Southampton Solent University

FACULTY OF BUSINESS, LAW, AND DIGITAL TECHNOLOGY

MSc Applied AI and Data Science
2021-2022
Talabi Hadizat Yetunde
"Prediction of survival of heart failure patients using machine learning models"

Supervisor : Dr Jarutas Andritsch
Date of submission : Sept. 2022

This report is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc Applied AI and Data Science.

Acknowledgement

My heartfelt gratitude goes to Allah (SWT) for His mercies, assistance, and grace throughout my life, especially during this MSc programme.

Dr Jarutas Andritsch, my project supervisor. I'd like to thank you for all your advice, support, and confidence in my abilities. I also appreciate your constant constructive criticism and gentle nudges to improve my academic research.

To my family, I am at a loss for words and will never be able to express how grateful I am for your support. My heartfelt gratitude goes to my incredible mother, Alhaja Shakirah Talabi, and my brother, Olawale Talabi, for their significant investment in my life—financially, spiritually, and emotionally. My daily prayer is to be a huge blessing to you both in return.

I am extremely grateful to my best friend, my purpose partner Abayomi Wasiu Kayode who has provided me with unwavering financial, emotional, moral, and academic support. I am deeply grateful for his help. Without his encouragement, I probably would not have even started this course.

Finally, a sincere thank you to all the outstanding academic researchers and practitioners whose work is cited in this thesis. You have given young academic researchers like me a solid foundation on which to build our careers because of your tremendous effort and contribution to research.

Abstract

Heart failure is one of the leading causes of death worldwide. Human survival, one of the functions the heart regulates, demands that the heart be safeguarded and aware of its danger. Therefore, the well-being of the heart determines whether humans will survive. A predictor tool is needed to address the issue of cardiovascular disease-related death events to reduce or prevent mortality and increase longevity. Data analytic techniques can be employed to identify patterns and relationships in patients' health records that are not currently recognised by clinical specialists. The predictive capabilities of Artificial Intelligence make it easy to accurately project future occurrences and results based on known records.

The aim of this study is to use machine learning to predict heart failure patient survival and to create a system that can predict heart failure patient survival using significant features The "Heart Failure Clinical Records dataset," which includes data on 299 hospitalised patients, is used in this study to analyse heart failure survivors. Using the Random Forest algorithm to determine the importance of features, analysis of this data demonstrated that the most important features were Age, Ejection fraction, Serum creatinine, Serum sodium, and follow-up time. The ensemble learning approach was utilised to maximise the benefit of multiple model predictions. K-nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree Classifier are the base models used (DTC). The five most crucial features in the dataset were used to train the models. A hard voting ensemble approach was used, and the evaluation resulted in an accuracy score of 88%. Whereas the recall and f1 score for the deceased (Not alive) is lower than anticipated (62 and 74), this could be attributed to the relatively small number of instances in this category, but overall, the model's prediction is within the acceptable range. Simple software that predicts patient survival based on user input was created using the pre-trained voting classifier model.

CONTENTS

TABLES

FIGURES

Acronyms

CHAPTER ONE

1.1 Introduction And Background

Circulation of blood throughout the human body is carried out by an organ, located in the front of the chest, slightly behind the left breastbone known as the "heart". This act makes this organ essential to human existence (Rahayu *et al*. 2020). One of the functions that the heart controls, human survival, necessitates that the heart be protected and aware of its damage. The heart uses the blood as a carrier to transport vital nutrients and fluid to other parts of the body. If it is unable to perform its function, the brain and other bodily organs stop working, and the person will die within a few minutes. Any condition that prevents the heart from functioning properly is known as cardiovascular disease.

It may be challenging to diagnose cardiovascular disease due to several contributing factors, such as hypertension, high cholesterol, diabetes, an irregular heartbeat, and numerous other conditions (CVD). The way symptoms of CVD present themselves can occasionally vary by gender. Male patients, for illustration, are more likely to experience chest pain, whereas female patients may also experience nausea, severe exhaustion, and shortness of breath. Numerous techniques have been developed by researchers to predict heart disease but doing so early on is not very effective for a number of reasons, including but not restricted to method accuracy, complexity, and execution time. Because of this, many lives can be saved with an accurate diagnosis, management, and treatment (Ishaq *et al*. 2021).

According to the WHO, the leading cause of death worldwide is cardiovascular disease (CVD It kills an estimated 17.9 million people each year (WHO 2021). It is also referred to as heart disease. At least 26 million people are affected by this escalating global pandemic (Savarese and Lund 2017). Various illnesses that affect the blood vessels and heart muscles collectively are referred to as heart disease (arteries and veins that transport blood to the heart, brain, and other

organs). Heart diseases increase healthcare spending while also reducing an individual's productivity (Rindhe *et al.* 2021). These severe conditions include aneurysms of the aorta, heart attacks, coronary heart disease, heart failure, ischemic heart disease, stroke, congenital heart disease, endocarditis, hypertension, peripheral artery disease, and rheumatic heart disease (WHO 2021).

Heart failure is one instance of a cardiovascular disease that possesses a high morbidity and mortality rate. It is a clinical syndrome brought on by structural and functional myocardial abnormalities that impair ventricular filling or blood ejection. The status of any heart disease in which, despite adequate ventricular filling, cardiac output is reduced, or the heart is incapable of pump blood quickly enough to meet the body's needs while maintaining normal function parameters (Chicco and Jurman 2020).
Medically, people who have heart failure have a syndrome that includes the typical signs and symptoms (such as pulmonary crackles, displaced apex beats, and elevated jugular venous pressure) brought on by a heart condition that is abnormal in both structure and function (Authors/Task Force Members *et al.* 2012). Shortness of breath at rest or during exertion are the main symptoms, along with fatigue and indications of fluid retention, such as swollen ankles. Altered heart structure or function could also exist. Heart failure can be challenging to diagnose, especially in the early stages. Many heart failure symptoms lack specificity and are therefore ineffective in distinguishing between heart failure and other health issues, even though symptoms cause patients to seek medical attention (King, Kingery and Casey 2012). The term "heart failure" describes a failing heart, which can result in insufficient oxygen delivery to all body organs. Oxygen is required by living cells because it is crucial to their metabolism. Multicellular organisms that are large, complex, and active require more energy for their metabolism than can be supplied by simple molecular diffusion of gases and nutrients in tissues (Carreau *et al.* 2011). Giving tissues and cells the right amount of oxygen is therefore essential.

According to data, heart failure is a major and expanding global public health concern. Heart failure will occur more frequently despite the incidence being

predicted to stay stable due to the ageing population and advancements in treatment (Savarese and Lund 2017). Heart failure Syndrome was first recognised as an impending epidemic about three decades ago. The annual mortality rate from heart failure ranges from 5% to 75% (Levy *et al*. 2006). Due to population growth and ageing, the overall number of heart failure patients is still rising in the modern world. However, it appears that the case mix for heart failure is altering. Alarmingly, the opposite trends have been seen in the relatively young, possibly due to an increase in obesity, while the incidence has decreased or even stabilised in some groups (Groenewegen *et al*. 2020).

However, other potential causes of HF include dysfunction of the pericardium, myocardium, endocardium, heart valves, or great vessels alone or in combination. Reduced left ventricular myocardial function is the most common cause of HF (Inamdar and Inamdar 2016). Other conditions like HIV, alcoholism or cocaine abuse, thyroid issues, an abundance of vitamin E in the body, radiation or chemotherapy are among the other causes of heart failure, in addition to coronary artery disease, excessive salt consumption, water retention, diabetes, obesity, inactivity, high blood pressure, and others (Ahmad *et al*. 2017). Some of the major pathogenic mechanisms causing HF include increased hemodynamic overload, ischemia-related dysfunction, ventricular remodelling, excessive neuro-humoral stimulation, abnormal myocyte calcium cycling, excessive or insufficient extracellular matrix multiplication, exacerbated apoptosis, and genetic mutations (Inamdar and Inamdar 2016). Heart failure can occur suddenly due to a variety of conditions, including postpartum cardiomyopathy, myocarditis, tachycardia, bradycardia, septic shock, heart blockage, anaemia, and sleep apnoea, among others. To determine whether heart failure can be reversed, alternative causes should be identified, treated, and monitored as soon as possible (King, Kingery and Casey 2012).

Heart failure (HF) is a significant contributor to low life expectancy, a leading cause of premature death, and a high incidence of hospitalisation (Gilbert *et al*. 2006). In almost all areas of human activity, including the healthcare sector, machine learning has recently gained significant traction, according to well-known studies (Oladimeji and Oladimeji 2020). Machine learning, a subset of

artificial intelligence, enables computers to study history and make predictions. The factors that might predict survival can be defined using machine learning to better manage these individuals. To implement proper management of heart failure patients, lower mortality, or increase patient survival, the most important clinical traits (or risk factors) that may cause heart failure must be identified beforehand.

Heart failure can be categorised based on the location of the deficit, when it first appeared and the heart's functional state. Biventricular, left ventricular or right ventricular deficits can all occur depending on where the deficit is located. Acute and chronic conditions can be distinguished based on when they first manifest. Clinically, heart failure with preserved ejection fraction (HFpEF) and heart failure with reduced ejection fraction (HFrEF) are the two main categories of HF based on the functional status of the heart (Pearse and Cowie 2014). In older adults and females, HFpEF is prevalent.

Four functional classes of HF are identified by the New York Heart Association (NYHA) functional classification as follows:

Class I: Physical activity limitations and common physical activity are unaffected by HF.

Class II: HF causes mild restrictions on physical activity; patients are comfortable when at rest, but routine exercise triggers HF symptoms.

Class III: Patients are comfortable at rest, but HF symptoms are triggered by less-than-normal activity. HF severely restricts physical activity.

Class IV: Patients with HF are unable to engage in any physical activity without experiencing HF symptoms, or they experience symptoms even when they are at rest. (Inamdar and Inamdar 2016).

Data mining techniques have been successfully applied in several prominent fields, including marketing and e-commerce. It is acknowledged in the health sector as well (Canlas 2009). Despite disparities and conflicts in approaches, the need for data mining in the health sector has increased. Data mining is the

process of examining data to uncover hidden details that can be used to make critical decisions for the future. It is very challenging, if not impossible, for humans to sort through the enormous volume of data stored in medical databases and find knowledge (Cheng, Wei and Tseng 2006). As a result, machine learning algorithms must be used to help medical professionals analyse data for better health policy, disease outbreak detection, and avoidable hospital deaths.

1.2 Statement of Problem

Investigating the factors that contribute to death in people suffering from heart failure. In the absence of a system that can detect the risk factors that influence patients' survival or a survival predictor that can serve as guidance for physicians and patients, the death rate will continue to rise.

1.3 Research Questions

This research looked for answers to the following specific questions considering the goal.

1. What current health conditions (variables) exist in heart failure patients that may indicate a high risk of death?
2. What are the factors/variables that ensure the survival of patients with heart failure?

1.4 Research Aim

The aim of this study is to use machine learning to predict the survival of heart failure patients and to develop a system that can predict the survival of heart failure patients based on pre-existing patient conditions.

1.5 Objectives

- Acquisition of relevant medical data.
- Conduct an exploratory analysis of the data.

• To conduct a feature correlation analysis.

• To determine which characteristics are most important in predicting a patient's survival with heart failure.

• Conduct a cluster analysis to identify categories and patterns.

• To develop a simple-to-use, publicly available medical data-based model for predicting the survival of heart failure patients.

• Performance testing of the model.

• Determine the best software platform for developing a web application.

1.6 Significance of The Study

Even though heart failure is among the leading causes of death, each patient's prognosis is different. To prevent death, it is crucial to pinpoint the most important characteristics. Most cardiovascular patient deaths happen when the illness is discovered later, which can result in heart failure. Therefore, early diagnosis will reduce mortality or improve patient survival. Both the management of known heart disease patients and the prevention of heart failure can be accomplished by a well-equipped clinician who can predict a patient's survival. Survival prediction entails the prevention of all diseases that can lead to heart failure. Heart disease-related medical procedures are well known to be pricy and time- and resource-intensive. Heart failure patients should be managed properly to conserve resources, time, and money.

1.7 Proposed Artefact

The suggested artefact is a web-based application that employs a trained classification model to predict the survival of patients with heart failure based on relevant variables.

1.8 Value Propositions

The availability of a system that predicts the survival of heart failure patients based on key characteristics will facilitate:

1. Helpful in assisting clinicians in predicting patient survival through the analysis of relevant features.
2. A decline in hospitalisation rates will result in less time, money, and wasted resources on healthcare.
3. Being able to predict a patient's likelihood of survival based on their overall risk profile will make it simpler to decide which patients require more intensive monitoring and treatment.
4. To choose the appropriate populations to test prospective new treatments on to gauge their impact on survival.
5. A knowledgeable, well-equipped physician with the ability to predict heart failure patients' survival will be better able to prevent mortality by providing high-quality medical care. This will result in better services and more advanced medical knowledge. Most patients can benefit from effective treatment that raises their chances of survival and quality of life, even though there is no known cure for heart failure.
6. A decline in heart failure-related deaths globally.

1.9 Research Project Specifications

This describes the functionality of the system, the programming language , tools and libraries utilized during this project. The next section describes the specifications.

1.9.1 Software Functionality

The table below describes the functional and non-functional features of the application.

Table 1: System Prerequisite

Functional	• Predicts the survival of heart failure patients using machine learning. • Clients should be able to input values for pre-existing conditions and predicts survival based on the inputs. • Display a survival prediction based on the inputs.
Non-functional	• Intuitive software • Easy to use GUI

1.9.2 Tools and Technologies

The following resources were used to execute the project:

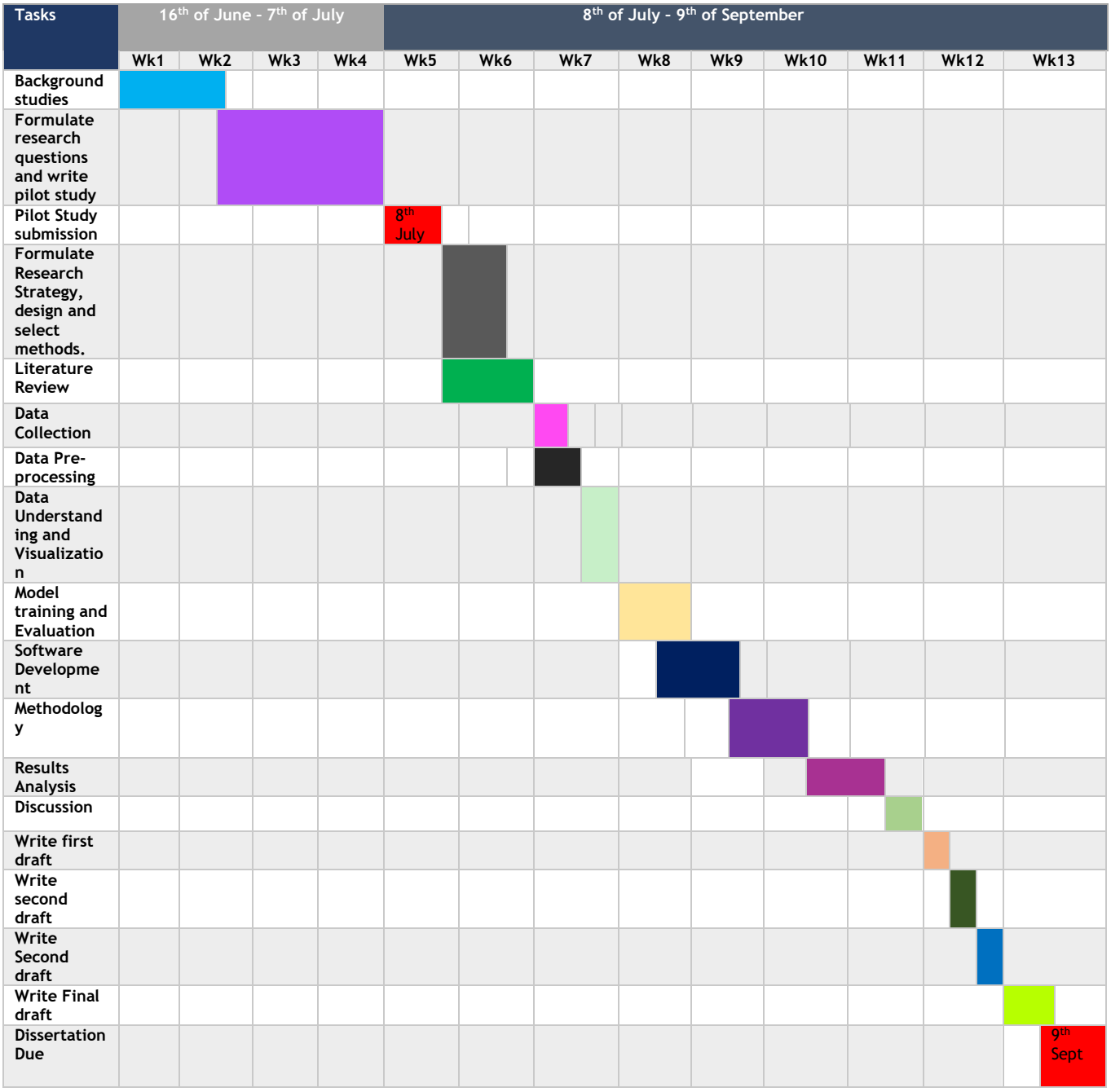
1) Pycharm: This is an Integrated Development Environment (IDE) with a wide range of features for writing, compiling, debugging, and monitoring resources. It also allows visualization and code mapping functions (Hu, Ma and Zhao 2018).
2) UCI machine learning repository: A repository to obtain a secondary dataset with the right variables that can provide the answers to the researched questions.
3) Python: A well-designed programming language that can be applied to real-world projects. Python is a simple, open source, powerful, portable language that supports a variety of other technologies (Srinath 2017)
4) Tableau: It is a visualisation tool that was used to implement clustering analysis to identify hidden patterns in the dataset and compare them to the expected classes.
5) Excel: This is a visualization tool that was utilized to achieve binary distribution of variables and bivariate analysis (Lilly and Miller 2021).
6) Python data loading and manipulation packages, such as:
    i. NumPy: A library that is used to process high-level mathematical functions and multidimensional arrays (Bressert 2012).

ii. Pandas: A package for efficient and intuitive structured data handling and processing.
iii. Matplotlib: To portray data in a visual format, such as graphs and plots, to gain understanding or effectively convey information.
iv. Seaborn: To produce detailed statistical graphs like violin diagrams and heatmap.
v. Scikit-Learn: Through a Python interface, it was used to create a powerful machine learning and statistical modelling tools for tasks like classification, regression, dimensionality reduction, predictive analytics, and a wide range of other uses. Python's Scikit-learn package incorporates a variety of cutting-edge machine learning techniques for both supervised and unsupervised problems. (Pedregosa *et al*. 2011).
vi. Joblib: It was used to save the trained model.
vii. Streamlit: This is a web application framework for creating and developing Python-based web applications that generate results and provide interactive experiences. Streamlit is also extremely quick and flexible to implement and deploy (Mitheran, Narayanan and Singaravelu 2022).
viii. Python Imaging Library: An image was read and included in the web application using the open-source Python Imaging Library (expansion of PIL), which is a language for image processing (Lindblad and Kinser 2013).

1.10 Project Plan and timetable

The Gantt chart that follows shows the timetable for the entire research and implementation process.

Table 2: Gantt chart for research implementation



1.11 Chapter Summary and Conclusion

This chapter described the scope of the ongoing research. The domain (medical) in which it's being carried out and the focused group (heart failure patients). Additionally, the rationale for conducting this research is on the survival of patients with heart failure. These have all been briefly discussed, hence the remaining chapters of this thesis will run in the order listed below.

1.12 Thesis Outline

The structure of this thesis is as follows:

Chapter 2 provides a concise overview of previous works and lays the groundwork for the research. The goal of this chapter is to understand the current state of the research (and to show that knowledge). Also, the various methods that have been established on this topic are also discussed in this chapter.

Chapter 3 focuses on the research methodology, as well as the conceptual framework and theoretical underpinnings of the current study. Also discussed are the methods and approaches employed in this study.

Chapter 4 centres on the outcomes of the methods and the techniques adopted in chapter three.

Chapter 5 provides detailed findings deduced from the analysis through discussion.

Finally, Chapter 6 will concentrate on the study's findings. This will cover the overall conclusions, the benefits that follow, potential study extensions, the study's limitations, and future research recommendations.

