Prediction system for Cardio-vascular diseases using data mining algorithms

T Mavetera

orcid.org/0000-0002-0905-3278

Dissertation accepted in fulfillment of the requirements for the degree

Master of Science in Computer Science

North West University

Supervisor: Professor BM Esiefarienrhe

Graduation Ceremony: May 2024

DECLARATION

I, TINOTENDA MAVETERA, hereby declare that this project titled, *"Cardio-vascular diseases prediction system using data mining techniques"* is my own work undertaken at North-West University, Mafikeng, for the degree of **Master's in Computer Science in the Faculty of Natural and Agricultural Sciences** and has not been submitted for the award of a degree to any other university. All the materials and sources used are duly acknowledged in the text and the final references.

2023/09/04

Signature Date

APPROVAL:

2023/09/05

Signature Date

SUPERVISOR:
Professor BM Esiefarienrhe
Department of Computer Science and Information Systems
North-West University
South Africa

DEDICATION

I dedicate this research to my late father

Professor Nehemiah Mavetera.

Thank you, father, for everything you have done for me.

ACKNOWLEDGEMENTS

I praise the Almighty God who granted me the strength, knowledge and understanding to commence and finish this study.

I thank my supervisor, Professor B.M Esiefarienrhe, for the incessant support in my Master's research. I appreciate your patience, motivation, and immense knowledge. Your guidance helped me in the writing of this dissertation. I would not have imagined a better supervisor.

I am indebted to my parents (Prof N Mavetera and Mrs C Mavetera) for being my role models. To my siblings Patience, Kudzai, Farai, and Kudzanai – thank you for the pressure to get this study done on time. I love you.

Finally, I thank my friends, Dr Princess Khumalo, Munashe Usada, Natasha Ngwenya, Wadzanai Dzoma and Takudzwa Mahachi for the encouragement when all seemed foggy.

ABSTRACT

Background: Cardiovascular disease (CVD) is a major global health concern, accounting for a significant proportion of morbidity and mortality worldwide. Early detection and prevention of CVD reduces the burden of this disease, and data mining techniques show great potential in predicting CVD risk. Data mining algorithms use large datasets to identify patterns and relationships that may not be immediately apparent, making them specifically suitable for analysing complex biomedical data.

Objective: This research is designed to determine the best performing machine learning algorithm based on the dataset and demonstrate how this algorithm could be used to predict cardiovascular disease and improve both diagnosis and treatment. The Study examined various data mining techniques and their performance in predicting CVD risk and explore the potential of combining different algorithms for improved prediction accuracy.

Methodology: In this research, Knowledge Discovery in Database with iterative and interactive techniques was used to identify patterns from the dataset. The dataset used in this research was obtained from UCI Machine Learning Repository.

Results: The results of this study demonstrate that three out of the four implemented algorithms exhibited excellent performance, yielding high classification accuracy for predicting cardiovascular disease. Specifically, Naïve Bayes achieved 85.25%, Decision Tree achieved 71.43%, while the Artificial Neural Network outperformed the other algorithms with a classification accuracy of 91.75%. A classification accuracy of 67% achieved by KNN cannot be said to be excellent. These findings suggest that the top three algorithms could be utilized effectively for predicting the presence of cardiovascular disease.

Conclusion: To conclude, this research demonstrates that machine learning could be applied to healthcare and assist medical practitioners to diagnose the presence of cardiovascular disease effectively and thus contribute to the development of efficient CVD prevention strategies that improve public health outcomes.

Keywords: Cardiovascular Disease, Machine Learning, Data Mining, Decision Tree, KNN, ANN Naïve Bayes

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AUC	Area Under the Curve.
ANN	Machine Learning
CRISP-DM	Cross Industry Process for Data Mining
CSV	Comma Separated Values
CVD	Cardiovascular Disease
DT	Decision Tree
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
KNN	Key Nearest Neighbor
MLPNN	Multi-Layer Perceptron Neural Network
UCI	University of California, Irvine
WHO	World Health Organisation.

CHAPTER ONE: INTRODUCTION

1.1 Background

Cardiovascular disease (CVD) is a significant global concern for public health, responsible for more deaths than any other disease (WHO, 2020). In the face of this global health crisis, the use of data mining and machine learning techniques has become increasingly important in the development of prediction systems for early diagnosis and treatment of CVD.

Heart disease has been perceived erroneously as a problem primarily afflicting developed countries. However, recent data suggests that the incidence of heart disease is also prevalent in developing countries where access to adequate and effective healthcare is often limited (Damtew, 2011). This trend is particularly concerning as it not only affects men, but also women, who are just as susceptible to heart disease, as mentioned by the Office on Women's Health (Bruxvoort, 2009).

Currently, data mining has emerged as a crucial field of study due to its capacity to establish significant insights from data (Tomar & Agarwal, 2013). There has been a rapid evolution in the field of data mining in recent years, with improvements in technology and the availability of large, complex data sets. ZDNET News, an online technology magazine, predicts that data mining could be one of the most ground-breaking developments of the century (Rachel, 2001). Additionally, data mining was selected as one of the 10 emerging technologies that could transform the world by The Technology Review Ten in 2001. As a result, many companies now recognise the significance of effectively analysing data from their customers, partners, and suppliers.

Data mining techniques, such as decision trees, neural networks, and clustering algorithms, have been applied in various fields to extract meaningful information from large and complex data sets (Ramageri, 2010). Machine learning, a subfield of artificial intelligence, has experienced notable expansion and has been utilized in different domains like healthcare, natural language processing, speech, and image recognition.

Within the healthcare sector, data mining and machine learning techniques have been used to analyse electronic health records, genetic and other types of biomedical data. These techniques have been used to identify risk factors for various diseases, predict patient outcomes, and improve

decision-making in clinical practice. The prediction of CVD has been improved through the application of data mining and machine learning techniques, aimed at enhancing early detection and treatment of this ailment.

In the medical industry, there are major diseases that can be predicted, diagnosed and analysed such as cancer Leukaemia, Lung, and Breast, Diabetes, Tuberculosis, HIV/AIDS, and the cardiovascular diseases. CVD is the primary cause of mortality across the globe, causing more deaths than any other disease. The World Health Organisation (WHO) estimates that 17.9 million deaths occurred due to CVD in 2016, accounting for 31% of all global deaths. CVD encompasses a range of conditions affecting the heart and blood vessels, including coronary heart disease, stroke, and hypertension. The aetiology of these diseases involves a complex interplay of genetic predisposition, lifestyle choices, and environmental influences (WHO, 2020).

