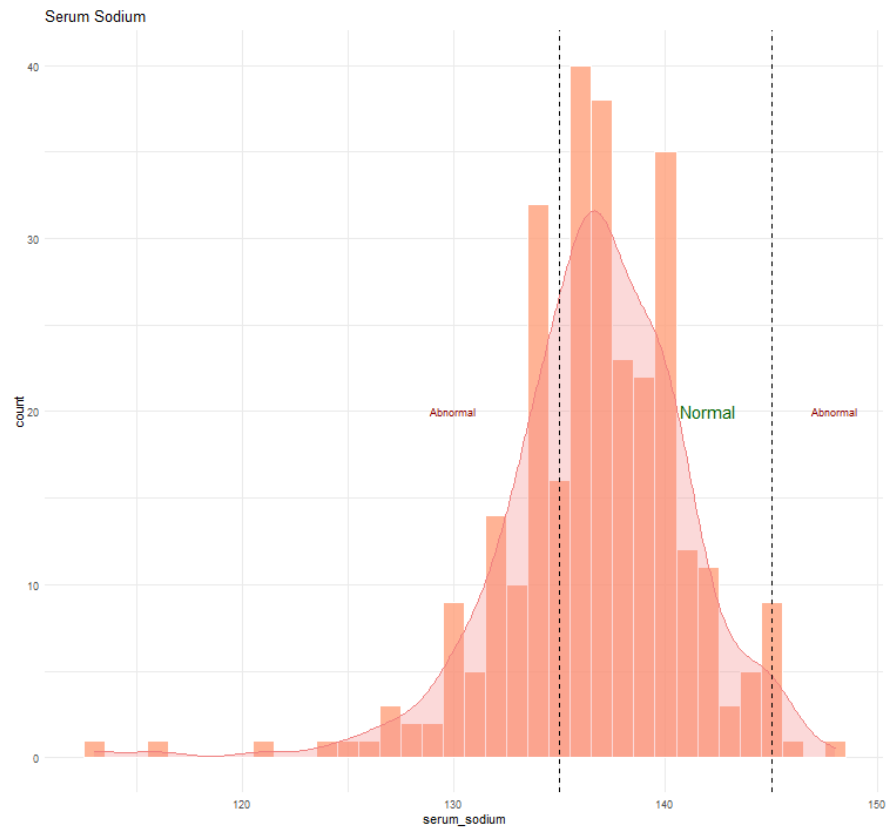


Figure A.5: Heart failure by serum_sodium distribution

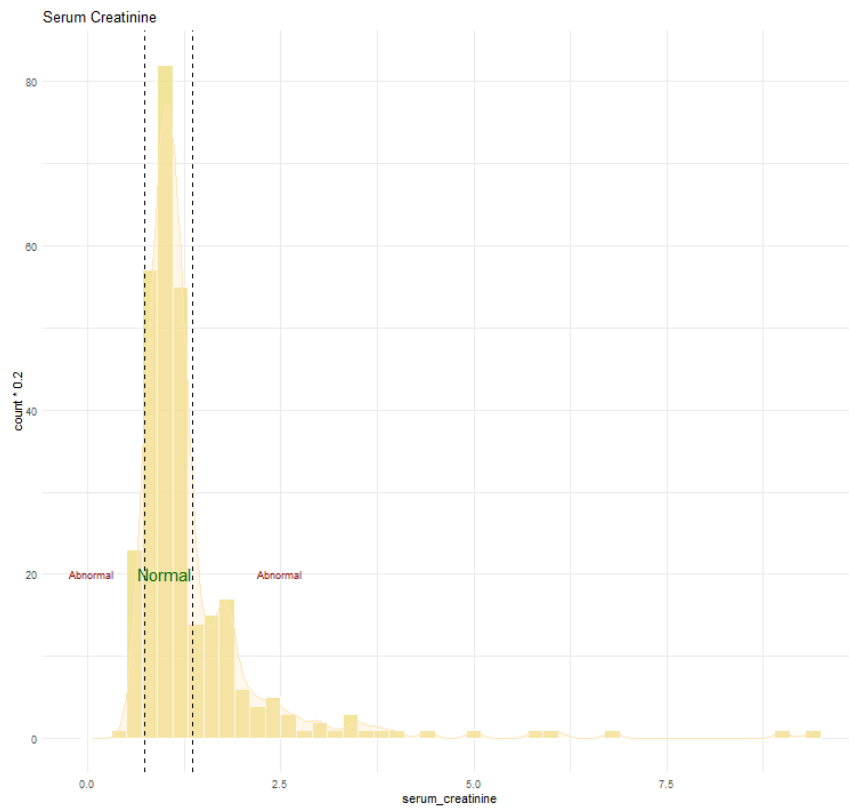
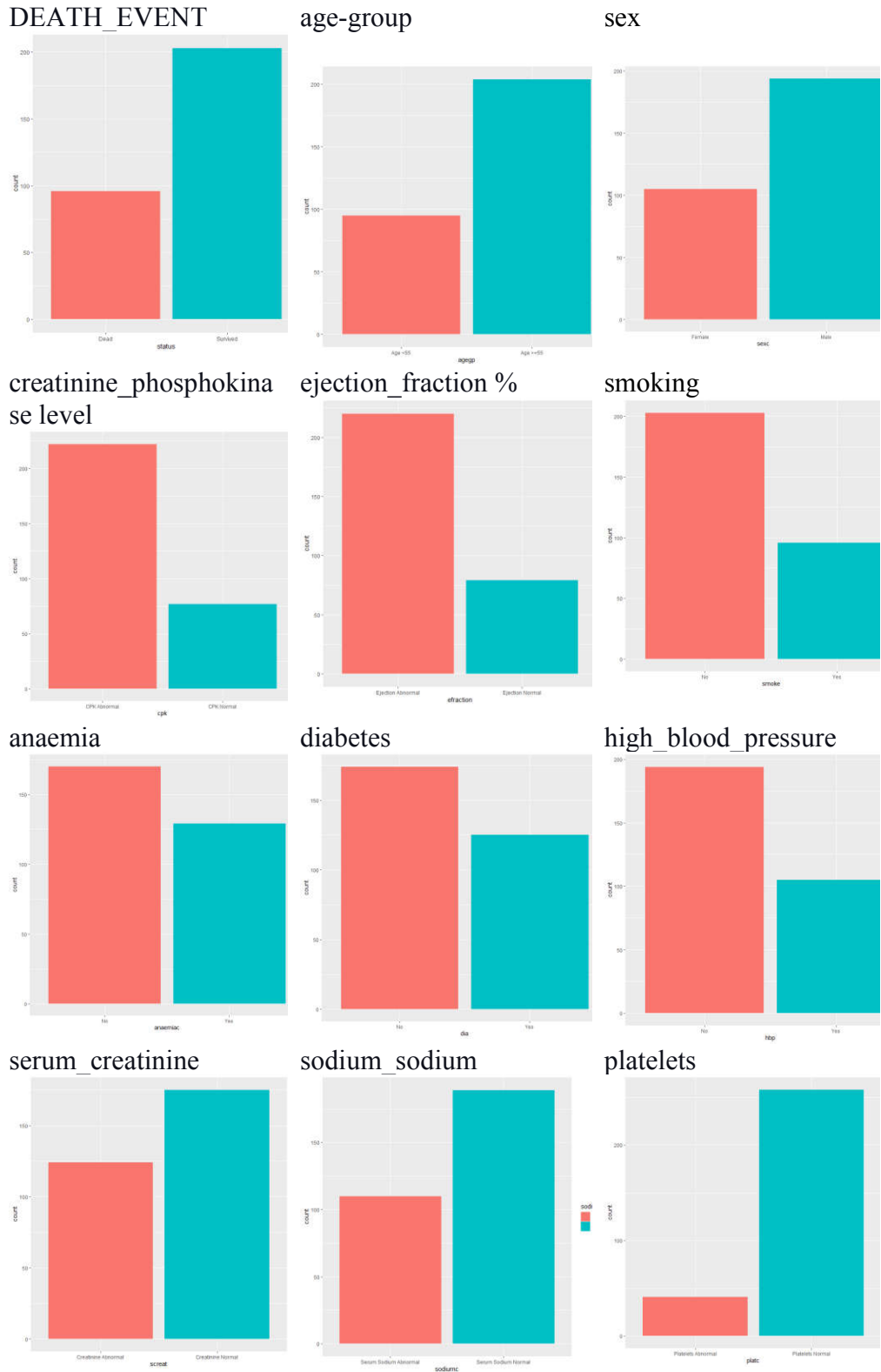


Figure A.6: Heart failure by serum_creatinine distribution

APPENDIX B

DATA DISTRIBUTION FOR CATEGORISE VARIBALES



APPENDIX C

ANALYSIS FOR ALL LINEAR REGRESSION MODELS

Table C.1: age ~ high_blood_pressure + diabetes + serum_creatinine

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
age						
intercept	69.10	21.95	0.00	58.46	1.39	<0.05
sex	2.05	1.64	0.21			
anaemia	1.66	1.41	0.24			
high_blood_pressure	2.37	1.43	0.10	2.31	1.42	0.10411
smoking	-0.18	1.64	0.91			
diabetes	-1.95	1.41	0.17	-2.23	1.37	0.10577
creatinine_phospokinase	0.00	0.00	0.30			
ejection_fraction	0.08	0.06	0.20			
platelets	0.00	0.00	0.59			
serum_creatinine	1.64	0.67	0.01	1.79	0.66	0.00685
serum_sodium	-0.10	0.16	0.53			
	Ajd R ² = 0.03149, p-value < 0.05			Ajd R ² = 0.033, p-value < 0.05		