CHAPTER TWO

2.1 Literature Review

It has frequently been reported that the mortality rate for heart failure patients who are discharged from the hospital is high. Numerous studies have investigated the factors that may lead to early death or hospital readmission in these patients. This study is not the first of its kind; other researchers have conducted studies that are comparable to this one. Using logistic regression, Bouvy et al. assessed the information from surveys on quality of life, drug use, physical examination (such as blood pressure), and history that independently contributed to mortality prediction. It was found that a patient's medical history and physical examination could predict an event that would result in death within 18 months (Bouvy *et al*. 2003).

In 2017, Ahmad et al. released a dataset of patient medical records from the Institute of Cardiology and Allied hospital in Faisalabad, Pakistan, that had been previously gathered. Age, ejection fraction, serum creatinine, sodium, anaemia, platelets, creatinine phosphokinase, blood pressure, gender, diabetes, and smoking status were all considered by the authors as they used Cox regression to estimate the mortality among heart failure patients. Using a Kaplan Meier plot to analyse the overall pattern of survival, it was found that mortality had a high impact in the early stages of the study and then steadily increased until its conclusion. The main risk factors for increased death events in heart failure patients were determined to be advancing age, renal dysfunction, high blood pressure, and a lower ejection fraction (Ahmad *et al*. 2017). It's worth noting that they've made the dataset open to the scholarly community.

For patients at high risk, Melillo et al. suggested an automatic classifier that would distinguish them from patients at low risk. With a 93.3 percent sensitivity and a 63.5 percent specificity in their study, the classification and regression tree (CART) performed better. They only looked at 12 low-risk and 34 high-risk patients. To examine the efficacy of their suggested method, a larger dataset is required (Melillo *et al*. 2013).

In 2018, Chala Beyene et al. suggested using data mining techniques to predict and analyse the occurrence of heart disease. The main goal is to predict the occurrence of heart disease to provide an early automatic diagnosis of the disease with a quick result. The proposed methodology is crucial in a healthcare organisation where specialists lack knowledge and skill. To ascertain whether a person has heart disease, various medical factors are taken into consideration, including age, gender, blood sugar levels, and heart rate. WEKA software was used to compute dataset analyses (Beyene and Kamat 2018).

Furthermore, in 2019, Zahid et al used gender-based models to observe the risk variables connected to each patient's likelihood of dying. It was concluded that there were similar patterns of survival in males and females. Hence, gender does not correlate to the deadly events of the patients (Zahid *et al*. 2019).

A clinical support system (CDSS) for heart failure analysis was examined by Guidi et al. They investigated the performance of various machine learning classifiers. Random forest and CART outperformed other methods with an accuracy of 87.6% (Guidi *et al*. 2014).

A system was suggested by Shah et al. to study various heart conditions as well as the main causes of fatalities. These included Decision Tree (DT), Naive Bayes (NB), RF, and KNN, among other supervised machine learning algorithms. Because their research goal is to create an accurate and effective system with fewer attributes, only 14 out of the 76 attributes were used. KNN excelled among four supervised machine learning classifiers. To enhance the classification outcomes, ensemble approaches could be used (Shah, Patel and Bharti 2020).

In 2020, Researchers used only two features, serum creatinine and ejection fraction, across ten machine learning techniques to predict patients with heart failure survival. Random Forest performed better, reaching a maximum accuracy of 74%. (Chicco and Jurman 2020). They also suggested additional machine learning-based research on different cardiovascular disease datasets containing more variables.

In a different study, Sri Rahayu et al. estimated the survival of heart failure patients using SMOTE (Synthetic Minority Over-Sampling Technique), data mining methods, and multiple classification algorithms (Chicco & Jurman, 2020). This was done to determine which algorithm would be the most effective to use in the dataset. It was determined that the Random Forest algorithm outperformed other algorithms with a maximum accuracy of 85.82 percent when pre-treatment resampling techniques were used (Rahayu *et al*. 2020). It was suggested that additional research be conducted by later researchers who will implement the pattern that is formed and develop software to predict the survival of patients with heart failure.

Stephan examined 268 ambulatory heart failure patients who were prospectively enrolled as part of the Studies Investigating Co-morbidities Aggravating Heart Failure (age 67.1 ±10.9 years, New York Heart Association class 2.3 ±0.6, left ventricular ejection fraction 39.3%, and 21% female) in a different study. In 47 people, muscle wasting was found to be a standalone predictor of mortality. (17.5 per cent) with ambulatory heart failure (von Haehling *et al*. 2020).

Another study, published in 2021, used SMOTE and effective data mining approaches to identify variables that could improve the accuracy of cardiovascular patient survival prediction. This was accomplished using a dataset of 299 heart failure survivors. Among the nine classification models used to predict survival were the Adaptive Boosting classifier (AdaBoost), Logistic Regression (LR), Stochastic Gradient classifier (SGD), Random Forest (RF), Gradient Boosting classifier (GBM), Extra Tree Classifier (ETC), Gaussian Naive Bayes classifier (G-NB), and Support Vector Machine (SVM). The experimental results show that ETC outperforms other models in predicting cardiac patient survival, with a 0.92 accuracy value when combined with SMOTE (Ishaq *et al*. 2021). Additionally, They also suggested employing multiple combinations of machine learning models in future studies to benefit from their synergistic effects and to improve feature selection methods to boost the effectiveness of machine learning models.

2.2 Research gap

The review of the literature revealed that numerous factors have been studied by researchers on various datasets, and various approaches have been taken into consideration for predicting the occurrence of death events in heart failure patients using machine learning. Some of the gaps in this area have been highlighted by a thorough analysis of the literature that is currently available on the prediction of survival in heart failure patients. Due to this, two gaps have been found that this research seeks to fill. Firstly, most of the existing studies have only concentrated on creating a model to forecast the survival of patients with heart failure. However, there has not been any software building for this study. Hence, this study will endeavour to address this gap by creating software for prediction. Secondly, of all the existing literature on survival prediction of heart failure patients, none of these studies has implemented the patterns formed in the dataset. Hence, this study will examine data to identify and put formed patterns into practice. Additionally, the suggested framework will not only create software but also conduct research using a variety of machine learning models and better feature selection methods to enhance model performance.

CHAPTER 3

3.0 Research Methodology

3.1 Introduction

This chapter provides a brief overview of the research methodology that would be used to carry out the project goals and objectives. In addition to what makes up a research approach, the various categories, the methodological philosophies that underpin the current study, and the justification for choosing the chosen approach are discussed.

3.2 Methodology

This relates to the research methodology that was employed. It is also referred to as the research's plan or strategy (Creswell 2014). On the other hand, research methods are simply all the methods used during a research project; as such, they can be regarded as a subset of research methodology. More comprehensive than research methods, the research methodology entails identifying, selecting and utilization of the most effective approach to address a research issue (s). The purpose of conducting a research project, a statement of the problem, identifying research questions, the method of hypothesis formulation, appropriate collection of the most suitable data, the method of data analysis, and then coming to conclusions that may be either generalisation for other theoretical formulations or solutions to the current problem and why it's been chosen are all considered to be components of research methodology (Kothari 2004). There are three holistic methods of research methodologies which include the quantitative approach, qualitative approach, and mixed methods of approach. These approaches are frequently used in carrying out research projects (Opoku, Ahmed and Akotia 2016). Quantitative approaches generate numerical data while qualitative approaches produce textual data and mixed approaches encompass both numerical and textual data. To gain a thorough understanding of the topic at hand, this project will employ hybrid research methods (Qualitative and Quantitative approaches). This methodology will include the gathering of

pertinent medical data, qualitative data assessment, exploratory data analysis, data visualisation, training of a predictive model, model evaluation, and software development.

3.3 Dataset Description

A dataset will undoubtedly serve as the research's main resource. Data will be gathered utilising a mixed method approach for improved comprehension. To help researchers better understand their research problem, a mixed-methods approach can combine the advantages of both quantitative and qualitative research (and the resulting data)(Creswell 2014). To predict the survival of heart failure patients, there is a need to collect a dataset. From the UC Irvine Machine Learning (UCI) repository website, a secondary dataset titled "Heart Failure Clinical Records" was retrieved as an excel file for use in this project. The dataset is made up of the medical records of 299 heart patients who were retrieved during the follow-up period collected at the Institute of Cardiology and Allied Hospital in Faisalabad Pakistan. Each patient profile contains 13 clinical features, including age, anaemia, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, platelet count, smoking and time. The dataset is a numeric one that was released to the public in 2020. Out of 299 records, 194 are men and 105 are women. There are 203 non-smokers and 96 smokers. 125 people have diabetes, compared to 174 who do not. There are 105 known hypertensive patients and 194 non-hypertensive patients. All the patients are over the age of 40. In the death-event class, 0 denotes being alive and 1 denotes being dead. The ejection fraction ranges from 14 to 80%. According to the American Heart Association, a borderline ejection fraction is between 41 and 50 percent, while a normal one ranges from 50 to 75 percent (Fletcher *et al*. 1990). Therefore, there are abnormal ejection fractions. Creatinine phosphate ranges between 23 and 7861mcg/L. Platelets count is between 25000 and 850,000. Serum creatinine ranges between 0.5 and 9.4. Serum sodium is between113 and 148(mEq/L). Table 3 provides a comprehensive description of the dataset.

Independent variables: Age, anaemia, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, platelet count, smoking and time

Dependent variable: Death_event

Table 3: Dataset Description

Index	Name	Description	Data Types	Instances/Range
1	Age	Patients' ages in years	Numeric Ratio Int64	40-95(years)
2	Anaemia	Reduction in haemoglobin or red blood cells	Categorical Nominal Int64	0-not anaemic 1-anaemic.
3	Creatinine phosphate	Blood CPK enzyme concentration (mcg/L)	Numeric Ratio	23-7861
4	Diabetes	Whether or not the patient has diabetes. (Boolean)	Categorical Nominal Int64	0-NO 1-YES.
5	Ejection fraction	Amount of blood leaving the heart during each contraction (percentage).	Numeric Ratio Int64	14-80.
6	High blood pressure	Whether or not the patient has hypertension (Binary).	Categorical Nominal int64	0-No 1-Yes
7	Platelets	Blood platelet content (Kiloplatelet/mL)	Numeric Interval Float	25000.00- 850000.00
8	Serum creatinine	Level of serum creatinine in the blood (mg/dL)	Interval Float	0.5-9.4
9	Serum sodium	Level of serum sodium in the blood (mEq/L)	Numeric Interval Int64	113-148
10	Sex	Either a man or a woman (Binary)	Categorical Nominal Int64	0-Female 1-Male
11	Smoking	Whether or not the patient smokes (Binary).	Categorical Nominal Int64	0-NO 1-YES
12	time	Period of follow up (days)	Numeric Interval Int64	4-285
13	Death_event	If the patient deceased during the follow-up period or not (Binary).	Categorical Nominal Int64	0-Alive 1-Deceased

3.4 Data Pre-processing

Data pre-processing is the art of organising and cleaning data so that it is well-formed. It is a technique for improving data quality and preparing data for analysis (Kamiran and Calders 2012). It is crucial to pre-process real-world data before using it because it is unclean, overly complex, and inaccurate. Before manipulation, it is important to comprehend the data to ensure that we have the correct data and gain insight into it.

3.5 Exploratory Data Analysis (EDA)

Exploratory data analysis, or EDA for short, is a term coined by John W. Tukey to describe the act of looking at data to see what it appears to say (Tukey 1977). It employs statistical and graphical methods of analysis to uncover important information in a dataset (Martinez, Martinez and Solka 2017). The analysis involves univariate, bivariate and multivariate data analysis.

3.5.1 Univariate Analysis

Important information regarding a single variable occurring independently to be analysed can be provided by descriptive statistics to identify odd values and inconsistent data (Park 2015). The central tendency of each variable (continuous variable distribution) was measured by its mean, median, and mode. Histograms and box plots were created to visualize and understand variable distribution i.e. compared to the theoretical standard normal distribution and to display outliers. Since a dataset's non-normality can produce invalid and unreliable results, checking for normality has been regarded as a crucial statistical step that should be performed before conducting other statistical analyses (Das and Imon 2016). Other statistical tests of normality were performed like skewness and kurtosis which were presented using a numerical approach. Numerical techniques offer unbiased ways to assess normality, whereas graphical techniques are simple to understand and intuitive (Lo *et al*. 1995).

Visualization was carried out in excel software to identify and understand binary variables distributions and bivariate analysis.

3.5.2 Bivariate Analysis

Bivariate analysis is the examination of two variables (usually the outcome variable and the independent variables) to determine the relationship between their values (Agresti 2010). Excel was used to conduct this analysis. Excel was utilised to complete this analysis. For univariate, bivariate, and multivariate distribution plotting, Excel is the most widely used and potent general-purpose spreadsheet software (Vidmar 2007).

Qualitative bivariate analysis was carried out between each binary independent variable and dependent variable (death-event). By using two categorical variables in a qualitative bivariate analysis, it is possible to see how the categories of the independent variable influence the values of a particular outcome variable (Agresti and Finlay 2009). Additionally, a bivariate analysis incorporating both quantitative and qualitative variables was performed (continuous independent variables and outcome). In this instance, a preliminary analysis was performed using the line graph to compare the mean values of a continuous variable to the outcome variable category (Shahabi, Ahmad and Khezri 2013).

3.5.3 Multivariate Analysis

Univariate and bivariate methods have historically been employed to analyse the data. In univariate analysis, a single variable is statistically tested, whereas, in bivariate analysis, two variables are statistically tested. Inherently multidimensional problems involving three or more variables necessitate the use of multivariate data analysis (Joseph *et al*. 2010). In this context, multivariate analyses such as principal components analysis, cluster analysis, and correlation analysis amongst independent features and dependent variable were carried out. These were implemented in python.

3.5.3.1 Correlation Analysis

Correlation analysis is required because machine learning models can underperform if fitted with highly correlated feature data. The strength of a relationship between variables is measured or evaluated using correlation analysis. The numerical index known as the coefficient (r or p(Rho)) whose values range from -1 to +1 indicates how closely

related variables are to one another and how much their variation affects each other (Gogtay and Thatte 2017). A correlation matrix and heatmap were used to visualise the relationship.

3.5.3.2 Principal Components Analysis

When working with fewer variables, relationships between them could appear evident, but when working with more variables, it becomes necessary to access these relationships more quickly. The variation of some axes in data sets with numerous variables may be high, while the variance of others may be low and hence can be disregarded. One can start with 20 original variables but end up with only three or four meaningful axes. This is known as dimensionality reduction of a dataset. Principal Components Analysis, or PCA, is the term for this method of rotating data so that each axis shows a decreasing distribution of variance (Holland 2008). The basic aim of this PCA is to find and analyse principal components in the dataset, remove redundancy and identify the most significant features for better analysis and increasing interpretability as well as minimising information loss (Maćkiewicz and Ratajczak 1993). The mechanism of action is by transforming input data into a space with components rather than features (Jolliffe and Cadima 2016). PCA was performed on this dataset to identify the number of possible principal components.

3.5.3.3 Clustering Analysis

Cluster analysis was used to identify meaningful hidden patterns and homogeneous groups in the set of variables. However, it is also used to confirm or compare with previously reported classes. Cluster analysis is a popular and rapidly growing method for analysing multivariate data (Kettenring 2006). It is used to identify homogenous groups in a dataset (Scott and Knott 1974). It is an unsupervised machine learning method that partitions the dataset into several meaningful subgroups (clusters) based on similar attributes. The partitioning is done so that intracluster differences are minimal and inter-cluster differences are maximized. Examples of clustering algorithms include K-means, K-meloids, Density-based spatial clustering of application with noise (DBSCAN), and OPTICS

(Abbas 2008). K-means method of clustering was adopted to achieve hidden patterns in the dataset. The algorithm works by splitting the data into k-clusters, with each cluster's centroid being the mean value of all the data points in it. K-means uses an iterative process to find cluster centres by minimising the distance between each cluster point and the cluster centre.K-means assumes all directions are equally important.

3.6 Feature Ranking/Selection

Feature Selection is an example of a dimensionality reduction(DR) technique which helps for more accurate categorization, compression, and visualisation of high-dimensional data by minimising undesirable features of high-dimensional environments (Van der Maaten 2007). It is a process of reducing the number of input variables when developing a predictive model. DR helps to reduce the overfitting of a model, enable visualization, eliminate feature redundancy, lower computational resources, and avoid slow development of model. Data visualisation makes it easier to spot and explain any hidden patterns that may be present in a dataset. Visualising the features, also makes it easier to gain a better understanding of the dataset. Regarding the dependent variable, it is critical to evaluate the significance of each feature for decision-making (Rogers and Gunn 2005). Feature importance describes this analysis, and it helps qualitatively to get more information about the dataset. The feature ranking method was applied using Random Forest (RF). Random Forest uses multiple decision trees to achieve this analysis. However, the topmost important attributes can then be selected to reduce dimensionality.

3.7 Process Diagram

The following chart describes the proposed framework for the research processes (figure 1).

Figure 1: Proposed Framework of the Research Process

3.8 Model Development

Model development is the act of evaluating the capacities, selecting, training, and deploying appropriate models to solve the defined problem. This development involves identifying a specific problem and associated method of success metrics, collecting a suitable dataset to address the problem, selecting the most suitable model to address the problem and applying of the model to

solve the problem (deployment). The first two steps (problem definition and data collection) have been sorted. However, the remaining two steps which are selecting the right algorithms and model fitting would be discussed below, along with projected success metrics.

3.8.1 Selecting algorithms

This is an important consideration that occurs before model fitting. It entails selecting the proper machine learning algorithms based on factors including the technical nature of the described problem, the appropriate theoretical model for the problem, the type of dataset and the algorithm with the best likely performance for the problem case dataset. Regression difficulties and classification challenges are two categories of supervised learning issues. When outputs are continuous, an issue is called a regression problem; when they are categorical, a problem is called a classification problem (Sen, Hajra and Ghosh 2020). Classification algorithms are algorithms that solve a classification problem. The necessity for a classification algorithm stems from the nature of the described topic, which is a classification problem based on the research's focus on "Survival prediction of heart failure patients". The research focus is a binary classification problem which involves only two classes. Examples of classification algorithms considered in this study are K-nearest neighbour (KNN), Support vector machine (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC).

The K-Nearest Neighbour Classifier (KNN) is perhaps the simplest classifier in the toolbox of machine learning techniques. Classification is accomplished by locating the closest neighbours to a query sample and utilising those neighbours to determine the class of the query (Cunningham and Delany 2021). However, it predicts based on the majority. It carries out similarities' metric known as Euclidean distance between trained dataset and test data concerning closest K-number of records, where k is the number of neighbours to be considered. It is suitable for the focused problem being a classification algorithm.

Support Vector Machine is a known supervised machine learning algorithm used for classification, as well as regression problems. It outperforms other classifiers

in terms of accuracy. However, it has a unique feature known as the kernel trick for separating classes that cannot be classified linearly (non-linear spaces). Its primary mode of action is to split data points into classes by creating a hyperplane with the greatest margin (Pradhan 2012). A hyperplane can be a point in one-dimensional space, a line in two-dimensional space, or a surface in three-dimensional space. It is appropriate for the focused issue, which is a classification problem "Survival prediction of heart failure patients". However, it is also well recognised to be highly effective on both small and large datasets with numerous features.

Logistic Regression (LR) is the most popular supervised machine learning algorithm for binary classification and proportional response datasets (Agresti 2018). However, its expansion to multi-class classification issues and naturally providing probabilities are two of the key benefits of LR (Karsmakers, Pelckmans and Suykens 2007). The logistic model is used to express a probability, which is always between 0 and 1 (binary). It also enables the quantification of the relationship between the occurrence of an event (qualitative dependent variable) and the factors that can influence it (independent variables) (El Sanharawi and Naudet 2013). logistic regression accepts input and predicts which class the input belongs to (output) using a sigmoid function. It uses three coefficients as parameters. The mechanism of action is by using **stochastic gradient descent** to estimate the best values of the coefficients based on the training dataset to generate the output and then transform using a logistic function to separate points into classes. It is suitable for the focused topic, which is a classification problem titled "Survival prediction of heart failure patients" which is a binary classification. It also could tackle complex nonlinear data (Wu and Li 2018).

Decision Tree Classifier (DTC) is a well-known supervised machine learning classification algorithm. The most important quality of DTC is its capacity to transform complex decision-making problems into simple processes, resulting in a solution that is easier to understand and analyse (Priyanka and Kumar 2020). The model is built in the form of a tree structure. It breaks down the dataset into smaller and smaller subsets by splitting while the corresponding tree is incrementally developed. DTC is made up of building blocks known as nodes.

There are three main nodes: internal nodes, leaf nodes, and root nodes. When a split yields no information gain, the tree is said to be in an unlearning state, and the node is classified as a leaf. However, the existence of a node, on the other hand, is determined by the state of its predecessor. It is appropriate for the targeted topic because it is a classification problem with the title "Survival prediction of heart failure patients".