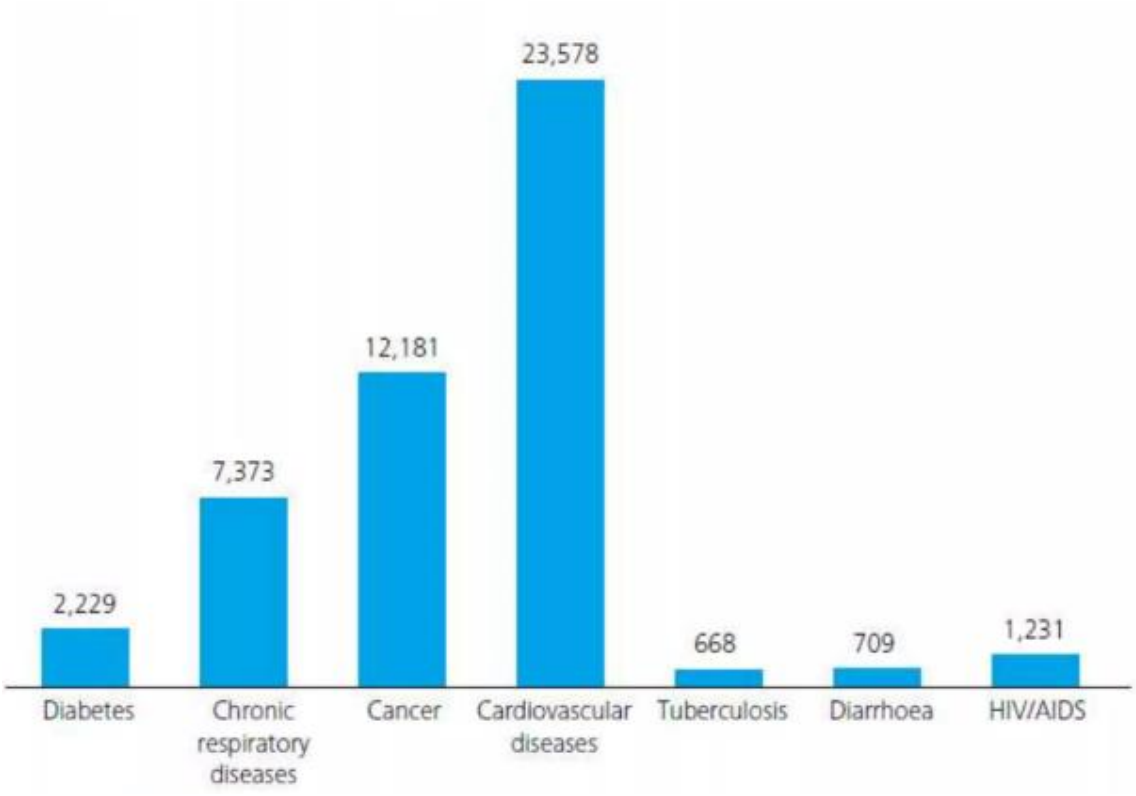


Figure 1. 1. Mortality Rate from Major Communicable and Non-Communicable Disease - Source (WHO), Number of deaths in 000's

Factors that contribute to the risk of developing CVD include hypertension, high cholesterol levels, smoking, obesity, diabetes, and a sedentary lifestyle. However, many of these risk factors are modifiable, meaning they can be changed through lifestyle changes and medical interventions.

Early identification of individuals at high risk for CVD is crucial for the implementation of preventive measures, such as lifestyle modification, medication, and surgery.

CVD is a concern to different countries. Lack of basic and primary healthcare in many developing countries, including South Africa, accounts for the increase in many people suffering from early detectable disease and treatments. According to Soni (2011), CVD is generally regarded as the primary cause for mortality in the adult population.

Over the last several years, numerous prediction models have been developed to identify individuals at high risk of CVD. These models have been based on traditional statistical techniques that have focused on a limited number of risk factors. Because these models have had limited success in identifying individuals at high risk for CVD, the emergence of data mining and machine learning algorithms has enabled the analysis of vast and intricate datasets, allowing for the detection of novel risk factors associated with CVD (Ramageri,2010).

1.2 Statement of the problem

Heart disease is a prevalent medical concern in South Africa, with nearly 215 deaths each day caused by CVD, according to the Heart and Stroke Foundation South Africa (Heart Foundation, 2019). While some hospitals employ decision support systems, many hospital information systems are primarily designed for patient billing and inventory management, with limited functionality for generating patient statistics, as noted by Palaniappan (2008). To address this gap, data mining techniques are applied in health systems to scientifically analyse data and employ predictive analytics in identifying inefficiencies and extracting patterns from the data which can reduce health and treatment costs (Dehkordi & Sajedi, 2019). Currently, clinical decisions are based on the intuition and experience of doctors, rather than leveraging the extensive data available. The inadequate functionality of hospital information systems for generating patient statistics could result in undesired bias, errors, and increased medical expenses, thereby potentially compromising the quality of patient care (Palaniappan, 2008).

1.3 Research aims and objectives

1.3.1 Research aim

The aim of this research is to determine the best performing data mining classification algorithm in the prediction of cardio-vascular disease.

1.3.2 Research questions

The main research question that guides this study is:

How can machine learning be deployed or used in determining the best performing data mining classification algorithm in the prediction of cardio-vascular disease?

The main research question is addressed in steps by answering the following sub- questions:

- **RQ1** What data could be used to identify the features associated with cardiovascular disease?
- **RQ2** How can the factors contributing to cardiovascular diseases be modelled using Unified Modelling Language tools for the purpose of mining?
- **RQ3** Which algorithms are best suited for the task of identification and diagnostic support for cardiovascular disease?
- **RQ4**: What prediction system could be developed based on the selected classification algorithm for the effective and early detection and prevention of CVD?

1.3.3 Research objectives

The objectives of this research are outlined subsequently as designed to:

- **RO1:** Identify the features causing cardiovascular disease from a dataset.
- **RO2:** Design cardio-vascular disease prediction system using Unified Modelling Language (UML) tools.
- **RO3:** Implement (RO2) using data mining techniques namely Naive Bayes, Decision Tree, K-Nearest Neighbour and Artificial Neural Network to interpret the results obtained.
- **RO4:** To develop a CVD prediction system based on the selected classification algorithm, which could be used to aid in the early prevention, detection, and treatment of CVD.

1.4 Research methodology

Researchers have developed so many methodologies that could resolve the current problem. According to Bell and Waters (2014), as a researcher, it is incumbent to carefully select an approach that is not only suitable for the specific problem addressed and aligned with the research objectives but is also capable of generating dependable outcomes.

For this study, the researcher intends to construct a predictive model that can classify cases of cardiovascular disease using data obtained from transthoracic echocardiography examinations. To achieve this objective, the researcher employs the Knowledge Discovery in Database (KDD) process model for data mining. As outlined by Fayyad et al. (1996), this model follows a nine-stage life cycle that is depicted in Figure 2.2 and is commonly utilized in data mining initiatives.

The KDD process model is incorporated into the data science research (DSR) methodology because DSR is geared towards creating specific research artefacts and the artefact depends on the domain in which it is created. Because this research entails data mining, the artefact created involves the use of an appropriate model suitable for data mining which is the KDD model.