Table C.2: log(age) ~ anaemia + high_blood_pressure + serum_creatinine

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
age						
intercept	4.18	0.36	<2e-16	4.02	0.02	< 2e-16
sex	0.03	0.03	0.27			
anaemia	0.03	0.02	0.21	0.03	0.02	0.15344
high_blood_pressure	0.04	0.02	0.08	0.04	0.02	0.0883
smoking	0.00	0.03	0.92			
diabetes	-0.03	0.02	0.27			
creatinine_phospokinase	0.00	0.00	0.26			
ejection_fraction	0.00	0.00	0.22			
platelets	0.00	0.00	0.43			
serum_creatinine	0.03	0.01	0.01	0.03	0.01	0.00603
serum_sodium	0.00	0.00	0.62			
	Ajd R ² = 0.0327, p-value < 0.05			Ajd R ² = 0.03314, p-value < 0.05		

Table C.3: $(age)^{0.02} \sim \text{anaemia} + \text{high_blood_pressure} + \text{serum_creatinine}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
age						
intercept	1.09	0.01	<2e-16	1.08	0.00	< 2e-16
sex	0.00	0.00	0.27			
anaemia	0.00	0.00	0.21	0.00	0.00	0.15403
high_blood_pressure	0.00	0.00	0.08	0.00	0.00	0.08868
smoking	0.00	0.00	0.93			
diabetes	0.00	0.00	0.27			
creatinine_phospokinase	0.00	0.00	0.26			
ejection_fraction	0.00	0.00	0.22			
platelets	0.00	0.00	0.43			
serum_creatinine	0.00	0.00	0.01	0.00	0.00	0.00603
serum_sodium	0.00	0.00	0.61			
	Ajd R ² = 0.03268, p-value < 0.05			Ajd R ² = 0.03309, p-value < 0.05		

Table C.4: $\text{creatinine_phospokinase} \sim \text{anaemia}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
creatinine_phospokinase						
intercept	-1252.00	1826.00	0.49	742.79	73.17	< 0.05
sex	175.70	133.90	0.19			
anaemia	-358.70	113.60	0.00	-373.05	111.40	0.00092
high_blood_pressure	-113.00	117.60	0.34			
smoking	-130.10	134.20	0.33			
diabetes	-18.93	115.90	0.87			
age	-4.95	4.82	0.30			
ejection_fraction	-3.28	4.84	0.50			
platelets	0.0003	0.00	0.63			
serum_creatinine	14.01	55.48	0.80			
serum_sodium	16.81	13.13	0.20			
	Ajd R ² = 0.02475, p-value > 0.05			Ajd R ² = 0.03314, p-value < 0.05		

Table C.5: $\log(\text{creatinine_phosphokinase}) \sim \text{anaemia}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
creatinine_phosphokinase						
intercept	5.18	2.12	0.02	5.88	0.08	< 2e-16
sex	0.08	0.16	0.59			
anaemia	-0.50	0.13	0.00	-0.51	0.13	9.27E-05
high_blood_pressure	-0.16	0.14	0.25			
smoking	-0.17	0.16	0.28			
diabetes	0.05	0.13	0.73			
age	-0.01	0.01	0.31			
ejection_fraction	-0.01	0.01	0.27			
platelets	0.0000	0.00	0.76			
serum_creatinine	-0.03	0.06	0.69			
serum_sodium	0.01	0.02	0.54			
	Ajd R ² = 0.03787, p-value > 0.05			Ajd R ² = 0.04703, p-value < 0.05		

Table C.6: $(\text{creatinine_phosphokinase})^{0.14} \sim \text{anaemia}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
creatinine_phosphokinase						
intercept	0.47	0.13	0.00	0.44	0.01	< 2e-16
sex	0.00	0.01	0.71			
anaemia	0.03	0.01	0.00	0.03	0.01	0.00014
high_blood_pressure	0.01	0.01	0.26			
smoking	0.01	0.01	0.32			
diabetes	0.00	0.01	0.66			
age	0.00	0.00	0.32			
ejection_fraction	0.00	0.00	0.23			
platelets	0.0000	0.00	0.75			
serum_creatinine	0.00	0.00	0.62			
serum_sodium	0.00	0.00	0.59			
	Ajd R ² = 0.0355, p-value > 0.05			Ajd R ² = 0.04448, p-value < 0.05		

Table C.7 : ejection_fraction ~ sex + serum_sodium

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
ejection_fraction						
intercept	-30.26	22.16	0.17	-22.62	20.81	0.27791
sex	-3.37	1.62	0.04	-3.56	1.40	0.0117
anaemia	-0.06	1.41	0.97			
high_blood_pressure	-0.23	1.43	0.87			
smoking	-0.30	1.64	0.85			
diabetes	-0.23	1.41	0.87			
age	0.07	0.06	0.20			
creatinine_phospokinase	0.00	0.00	0.50			
platelets	0.0000	0.00	0.38			
serum_creatinine	0.14	0.68	0.84			
serum_sodium	0.47	0.16	0.00	0.46	0.15	0.00261
	Ajd R ² = 0.02909, p-value < 0.05			Ajd R ² = 0.04515, p-value < 0.05		