3.8.1.1 Ensemble Learning

Ensemble learning employs multiple machine learning algorithms (series of base classifiers) to generate a more robust prediction than a single technique can acquire, and the result is assigned to the appropriate class using a majority voting mechanism (Dogan and Birant 2019). It also examines the individual performances of each classifier in the ensemble and modifies their contributions to class prediction. This technique provides a synergistic effect among individual models and reduces their tendency to memorize noise (overfitting). For this approach all models are trained on the same dataset and prediction is done separately. The final prediction is based on a voting classifier that acts as the meta-model. The models are proficient in many ways, therefore for the sake of this research, a voting classifier is taken into consideration, and the best findings are chosen.

3.8.2 Model Fitting

Each model was fitted to the dataset for training. Model fitting is the use of mathematical and statistical approaches to a set of simulation input/output data to (i) estimate the model's parameter values and (ii) evaluate the estimated parameter values regarding the data set using quantitative criteria (Kleijnen and Sargent 2000). It means performing numerical optimisation i.e., looking for minimum cost function in parameter space. Prior to fitting, it is crucial to specify/instantiate the type or form of the model. The requirement for fitting is performing optimization, availability of problem data, model parameters and hyperparameters.

The dataset was split into train and test split using the quantitative approach known as the holdout method. It involves splitting the dataset into a ratio of 70/30, where 70% is the train set and 30% serves as test data (to evaluate the model performance). For classification tasks, this ratio is used in several publications and helps prevent overfitting (Yousaf *et al*. 2020).

3.8.3 Model Tuning

Many contemporary regression and classification models can simulate complex relationships and are quite adaptive. A collection of configuration variables called hyper-parameters, which can assist each model in identifying predicted patterns in the data, are mostly used to govern each model's capacity for adaptation (Kuhn and Johnson 2013). The setting of these parameters is known as model fitting or hyper-parameter optimization. However, this tuning setting assists the model to achieve low predictive errors such as overfitting and underfitting when exposed to unseen data. Overfit models achieve excellent prediction on train data and low prediction on unseen data. While underfit models achieve low prediction on both trained data and unseen data. Therefore, it is important to tune the model to identify the best hyperparameters before fitting to achieve a good predictive performance. Hence, without using a methodical technique to tune the model, it would be difficult to identify the overfitting/underfitting error until the model is subjected to a new sample of data. In this study, the "**grid search**" method of hyperparameter optimization was taken into consideration. This search considers several combinations of hyperparameters, computes the performance metric using cross-validation and selects the one with the lowest error rating.

3.8.4 Model Evaluation

The main concern in classification issues is good classification accuracy. A performance evaluation was conducted to determine whether each model had truly learned and was effective at classifying the problem data sets. The evaluation involves how accurately the model classified the test (unseen) data

using an accuracy score. This was carried out after fitting the model to the trained data. The prediction accuracy was obtained using a **confusion matrix** which contains information about actual and predicted classifications (Kohavi and Provost 1998). Estimates of classification model possibilities are generated using the confusion matrix expressions such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The total of true positive (TP) and true negative events are those that were correctly classified (TN). The total of false positives (FPs) and false negatives constitute events that were incorrectly categorised (FN) (Vujović 2021).

Accuracy, precision, recall, and F1 score, are some of the most commonly used binary classification measures (Sokolova and Lapalme 2009). The success metric to evaluate the classification model are described in the table below (Table 2).

Table 4: Success Evaluation Metrics

Metrics	Values
Accuracy	>80%
Precision	>80%
Recall	>80%
F1 score	>80%

Accuracy is the overall percentage of how often the model predictions are correct.ie fraction of all instances predicted correctly (Lipton, Elkan and Narayanaswamy 2014).

$$\text{Accuracy} = TP+TN\ /(TP+TN+FP\ +FN)$$

Precision is the percentage of relevant cases among the predicted instances by the model i.e., how often the model prediction is correct.

$$\text{Precision} = TP/(TP+FP)$$

Recall is the ability of the model to correctly predict the relevant instances.the ie proportion of relevant cases predicted.

Recall = TP/(TP+FN)

F1 score is the simple harmonic mean of the recall and precision values of models. It is the measure of incorrectly classified occurrences (Tatbul *et al*. 2018).

F1 score = 2x ((Precision x Recall) /(Precision + Recall))

Both accuracy and f1 score are significant metrics for assessing classifier performance. The F1 score, on the other hand, is extensively used to assess the effectiveness of a binary classifier when one class is rare (imbalance data), which is common in real-world data (Lipton, Elkan and Narayanaswamy 2014).

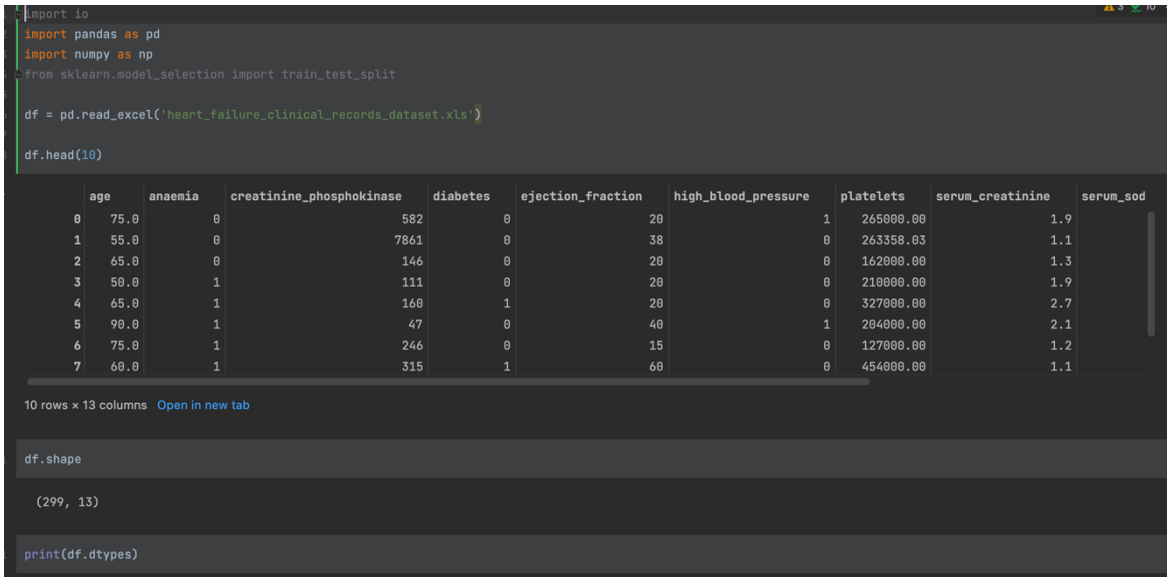
3.9 Software Design

The software requirement lifecycle was utilised to aid in the creation of the end system, which is the GUI (Graphics user interface). The Software Development Life Cycle (SDLC) is a technique for developing or maintaining software systems that are either descriptive or prescriptive of how software is or should be built (Ragunath *et al*. 2010). It usually encompasses several stages, ranging from preliminary development analysis to post-development software testing and evaluation. Various developers employ different SDLC approaches. The traditional development method of the SDLC process was applied in the study which encompasses four stages (Leau *et al*. 2012). Stage one started with the creation of the requirement and overall project plan. Once the criteria are established, stage two which is the design and architectural planning process begin. This stage involves creating a robust workable diagram to gain a more focused system understanding before implementation, choosing the programming language and platform. During stage three, which is implementation and coding, the user interface was created, and all feature designs were created. The system was implemented using python and streamlit platform. The last stage is testing, which is the assessment of the final design before its release to users. validating that all features and functions are functioning as intended.

3.10 Research Implementation

Implementation is the action that must originate after any conceptualization that leads to actualization. therefore, the step-by-step execution of the research plan, methods, and ideas can be seen in this section which illustrates screenshots of model/ application(code) as well as descriptions.

Statistical check to evaluate normality involves checking for skewness values, using of a histogram for visual evaluation and checking for kurtotic values of variables.



```
import io
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

df = pd.read_excel('heart_failure_clinical_records_dataset.xls')

df.head(10)
```

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sod
0	75.0	0	582	0	20	1	265000.00	1.9	
1	55.0	0	7861	0	38	0	263358.03	1.1	
2	65.0	0	146	0	20	0	162000.00	1.3	
3	50.0	1	111	0	20	0	210000.00	1.9	
4	65.0	1	160	1	20	0	327000.00	2.7	
5	90.0	1	47	0	40	1	204000.00	2.1	
6	75.0	1	246	0	15	0	127000.00	1.2	
7	60.0	1	315	1	60	0	454000.00	1.1	

10 rows × 13 columns Open in new tab

```
df.shape
```

```
(299, 13)
```

```
print(df.dtypes)
```

Figure 2: Importing libraries and showing dimensions on a screenshot

To make sure the programme functions, the necessary libraries were first imported into the integrated development environment (PyCharm). Some of the imported libraries are shown in Figure 2. The pandas' package was used to load the excel file into a dataframe and perform a qualitative analysis on it. The first ten rows were displayed using the pandas head method. The data's dimensions are 299 rows and 13 columns.

```
df.isna()
nun = df[df.isna().any(axis=1)]
print('\nTotal number of empty rows \n:', len(nun))
```

```
Total number of empty rows
: 0
```

```
#calculating missing values in the dataset
df.isnull().sum()
```

	data
age	0
anaemia	0
creatinine_phosphokinase	0
diabetes	0
ejection_fraction	0
high_blood_pressure	0
platelets	0
serum_creatinine	0
serum_sodium	0

Length: 13, dtype: int64 Open in new tab

Figure 3: Screenshot demonstrating no missing values

When missing values were evaluated using the pandas isna() method, it was found that there were none, so there is no need for imputation (replacing missing values). Additionally, no encoding is required because the dataset's attributes are all numerical (figure 3).

3.10.1 Measures of central tendencies on all variables

```
df.describe()
```

	age	anaemia	cr_ph	diabetes	ej_fr	hbp	platelets	ser_cr	ser_na	sex	smoking	time
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000	299.00000	299.0
mean	60.833893	0.431438	581.839465	0.418060	38.083612	0.351171	263358.029264	1.39388	136.625418	0.648829	0.32107	130.2
std	11.894809	0.496107	970.287881	0.494067	11.834841	0.478136	97804.236869	1.03451	4.412477	0.478136	0.46767	77.6
min	40.000000	0.000000	23.000000	0.000000	14.000000	0.000000	25100.000000	0.50000	113.000000	0.000000	0.00000	4.0
25%	51.000000	0.000000	116.500000	0.000000	30.000000	0.000000	212500.000000	0.90000	134.000000	0.000000	0.00000	73.0
50%	60.000000	0.000000	250.000000	0.000000	38.000000	0.000000	262000.000000	1.10000	137.000000	1.000000	0.00000	115.0
75%	70.000000	1.000000	582.000000	1.000000	45.000000	1.000000	303500.000000	1.40000	140.000000	1.000000	1.00000	203.0
max	95.000000	1.000000	7861.000000	1.000000	80.000000	1.000000	850000.000000	9.40000	148.000000	1.000000	1.00000	285.0

Figure 4: A screenshot of the central tendencies of features.

For easier comprehension, the dataset's basic statistics, including count, minimum, maximum, mean, and standard deviation for each column were

obtained using the pandas describe() method (figure 4). There are 299 fields in total, so the count of 299 confirms there are no missing values.

3.10.2 Measures of the variability of continuous variables

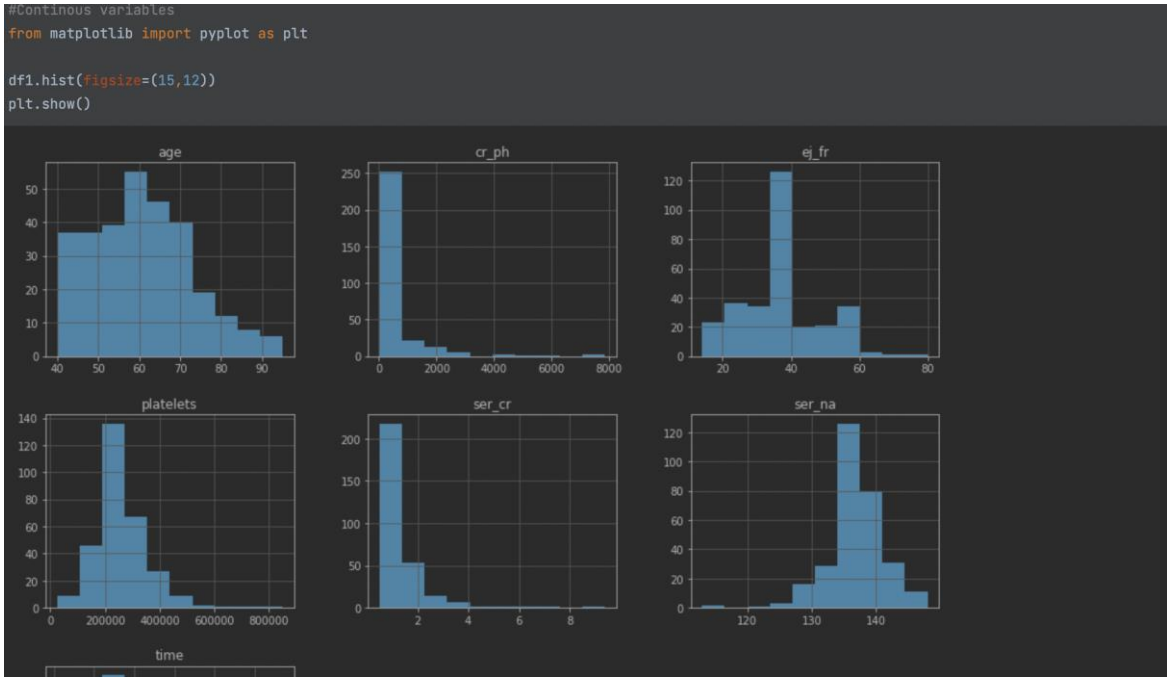


Figure 5: Histogram displaying the skewness and distribution of continuous variables

Figure 5 above depicts the outcome of the normality check on all continuous variables using histogram, which reveals that certain variables are positively skewed while others are negatively skewed, indicating abnormal distribution. Figure 6 shows the skewed values of all continous variables.

```
df1.skew()

age           0.423062
cr_ph         4.463110
ej_fr         0.555383
platelets     1.462321
ser_cr        4.455996
ser_na       -1.048136
time          0.127803
dtype: float64
```

Figure 6: Screenshot of skewed values of continuous variables

```
df1.kurt()

age           -0.184871
cr_ph         25.149046
ej_fr          0.041409
platelets      6.209255
ser_cr        25.828239
ser_na         4.119712
time          -1.212048
dtype: float64
```

Figure 7: Screenshot showing the kurtotic values of continuous variables

Figure 7 demonstrates that while certain variables are leptokurtic with high kurtotic values, others are platykurtic (low centre area, lighter and shorter tails) which indicates the presence of outliers. Variables with high kurtotic values have more outliers, whereas variables with low kurtosis have fewer outliers (Qiu, Murphy and Suter 2020). kurtotic values of variables were determined to assess the degree of flatness and peakedness of a distribution. The likelihood that the values are regularly distributed increases with the distance from zero.

```
from plotly.subplots import make_subplots
import plotly.graph_objects as go

fig = make_subplots(rows=4, cols=2,
                    # Set the subplot titles
                    subplot_titles= ['age','cr_ph','ej_fr', 'platelets','ser_cr','ser_na','time'])
#create boxplot visualization of numeric columns
fig.add_trace(go.Box(x=df.age, name='', showlegend=False), row=1, col=1)
fig.add_trace(go.Box(x=df.cr_ph, name='', showlegend=False), row=1, col=2)
fig.add_trace(go.Box(x=df.ej_fr, name='', showlegend=False), row=2, col=1)
fig.add_trace(go.Box(x=df.platelets, name='', showlegend=False), row=2, col=2)
fig.add_trace(go.Box(x=df.ser_cr, name='', showlegend=False), row=3, col=1)
fig.add_trace(go.Box(x=df.ser_na, name='', showlegend=False), row=3, col=2)
fig.add_trace(go.Box(x=df.time, name='', showlegend=False), row=4, col=1)

fig.update_layout(height=800, width=1000)
fig.show()
```

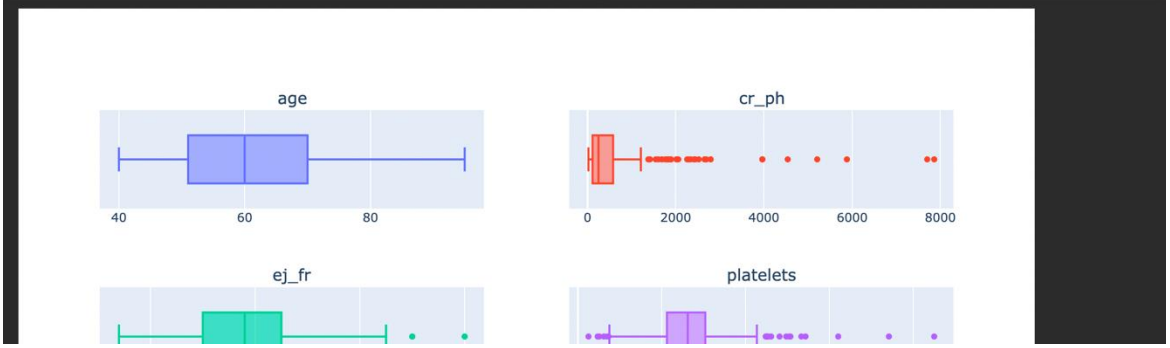


Figure 8: Screenshot displaying box plots of continuous variables with codes

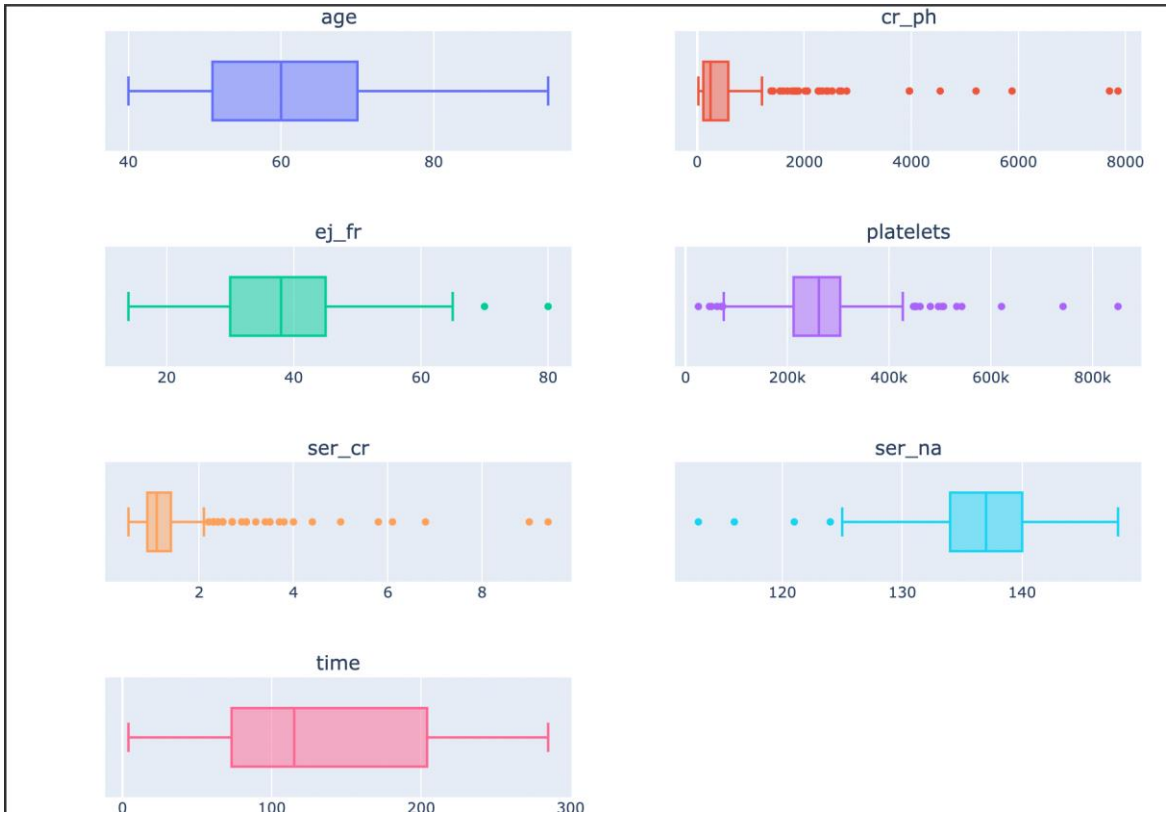


Figure 9: A screenshot showing box plots for continuous variables

Furthermore, Plotly library was used to obtained boxplots on the continuous variables which revealed that they all have outliers except age and time (figure 6). There are more outliers in the serum creatinine (ser_cr) ,and platelets variables than others (Figure 8 and 9). Outliers can diminish normality, weaken the validity of statistical tests, and raise error variance (Osborne and Overbay 2004).Therefore there is need to remove the outliers to avoid errors. Further exploration was done by computing bivariate and multivariate analysis.

```
import seaborn as sns
plt.figure(figsize=(20,19))
features = ['age','cr_ph','ej_fr','platelets','ser_cr','ser_na', 'time']
for n, data in enumerate(features):
    ax = plt.subplot(3, 3, n + 1)
    sns.kdeplot(data = df1, x = data, shade=True)
    plt.grid('on')

plt.subplots_adjust(wspace=0.35, bottom=0.2, hspace=0.4)
plt.suptitle("Kernel DEnsity Plots")
plt.show()
```