The primary objective of the study is to determine the best performing machine learning algorithm based on the dataset and demonstrate how this algorithm could be used to predict cardiovascular disease and improve both diagnosis and treatment. To attain this aim, a comprehensive review of the relevant literature was conducted, focusing on data mining studies conducted in the context of diagnosing cardiovascular diseases. Therefore, four classifiers (Naïve Bayes, Decision Tree, K-Nearest Neighbour and Artificial Neural Network) are selected to build various classifiers. The construction of a classifier varies across different techniques, resulting in different behaviours and outcomes for each method.

To emphasize the practical feasibility of this approach, Google Collaboratory is used to implement the chosen classifiers. The algorithms' performance is evaluated using the leave one out cross-validation (holdout) model in all experiments. This involves dividing the original dataset into training and testing sets, with one observation being held out during each validation cycle while the rest of the observations act as the training set to constructing the classifier model. This process is reiterated until all observations have acted as the test observations. Throughout the validation

process, performance errors are computed at each iteration, and upon completion, the mean performance is ultimately determined.

This study contributes to practice as a relevant problem has been identified, which is the problem of predicting the likelihood of experiencing heart disease or cardiovascular disease using data mining algorithms. An understanding of the topic and problem is deepened through literature reviews of related topics. The theoretical connections and research contributions are clarified throughout this research.

1.5 Subset Data for Experiment

1.5.1 Dataset used

The dataset used for this study is open and publicly available from UCI (University of California, Irvine, CA) Machine Learning Repository and is accessible via http://archive.ics.uci.edu/ml/datasets/Heart+Disease (Aha, 1988). These datasets share the same format for instances and attributes. They both contain a total of 76 raw attributes, including the predicted attribute, but only 14 of these attributes are relevant to this study. The Cleveland Clinic Foundation dataset comprises 303 patient records, while the Hungarian Institute of Cardiology dataset has 294 patient records. By integrating both datasets, the total number of instances in the dataset is 597.

1.6 Ethical considerations

The database used is publicly available and do not have any ethical restriction attached to it. For verification purposes, the database is available at:

http://archive.ics.uci.edu/ml/datasets/Heart+Disease

1.7 Dissemination of Results and contribution to knowledge

The following methods were used to disseminate the study's findings:

- A presentation for the School of Computer Science and Information Systems at the North-West University.

- An article has been submitted to Researchfora, a journal of Information systems and information and it is under peer review.
- A paper has been presented at an international conference on medical and health science, Rome, Italy and the paper will be published at the conference proceedings.
- The research document shall be deposited in the libraries of pertinent organizations, enabling interested readers to access and utilize it.

1.8 Motivation for the study

The motivation for applying the datamining algorithms in medical diagnosis are elaborated as follows:

1) The results of this research offered the deep meaning and knowledge embedded in databases such that hospitals could archive their data and make them available to research.
2) This research contributes to policy makers in healthcare institutions by enhancing patient safety and reducing death attributed to heart disease.
3) Medical organisations both private and public could apply this research to avoid medical errors through data mining of their medical records, where they can find useful information that can assist them in medical breakthrough and efficient services to patients.
4) This research will be adopted by Software developers to enable hospitals to make early detection of CVD possible thereby reducing deaths and exerting positive impact on patient life.

1.9 Structure of the study

This research is presented in six chapters which are outlined as follows.

1: Introduction
2: Literature Review
3: Research methodology
4: Data preparation and mining
5: Results and discussion
6: Conclusions, recommendation, and future work

In this introductory chapter of the study, a concise summary of the research is presented, comprising an outline of the research problem tackled, the study's goals, the research methodology utilised, as well as the contributions of the study.

Chapter two of the thesis reviews pertinent literature regarding data mining and heart disease. This section grounds readers in the techniques and vocabulary used in subsequent parts of the thesis. Additionally, a few studies that explore similar are also reviewed to consolidate the trajectory of the study and equally avoid replication.

The third chapter provides a comprehensive explanation of the research methodology, elaborating on the algorithms and performance measures utilised for both model development and performance evaluation.

Chapter four of the thesis focuses on data preparation, including the various data mining transformations that were applied to the data.

The fifth chapter focuses on the results after the performing the data mining processes and the discussion of the results.

The sixth chapter and final chapter functions as a conclusion. It offers a synopsis of the research, including the methodology employed and the experimental results obtained. Based on the study's findings, general conclusions are drawn, and recommendations for future research in this field are proffered, highlighting potential avenues for exploration that were not addressed in the current study. Overall, this final chapter ties together the various threads of the research and provides a comprehensive estimation of the contributions to the field.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Significant research has been conducted on the diagnosis of heart disease. Many risk factors connected with cardiovascular diseases have been extensively studied, including age, gender, chest pain, blood pressure, cholesterol, blood glucose, family history, obesity, and physical inactivity. Understanding of these risk factors has made it easier for healthcare professionals to diagnose a patient's heart disease (Chaurasia & Pal, 2014). In this regard, researchers have applied a variety of data mining techniques, association rule learning techniques, clustering, and classification algorithms to extract key patterns for predicting heart disease. This chapter first introduces various concepts necessary to understanding this research and then looks at specific articles on the diagnosis of CVD using data mining techniques. These are evaluated relative to the goals set for this research to identify the similarities and indicate the differences such that a clear-cut trajectory is mapped for the current study.

2.2 Data Mining

Data mining emerged during the mid-1990s and is recognized as a potent tool for uncovering patterns and valuable insights from datasets. This technology has the capacity to clarify previously undisclosed patterns and generate insights from extensive datasets. data mining is therefore a valuable resource in the domain of data analysis, enabling the extraction of meaningful and relevant information from vast amounts of data, which was not possible before its inception in the 1990s (Kerssens, 2019). This makes data mining a powerful tool for businesses and researchers alike, enabling them to chart new insights and make informed decisions based on the patterns extracted from massive datasets. In summary, data mining has revolutionized the way we analyse data and remains a valuable tool in the field of data science. The potential of data mining is only getting attention now since its inception two decades ago. Several research studies emphasise that the utilisation of Data Mining techniques enables data holders to analyse and uncover unexpected correlations within their data, ultimately contributing to informed decision-making (Kolling ,2021; Hand, 2001).

Beneath the multitude of data collected, organised, and stored by different devices, systems, and entities, there exist valuable patterns and hidden information. By analysing and manipulating these

data, crucial insights are obtained, which can greatly influence everyday business decisions in the future. So, underneath the layers of data, lies the potential for valuable discoveries that could radically transform commercial choices. All of this is made possible by data mining from databases and using such analysis to solve issues in data processing, pattern discovery, and relationship understanding (Witten & Frank, 2002).

Han (2006) define DM in a simplest form as the capacity to identify patterns in the available data. Data collection is the initial step of the conventional data mining application workflow, which is followed by data pre-processing and an analytical processing/algorithm phase. Figure 2.1 depicts this flow, with the construction blocks corresponding to various designs for application solutions.