Table C.8: log(ejection_fraction) ~ age + serum_sodium + sex

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
ejection_fraction						
intercept	1.52	0.59	0.01	1.56	0.57	0.00655
sex	-0.09	0.04	0.05	-0.09	0.04	0.01392
anaemia	-0.01	0.04	0.85			
high_blood_pressure	-0.01	0.04	0.72			
smoking	0.00	0.04	0.93			
diabetes	0.00	0.04	0.90			
age	0.00	0.00	0.12	0.00	0.00	0.1308
creatinine_phospokinase	0.00	0.00	0.65			
platelets	0.0000	0.00	0.33			
serum_creatinine	0.00	0.02	0.92			
serum_sodium	0.01	0.00	0.00	0.01	0.00	0.00049
	Ajd R ² = 0.03655, p-value < 0.05			Ajd R ² = 0.05519, p-value < 0.05		

Table C.9: (ejection_fraction)^{0.26} ~ age + serum_sodium + sex

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
ejection_fraction						
intercept	1.23	0.40	0.00	1.25	0.38	0.00112
sex	-0.06	0.03	0.04	-0.06	0.03	0.01222
anaemia	0.00	0.03	0.88			
high_blood_pressure	-0.01	0.03	0.76			
smoking	0.00	0.03	0.91			
diabetes	0.00	0.03	0.96			
age	0.00	0.00	0.14	0.00	0.00	0.13762
creatinine_phospokinase	0.00	0.00	0.60			
platelets	0.0000	0.00	0.35			
serum_creatinine	0.00	0.01	0.97			
serum_sodium	0.01	0.00	0.00	0.01	0.00	0.00072
	Ajd R ² = 0.03457, p-value < 0.05			Ajd R ² = 0.05347, p-value < 0.05		

Table C.10: serum_creatinine ~ sex + serum_sodium

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
serum_creatinine						
intercept	6.77	1.90	0.00	6.43	1.85	0.00058
sex	0.00	0.14	1.00			
anaemia	0.10	0.12	0.44			
high_blood_pressure	-0.03	0.13	0.81			
smoking	-0.07	0.14	0.63			
diabetes	-0.11	0.12	0.37			
age	0.01	0.01	0.01	0.01	0.00	0.00798
creatinine_phospokinase	0.00	0.00	0.80			
platelets	0.0000	0.00	0.80			
ejection_fraction	0.00	0.01	0.84			
serum_sodium	-0.04	0.01	0.00	-0.04	0.01	0.00139
	Ajd R ² = 0.03228, p-value < 0.05			Ajd R ² = 0.05209, p-value < 0.05		

Table C.11: $\log(\text{serum_creatinine}) \sim \text{age} + \text{serum_sodium}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
serum_creatinine						
intercept	3.09	0.80	0.00	3.16	0.79	7.45E-05
sex	0.01	0.06	0.81			
anaemia	0.00	0.05	0.96			
high_blood_pressure	-0.06	0.05	0.22			
smoking	-0.06	0.06	0.33			
diabetes	-0.03	0.05	0.59			
age	0.01	0.00	0.00	0.01	0.00	8.88E-05
creatinine_phospokinase	0.00	0.00	0.67			
platelets	0.0000	0.00	0.90			
ejection_fraction	0.00	0.00	0.20			
serum_sodium	-0.02	0.01	0.00	-0.03	0.01	9.67E-06
	Ajd R ² = 0.09651, p-value < 0.05			Ajd R ² = 0.1072, p-value < 0.05		

Table C.12: $(\text{serum_creatinine})^{-0.99} \sim \text{age} + \text{ejection_fraction} + \text{high_blood_pressure} + \text{serum_sodium}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
serum_creatinine						
intercept	-0.9304	0.5496	0.0915	-0.9502	0.5387	0.0788
sex	-0.0179	0.0413	0.6657			
anaemia	0.0100	0.0355	0.7787			
high_blood_pressure	0.0623	0.0362	0.0864	0.0601	0.0356	0.0924
smoking	0.0492	0.0413	0.2349			
diabetes	0.0010	0.0356	0.9785			
age	-0.0069	0.0015	0.0000	-0.0069	0.0014	0.0000
creatinine_phospokinase	0.0000	0.0000	0.4661			
platelets	0.0000	0.0000	0.9375			
ejection_fraction	0.0030	0.0015	0.0418	0.0030	0.0015	0.0421
serum_sodium	0.0152	0.0040	0.0002	0.0156	0.0039	0.0001
	Ajd R ² = 0.1215, p-value < 0.05			Ajd R ² = 0.1337, p-value < 0.05		