Kernel DEnsity Plots

Figure 10: kernel density plots of continuous variables on a screenshot

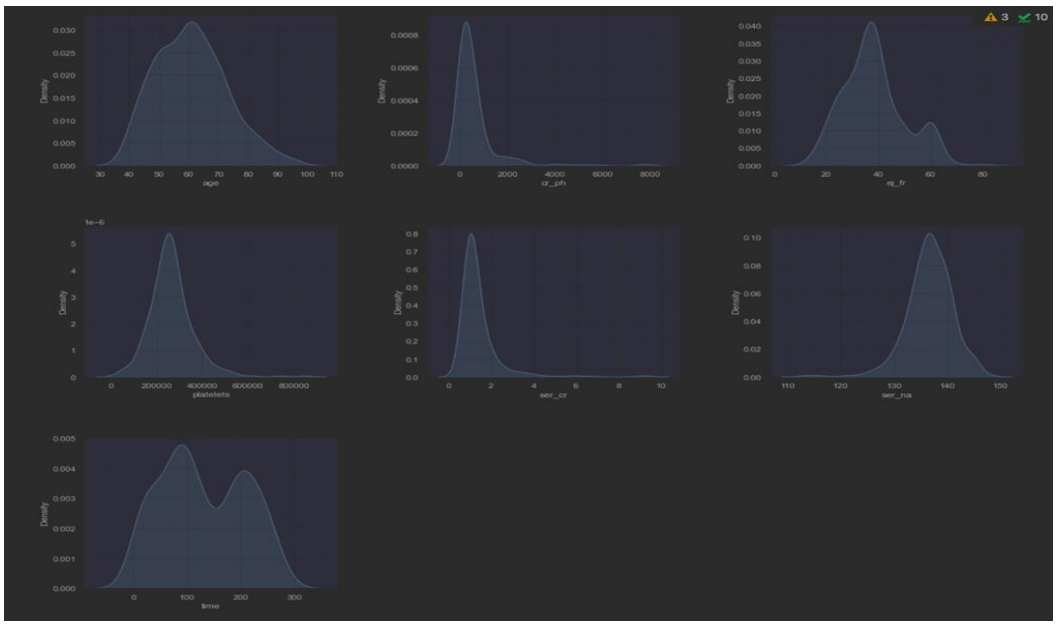


Figure 11: Screenshot -2 displaying kernel density plots of continuous variables

The density plots above (Figure 10 and 11) confirm the presence of outliers and the level of skewness in the variables. It also shows that some variables, such as ejection fraction (ej fr) and time, are bimodal (have two peaks), whereas others only have one. As a result, the dataset was normalised using the **L2 method** of normalisation, because they do not have a Gaussian distribution (Appendice). L2 method of normalisation leverages more features and the distance between points remains the same (Vafaei, Ribeiro and Camarinha-Matos 2018).

The heatmap below (figure 12) shows the relationship or strength between the independent variables and the dependent variable 'death events' (status). On the heatmap above, values closer to -1 and +1 indicate significant/strong correlations, whereas values closer to zero indicate weak correlations (Evans and Basu 2013).

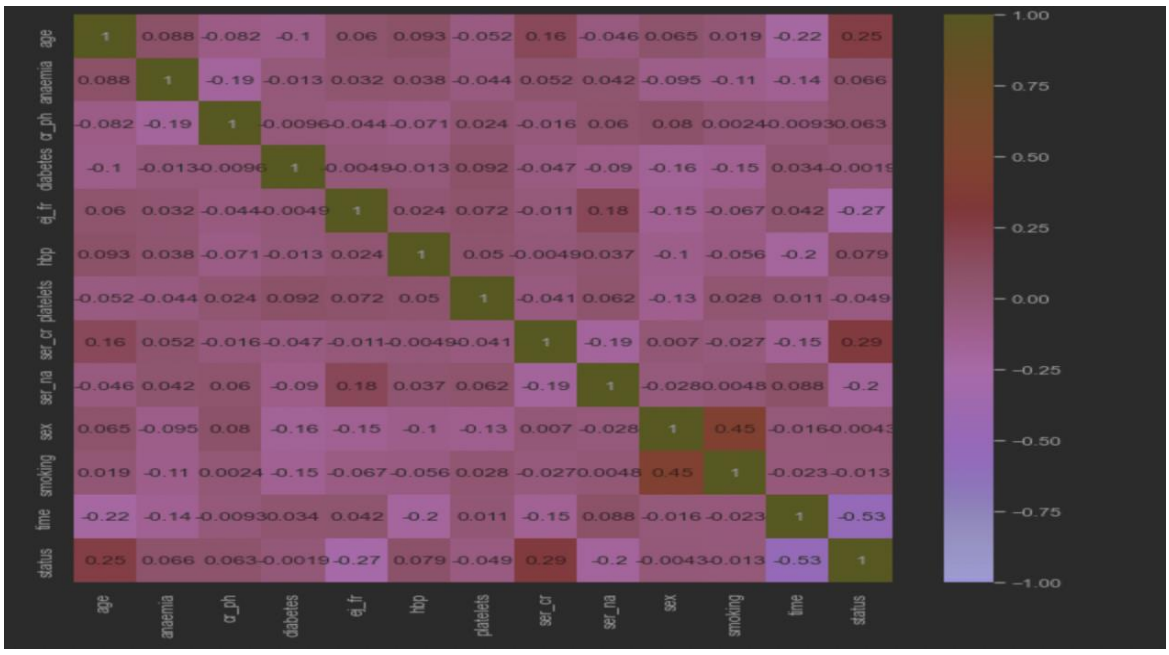


Figure 12:Correlation analysis result with seaborn and matplotlib.

```
corr[abs(corr['status']) > 0.2] ['status']
```

```
age       0.253729
ej_fr    -0.268603
ser_cr    0.294278
time     -0.526964
status    1.000000
Name: status, dtype: float64
```

Figure 13:Screenshot displaying Variables with a correlation above 0.2

The figure above (figure 13) shows the variables with a correlation above 0.2. It can be inferred that 'time' has the strongest correlation to the dependent variable 'death events' (status) when compared to the other variables (age, ej_fr, and ser_cr), whereas age has the lowest correlation value. Age, ej_fr, ser_cr, and time can be placed in the variables' strength order, from weakest to strongest.

```
from numpy import set_printoptions
from sklearn.preprocessing import Normalizer
# #Splitting into dependent and independent data
X_data = df.drop(['status'], axis=1) #independent variables
Y_data = df.pop('status')

#normalised the independent data
#l2 leverages more features and distances between points remain the same
df_normaliser=Normalizer(norm='l2').fit(X_data)
my_normalized_data =df_normaliser.transform(X_data)
set_printoptions(precision=2)
print('\n My Normalised data:\n', my_normalized_data[200:204])
#dataset are now within same ranges
```

```
My Normalised data:
[[8.63e-04 1.37e-05 2.42e-02 0.00e+00 6.16e-04 0.00e+00 1.00e+00 9.59e-06
  1.88e-03 1.37e-05 0.00e+00 2.55e-03]
 [1.19e-04 0.00e+00 8.17e-04 2.65e-06 1.59e-04 2.65e-06 1.00e+00 2.65e-06
  3.61e-04 2.65e-06 0.00e+00 4.93e-04]
 [3.18e-04 0.00e+00 4.41e-04 0.00e+00 2.73e-04 4.55e-06 1.00e+00 4.09e-06
  6.27e-04 4.55e-06 0.00e+00 8.45e-04]
 [2.83e-04 0.00e+00 2.78e-04 0.00e+00 1.18e-04 4.72e-06 1.00e+00 1.65e-05
  6.42e-04 4.72e-06 4.72e-06 8.82e-04]]
```

Figure 14:Screenshot showing splitting and normalization of dataset

Figure 14 shows splitting of dataset into dependent and independent variables as well normalisation of data. The dataset was normalised using the **L2 method** of normalisation, because they do not have a Gaussian distribution . L2 method of

normalisation leverages more features and the distance between points remains the same (Vafaei, Ribeiro and Camarinha-Matos 2018)

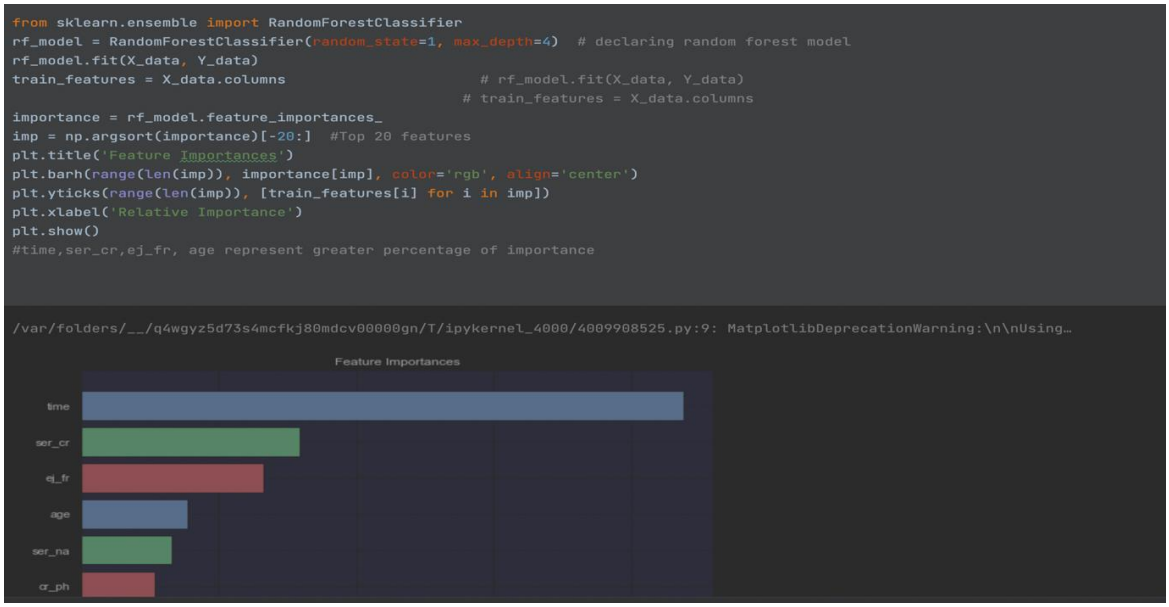


Figure 15:Screenshot displaying feature importance

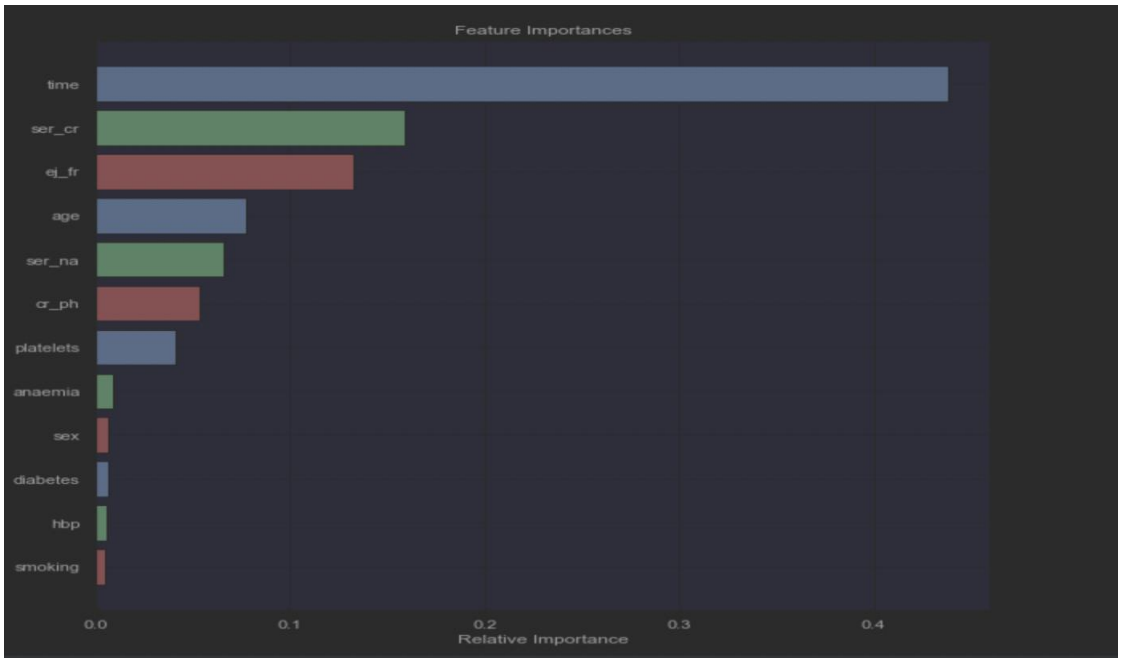


Figure 16: Screenshot 2 demonstrates the significance of a feature.

Feature Analysis was performed with RandomForestClassifier in python to identify the important variables in the dataset. About 45% of the entire dataset was found to be contributed by the time. Time was found to be the most crucial variable,

followed by serum creatinine, ejection fraction and age(Figure 15 and 16). And while some features were dropped due to low importance, these five were chosen to reduce dimensionality.

```
from numpy import set_printoptions
from sklearn.preprocessing import Normalizer
# #Splitting into dependent and independent data
X_data = df[['age', 'ej_fr', 'ser_cr', 'ser_na','time']]
# X_data = df.drop(['status','anaemia','cr_ph','diabetes','hbp','platelets','sex','smoking'
Y_data = df.pop('status')

#normalised the independent data
#l2 leverages more features and distances between points remain the same
df_normaliser=Normalizer(norm='l2').fit(X_data)
my_normalized_data =df_normaliser.transform(X_data)
set_printoptions(precision=2)
print('\n My Normalised data:\n', my_normalized_data[200:204])
#dataset are now within same ranges
```

```
My Normalised data:
[[0.26 0.18 0.   0.56 0.76]
[0.19 0.25 0.   0.56 0.77]
[0.28 0.24 0.   0.55 0.75]
[0.25 0.1  0.01 0.57 0.78]]
```

Figure 17: Screenshot of topmost feature splitting and normalisation

Figure 17 shows splitting the dataset into dependent and independent variables (with the topmost important variables selected) as well as normalisation.

```
# Spliiting dataset into train and test
from sklearn.model_selection import train_test_split

# Split dataset into training set and test set
X_train, X_test, Y_train, Y_test = train_test_split(X_data,
                                                    Y_data,
                                                    test_size=0.3,  # 70% training and 30% test
                                                    random_state=1)
#Checking the  number of test and train data
print('\n The total of training dataset:', X_train.shape)
print('\n The total of test dataset:', X_test.shape)
print(Y_test.shape)
```

```
The total of training dataset: (209, 5)

The total of test dataset: (90, 5)
(90,)
```

Figure 18: Screenshot of dataset splitting into train and test sets

Figure 18 shows the splitting of the dataset into train and test set (hold out method) using the train_test_split function from scikit-learn package. It also shows the dimensions of each set.

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
#importing the models
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import VotingClassifier
from sklearn.svm import SVC

svm =SVC(kernel='poly', max_iter=-1, degree=3, probability=True)
dt = DecisionTreeClassifier(criterion='entropy', max_depth=3, min_samples_leaf=0.05, min_samples_split=2,random_state=1)
kn =KNeighborsClassifier(n_neighbors=8,leaf_size=30,metric='minkowski', p=2)
lr =LogisticRegression(intercept_scaling= '1',max_iter=1000,multi_class= 'auto', penalty='l2',random_state=1, solver='newton-cg',tol=0.0001, verbose= 0,
 warm_start=False)

class_list = [('DecisionTreeClassifier:',dt),('Supportvector:',svm),('LogisticRegression:',lr),('KNeighborsClassifier:',kn)]

#iteration
for clsf_name,clsf in class_list:
    clsf.fit(X_train,Y_train)
    Y_pred = clsf.predict(X_test)
    print('\n{:s} :{:.3f}'.format(clsf_name, accuracy_score(Y_test, Y_pred)))

votingc = VotingClassifier(estimators=class_list, voting='soft')
votingc.fit(X_train, Y_train)

#predict test label
Y_pred_vc = clsf.predict(X_test)
```

Figure 19: Screenshot of base model importation and dataset fitting

Figure 19 shows importation of algorithms , tuning and fitting of base models to train set as well as testing the models with the test set of the dataset (30% of the data).

```
#predict test label
Y_pred_vc = clsf.predict(X_test)


 DecisionTreeClassifier: :0.844

 Supportvector: :0.856

 LogisticRegression: :0.833

 KNeighborsClassifier: :0.878


print('\n voting classifier {:.3f}'. format(accuracy_score(Y_test, Y_pred_vc)))


  voting classifier 0.878
```

Figure 20: Screenshot of the accuracy of base models

The accuracy of each base models was checked with metrics function of scikit learn package (Figure 20). It also displays the accuracy of the voting classifier.

```
#confusion matrix
matrix_info = confusion_matrix(Y_test, Y_pred_vc)
print('\n Confusion matrix on heart failure data\n', matrix_info, '\n')

class_report = classification_report(Y_test, Y_pred_vc)
print('Classification report:\n ', class_report)
```

```
  [[63  1]
  [10 16]]

 Classification report:
               precision    recall  f1-score   support

           0       0.86      0.98      0.92        64
           1       0.94      0.62      0.74        26

    accuracy                           0.88        90
   macro avg       0.90      0.80      0.83        90
weighted avg       0.89      0.88      0.87        90
```

Figure 21: Screenshot of the voting classifier's confusion matrix

The classification report of the voting classifier was viewed using confusion matrix (figure 21).

```
#MAKING MODEL PERSISTENT FOR USE

#VIEWING THE TEST DATASET
print(X_test.tail(5).join(Y_test.tail(5)))
```

```
      age  ej_fr  ser_cr  ser_na  time  status
122  60.0     38    0.75     140    95       0
246  55.0     25    1.10     138   214       1
278  50.0     30    0.70     136   246       0
251  55.0     35    0.80     143   215       0
19   48.0     55    1.90     121    15       1
```

```
#SAVING MODEL
my_model = 'dissert_model.sav'
jb.dump(votingc, my_model)
```

```
['dissert_model.sav']
```

```
#TESTING THE MODEL EFFECT

load_my_model = jb.load(my_model)
results = load_my_model.score(X_test, Y_test)
print('\n This is the result of the persistent model\n', results)
```

```
 This is the result of the persistent model
 0.8666666666666667
```

Figure 22: Screenshot of how the model was saved

The trained model was saved using joblib package and the persistent model was evaluated, as can be seen in figure 22.

3.10.2 Application Code interpretation

```
import numpy as np
import streamlit as st
import joblib as jb
from PIL import Image

# Creating title
st.title("A Survival Prediction System")
st.text("This system uses five(5) inputs to predicts the survival of heart failure patients")
image = Image.open('new.jpg')
st.image(image, width=700)



# sex = st.selectbox('Sex', options=['Female', 'Male'])
age = st.number_input('Age (years)', min_value=0)  # min_value=40, max_value=95)
ej_fr = st.number_input('Ejection Fraction (%)', min_value=0)  # min_value=14, max_value=80)
ser_cr = st.number_input('Serum Creatinine (mg/dl)', min_value=0.5, max_value=9.4)
ser_na = st.number_input('Serum Sodium(mEq/L)', min_value=0)  # min_value=113, max_value=148)
time = st.number_input('Time (days)', min_value=0)

st.write('The user inputs are {}'.format([age, ej_fr, ser_cr, ser_na,time]))


def transform():
    age_n = float(age)
    ej_fr_n = float(ej_fr)
    ser_cr_n = float(ser_cr)
    ser_na_n = float(ser_na)
    patient = [age_n, ej_fr_n, ser_cr_n, ser_na_n, time]
    return patient
```

Figure 23: Screenshot displaying the first part of the application

Figures 23 and 24 show the implementation of the application. The first section displays all libraries used, input boxes as well as all the functions utilized. The application contains three(3) functions which are transform, prediction and status. The transform function takes no parameters, converts all values inputted by user into floats and returns the float values. Prediction function takes a parameter (float values) , calls the saved model, make prediction based on the parameter and returns the prediction .The status function takes to parameter, calls the other two functions, and displays prediction to user.