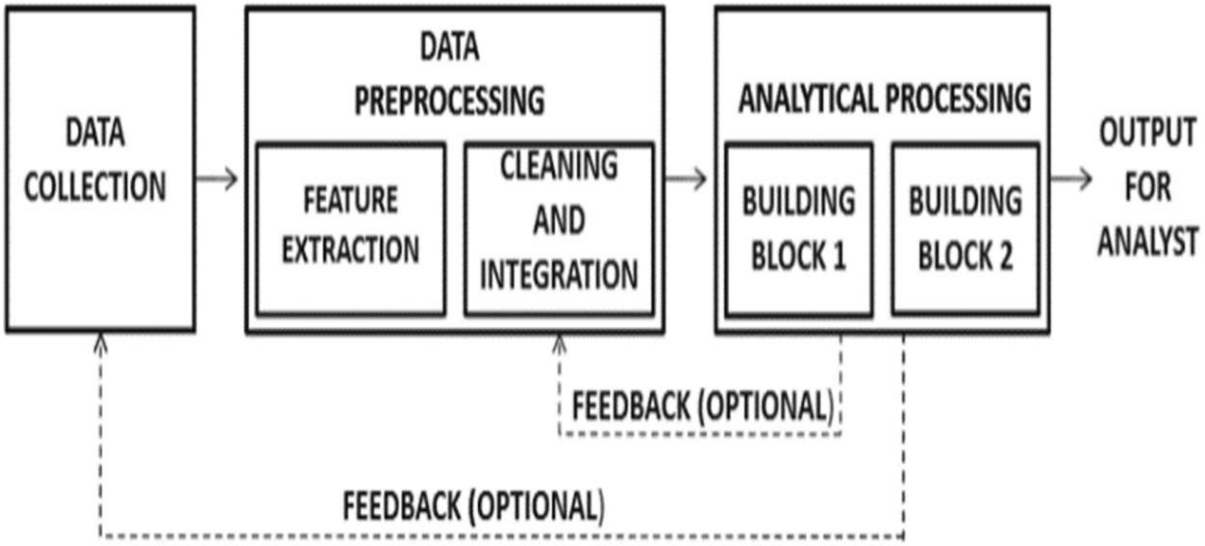


Figure 2. 1: Data Mining Flow (Aggarwal, 2015)

Figure 1 shows that in the first phase, different tools are employed to gather and store the data in the data collection phase. The importance of good quality data has a positive impact on the mining process. The second phase involves transforming the data into a format that is compatible with data mining algorithms, while the third phase entails the analytical processing, where the analyst uses their experience and expertise to design an effective method for analysis (Aggarwal, 2015).

DM (Data Mining) and KDD (Knowledge Discovery in Databases) are often conflated as interchangeable terms, although some researchers view them as distinct concepts. This is because DM is recognized as a critical stage within the broader KDD process (Fayyad et al, 1996). Fayyad

et al (1996) propose that the knowledge discovery process can be divided into a series of structured stages. The initial stage involves data selection, which entails the collection of data from numerous sources. The second stage in the data analysis process is pre-processing, which encompasses cleaning and preparing the selected data for further analysis. In the third phase, known as data transformation, the data is converted into a format that is suitable for subsequent processing. Subsequently, in the fourth stage, the transformed data is subjected to an appropriate data mining method to extract valuable insights. The ultimate phase, referred to as evaluation, involves evaluating the efficiency of the mining technique used and gauging the overall effectiveness of the entire process. These stages are depicted in Figure 2.2.

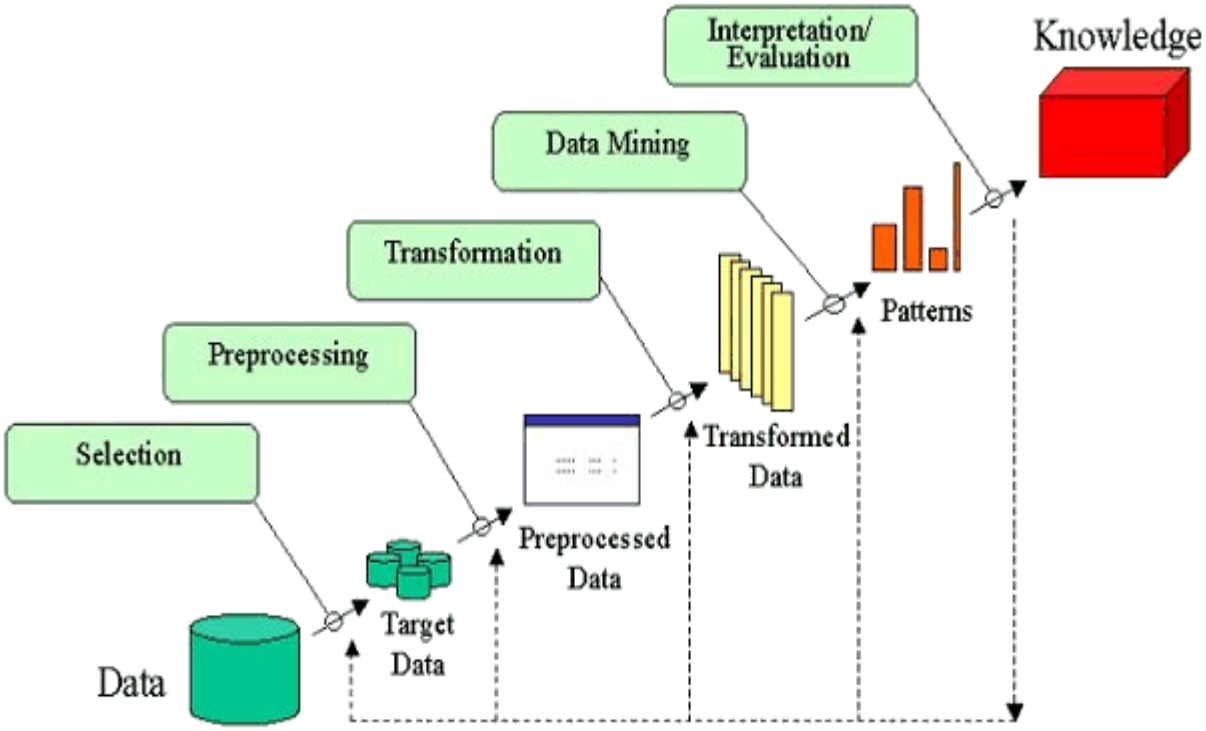


Figure 2. 2: KDD Process (Fayyad et al, 1996)

2.3 Knowledge Discovery in Database in Data Mining

KDD is a one of the mostly applied data mining process models and the one adopted for this study. The KDD is incorporated within the data science research (DSR) methodology as a specific data mining model for knowledge extraction and a process for detecting genuine, new, potentially helpful, and intelligible patterns in data and is non-trivial. Among these process models, there is CRISP-DM (Cross-Industry standard Process for Data Mining) and SEMMA (Sample, Explore, Modify, Model and Assess).