Table C.13: serum_sodium ~ diabetes + ejection_fraction + serum_creatinine

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
serum_sodium						
intercept	135.40	1.77	< 2e-16	135.67	0.93	< 2e-16
sex	-0.08	0.60	0.90			
anaemia	0.56	0.52	0.28			
high_blood_pressure	0.31	0.53	0.55			
smoking	0.07	0.60	0.90			
diabetes	-0.93	0.52	0.07	-0.87	0.50	0.08228
age	-0.01	0.02	0.53			
creatinine_phospokinase	0.00	0.00	0.20			
platelets	0.0000	0.00	0.41			
ejection_fraction	0.06	0.02	0.00	0.06	0.02	0.00216
serum_creatinine	-0.79	0.24	0.00	-0.82	0.24	0.00071
	Ajd R ² = 0.05622, p-value < 0.05			Ajd R ² = 0.06608, p-value < 0.05		

Table C.14: log(serum_sodium) ~ diabetes + ejection_fraction + serum_creatinine

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
serum_sodium						
intercept	4.91	0.01	< 2e-16	4.91	0.01	< 2e-16
sex	0.00	0.00	0.95			
anaemia	0.00	0.00	0.28			
high_blood_pressure	0.00	0.00	0.54			
smoking	0.00	0.00	0.94			
diabetes	-0.01	0.00	0.06	-0.01	0.00	0.07036
age	0.00	0.00	0.53			
creatinine_phospokinase	0.00	0.00	0.21			
platelets	0.0000	0.00	0.39			
ejection_fraction	0.00	0.00	0.00	0.00	0.00	0.00233
serum_creatinine	-0.01	0.00	0.00	-0.01	0.00	0.00063
	Ajd R ² = 0.06704, p-value < 0.05			Ajd R ² = 0.06608, p-value < 0.05		

Table C.15: (serum_sodium)² ~ diabetes + ejection_fraction
+serum_creatinine

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
serum_sodium						
intercept	18360	475.80	< 2e-16	18417	249.77	< 2e-16
sex	-30.90	161.60	0.85			
anaemia	151.90	138.80	0.27			
high_blood_pressure	82.22	141.60	0.56			
smoking	25.49	161.80	0.87			
diabetes	-240.00	138.70	0.08	-224.99	134.65	0.09579
age	-3.71	5.80	0.52			
creatinine_phosphokinase	0.09	0.07	0.19			
platelets	0.0006	0.00	0.42			
ejection_fraction	17.31	5.74	0.00	17.48	5.62	0.00203
serum_creatinine	-210.30	65.57	0.00	-217.11	64.31	0.00083
	Ajd R ² = 0.05514, p-value < 0.05			Ajd R ² = 0.06487, p-value < 0.05		

Table C.16: platelets ~ diabetes + sex + smoking

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
platelets						
intercept	118057.12	185374.80	0.52	270438.00	11245.00	<2e-16
sex	-30806.88	13510.95	0.02	-33203.00	13168.00	0.0122
anaemia	-8192.72	11721.25	0.49			
high_blood_pressure	9171.64	11945.59	0.44			
smoking	22882.12	13582.82	0.09	23602.00	13440.00	0.0801
diabetes	16944.30	11725.06	0.15	16470.00	11532.00	0.1543
age	-267.37	489.46	0.59			
serum_sodium	1107.75	1335.73	0.41			
creatinine_phosphokinase	2.89	5.98	0.63			
ejection_fraction	430.42	490.93	0.38			
serum_creatinine	-1423.07	5632.45	0.80			
	Ajd R ² = 0.01057, p-value < 0.05			Ajd R ² = 0.02131, p-value < 0.05		

Table C.17: $\log(\text{platelets}) \sim \text{sex} + \text{smoking}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
platelets						
intercept	12.10	0.76	<2e-16	12.47	0.04	< 2e-16
sex	-0.12	0.06	0.03	-0.14	0.05	0.00905
anaemia	-0.02	0.05	0.63			
high_blood_pressure	0.07	0.05	0.14			
smoking	0.09	0.06	0.12	0.09	0.06	0.12262
diabetes	0.05	0.05	0.33			
age	0.00	0.00	0.50			
serum_sodium	0.00	0.01	0.63			
creatinine_phosphokinase	0.00	0.00	0.65			
ejection_fraction	0.00	0.00	0.44			
serum_creatinine	-0.01	0.02	0.60			
	Ajd R ² =0.008643, p-value < 0.05			Ajd R ² = 0.01676, p-value < 0.05		