```
def prediction(x):
    loaded_model = jb.load('dissert_model.sav')
    patient_value = np.array(x).reshape(1, -1)
    predict = loaded_model.predict(patient_value)
    return predict


def status():
    retrieved = transform()
    predicted = prediction(retrieved)
    st.subheader("Prediction")
    if predicted == 0:
        st.write("We predict the patient to be: Alive")
    elif predicted == 1:
        st.write(" We predict the patient : Not alive")
    else:
        st.write("Please check your data input")


status()
```

Figure 24: Screenshot showing. the second part of the application

3.11 Summary

This chapter has provided a research methodology (Actions to be taken to aid knowledge discovery) as well as the researcher's methodological stance in the context of the current study. Additionally, it provided all the procedures and methods used during each stage, including data collection, pre-processing, exploratory data analysis, distribution, algorithm selection, model development, software development, and GUI design.

The overall methodology used in this study is a mixed methodological approach, which includes both quantitative and qualitative methods. Kurtotic and skewness were employed to quantify dataset variability while statistical analysis was utilised to evaluate measures of central tendency on the data collected. The dataset was explored using a univariate, bivariate and multivariate method of analysis. Suitable algorithms were selected after so much research on the problem case which is “classification”. The model was developed based on the best algorithm that met the required success metrics.

CHAPTER 4

4.0 Result Interpretation

4.1 Introduction

This section presents the findings of the quantitative data analysis carried out on discrete variables like sex, anaemia, smoking, diabetes, and HBP as well as continuous variables like age, time, platelets, ejection fraction, and serum creatinine. The relationship between the dependent variable (status) and the independent variables is also provided in detail. Additionally, this chapter offers thorough insights drawn from the analysis.

4.2 Binary/Demographic variable distribution

4.2.1 Gender distribution of participants

Figure 25 shows the count of male patients/participants as 194 while female is 105.

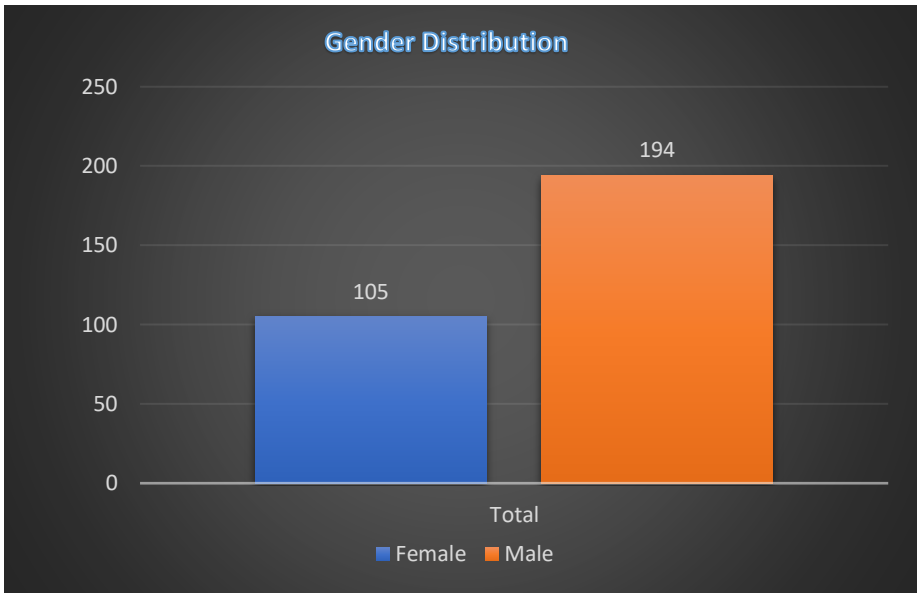


Figure 25: Gender Distribution of participants

4.2.2 Smoking /status distribution of participants

Figure 26 below shows the smoking status distribution of all the patients. Non-smokers make up most patients.

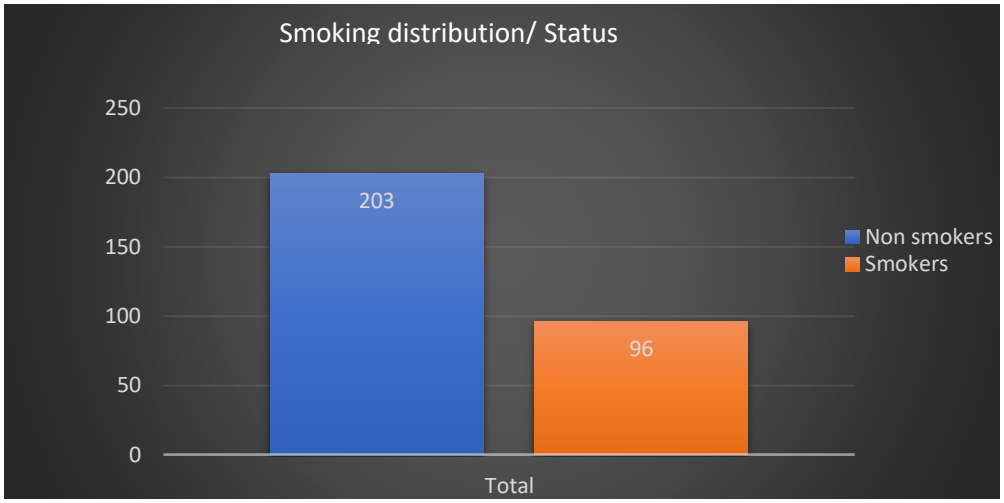


Figure 26: Smoking Status Distribution of Participants

4.2.3 Diabetes /status distribution of participants

The highest population of patients (174 participants) are not diabetics while diabetic patients are 125 by count.

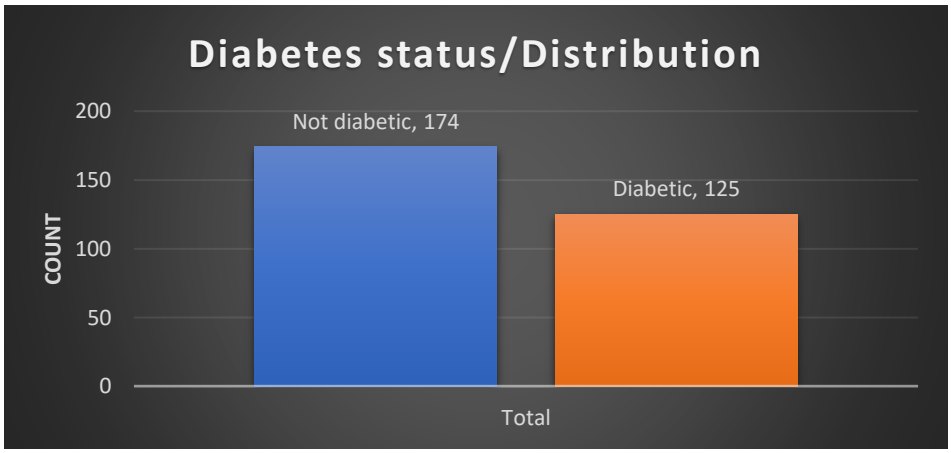


Figure 27: Diabetes Status Distribution

4.2.4 High blood pressure(HBP) status distribution of participants

The highest population of patients (194 participants) are non-hypertensives,

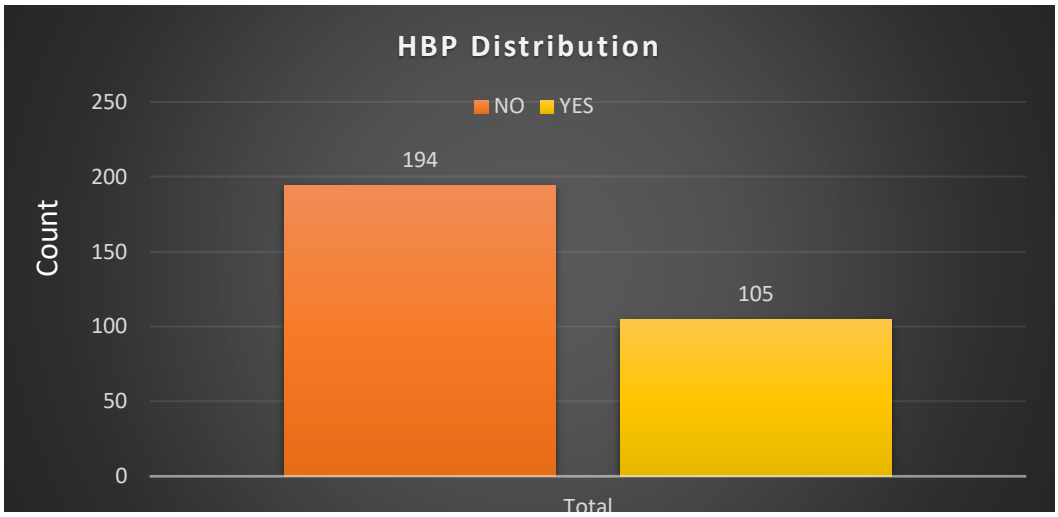


Figure 28: HBP Status Distribution

4.2.5 Anaemia status distribution of participants

The majority of patients (170 participants) are not anaemic, compared to 129 who are.

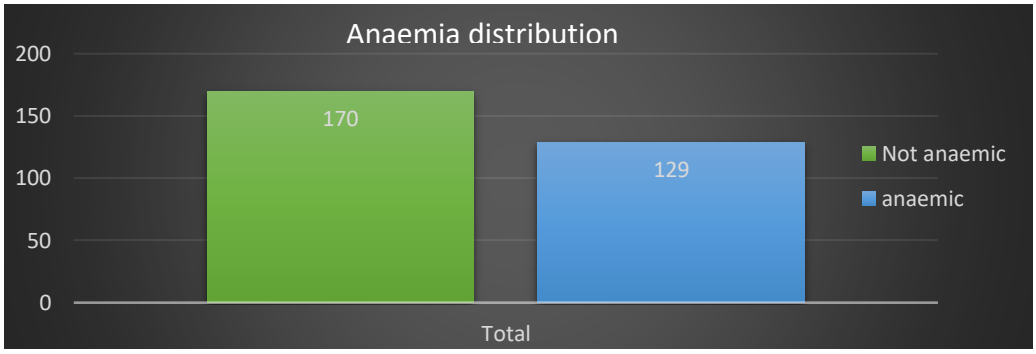


Figure 29: Anaemia Status Distribution of Participants.

4.2.6 Distribution of participants' death event (status)

In comparison to patients who passed away, a higher percentage of patients are still alive; this category represents 67.9% of the population following their follow-up time while 32.1% represented those who are deceased (figure 30).

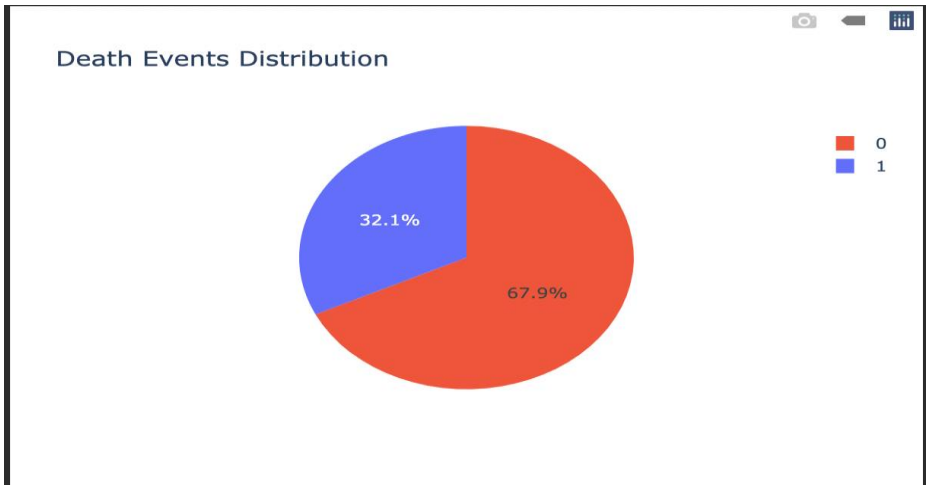


Figure 30: Death Events Distribution of Participants

4.3 Bivariate Analysis

According to the line chart below (figure 31), the average age of individuals who are still alive is 59, while the average age of those who have died is 65.

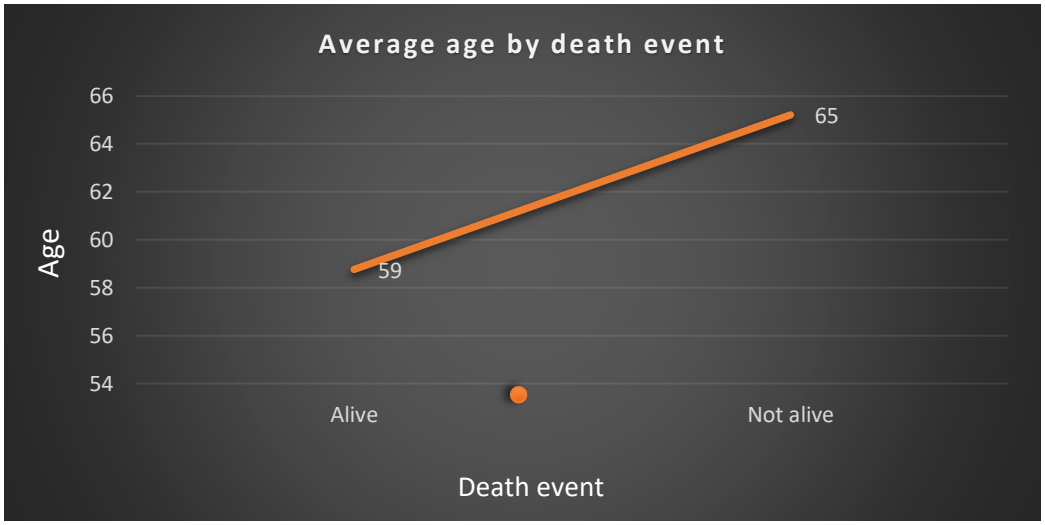


Figure 31: Average age by death events

The figure below (figure 32) shows the smoking status of patients with the dependent variable (death event). The bar chart shows that the number of non-smokers in category 1 (Alive) decreased from 137 to 66 in category 2 (Not alive).

Similarly, the number of smokers decreased exponentially from 66 (Alive) to 30 in category 2 (Not alive).

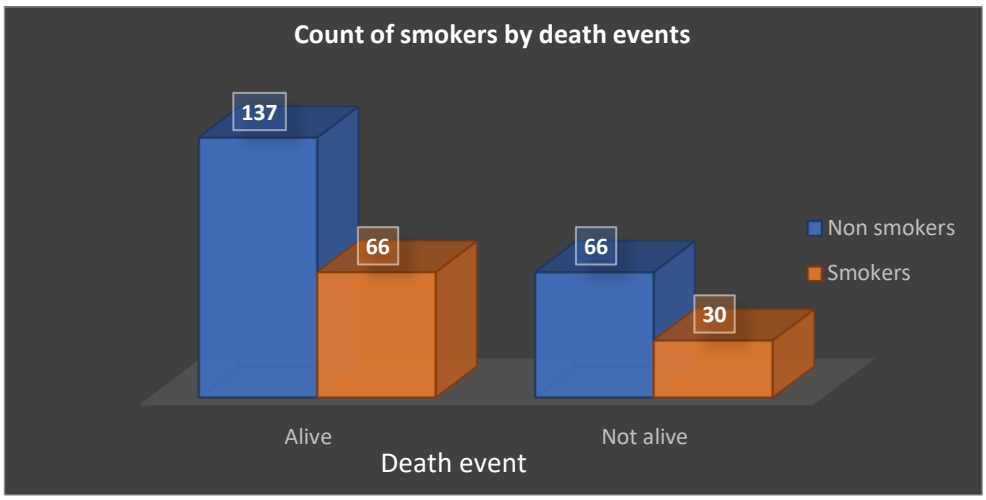


Figure 32: Smoking status by death events

The figure below (figure 33) shows the diabetic state of patients with the dependent variable (death event). The bar chart shows that the number of non-diabetics in category 1 (Alive) declined from 118 to 56 in category 2 (Not alive). The number of diabetics also dropped dramatically from 85 (Alive) to 40 in category 2 (Not alive).

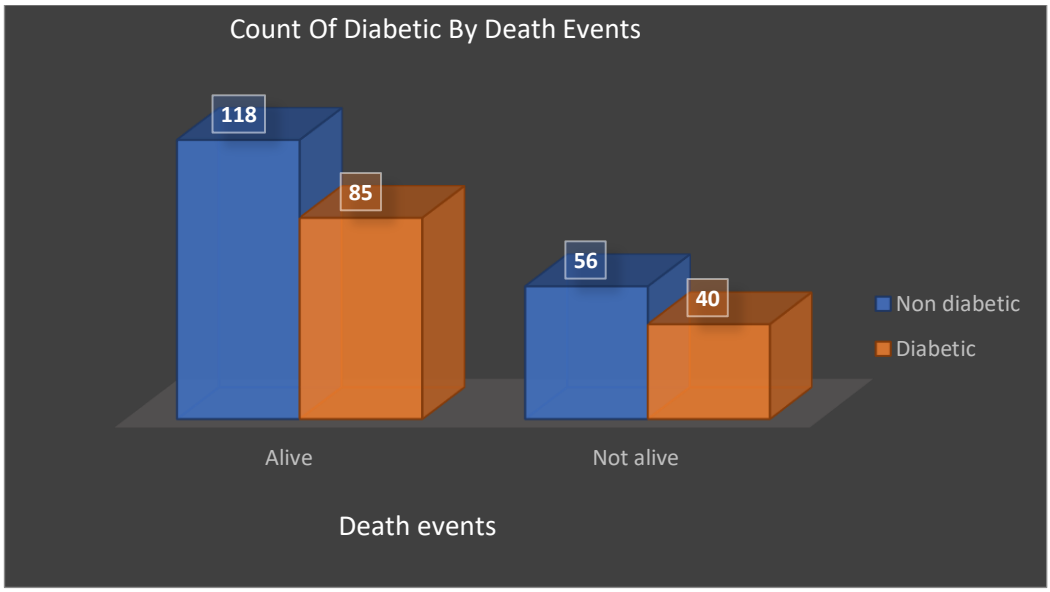


Figure 33: Diabetes status by death events

The figure below (Figure 34) depicts the hypertensive status of patients with the dependent variable (death event). According to the line chart, the number of hypertensives in category 1 (Alive) has decreased from 66 to 39 in category 2 (Not alive). Non-hypertensives decreased substantially from 137 (Alive) to 57 in category 2 (Not alive).

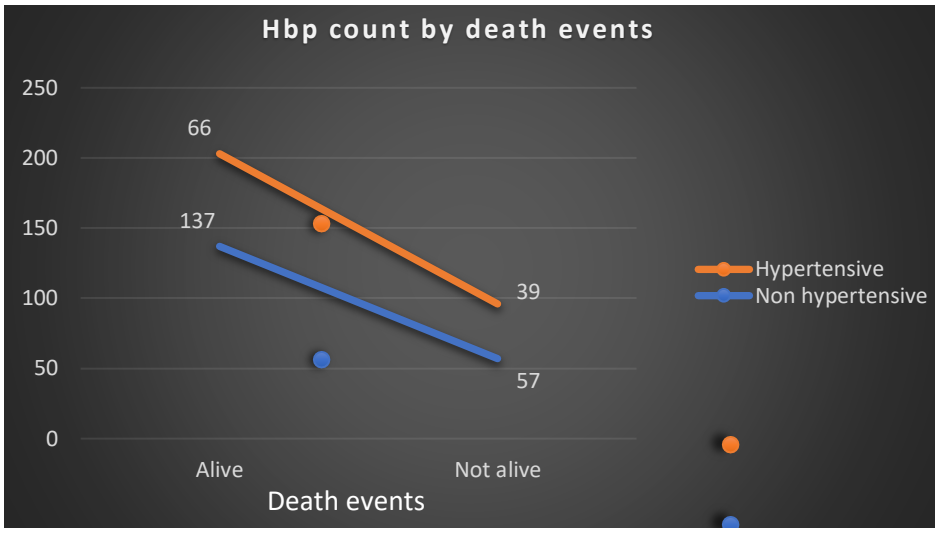


Figure 34: HBP count by death events

The figure below (figure 35) shows the patients' gender with the dependent variable (death event). The line chart shows that the number of females in category 1 (Alive) declined from 71 to 34 in category 2 (Not alive). The number of males also dropped dramatically from 132 (Alive) to 62 in category 2 (Not alive).

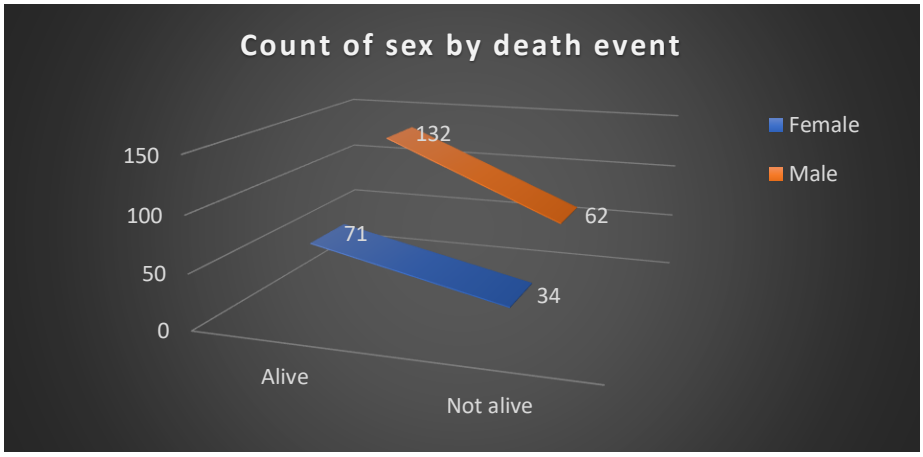


Figure 35: Gender by death_events

The figure below (figure 36) shows the patients' anaemic status with the dependent variable (death event). The line chart shows that the number of not anaemic individuals in category 1 (Alive) declined from 120 to 50 in category 2 (Not alive). The number of anaemic also dropped dramatically from 83 (Alive) to 46 in category 2 (Not alive).

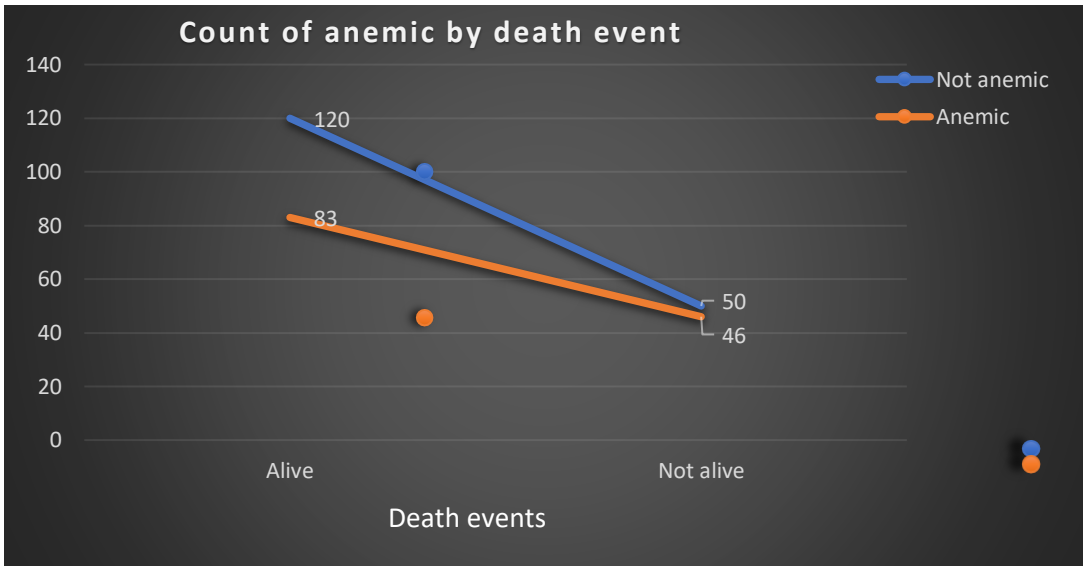


Figure 36: Counts of anaemia by death_events

The figure below (figure 37) shows the average platelet value of individuals with the dependent variable (death event). The line chart shows that the average

platelet value in category 1 (Alive) declined from 266657 to 256381 in category 2 (Not alive).

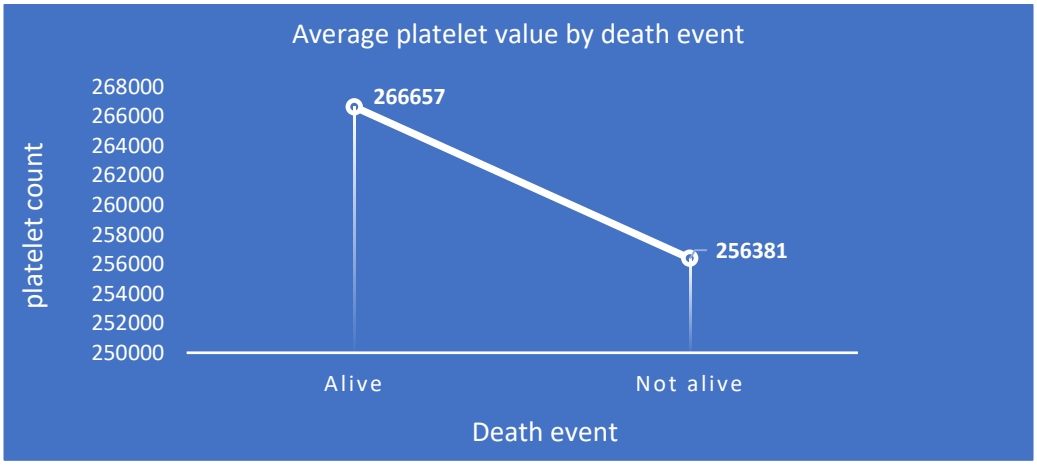


Figure 37: Average platelet value by death event

The figure below (figure 38) shows the maximum serum creatinine (ser_cr) value of individuals with the dependent variable (death event). The line chart shows that the maximum ser_cr value in category 1 (Alive) rose from 6 to 9 in category 2 (Not alive).

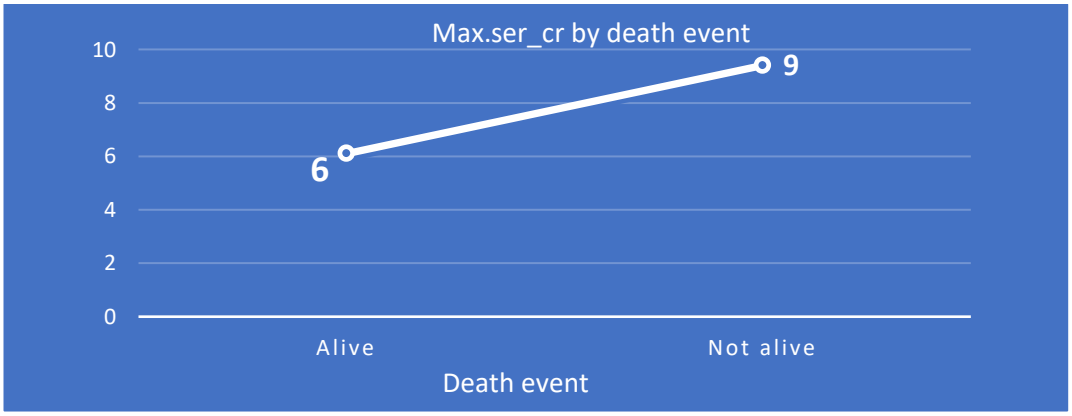


Figure 38: Maximum serum creatinine by death event

The figure below (figure 39) shows the average serum sodium (ser_na) value of individuals with the dependent variable (death event). The line chart shows that

the average ser_na value in category 1 (Alive) decreased from 137 to 135 in category 2 (Not alive).

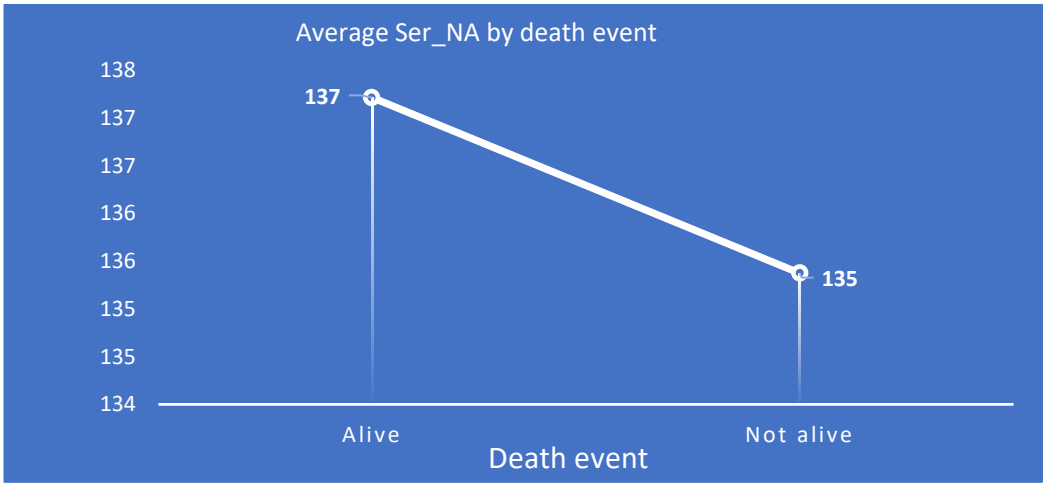


Figure 39: Average Serum sodium by death event

As seen in the line chart below (figure 40), the average creatinine phosphokinase(cr_ph) of participants alive is 540 while deceased is 670. This depicts that those who had high levels of cr_ph died more frequently than those who had low levels.

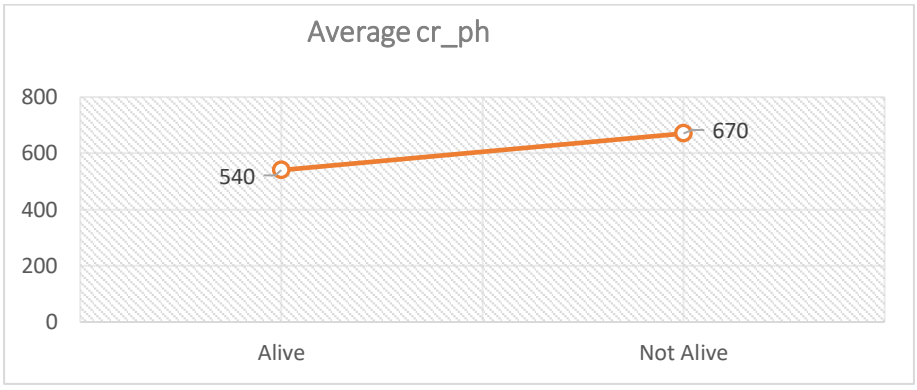


Figure 40: Average creatinine phosphokinase

The figure below (figure 41) shows the relationship between ejection fraction(ej_fr) variable and dependent variable(death events).The maximum ejection fraction (ej_fr) of individuals who are living is 80, whereas for those who are dead is 70, as seen in the line chart above.

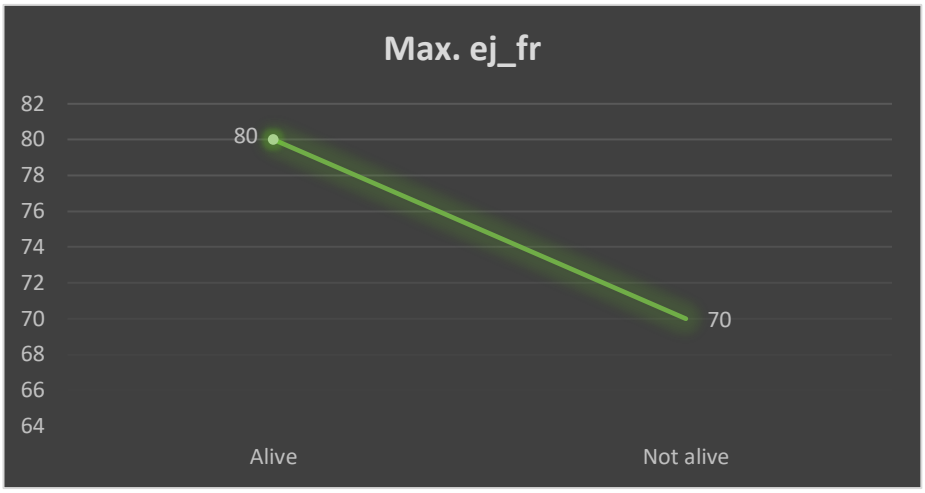


Figure 41: Screenshot showing Maximum ejection fraction by death event

As seen in the line chart below (figure 42), the average follow-up time of HF patients in the first category 'alive' is 158 days while the second category 'Not alive' is 71days..

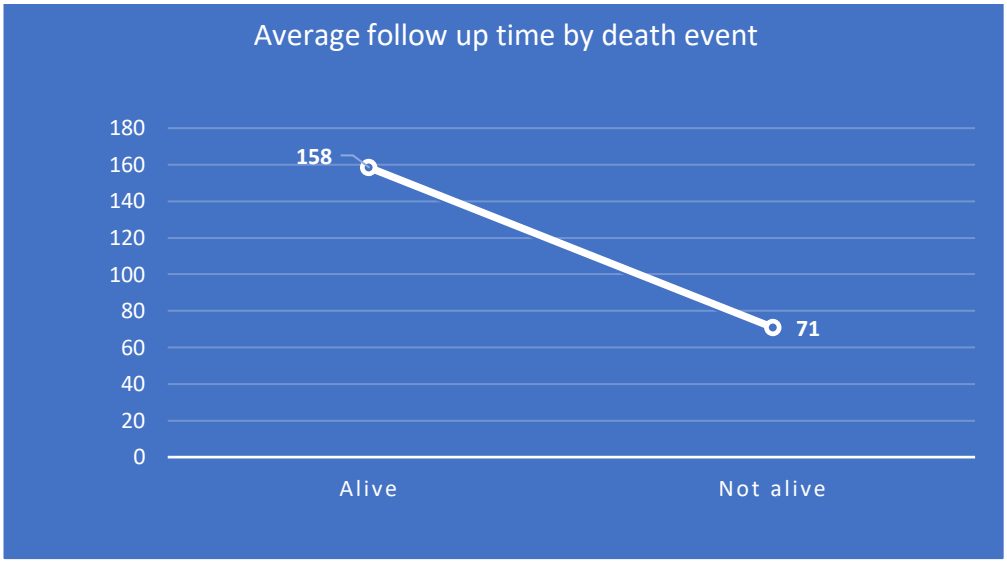


Figure 42: Average follow up time by death event

Table 5: Bivariate analysis results of variables with the dependent variable.

Independent variables	Death events	
	Alive	Not alive
Average Age	59	65
Smoking	137 Non smokers 66 Smokers	66 Non-smokers 30 Smokers
Diabetes	118 Not diabetic 85 Diabetic	56 Not diabetic 40 Diabetic
HBP	66 Hypertensives 137 Not hypertensives	39 Hypertensives 57 Not hypertensives
Sex	132 Males 71 Females	62 Males 34 Females
Anaemia	120 Not anaemic 83 Anaemic	50 Not anaemic 46 Anaemic
Average platelet	266657	256381
Max. ser_cr	6	9
Average ser_na	137	135
Average cr_ph	540	670
Maximum ej_fr	80	70
Average follow up time	158	71

Table 5 shows the overall results of the bivariate analysis performed on the dataset.

4.4 Multivariate Analysis

4.4.1 Correlation Analysis

Table 6: Features with correlation values above 0.2

	Features	Correlation values
1	Time	-0.53
2	Age	0.25
3	Ejection Fraction	-0.27
4	Serum creatinine	0.29
5	Serum sodium	0.20

The figure below (figure 27) shows the variables with a correlation above 0.2. It can be inferred that 'time' has the strongest correlation to the dependent variable 'death events' (status) when compared to the other variables.

4.4.2 Clustering

K-means clustering was performed with Tableau, which generated two clusters based on continuous variables such as creatinine phosphokinase, platelet, ejection fraction, serum creatinine, serum sodium, age, and time. Patients in cluster 1 patients can be seen as older, high values of serum phosphokinase, lower ejection fraction, higher serum serum creatinine and lesser follow up time compared with the second cluster.

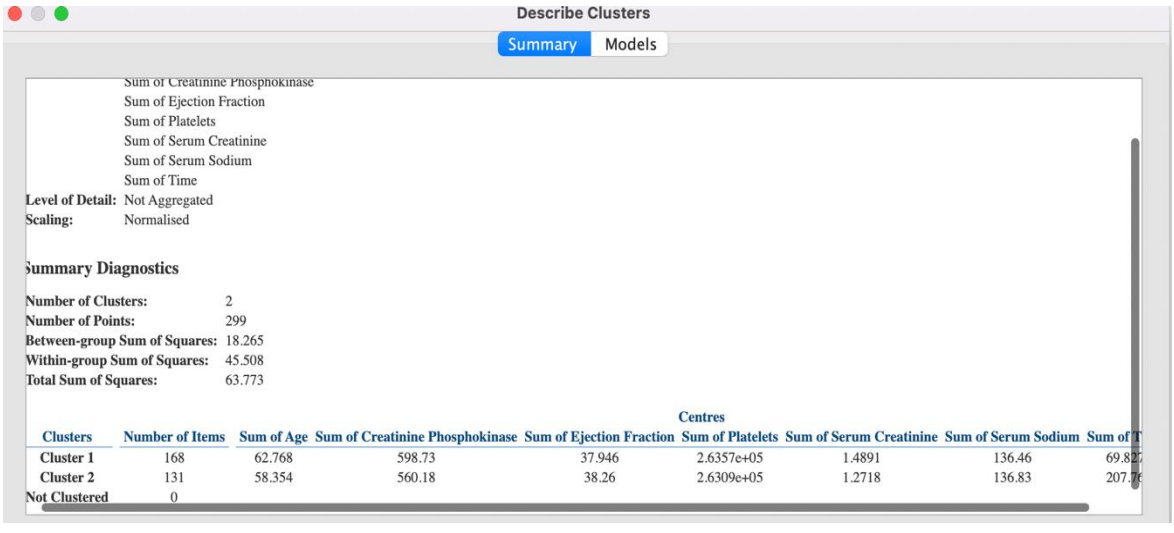
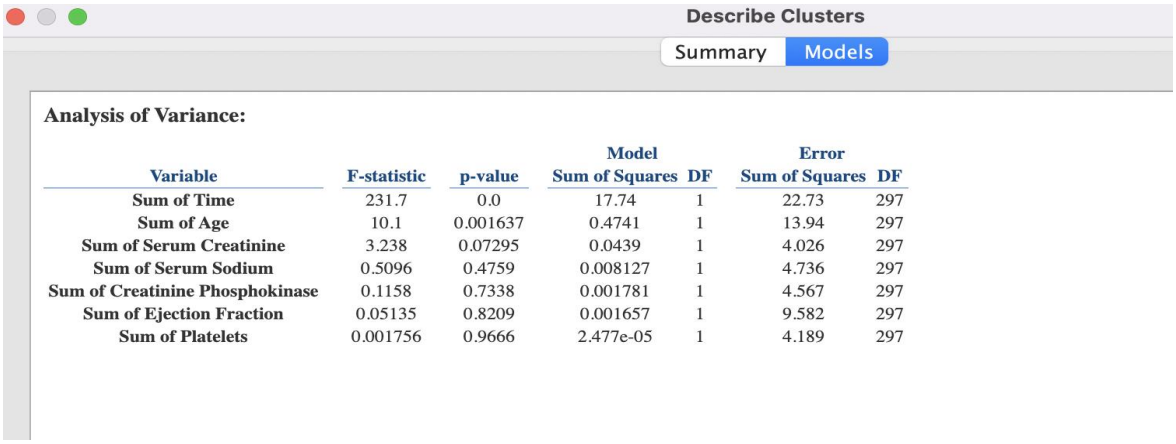


Figure 43: Brief description of Tableau clusters

The figure below (figure 29) shows the order of significance of variables to the clusters.



Variable	F-statistic	p-value	Model Sum of Squares	Model DF	Error Sum of Squares	Error DF
Sum of Time	231.7	0.0	17.74	1	22.73	297
Sum of Age	10.1	0.001637	0.4741	1	13.94	297
Sum of Serum Creatinine	3.238	0.07295	0.0439	1	4.026	297
Sum of Serum Sodium	0.5096	0.4759	0.008127	1	4.736	297
Sum of Creatinine Phosphokinase	0.1158	0.7338	0.001781	1	4.567	297
Sum of Ejection Fraction	0.05135	0.8209	0.001657	1	9.582	297
Sum of Platelets	0.001756	0.9666	2.477e-05	1	4.189	297

Figure 44: Analysis of variance of clusters on Tableau

4.5 Model Evaluation Result

Table 7: Algorithm Evaluation Result

Algorithms	Accuracy(%)	Precision (%)	Recall(%)	F1 Score(%)
DecisionTree Classifier	85	Alive =91 Not Alive=67	Alive =89 Not Alive =71%	Alive =90 Not Alive =69
SVC	86	Alive =86 Not Alive =84	Alive =95 Not Alive= 62	Alive =90 Not Alive=71
KNN	88	Alive =86 Not Alive =94	Alive =98 Not Alive =62	Alive =92 Not Alive =74
Logistic Regression	83	Alive =87 Not Alive =74	Alive =91 Not Alive =65	Alive =89 Not Alive =69
Voting Classifier	88	Alive =86 Not Alive =94	Alive =98 Not Alive =62	Alive =92 Not Alive =74

4.6 Software Output

The pre-trained voting classifier model was used to develop a simple Streamlit application that predicts the results based on user input. The screenshots of the application's user interface are shown in Figures 31 and 32. The (+ and -) buttons on the right-hand side of the interface allow the user to change numerical values, or they can manually enter the values.

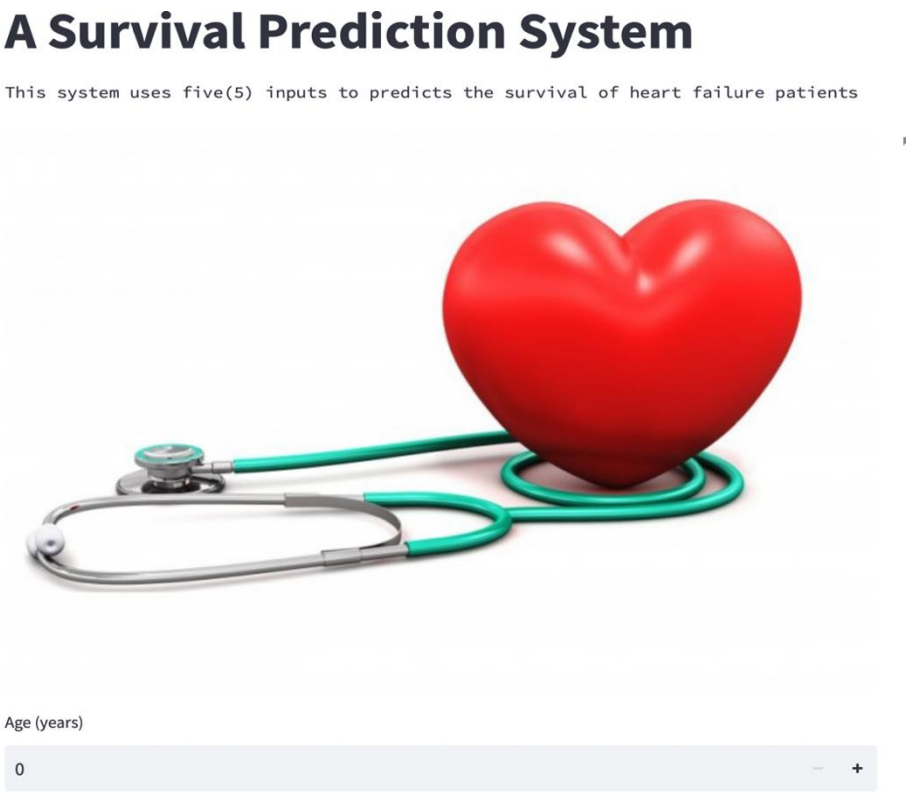


Figure 45: First Screenshot of the User Interface

Figure 46: Second user interface screenshot

CHAPTER FIVE

5.0 Discussions

5.1 Introduction

In this chapter, the findings and unexpected findings of this thesis are discussed (the events are described, followed by a discussion of their possible causes).

5.2 Binary Distribution

It was discovered that the majority of patients/participants are male, which may be due to the underrepresentation of women in clinical trials and women are less likely than men to seek out evidence-based medical care, according to community-based studies of HF patients (Stein *et al*. 2013). Moreover, American Heart Association estimates that there were slightly more males with heart failure (HF) than women (Go *et al*. 2013).

Non-smokers make up most patients; this may be a consequence of inadequate knowledge of the onset and prognosis of HF due to smoking (Conard *et al*. 2009) or the dataset is not the perfect representation of HF patients. However, Smoking is a substantial risk factor for heart failure (HF) (Krumholz *et al*. 1997).

The highest population of patients (174 participants) are not diabetics; this could be due to an unidentified contribution of diabetes to the burden of heart failure (HF). Furthermore, a small percentage of hospitalised HF patients are known to be diabetic, and this fraction has been increasing over the last decade (Echouffo-Tcheugui *et al*. 2016). Diabetes, on the other hand, is one of the conditions attributed to heart failure (Baker 2002).

Based on known measurements of health markers, elevated blood pressure is defined as blood pressure of 140/90 mm /Hg or greater on both readings (systolic and diastolic). The highest population of patients (194 participants) are non-hypertensives, which may be related to the fact that the risk of heart failure is greater than in normotensive persons as compared to hypertensives (Kannel *et al*. 1999).

The majority of patients (170 participants) are not anaemic, compared to 129 who are. Congestive heart failure (CHF) frequently results in anaemia, which is indicated by a haemoglobin level of less than 12 g/dL (Silverberg *et al*. 2006). Additionally, it has been recognised as one of the neurohormonal abnormalities in the pathophysiology and progression of congestive heart failure (CHF) (Alexandrakis and Tsirakis 2012).

In comparison to patients who passed away, a higher percentage of patients are still alive; this category represents 67.9% of the population following their follow-up time while 32.1% represented those who are deceased. There are consequently fewer occurrences of individuals passing away. This dataset may be seen to be unbalanced. Unbalanced data sets occur when at least one class is represented by a small number of training instances (referred to as the minority class) while other classes constitute the majority (Ganganwar 2012).

5.3 Bivariate Analysis

The average age of individuals who are still alive is 59, while the average age of those who have died is 65. This shows that older patients with HF died at a higher rate than younger people. Accordingly, older patients have a worse chance of surviving than younger ones and are more prone to having additional ailments that can complicate heart failure. The incidence of congestive heart failure increased dramatically with age. Age-related changes in clinical outcomes in HF patients are progressively worse as patients get older (Rich *et al*. 2001). This could be the tenable explanation for the low rate of survival among the elderly.