SEMMA was developed by the SAS Institute as a method for carrying out a data mining project. The model assumes a five-step cycle for the process. The SEMMA process is simple to comprehend and implement, enabling for the orderly development and upkeep of DM projects. According to Damtew (2011), the SEMMA process offers a structure for innovation, development, and progress, aiding in the proposal of business solutions and the identification of corporate objectives for data mining.

In 1996, analysts from DaimlerChrysler, SPSS, and NCR (National Cash Register) collaborated to create the CRISP-DM model, which consists of six stages in a cycle. CRISP-DM is a thoroughly documented model, with each stage carefully planned, designed, and executed (Damtew, 2011).

Fayyad et al. (1996) proposed the KKD process model, which has five key processes that run in iterative mode. Data is the first application in the first stage and extracting meaningful knowledge as the last stage as shown in Figure 2.1. The KDD model steps are summarized below:

- Selection

  This step requires comprehending the objectives and clearly defining the problem addressed. It involves identifying the available data and obtaining additional data as necessary, consolidating all relevant data into a unified dataset, and selecting the attributes used in the knowledge discovery process. This stage is crucial in the Knowledge Discovery in Databases (KDD) process, as it is where data mining algorithms learn and extract knowledge from the available data. Without data, mining would not be possible, making this stage foundational in the KDD process.

- Pre-processing

  In this step, data readiness is enhanced. The significance of this stage cannot be overstated, as it is one of the pivotal stages in the KDD process. This stage holds immense importance in extracting valuable insights and knowledge from large datasets. It involves data cleaning for anomalies such as managing missing values and eliminating noise and outliers. Data pre-processing is also called data cleaning because it removes information that is not necessary to a specific project. The Figure 2.3 illustrates how data pre-processing works.

Figure 2. 3: Data pre-processing Stages

UCI Cleveland Dataset is the selected dataset. The data contains irrelevant and missing values. Data cleaning involves eliminating erroneous, incomplete, and imprecise data, while also filling in missing values. In the process of cleaning, decision on how to handle missing values and outliers are implemented. Understanding the background and derived attributes, the data is then integrated into sources and the result is stored in new tables and records. The last step is formatting data, wherein purely syntactical modifications are made to meet the requirements for compatibility with the specific modelling tool.

- Transformation

  In this stage, the focus is on enhancing the quality of data utilized for data mining through the preparation and development of improved datasets. At this stage, the data undergoes a conversion process to match the specific format required by the mining algorithm. The process of transforming the data is divided into two main components: data mapping and code generation. The act of allocating items from a source base to a destination base to record transformations is known as data mapping, whereas code generation is the process of producing the transformation program itself. The process of selecting attributes included in the KDD project can have a significant impact on the project's success, and this step is typically dependent on the specifics of the project undertaken. For instance, in medical examinations, the ratio of attributes may be more critical than the individual attributes themselves. Therefore, it is essential to consider carefully the most relevant attributes included in the project to achieve the desired outcomes (Damtew, 2011).

- Data Mining

  Choosing the appropriate data modelling technique is critical at this stage and depends on the study's objectives. For instance, to develop a predictive model for detecting cardiovascular disease, a classification technique is used. Classification algorithms like Neural Network, K-Nearest Neighbor, Naive Bayes, and Decision Trees are considered for this purpose. To evaluate the performance of the selected classification methods, experiments were conducted on the entire dataset using Google Collaboratory software.

- Interpretation and Evaluation.

  Models were reviewed in this step to see how well they met the data mining goals established in the previous step. The process involves several stages, including comprehending the data, ascertaining the novelty and significance of new information, interpreting medical implications of the results, and assessing their relevance to the project objective. Evaluation metrics such as classification accuracy, area under the receiver operating characteristic (ROC) curve, and a confusion matrix table were utilized to assess the performance of the classification methods. The accepted models were preserved, and the entire knowledge discovery process for the disapproved models was updated to detect failures and misleading phases (Damtew, 2011).

2.4 Cardiovascular Disease

2.4.1 What is cardiovascular disease?

Cardiovascular disease (CVD) encompasses a wide range of conditions that affect the functioning of the heart. It is a major global health concern. Before discussing CVD, it is important to understand the anatomy and function of the heart. The heart is a muscular organ that is divided into four chambers, two on each side separated by valves. These chambers are called the atria and the ventricles. The atria receive blood and the ventricles contract to eject it from the heart. The heart's right portion circulates blood that lacks oxygen to the lungs, where the red blood cells take up oxygen. After oxygenation, the blood is transported back to the left atrium and ventricle, which expels it to nourish the body's tissues and organs. The oxygen in the blood provides energy and is essential for maintaining good health (Mosca, 2007).

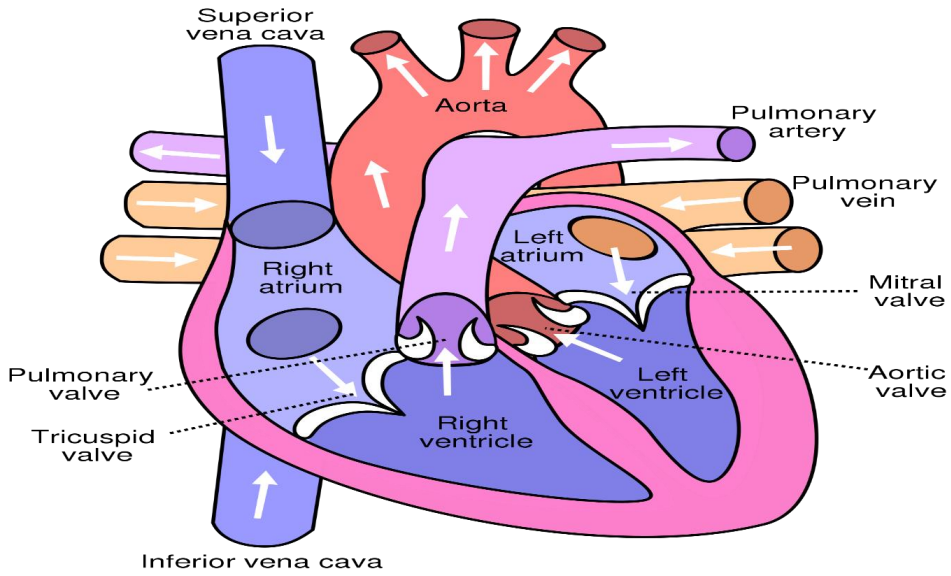


Figure 2. 4: The human heart (Wikimedia Commons, 2003)

Cardiovascular disease represents a diverse range of diseases, conditions, and disorders. The heart can be affected by various types malfunctions, a significant number of which are related to coronary artery disease. Atherosclerosis is a medical condition characterised by the formation of plaque on the inner walls of arteries. These arteries are responsible for carrying blood to the heart and various parts of the body. Plaque comprises several substances including calcium, cholesterol, and fat.