Table C.18: $(\text{platelets})^{0.51} \sim \text{sex} + \text{smoking}$

Dependent variable	Without Stepwise Function			With Stepwise Function		
	Estimation	Std. Err	p value	Estimation	Std. Err	p value
platelets						
intercept	420.87	190.77	0.03	554.12	9.75	< 2e-16
sex	-31.90	13.90	0.02	-36.54	13.50	0.00719
anaemia	-7.55	12.06	0.53			
high_blood_pressure	13.83	12.29	0.26			
smoking	22.74	13.98	0.10	21.91	13.80	0.1135
diabetes	15.12	12.07	0.21			
age	-0.29	0.50	0.57			
serum_sodium	0.91	1.37	0.51			
creatinine_phosphokinase	0.00	0.01	0.62			
ejection_fraction	0.43	0.51	0.39			
serum_creatinine	-2.39	5.80	0.68			
	Ajd R ² =0.01044, p-value > 0.05			Ajd R ² = 0.01814, p-value < 0.05		

Appendix E

1st Logistics Regression model

```
> car::vif(model_logit)
              age          anaemia creatinine_phosphokinase          diabetes          ejection_fraction
              1.136052          1.217166          1.105332          1.087607          1.152821
high_blood_pressure          platelets          serum_creatinine          serum_sodium          sex
              1.120276          1.124176          1.107929          1.069927          1.373699
              smoking          time
              1.368931          1.276122
> DescTools::PseudoR2(model_logit, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.4511851 0.3531671
> performance::performance_hosmer(model_logit, n_bins = 15)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 7.906
df: 13
p-value: 0.850

Summary: model seems to fit well.
> lmtest::lrtest(model_logit)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
serum_sodium + sex + smoking + time
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 13 -72.789
2 1 -132.629 -12 119.68 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> car::vif(step.mod)
              age ejection_fraction serum_creatinine          serum_sodium          time
              1.074863          1.147802          1.055757          1.033250          1.122938
> DescTools::PseudoR2(step.mod, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.4370960 0.3918569
> performance::performance_hosmer(step.mod, n_bins = 8)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 7.916
df: 6
p-value: 0.244

Summary: model seems to fit well.
> lmtest::lrtest(step.mod)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium +
time
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 6 -74.657
2 1 -132.629 -5 115.94 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> caret::confusionMatrix(predict_logit, test.heart$DEATH_EVENT, positive = '1')
Confusion Matrix and Statistics

              Reference
Prediction 0 1
0 51 8
1 9 20

Accuracy : 0.8068
95% CI : (0.7088, 0.8832)
No Information Rate : 0.6818
P-value [Acc > NIR] : 0.006377

Kappa : 0.559

Mcnemar's Test P-Value : 1.000000

Sensitivity : 0.7143
Specificity : 0.8500
Pos Pred Value : 0.6897
Neg Pred Value : 0.8644
Prevalence : 0.3182
Detection Rate : 0.2273
Detection Prevalence : 0.3295
Balanced Accuracy : 0.7821

'Positive' class : 1
```

Appendix F

2nd Logistics Regression Model

```
> car::vif(model_logit_upsampling)
      age      anaemia creatinine_phosphokinase      diabetes      ejection_fraction
high_blood_pressure 1.200887      1.166800      1.131459      1.151015      1.318866
      1.153542      1.227976      1.154071      serum_sodium      sex
      smoking      time
      1.434234      1.351501
> DescTools::PseudoR2(model_logit_upsampling, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.4892499 0.4236728
> performance::performance_hosmer(model_logit_upsampling, n_bins = 15)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 12.024
df: 13
p-value: 0.526

Summary: model seems to fit well.
> lmtest::lrtest(model_logit_upsampling)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
serum_sodium + sex + smoking + time
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 13 -101.25
2 1 -198.24 -12 193.98 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> car::vif(step_mod_upsampling)
      age ejection_fraction serum_creatinine      serum_sodium      time
1.128544      1.289818      1.083515      1.085613      1.263661
> DescTools::PseudoR2(step_mod_upsampling, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.4815032 0.4512368
> performance::performance_hosmer(step_mod_upsampling, n_bins=8)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 31.684
df: 6
p-value: 0.000

Summary: model does not fit well.
> lmtest::lrtest(step_mod_upsampling)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + serum_sodium +
time
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 6 -102.79
2 1 -198.24 -5 190.91 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> caret::confusionMatrix(predict_step_mod_upsampling, test_heart$DEATH_EVENT, positive = '1')
Confusion Matrix and Statistics

      Reference
Prediction 0 1
      0 46 6
      1 14 22

      Accuracy : 0.7727
      95% CI : (0.6711, 0.8553)
      No Information Rate : 0.6818
      P-value [Acc > NIR] : 0.0401

      Kappa : 0.5133

      McNemar's Test P-value : 0.1175

      Sensitivity : 0.7857
      Specificity : 0.7667
      Pos Pred Value : 0.6111
      Neg Pred Value : 0.8846
      Prevalence : 0.3182
      Detection Rate : 0.2500
      Detection Prevalence : 0.4091
      Balanced Accuracy : 0.7762

      'Positive' class : 1
```