Looking at the smoking status of patients with the dependent variable (death event). It shows that the number of non-smokers in category 1 (Alive) decreased from 137 to 66 in category 2 (Not alive). Similarly, the number of smokers decreased exponentially from 66 (Alive) to 30 in category 2 (Not alive). This illustrates that smoking status has little bearing on the patient's chance of surviving.

The number of non-diabetics in category 1 (Alive) declined from 118 to 56 in category 2 (Not alive). The number of diabetics also dropped dramatically from 85 (Alive) to 40 in category 2 (Not alive). This demonstrates that a patient's diabetes condition has minimal impact on his or her chances of survival.

Furthermore, the number of hypertensives in category 1 (Alive) has decreased from 66 to 39 in category 2 (Not alive). Non-hypertensives decreased substantially from 137 (Alive) to 57 in category 2 (Not alive). This illustrates that a patient's hypertensive status has little effect on his or her odds of survival.

Additionally, the number of females in category 1 (Alive) declined from 71 to 34 in category 2 (Not alive). The number of males also dropped dramatically from 132 (Alive) to 62 in category 2 (Not alive). Males outnumber females in the 'not alive' group, implying that men with HF have a poorer rate of survival. This demonstrates that a patient's gender influences the likelihood of survival. Accordingly, it may be argued that the female gender is an independent predictor of death events. This finding is opposed by Zahid et al who stated that gender does not correlate to the deadly events of HF patients (Zahid *et al*. 2019).

The number of not anaemic individuals in category 1 (Alive) declined from 120 to 50 in category 2 (Not alive). The number of anaemics also dropped dramatically from 83 (Alive) to 46 in category 2 (Not alive). Non-anaemic patients outnumber anaemic in the 'not alive' group. This shows that the likelihood of survival is unaffected by the anaemic condition of a patient. However, anaemia has been linked to a lower chance of survival in all patient populations (Lindenfeld 2005). This explains the cause of death in the 46 HF patients identified as having HF comorbidities such as decreased renal function, among other things. Furthermore, the 50 individuals who died without being anaemic may have had other HF issues unrelated to anaemia.

The average platelet value in category 1 (Alive) declined from 266657 to 256381 in category 2 (Not alive). This implies that individuals with low platelet counts have a reduced likelihood of survival. This demonstrates that a patient's platelet count has an impact on their chance of survival.

The maximum serum creatinine value in category 1 (Alive) rose from 6 to 9 in category 2 (Not alive). This implies that individuals with serum creatinine above 6 have a reduced likelihood of survival. This reveals that a patient's serum creatinine level affects his or her chances of survival.

The average serum sodium value in category 1 (Alive) decreased from 137 to 135 in category 2 (Not alive). This implies that there is minimal variance between the two values. This reveals that a patient's serum sodium level has a negligible impact on chances of survival.

The average creatinine phosphokinase(cr_ph) of participants alive is 540 while deceased is 670. This depicts that those who had high levels of cr_ph died more frequently than those who had low levels.

This signifies that HF patients with greater cr_ph levels have a poorer chance of survival than patients with lower values.

The maximum ejection fraction (ej_fr) of individuals who are living is 80, whereas for those who are dead is 70, as seen in the line chart above. This shows that people with a lower percentage of ejection fraction died at a higher rate than those with a higher percentage. This means that HF patients with ejection fraction < 70 have a lesser likelihood of surviving than those with a higher percentage.

Also, the average follow-up time of HF patients in the first category 'alive' is 158 days while the second category 'Not alive' is 71days. This demonstrates that people who had shorter follow-up days than those who had longer follow-up days died more frequently.

5.4 Multivariate Analysis

The relationship or strength between the independent variables and the dependent variable 'death events' (status) shows ‘time’ has the strongest correlation to the dependent variable ‘death events’ (status) when compared to the other variables, whereas age has the lowest correlation value. Serum sodium,

age, ejection fraction, serum creatinine, and time can be placed in the variables' strength order, from weakest to strongest.

Correlation analysis reveals that values closer to -1 and +1 indicate significant/strong correlations, whereas values closer to zero indicate weak correlations (Evans and Basu 2013).

K-means clustering was performed with Tableau, which generated two clusters based on continuous variables such as cr_ph, platelet, ej_fr, ser_cr, ser_na, age, and time. After including dichotomous variables such as sex, anaemia, smoking, diabetes, and HBP, the direction of the clusters remained unchanged from 2. It has been discovered that the dichotomous variables have low variance, which automatically makes them the strongest in determining clusters (zero p-values). As a result, dichotomous variables were removed from the clusters to allow for a better understanding of other variables. The p-value is a probability that measures the significance and values <0.5 are considered significant (Liu *et al.* 2008). It can be interpreted that time is highly significant in cluster formation(value=0.0), whereas serum sodium is the least significant.

5.5 Model Evaluation

The results of model evaluation on the "Heart Failure Clinical Records" dataset after training the models shows that all models meet the required accuracy. However, other important factors need to be considered such as standard variation to avoid overfitting or underfitting. The lowest standard deviation is for DecisionTreeClassifier (0.078), followed by KNN (0.106), SVC (0.119), and LR (0.115). This implies that the models might overfit (due to high variance) except DecisionTreeClassifier. Hence, the reason for adopting the ensemble approach. The hard voting ensemble technique was used, and the results are shown in Table 7. The accuracy score for the voting classifier is 88%, and the F1 scores for the two possibilities are fairly within a reasonable range when compared to the success metrics. Whereas the recall and f1 score for the deceased (Not alive) is lower than expected (62 and 74), this could be attributed to the lower number of instances in this category, but overall, the model's prediction is within the acceptable range.

5.6 Performance Evaluation Against Benchmark Studies

Models developed by Ishaq et al. (2021) and Rahayu et al. (2020), both of which used the same dataset of "Heart Failure Clinical Records", were used as benchmark models to compare the performance of the models from this study. Nine classification models were used in Ishaq et al. (2021) research: the Decision Tree (DT), Adaptive Boosting Classifier (AdaBoost), Logistic Regression (LR), Stochastic Gradient Classifier (SGD), Random Forest (RF), Gradient Boosting Classifier (GBM), Extra Tree Classifier (ETC), Gaussian Naive Bayes classifier (G-NB), and Support Vector Machine (SVM). ETC performs better than other models, according to the experimental findings, with an accuracy value of 0.92% for prediction of heart patient survival.

Rahayu et al.(2020) study proposed SMOTE (Synthetic Minority Over-Sampling Technique), as well as data mining methods and multiple classification algorithms, to estimate patient survival. Artificial neural networks, Decision Trees, KNN, Support Vector Machines (SVM), Random Forest algorithms, and Naive Bayes were all used in the study. The accuracy produced by the SMOTE method used in the random forest is 85.82% higher than that of other algorithms.

A hard voting ensemble approach was used in this study, and the evaluation yielded an accuracy score of 0.88%. This method outperforms the Rahayu et al study (85.82% accuracy) but underperforms the Ishaq et al study (0.92% accuracy). However, it remains within an acceptable range.

CHAPTER SIX

6.0 Research Conclusion, Research Limitations and Suggestions for Future Research.

6.1 Introduction

The present study aimed to answer centred around identifying the pre-existing health factors of heart failure patients and how these factors impact their chance of survival. In addition to the goals, the current study also aimed to develop a system that can predict the survival of patients with heart failure by incorporating the research's findings.

This allows for a concise discussion of the findings as well as the study's limitations, suggestions for future research, and recommendations for future researchers.

6.2 Research Questions Reiterated

The following specific questions were the focus of this research in relation to the goal.

1. What current health conditions (variables) exist in heart failure patients that may indicate a high risk of death?
2. What are the factors/variables that ensure the survival of patients with heart failure.

6.3 Overall Research Conclusions - Findings Summary

The study's goal is to use machine learning to predict heart failure patient survival and to create a system that can predict heart failure patient survival. Analysing health data has revealed factors such as age, gender, platelet counts, serum creatinine, creatinine phosphokinase, and follow-up time to influence the survival of HF patients. However, to improve the performance of machine learning models, RandomForest algorithm was employed for feature selection, and the topmost important features were identified as age, ejection fraction,

serum creatinine, serum sodium and follow-up time. The patterns formed in the dataset were identified using clustering analysis. And it gave rise to two clearly defined clusters. The two categories in the independent variable (death_event) are confirmed by these clusters.

This study was performed using a combination of machine learning models to benefit from their advantages (Ensemble learning). In this case, the machine learning models used are K-nearest neighbour (KNN), Support vector machine (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC). The five most crucial features in the dataset were used to train the models. A hard voting ensemble approach was used, and the evaluation resulted in an accuracy score of 0.88%. Whereas the recall and f1 score for the deceased (Not alive) is lower than anticipated (0.62 and 0.74), this could be attributed to the relatively small number of instances in this category, but overall, the model's prediction is within the acceptable range. Simple software that predicts patient survival based on user input was created using the pre-trained voting classifier model.

6.4 Resulting Advantages

Through the analysis of relevant factors, this work has the potential to enhance the healthcare system and serve as a valuable tool for healthcare professionals in predicting the survival of heart failure patients. Additionally, it will help to decrease the hospitalisation rate, which will save time, money, and resources that could be used for healthcare. Furthermore, it will assist medical practitioners in selecting a population to test and evaluate the efficacy of a potential new treatment. By providing high-quality medical care and concentrating on the major risk factors that can result in death, physicians will be able to identify patients who require intensive monitoring and treatment. Hence, increasing the survival rate of HF patients.

6.5 Study limitations and recommendations for further research

Further research that can be carried out by subsequent researchers should incorporate the use of a larger dataset. Furthermore, the dataset only identified 12 features that can influence the survival of HF patients. Therefore, future

research on the survival of HF patients should consider more variables. Also, the participants in the dataset were from Pakistan, rendering it underrepresented globally. Data should be collected from various geographical locations around the world. Furthermore, the percentage of females in the sample in this study is 35.1%, which is relatively lower than the percentage of male participants (64.9%), and future research can ensure that a more equal ratio is included in the study, increasing the robustness of the study results and bringing forth new findings.

7.0 References

ABBAS, O.A., 2008. Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT),* 5(3),

AGRESTI, A., 2010. *Analysis of ordinal categorical data.* John Wiley & Sons

AGRESTI, A., 2018. *An introduction to categorical data analysis.* John Wiley & Sons

AGRESTI, A. and B. FINLAY, 2009. *Statistical methods for the social sciences.*

AHMAD, T. *et al.,* 2017. Survival analysis of heart failure patients: A case study. *PLoS ONE,* 12(7), e0181001

ALEXANDRAKIS, M.G. and G. TSIRAKIS, 2012. Anemia in heart failure patients. *International Scholarly Research Notices,* 2012

AUTHORS/TASK FORCE MEMBERS *et al.,* 2012. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European heart journal,* 33(14), 1787-1847

BAKER, D.W., 2002. Prevention of heart failure. *Journal of cardiac failure,* 8(5), 333-346

BEYENE, C. and P. KAMAT, 2018. Survey on prediction and analysis the occurrence of heart disease using data mining techniques. *International Journal of Pure and Applied Mathematics,* 118(8), 165-174

BOUVY, M.L. *et al.,* 2003. Predicting mortality in patients with heart failure: a pragmatic approach. *Heart,* 89(6), 605-609

BRESSERT, E., 2012. SciPy and NumPy: an overview for developers.

CANLAS, R.D., 2009. Data mining in healthcare: Current applications and issues. *School of Information Systems & Management, Carnegie Mellon University, Australia,*

CARREAU, A. *et al.,* 2011. Why is the partial oxygen pressure of human tissues a crucial parameter? Small molecules and hypoxia. *Journal of Cellular and Molecular Medicine,* 15(6), 1239-1253

CHENG, T., C. WEI and V.S. TSENG, 2006. Feature selection for medical data mining: comparisons of expert judgment and automatic approaches. *19th IEEE symposium on computer-based medical systems (CBMS'06).* IEEE, pp.165-170

CHICCO, D. and G. JURMAN, 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), 1-16

CONARD, M.W. *et al.,* 2009. The impact of smoking status on the health status of heart failure patients. *Congestive heart failure,* 15(2), 82-86

CRESWELL, J.W., 2014. Qualitative, quantitative and mixed methods approaches.