Heart disease can take various forms, but there exists a shared group of patent risk factors that determine if someone is likely to develop the condition. These risk factors comprise age, gender, high blood pressure, diabetes, elevated levels of cholesterol (hypercholesterolemia, hyperlipidaemia), obesity, and a sedentary way of life (Hajar, 2017).

2.4.2 Major Types of Cardiovascular Diseases

The researcher in this study has opted to focus on three commonly occurring types of heart disease. The following paragraphs delve into the details of these prevalent types. It should be noted that there exist numerous other varieties of heart disease beyond those covered here.

- **Coronary Heart Diseases**

In the realm of heart disease, this form is the most prevalent and is typically identified as the primary cause of heart attacks. Within the medical community, it is commonly referred to as ischemic heart disease. The term "coronary heart disease" describes damage to the heart that results from a reduction in blood flow, and in this case, the blood arteries that supply the heart muscles with blood narrow due to the accumulation of fatty deposits on their inner walls. A narrowing of the arteries that provide blood to the muscles of the heart occurs, resulting in a decrease in the amount of blood that can flow through them. This, in turn, causes a type of pain referred to as angina (Mozaffarian, 2015).

- **Congenital Heart Disease**

  The phrase "congenital" or "hereditary" heart disease refers to heart disease that runs in families. This type of heart disease is regarded as congenital since it is largely unavoidable. Any complication with the structure or operation of the heart resulting from improper heart development before birth is referred to as congenital heart disease. The most prevalent birth abnormality is congenital heart disease. Compared to other birth abnormalities, it is the cause of more infant fatalities in the first year of life (Damtew, 2011).

- **Congestive Heart Failure**

  When the heart is unable to effectively pump blood to the other organs in the body, it may result in a medical condition referred to as congestive heart failure. Heart problems and narrowed arteries frequently lead to congestive heart failure. A heart that has congestive heart failure functions far less effectively than it should and may cause other problems. Regular signs include oedema, shortness of breath, and kidney complications, which might result in a surprising weight increase. Congestive heart failure can be brought on by anything, even high blood pressure and alcohol addiction (Damtew, 2011).

2.5 Data Mining vs. Machine Learning

Patterson et al. (2017) opined that there is another crucial concept associated with DM which is machine learning (ML). Since ML reflects the methods, more precisely the algorithms utilised throughout the DM process to extract the structural description from the raw data, it differs from DM. Artificial intelligence (AI) involves machines imitating human-like intelligent behaviour and encompasses the subfield of machine learning (ML). AI systems are employed to execute intricate tasks in a manner akin to the approach humans take when tackling problem-solving.

Machine learning (ML) algorithms refer to the methods employed by data scientists to identify patterns in data, depending on the approach used for learning from the data, which can be supervised or unsupervised. The most common approach is supervised learning, where the training data includes labelled responses with known correct values of the target variable that needs to be predicted. Once the algorithm learns from the training dataset, it can make predictions on new data.

A machine learning system can have one of three functions: descriptive, predictive, or prescriptive. Descriptive machine learning uses data to explain what happened in the past, predictive machine learning uses data to forecast what could happen in the future, and prescriptive machine learning uses data to suggest what actions should be taken based on the predicted outcome (Brown, 2022). ML is categorised into two, Unsupervised ML and Supervised ML. In unsupervised ML, a program scans unlabelled data for patterns. Unsupervised ML discover hidden patterns or data groupings without human intervention (Brown 2022; Mosca 2007). Supervised ML uses labelled data sets to train models, which enables the models to improve their accuracy over time. Supervised ML is the most common type used. Figure 2.4 provides an overview of various data mining techniques and algorithms.

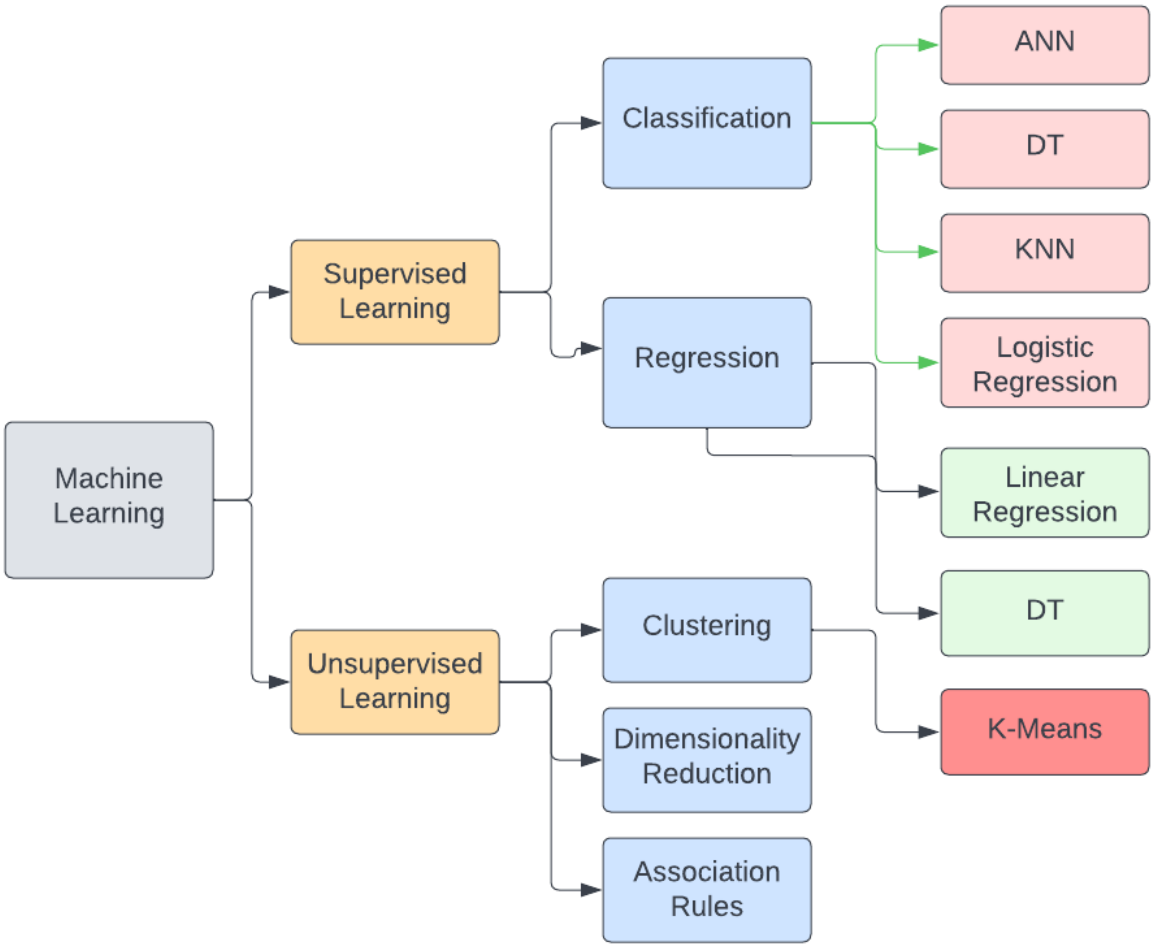


Figure 2. 5: Models and Algorithms used in Machine Learning

2.6 Data Mining Classification in Health Care

Medical decision making (DM) in healthcare is considered a crucial but complex task that requires precision. The healthcare industry utilises DM to solve practical health problems in the diagnosis and treatment of diseases. In healthcare, classification techniques are the preferred algorithms since they can forecast a patient's medical condition by categorising patient records and identifying the category that corresponds to a new patient record. This method is commonly used to categorize patient records in healthcare (Ahmad et al, 2015).