Appendix G

3rd Logistics Regression Model

```
> car::vif(model_logit2)
      age      anaemia creatinine_phosphokinase      diabetes      ejection_fraction
high_blood_pressure      platelets      serum_creatinine      serum_sodium      sex
      smoking
      1.338031
> DescTools::PseudoR2(model_logit2, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.2877283 0.1972502
> performance::performance_hosmer(model_logit2, n_bins = 14)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 13.748
df: 12
p-value: 0.317

Summary: model seems to fit well.
> lmtest::lrtest(model_logit2)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
serum_sodium + sex + smoking
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 12 -94.468
2 1 -132.629 -11 76.322 7.554e-12 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> car::vif(step.mod2)
      age      creatinine_phosphokinase      ejection_fraction      serum_creatinine      serum_sodium
      1.103404      1.030319      1.076110      1.040961      1.020898
> DescTools::PseudoR2(step.mod2, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.2802093 0.2349703
> performance::performance_hosmer(step.mod2, n_bins = 8)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 12.231
df: 6
p-value: 0.057

Summary: model seems to fit well.
> lmtest::lrtest(step.mod2)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
serum_creatinine + serum_sodium
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 6 -95.465
2 1 -132.629 -5 74.328 1.285e-14 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> caret::confusionMatrix(predict_step.mod2, test.heart$DEATH_EVENT, positive = '1')
Confusion Matrix and Statistics

      Reference
Prediction 0 1
0 52 18
1 8 10

      Accuracy : 0.7045
      95% CI : (0.5978, 0.7971)
      No Information Rate : 0.6818
      P-Value [Acc > NIR] : 0.37046

      Kappa : 0.2474

      McNemar's Test P-value : 0.07756

      Sensitivity : 0.3571
      Specificity : 0.8667
      Pos Pred Value : 0.5556
      Neg Pred Value : 0.7429
      Prevalence : 0.3182
      Detection Rate : 0.1136
      Detection Prevalence : 0.2045
      Balanced Accuracy : 0.6119

      'Positive' class : 1
```

Appendix H

4th Logistics Regression Model

```
> car::vif(model_logit_upsampling2)
      age      anaemia creatinine_phosphokinase      diabetes      ejection_fraction
1.191468      1.145457      1.166801      1.086013      1.104292
high_blood_pressure      platelets      serum_creatinine      serum_sodium      sex
1.092328      1.160263      1.145795      1.111091      1.440661
smoking
1.454806
> DescTools::PseudoR2(model_logit_upsampling2, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.3077514 0.2472187
> performance::performance_hosmer(model_logit_upsampling2, n_bins = 14)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 16.627
df: 12
p-value: 0.164

Summary: model seems to fit well.
> lmtest::lrtest(model_logit_upsampling2)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
serum_sodium + sex + smoking
Model 2: DEATH_EVENT ~ 1
#DF LogLik Df Chisq Pr(>Chisq)
1 12 -137.23
2 1 -198.24 -11 122.02 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> car::vif(step_mod_upsampling2)
      age creatinine_phosphokinase      diabetes      ejection_fraction      high_blood_pressure
1.188984      1.060543      1.072488      1.099002      1.032448
platelets      serum_creatinine      serum_sodium
1.141890      1.094696      1.108778
> DescTools::PseudoR2(step_mod_upsampling2, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.3068098 0.2614103
> performance::performance_hosmer(step_mod_upsampling2, n_bins = 11)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 23.479
df: 9
p-value: 0.005

Summary: model does not fit well.
> lmtest::lrtest(step_mod_upsampling2)
Likelihood ratio test

Model 1: DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
high_blood_pressure + platelets + serum_creatinine + serum_sodium
Model 2: DEATH_EVENT ~ 1
#DF LogLik Df Chisq Pr(>Chisq)
1 9 -137.42
2 1 -198.24 -8 121.64 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> caret::confusionMatrix(predict_step_mod_upsampling2, test.heart$DEATH_EVENT, positive = '1')
Confusion Matrix and Statistics

      Reference
Prediction 0 1
0 47 14
1 13 14

      Accuracy : 0.6932
      95% CI : (0.5858, 0.7871)
      No Information Rate : 0.6818
      P-Value [Acc > NIR] : 0.46

      Kappa : 0.2861

      McNemar's Test P-value : 1.00

      Sensitivity : 0.5000
      Specificity : 0.7833
      Pos Pred Value : 0.5185
      Neg Pred Value : 0.7705
      Prevalence : 0.3182
      Detection Rate : 0.1591
      Detection Prevalence : 0.3068
      Balanced Accuracy : 0.6417

      'Positive' class : 1
```