CUNNINGHAM, P. and S.J. DELANY, 2021. K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR),* 54(6), 1-25

DAS, K.R. and A. IMON, 2016. A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics,* 5(1), 5-12

DOGAN, A. and D. BIRANT, 2019. A weighted majority voting ensemble approach for classification. *2019 4th International Conference on Computer Science and Engineering (UBMK).* IEEE, pp.1-6

ECHOUFFO-TCHEUGUI, J.B. *et al.,* 2016. Temporal trends and factors associated with diabetes mellitus among patients hospitalized with heart failure: findings from Get With The Guidelines–Heart Failure registry. *American Heart Journal,* 182, 9-20

EL SANHARAWI, M. and F. NAUDET, 2013. Understanding logistic regression. *Journal francais d'ophtalmologie,* 36(8), 710-715

EVANS, J.R. and A. BASU, 2013. *Statistics, data analysis and decision modeling.* 5th ed. Boston: Pearson

FLETCHER, G.F. *et al.,* 1990. Exercise standards. A statement for health professionals from the American Heart Association. *Circulation,* 82(6), 2286-2322

GANGANWAR, V., 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering,* 2(4), 42-47

GILBERT, R.E. *et al.,* 2006. Heart failure and nephropathy: catastrophic and interrelated complications of diabetes. *Clinical Journal of the American Society of Nephrology,* 1(2), 193-208

GO, A.S. *et al.,* 2013. Heart disease and stroke statistics—2013 update: a report from the American Heart Association. *Circulation,* 127(1), e6-e245

GOGTAY, N.J. and U.M. THATTE, 2017. Principles of correlation analysis. *Journal of the Association of Physicians of India,* 65(3), 78-81

GROENEWEGEN, A. *et al.,* 2020. Epidemiology of heart failure. *European journal of heart failure,* 22(8), 1342-1356

GUIDI, G. *et al.,* 2014. A machine learning system to improve heart failure patient assistance. *IEEE journal of biomedical and health informatics,* 18(6), 1750-1756

HOLLAND, S.M., 2008. Principal components analysis (PCA). *Department of Geology, University of Georgia, Athens, GA,* , 30602-32501

HU, Q., L. MA and J. ZHAO, 2018. DeepGraph: A PyCharm tool for visualizing and understanding deep learning models. *2018 25th Asia-Pacific Software Engineering Conference (APSEC).* IEEE, pp.628-632

INAMDAR, A.A. and A.C. INAMDAR, 2016. Heart failure: diagnosis, management and utilization. *Journal of clinical medicine,* 5(7), 62

ISHAQ, A. *et al.,* 2021. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE access,* 9, 39707-39716

JOLLIFFE, I.T. and J. CADIMA, 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 374(2065), 20150202

JOSEPH, F. *et al.,* 2010. *Multivariate data analysis.* Pearson Prentice Hall

KAMIRAN, F. and T. CALDERS, 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems,* 33(1), 1-33

KANNEL, W.B. *et al.,* 1999. Profile for estimating risk of heart failure. *Archives of Internal Medicine,* 159(11), 1197-1204

KARSMAKERS, P., K. PELCKMANS and J.A. SUYKENS, 2007. Multi-class kernel logistic regression: a fixed-size implementation. *2007 International Joint Conference on Neural Networks.* IEEE, pp.1756-1761

KETTENRING, J.R., 2006. The practice of cluster analysis. *Journal of classification,* 23(1), 3-30

KING, M., J.E. KINGERY and B. CASEY, 2012. Diagnosis and evaluation of heart failure. *American Family Physician,* 85(12), 1161-1168

KLEIJNEN, J.P. and R.G. SARGENT, 2000. A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research,* 120(1), 14-29

KOHAVI, R. and F. PROVOST, 1998. Confusion matrix. *Machine Learning,* 30(2-3), 271-274

KOTHARI, C.R., 2004. *Research methodology: Methods and techniques.* New Age International

KRUMHOLZ, H.M. *et al.,* 1997. Readmission after hospitalization for congestive heart failure among Medicare beneficiaries. *Archives of Internal Medicine,* 157(1), 99-104

KUHN, M. and K. JOHNSON, 2013. Over-fitting and model tuning. *Applied predictive modeling.* Springer, pp.61-92

LEAU, Y.B. *et al.,* 2012. Software development life cycle AGILE vs traditional approaches. *International Conference on Information and Network Technology.* pp.162-167

LEVY, W.C. *et al.,* 2006. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation,* 113(11), 1424-1433

LILLY, B. and A.J. MILLER, 2021. Using an excel visual to illustrate univariate and bivariate variability. *Teaching Statistics,* 43(2), 79-84

LINDBLAD, T. and J.M. KINSER, 2013. NumPy, SciPy and Python Image Library. *Image Processing using Pulse-Coupled Neural Networks.* Springer, pp.35-56

LINDENFELD, J., 2005. Prevalence of anemia and effects on mortality in patients with heart failure. *American Heart Journal,* 149(3), 391-401

LIPTON, Z.C., C. ELKAN and B. NARAYANASWAMY, 2014. Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892,*

LIU, Y. *et al.,* 2008. Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association,* 103(483), 1281-1293

LO, S.K. *et al.,* 1995. Non-significant in univariate but significant in multivariate analysis: a discussion with examples. *Changgeng Yi Xue Za Zhi,* 18(2), 95-101

MAĆKIEWICZ, A. and W. RATAJCZAK, 1993. Principal components analysis (PCA). *Computers & Geosciences,* 19(3), 303-342

MARTINEZ, W.L., A.R. MARTINEZ and J.L. SOLKA, 2017. *Exploratory data analysis with MATLAB®.* Chapman and Hall/CRC

MELILLO, P. *et al.,* 2013. Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE journal of biomedical and health informatics,* 17(3), 727-733

MITHERAN, S., R.T. NARAYANAN and R. SINGARAVELU, 2022. User-Friendly Waveguide Mode Visualizer [Educator's Corner]. *IEEE Microwave Magazine,* 23(5), 96-100

OLADIMEJI, O.O. and O. OLADIMEJI, 2020. Predicting survival of heart failure patients using classification algorithms. *JITCE (Journal of Information Technology and Computer Engineering),* 4(02), 90-94

OPOKU, A., V. AHMED and J. AKOTIA, 2016. Choosing an appropriate research methodology and method. *Research Methodology in the Built Environment.* Routledge, pp.32-49

OSBORNE, J.W. and A. OVERBAY, 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation,* 9(1), 6

PARK, H.M., 2015. Univariate analysis and normality test using SAS, Stata, and SPSS.

PEARSE, S.G. and M.R. COWIE, 2014. Heart failure: classification and pathophysiology. *Medicine,* 42(10), 556-561

PEDREGOSA, F.*, et al.*, 2011. *Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot Edouard Duchesnay.* , pp.2825

PRADHAN, A., 2012. Support vector machine-a survey. *International Journal of Emerging Technology and Advanced Engineering,* 2(8), 82-85

PRIYANKA and D. KUMAR, 2020. Decision tree classifier: A detailed survey. *International Journal of Information and Decision Sciences,* 12(3), 246-269

QIU, W., W.J. MURPHY and A. SUTER, 2020. Kurtosis: a new tool for noise analysis. *Acoust Today,* 16(4), 39-47

RAGUNATH, P.K. *et al.,* 2010. Evolving a new model (SDLC Model-2010) for software development life cycle (SDLC). *International Journal of Computer Science and Network Security,* 10(1), 112-119

RAHAYU, S. *et al.,* 2020. Prediction of survival of heart failure patients using random forest. *Jurnal Pilar Nusa Mandiri,* 16(2), 255-260

RICH, M.W. *et al.,* 2001. Effect of age on mortality, hospitalizations and response to digoxin in patients with heart failure: the DIG study. *Journal of the American College of Cardiology,* 38(3), 806-813

RINDHE, B.U. *et al.,* 2021. Heart Disease Prediction Using Machine Learning. *Heart Disease,* 5(1),

ROGERS, J. and S. GUNN, 2005. Identifying feature relevance using a random forest. *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection".* Springer, pp.173-184

SAVARESE, G. and L.H. LUND, 2017. Global Public Health Burden of Heart Failure. *Cardiac failure review,* 3(1), 7-11

SCOTT, A.J. and M. KNOTT, 1974. A cluster analysis method for grouping means in the analysis of variance. *Biometrics,* , 507-512

SEN, P.C., M. HAJRA and M. GHOSH, 2020. Supervised classification algorithms in machine learning: A survey and review. *Emerging technology in modelling and graphics.* Springer, pp.99-111

SHAH, D., S. PATEL and S.K. BHARTI, 2020. Heart disease prediction using machine learning techniques. *SN Computer Science,* 1(6), 1-6

SHAHABI, H., B.B. AHMAD and S. KHEZRI, 2013. Evaluation and comparison of bivariate and multivariate statistical methods for landslide susceptibility mapping (case study: Zab basin). *Arabian journal of geosciences,* 6(10), 3885-3907

SILVERBERG, D.S. *et al.,* 2006. The interaction between heart failure and other heart diseases, renal failure, and anemia. *Seminars in nephrology.* Elsevier, pp.296-306

SOKOLOVA, M. and G. LAPALME, 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management,* 45(4), 427-437

SRINATH, K.R., 2017. Python–the fastest growing programming language. *International Research Journal of Engineering and Technology,* 4(12), 354-357

STEIN, G.Y. *et al.,* 2013. Gender-related differences in hospitalized heart failure patients. *European journal of heart failure,* 15(7), 734-741

TATBUL, N. *et al.,* 2018. Precision and recall for time series. *Advances in neural information processing systems,* 31

TUKEY, J.W., 1977. *Exploratory data analysis.* Reading, MA

VAFAEI, N., R.A. RIBEIRO and L.M. CAMARINHA-MATOS, 2018. Data normalisation techniques in decision making: case study with TOPSIS method. *International journal of information and decision sciences,* 10(1), 19-38

VAN DER MAATEN, L., 2007. An introduction to dimensionality reduction using matlab. *Report,* 1201(07-07), 62

VIDMAR, G., 2007. Statistically sound distribution plots in Excel. *Advances in Methodology and Statistics,* 4(1), 83–98

VON HAEHLING, S. *et al.,* 2020. Muscle wasting as an independent predictor of survival in patients with chronic heart failure. *Journal of cachexia, sarcopenia and muscle,* 11(5), 1242-1249

VUJOVIĆ, ŽĐ, 2021. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications,* 12(6), 599-606

WHO, 2021. *Cardiovascular diseases(CVDs)* [viewed June 21, 2022]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

WU, L. and M. LI, 2018. Applying the CG-logistic regression method to predict the customer churn problem. *2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS).* IEEE, pp.1-5

YOUSAF, A. *et al.,* 2020. Emotion recognition by textual tweets classification using voting classifier (LR-SGD). *IEEE Access,* 9, 6286-6295

ZAHID, F.M. *et al.,* 2019. Gender based survival prediction models for heart failure patients: A case study in Pakistan. *PloS one,* 14(2), e0210602

8.0 Appendices

8.1 Appendix A-Ethics Form

Ethical clearance for research and innovation projects

Project status

Status

Approved

Actions

Date	Who	Action	Comments
15:24:00 04 July 2022	Jarutas Andritsch	Supervisor approved	
15:14:00 04 July 2022	Hadizat Talabi	Principal investigator submitted	

Get Help

Ethics release checklist (ERC)

Project details

Project name:	Prediction of survival of heart failure patients using machine learning
Principal investigator:	Hadizat Talabi
Faculty:	Faculty of Business, Law and Digital Technologies
Level:	Postgraduate
Course:	Dissertation
Unit code:	COM726
Supervisor name:	Jarutas Andritsch
Other investigators:	

Checklist

Question	Yes	No
Q1. Will the project involve human participants other than the investigator(s)?	○	◉
Q1a. Will the project involve **vulnerable participants** such as children, young people, disabled people, the elderly, people with declared mental health issues, prisoners, people in health or social care settings, addicts, or those with learning difficulties or cognitive impairment either contacted directly or via a **gatekeeper** (for example a professional who runs an organisation through which participants are accessed; a service provider; a care-giver; a relative or a guardian)?	○	◉
Q1b. Will the project involve the use of **control groups** or the **use of deception**?	○	◉
Q1c. Will the project involve any **risk to the participants' health** (e.g. intrusive intervention such as the administration of drugs or other substances, or vigorous physical exercise), or involve psychological stress, anxiety, humiliation, physical pain or discomfort to the investigator(s) and/or the participants?	○	◉
Q1d. Will the project involve **financial inducement** offered to participants other than reasonable expenses and compensation for time?	○	◉
Q1e. Will the project be carried out by individuals unconnected with the University but who wish to use staff and/or students of the University as participants?	○	◉
Q2. Will the project involve sensitive materials or topics that might be considered offensive, distressing, politically or socially sensitive, deeply personal or in breach of the law (for example criminal activities, sexual behaviour, ethnic status, personal appearance, experience of violence, addiction, religion, or financial circumstances)?	○	◉
Q3. Will the project have detrimental impact on the environment, habitat or species?	○	◉
Q4. Will the project involve living animal subjects?	○	◉
Q5. Will the project involve the development for export of 'controlled' goods regulated by the Export Control Organisation (ECO)? (This specifically means military goods, so called dual-use goods (which are civilian goods but with a potential military use or application), products used for torture and repression, radioactive sources.) Further information from the Export Control Organisation	○	◉
Q6. Does your research involve: the storage of records on a computer, electronic transmissions, or visits to websites, which are associated with terrorist or extreme groups or other security sensitive material? Further information from the Information Commissioners Office	○	◉

Declarations

I/we, the investigator(s), confirm that:

☑ The information contained in this checklist is correct.

☑ I/we have assessed the ethical considerations in relation to the project in line with the University Ethics Policy.

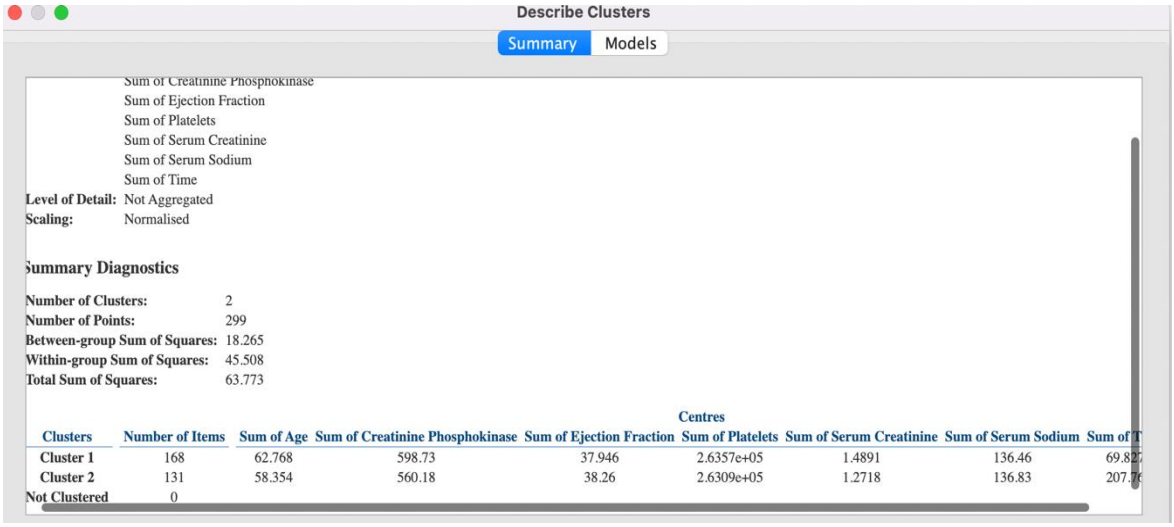
☑ I/we understand that the ethical considerations of the project will need to be re-assessed if there are any changes to it.

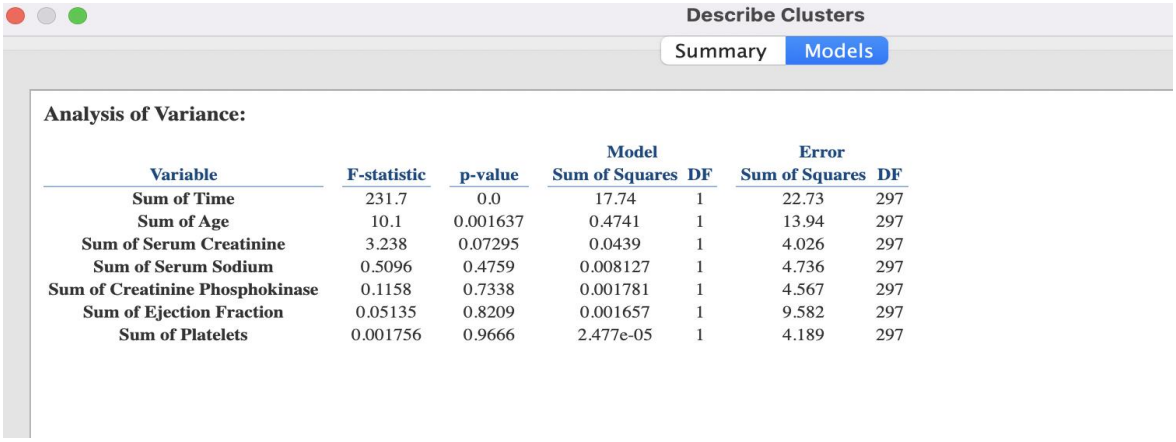
☑ I/we will endeavor to preserve the reputation of the University and protect the health and safety of all those involved when conducting this research/enterprise project.

☑ If personal data is to be collected as part of my project, I confirm that my project and I, as Principal Investigator, will adhere to the General Data Protection Regulation (GDPR) and the Data Protection Act 2018. I also confirm that I will seek advice on the DPA, as necessary, by referring to the Information Commissioner's Office further guidance on DPA and/or by contacting information.rights@solent.ac.uk. By Personal data, I understand any data that I will collect as part of my project that can identify an individual, whether in personal or family life, business or profession.

☑ I/we have read the prevent agenda.

8.2 Appendix B-Tableau Clustering Analysis





8.3 Appendix C- Model Development