Chang et al. (2009) utilized an integrated decision tree model to diagnose skin conditions in adults and children. The primary objective of their study was to examine the results of five experiments centred around six common skin diseases. Their goal was to develop an accurate predictive model

for dermatology. By combining decision tree and neural network classification methods, they discovered that neural networks achieved a high predictive accuracy rate of 92.62% in diagnosing skin diseases.

Another study by Das et al. (2009) suggested an intelligent medical decision-making support system that used SAS software to diagnose heart diseases. The proposed system relied mainly on the neural network approach. They conducted experiments on data obtained from the Cleveland heart disease database and found that neural networks had an 89.01% accuracy rate in diagnosing heart disease.

Furthermore, Curiac et al. (2017) conducted an experiment to examine psychiatric patient data by utilizing Bayesian Belief Network (BBN) to determine the primary factors and correlations of psychiatric diseases. They gathered real data from Lugoj Municipal Hospital and found that BBN was a crucial component in the medical decision-making process for predicting the likelihood of psychiatric patients based on identified symptoms.

In a paper by Meena et al. (2019), a decision support system was used to support the association rules more than decision trees, even though many models are well-supported by the latter. The research paper concentrates on predicting anaemia in children and exploring the connection between a mother's health and diet during pregnancy and its influence on her child's anaemic condition. The study employs decision tree and association rule mining techniques to analyse the data and extract valuable insights.

Tayefi et al. (2017) established that the prevalence rate of hypertension in the population was 32%. They used two decision tree models, where model one (1) had an accuracy of 73%, a sensitivity of 63%, a specificity of 77%, and an area under the ROC of 0.72. On the other hand, model II had corresponding values of 70%, 61%, 74%, and 0.68, respectively. Amina et al., (2019) developed a heart disease prediction model utilising various data mining techniques, such as KNN, DT, NB, Logistic Regression, Support Vector Machine, Neural Network, and Vote. Among these techniques, the best performing one, called "Vote," achieved an accuracy rate of 87.4% in predicting heart disease. Significant features were identified and utilised in developing the model.

Chern et al. (2020) introduced a method called DTEVCA that updates classification rules utilising new data to rectify bias and ensure the consistency of an e-visit system. This approach has a broad

range of applications, including new services like products or policies, beyond e-visit services. The study concentrated on e-visit qualified clinic visits, utilising medical records and demographic data of both clinics and patients.

Chung et al. (2020), developed an algorithm that aims to optimize the integration of data by leveraging the unique advantages of each algorithm to improve the overall performance of the system. The study focused on developing an ambient context-based model for health risk assessment, and the proposed algorithm utilised a Deep Neural Network as the primary analytical tool.

2.7 Data mining algorithms and techniques modelling in heart diseases

Data mining encompasses a range of algorithms and techniques, including regression, clustering, classification, association rules, and genetic algorithms, that extract significant patterns for predicting heart diseases with high accuracy. Among these techniques, classification is the most employed approach, which utilizes a set of pre-classified examples to construct a model that can classify a vast number of records (Ahmed and Hannan, 2012). The process of data classification consists of two main stages: learning and classification. In the learning phase, the classification algorithm examines the training data to identify patterns and relationships. The algorithm learns from the provided data to create classification rules. In the classification phase, the algorithm applies these learnt rules to the test data to estimate the accuracy and effectiveness of the classification process (Ahmed and Hannan, 2012).

Palaniappan and Awan (2008) developed a method of categorizing an Intelligent Heart Disease Prediction System (IHDPS) through the application of data mining techniques such as Naïve Bayes and Neural Network. The medical data used in the study included patient's age, gender, blood pressure, and blood sugar levels, which enabled them to forecast the possibility of heart disease among patients. The IHDPS is a reliable system that is easy to use and accessible online.

Anbarasi, Anupriya, and Iyenga (2010) introduced an approach called Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm (GA) in 2010. The GA is a type of optimisation technique that mimics natural genetics and selection. The authors developed a system for predicting cardiovascular disease that uses a reduced number of attributes to accurately determine the presence of heart disease. They accomplished this by using feature-based

subset selection, where only six attributes were utilised instead of the original 13 attributes. The authors employed three classifiers, namely DT, Classification with clustering, and NB, to diagnose patients with heart disease, and the Weka data mining tool was used to analyse the data for the experiments. The findings of the experiments verified that DT had the highest accuracy and the shortest construction time compared to NB and Classification via clustering.

Chen et al. (2011) introduced the HDPS: Heart Disease Prediction System in 2011, which employed a single data mining technique, the artificial neural network (ANN), to classify heart disease based on 13 different attributes. The dataset used in the study was obtained from the UCI machine learning repository and comprised 303 instances.

Dangare and Apte utilized a data mining neural network approach for predicting heart disease with a reported accuracy of approximately 100% (Dangare & APte, 2012). The system employed a Multilayer Perceptron Neural Network (MLPNN) with backpropagation algorithm (BP). The MLPNN is one of the essential models in neural networks, consisting of multiple levels of interconnected nodes. The backpropagation algorithm is a commonly utilized method for determining the disparity between predicted and actual values. It functions by traversing from the output nodes to the preceding layer of nodes, facilitating computation of this difference. The research was conducted utilising the WEKA data mining tool, and the dataset used contained 573 records that were divided into two parts for training and testing. To enhance the accuracy of the prediction, a total of 15 attributes were employed.

In 2013, Soni suggested a heart disease prediction system that was intelligent and efficient, using Weighted Associative Classifiers (WAC) (Soni, 2011). The WAC assigned varying weights to the attributes, depending on their predictive potential. The system was executed using the JAVA platform, and the benchmark data was derived from the publicly accessible UCI repository. The dataset employed in the trials consisted of 303 entries and 14 diverse attributes, and the results confirmed an accuracy rate of approximately 80% for WAC.

In 2014, Chaurasia and Pal conducted experiments using various data mining approaches to predict heart diseases (Chaurasia & Pal, 2014). The research utilised the Weka tool and the findings were compared using 10 cross-validations both with and without bagging. Bagging employs Bootstrap aggregation to enhance the classification accuracy. Three techniques, namely Naïve Bayes, J48 Decision Tree, and Bagging, were utilised and compared in the study. The dataset was obtained

from the Hungarian Institute of Cardiology, and it had 76 original attributes, but only 11 were selected for the experiments. The trials verified that bagging had the highest accuracy of 85.03%, whereas J48 Decision Tree and Naïve Bayes had accuracy rates of 84.35% and 82.31%, respectively.