Appendix I

5th Logistics Regression Model

```
> car::vif(model_logit5)
      anaemia      diabetes high_blood_pressure      sex      smoking      plat
1.059010      1.101967      1.124876      1.400586      1.342879      1.180848
sodiumn      cpkn      efractionn      screatn
1.121574      1.081543      1.078973      1.133983
> DescTools::PseudoR2(model_logit5, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.14637437 0.03609626
> performance::performance_hosmer(model_logit5, n_bin=14)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 13.032
df: 12
p-value: 0.367

Summary: model seems to fit well.
> lmtest::lrtest(model_logit5)
Likelihood ratio test

Model 1: DEATH_EVENT ~ anaemia + diabetes + high_blood_pressure + sex +
smoking + plat + sodiumn + cpkn + efractionn + screatn +
agegpn
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 12 -113.19
2 1 -132.63 -11 38.88 5.552e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> car::vif(step.mod5)
      sodiumn efractionn      screatn      agegpn
1.050780      1.021020      1.052644      1.000833
> DescTools::PseudoR2(step.mod5, c("McFadden", "McFaddenAdj"))
McFadden McFaddenAdj
0.12594589 0.08824668
> performance::performance_hosmer(step.mod5, n_bins = 7)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 3.847
df: 5
p-value: 0.572

Summary: model seems to fit well.
> lmtest::lrtest(step.mod5)
Likelihood ratio test

Model 1: DEATH_EVENT ~ sodiumn + efractionn + screatn + agegpn
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 5 -115.92
2 1 -132.63 -4 33.408 9.854e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> caret::confusionMatrix(predict_step.mod5, test.heart$DEATH_EVENT, positive = '1')
Confusion Matrix and Statistics

      Reference
Prediction 0 1
          0 52 21
          1 8 7

      Accuracy : 0.6705
      95% CI : (0.5621, 0.767)
      No Information Rate : 0.6818
      P-value [Acc > NIR] : 0.63881

      Kappa : 0.1332

      McNemar's Test P-value : 0.02586

      Sensitivity : 0.25000
      Specificity : 0.86667
      Pos Pred Value : 0.46667
      Neg Pred Value : 0.71233
      Prevalence : 0.31818
      Detection Rate : 0.07955
      Detection Prevalence : 0.17045
      Balanced Accuracy : 0.55833

      'Positive' Class : 1
```


Appendix J

6th Logistics Regression Model

```
> car::vif(model_logit_upsampling3)
      anaemia      diabetes high_blood_pressure      sex      smoking      plat
1.064576      1.070226      1.172043      1.506000      1.371974      1.183760
      sodium      cpkn      efractionn      screatn      agegpn
1.107216      1.079497      1.098723      1.183141      1.057840
> DescTools::PseudoR2(model_logit_upsampling3, c("McFadden", "McFaddenAdj"))
      McFadden McFaddenAdj
0.1693833      0.1088506
> performance::performance_hosmer(model_logit_upsampling3, n_bins = 14)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 17.571
df: 12
p-value: 0.129

Summary: model seems to fit well.
> lmtest::lrtest(model_logit_upsampling3)
Likelihood ratio test

Model 1: DEATH_EVENT ~ anaemia + diabetes + high_blood_pressure + sex +
smoking + plat + sodium + cpkn + efractionn + screatn +
agegpn
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 12 -164.66
2 1 -198.24 -11 67.157 4.225e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> car::vif(step_mod_upsampling3)
high_blood_pressure      plat      sodium      efractionn      screatn      agegpn
1.131011      1.115490      1.085386      1.073004      1.093377      1.009283
> DescTools::PseudoR2(step_mod_upsampling3, c("McFadden", "McFaddenAdj"))
      McFadden McFaddenAdj
0.1615680      0.1262573
> performance::performance_hosmer(step_mod_upsampling3, n_bins = 9)
# Hosmer-Lemeshow Goodness-of-Fit Test

Chi-squared: 13.841
df: 7
p-value: 0.054

Summary: model seems to fit well.
> lmtest::lrtest(step_mod_upsampling3)
Likelihood ratio test

Model 1: DEATH_EVENT ~ high_blood_pressure + plat + sodium + efractionn +
screatn + agegpn
Model 2: DEATH_EVENT ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 7 -166.21
2 1 -198.24 -6 64.058 6.715e-12 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> caret::confusionMatrix(predict_step.mod6, test.heart$DEATH_EVENT, positive = '1')
Confusion Matrix and Statistics

      Reference
Prediction 0 1
      0 41 16
      1 19 12

      Accuracy : 0.6023
      95% CI : (0.4923, 0.7051)
      No Information Rate : 0.6818
      P-value [Acc > NIR] : 0.9549

      Kappa : 0.1088

      McNemar's Test P-value : 0.7353

      Sensitivity : 0.4286
      Specificity : 0.6833
      Pos Pred Value : 0.3871
      Neg Pred Value : 0.7193
      Prevalence : 0.3182
      Detection Rate : 0.1364
      Detection Prevalence : 0.3523
      Balanced Accuracy : 0.5560

      'Positive' Class : 1
```