```
# Spliiting dataset into train and test
from sklearn.model_selection import train_test_split

# Split dataset into training set and test set
X_train, X_test, Y_train, Y_test = train_test_split(X_data,
                                                    Y_data,
                                                    test_size=0.3,  # 70% training and 30% test
                                                    random_state=1)
#Checking the  number of test and train data
print('\n The total of training dataset:', X_train.shape)
print('\n The total of test dataset:', X_test.shape)
print(Y_test.shape)
```

```
 The total of training dataset: (209, 5)

 The total of test dataset: (90, 5)
(90,)
```

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
#importing the models
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import VotingClassifier
from sklearn.svm import SVC

svm =SVC(kernel='poly', max_iter=-1, degree=3, probability=True)
dt = DecisionTreeClassifier(criterion='entropy', max_depth=3, min_samples_leaf=0.05, min_samples_split=2,random_state=1)
kn =KNeighborsClassifier(n_neighbors=8,leaf_size=30,metric='minkowski', p=2)
lr =LogisticRegression(intercept_scaling= '1',max_iter=1000,multi_class= 'auto', penalty='l2',random_state=1, solver='newton-cg',tol=0.0001, verbose= 0,
 warm_start=False)

class_list = [('DecisionTreeClassifier:',dt),('Supportvector:',svm),('LogisticRegression:',lr),('KNeighborsClassifier:',kn)]

#iteration
for clsf_name,clsf in class_list:
    clsf.fit(X_train,Y_train)
    Y_pred = clsf.predict(X_test)
    print('\n{:s} :{:.3f}'.format(clsf_name, accuracy_score(Y_test, Y_pred)))

votingc = VotingClassifier(estimators=class_list, voting='soft')
votingc.fit(X_train, Y_train)

#predict test label
Y_pred_vc = clsf.predict(X_test)
```

Looks like you're using NumPy

```
DecisionTreeClassifier: :0.844

Supportvector: :0.856

LogisticRegression: :0.833

KNeighborsClassifier: :0.878
```

```
print('\n voting classifier {:.3f}'. format(accuracy_score(Y_test, Y_pred_vc)))
```

```
 voting classifier 0.878
```

```
#MAKING MODEL PERSISTENT FOR USE

#VIEWING THE TEST DATASET
print(X_test.tail(5).join(Y_test.tail(5)))
```

	age	ej_fr	ser_cr	ser_na	time	status
122	60.0	38	0.75	140	95	0
246	55.0	25	1.10	138	214	1
278	50.0	30	0.70	136	246	0
251	55.0	35	0.80	143	215	0
19	48.0	55	1.90	121	15	1

```
#SAVING MODEL
my_model = 'dissert_model.sav'
jb.dump(votingc, my_model)
```

```
['dissert_model.sav']
```

```
#TESTING THE MODEL EFFECT

load_my_model = jb.load(my_model)
results = load_my_model.score(X_test, Y_test)
print('\n This is the result of the persistent model\n', results)
```

```
  This is the result of the persistent model
  0.8666666666666667
```

```
x= 60.0,38,0.75,140,95
client_data = np.array(x).reshape(1,-1)
forte =load_my_model.predict(client_data)
print('This is the prediction')
print(forte)
```

```
 This is the prediction
 [0]
```

8.4 Appendix D- Application development

```python
import numpy as np
import streamlit as st
import joblib as jb
from PIL import Image

# Creating title
st.title("A Survival Prediction System")
st.text("This system uses five(5) inputs to predicts the survival of heart failure patients")
image = Image.open('new.jpg')
st.image(image, width=700)



# sex = st.selectbox('Sex', options=['Female', 'Male'])
age = st.number_input('Age (years)', min_value=0)  # min_value=40, max_value=95)
ej_fr = st.number_input('Ejection Fraction (%)', min_value=0)  # min_value=14, max_value=80)
ser_cr = st.number_input('Serum Creatinine (mg/dl)', min_value=0.5, max_value=9.4)
ser_na = st.number_input('Serum Sodium(mEq/L)', min_value=0)  # min_value=113, max_value=148)
time = st.number_input('Time (days)', min_value=0)

st.write('The user inputs are {}'.format([age, ej_fr, ser_cr, ser_na,time]))


def transform():
    age_n = float(age)
    ej_fr_n = float(ej_fr)
    ser_cr_n = float(ser_cr)
    ser_na_n = float(ser_na)
    patient = [age_n, ej_fr_n, ser_cr_n, ser_na_n, time]
    return patient
```

```python
def prediction(x):
    loaded_model = jb.load('dissert_model.sav')
    patient_value = np.array(x).reshape(1, -1)
    predict = loaded_model.predict(patient_value)
    return predict


def status():
    retrieved = transform()
    predicted = prediction(retrieved)
    st.subheader("Prediction")
    if predicted == 0:
        st.write("We predict the patient to be: Alive")
    elif predicted == 1:
        st.write(" We predict the patient : Not alive")
    else:
        st.write("Please check your data input")


status()
```