Vijayarani and Muthulakshmi (2013) carried out research on an Efficient Classification Tree Technique for Heart Disease Prediction. The objective of the study was to compare and evaluate the performance of different classification tree algorithms using a heart disease dataset. The algorithms tested in the study comprised DT, Random Forest, and LMT Tree algorithm. The investigation sought to analyse the effectiveness of these techniques in data mining for predicting heart disease.

In March 2014, Saravanakumar et al., utilised the Frequent Feature Selection approach to predict heart disease. This method employed a fuzzy measure and relevant nonlinear integral, which resulted in excellent performance. The team employed the WEKA data mining tool and used 1 000 records with eight (8) attributes that were assigned weights to predict the likelihood of heart attack. The mining process consisted of multiple stages, including the Weighted Support approach, which utilised the relative weight of each transaction in each record, and the Maximal Frequent Item set Algorithm (MAFIA), which combined novel and traditional algorithmic constructs to extract frequent item sets. At the conclusion of the process, the team computed the weight of significance for each pattern. This research explores the effectiveness of the Frequent Feature Selection method for predicting heart disease using these techniques.

Data mining techniques might be used to forecast cardiac disease (David & Belcy, 2018). They developed a prediction system by comparing three data mining categorisation algorithms: Random Forest, Decision Tree, and Nave Bayes. The goal was to examine and forecast the possibility of cardiac disease. The researchers utilized a heart disease benchmark dataset from the UCI machine learning library to evaluate the algorithms' performance. In predicting heart disease, the Random Forest algorithm achieved the highest precision of 81% when compared to the other methods.

Anitha and Sridevi (2019) proposed a comparison of three supervised machine learning algorithms, K-Nearest Neighbor, Naive Bayes, and SVM, using a heart disease dataset to determine whether a patient has heart disease, which would aid in diagnosis. In their study, they utilised R programming to implement the classification techniques. The findings of the research

indicated that the Naive Bayes algorithm had the highest accuracy in predicting the disease, at 86.6%. The SVM algorithm had a slightly lower accuracy of 77.7%, while KNN had the lowest accuracy of 76.67%.

Tarawneh et al. (2019) introduced a hybrid system for predicting heart disease that integrates multiple approaches into one algorithm. The researchers used data that had 303 instances and 14 features. They were able to reduce this to 12 features to compute it without loss in accuracy. The results showed that this composite model achieved a diagnostic accuracy of 89.2% using data mining techniques such as Naive Bayes, SVM, K-Nearest Neighbour, Neural Network, J4.8, and Random Forest. Among these techniques, Naive Bayes and SVM had the highest accuracy on the complete dataset. KNN performed the least in their system. They recommended the hybridisation approach because each individual technique t has its own capabilities that contribute to the prediction of heart disease and combining these techniques would result in the combination of all these capabilities across all features.

Bashir et al. (2019) explored the use of data science in predicting heart disease in the medical domain, focusing on feature selection techniques, and demonstrating improved accuracy in the various data mining techniques on heart disease datasets (Decision Tree, Logistic Regression, Logistic Regression SVM, Naive Bayes, and Random Forest). They found that logistic regression was the most effective feature selection tool for predicting heart disease in terms of level of precision.

Khan et al. (2019) proposed the Intelligent Heart Diseases Diagnosis Algorithm (IHDDA), which utilises heart signals, such as ECG graphs, to enable quick and precise diagnosis of heart diseases. The algorithm was developed to minimize the amount of time required for diagnosis while maintaining accuracy. The researchers tested the proposed algorithm on a supercomputer cluster using 300 patients with various heart conditions. The findings indicated that the IHDDA detected heart problems within 1.5 seconds at 97% accuracy rate. The algorithm required minimal physician effort and provided the highest diagnostic yield.

Krishman et al. (2019) proposed the application of two supervised data mining algorithms to estimate the probability of heart disease in patients. The researchers used the Naive Bayes Classifier and Decision Tree Classifier classification models to evaluate the dataset and determine which algorithm was the most accurate. The study results confirmed that the Decision Tree model

accurately predicts heart disease in patients 91% of the time, while the Naive Bayes classifier had an accuracy rate of 87%. The framework that has been created, in conjunction with the algorithm for machine learning classification, may have the potential to be expanded for the purpose of predicting or diagnosing additional diseases.

Subhadra, et al. (2019) presented a diagnostic system that utilizes the Multilayer Perceptron Neural Network for predicting heart disease. The researchers used 14 significant attributes, as suggested by medical literature, to diagnose heart disease. To improve prediction accuracy, they applied the backpropagation algorithm to train the data and iteratively compare the parameters. According to the findings, the diagnostic system demonstrated greater efficiency in forecasting the degree of risk associated with heart disease compared to alternative methods. The findings were tabulated, demonstrating the system's capability in predicting heart disease risk levels effectively. The decision-making system in this study, which utilised a multi-layered perceptron and an improved algorithm, was found highly effective. Through the partitioning of the training dataset into several subsets, the system attained an accuracy rate of 82.8% within a running time of 5.97 seconds. The prediction system developed in this study was even more accurate, achieving an accuracy rate of 93.39% by using 5 neurons in the hidden layer, and with a faster running time of only 3.86 seconds. The outcomes signify the effectiveness of the system and its capability to deliver precise forecasts promptly.

In a study by Alkhafaji et al. (2020), it was suggested that, in order to accurately predict heart disease and make appropriate treatment decisions for patients, it is necessary to collect data using three techniques with acceptable precision. The researchers conducted a review of classification requirements and used 19 traits related to cardiovascular diagnosis, ultimately narrowing the focus to the 10 most relevant traits. The findings indicated that the decision tree algorithm demonstrated the highest efficiency and accuracy, achieving 98.85% accuracy. The Bayesian Classifier also performed well, achieving 98.16% accuracy, while the neural network had the lowest accuracy at 91.31%.

In a study by Ramotra et al (2020) a model using the WEKA teaching tool was proposed to estimate heart diseases. The database used in the study contained 303 data records and 76 features. After re-processing the database and eliminating missing values, 297 data records with 13 input attributes were used for analysis. The research employed data mining strategies to detect patterns

and details in the dataset, and different classifiers such as Decision Tree, Naive Bayes, Support Vector Machines, and Artificial Neural Networks were employed to forecast outcomes using both WEKA and SPSS Modeler tools. The outcomes were evaluated using sensitivity, specificity, precision, and accuracy measurements, and it was confirmed that the Naive Bayes classifier produced the highest accuracy of 85.39% using WEKA, while the SVM classifier produced the highest accuracy of 85.87% using SPSS Modeler.

In a study by Vineet et al (2020) the goal was to achieve the best possible outcomes using neural networks. To accomplish this, several models were specified, and their performance metrics were gathered. Subsequently, the results of these models were compared against each other to identify the best-performing one. The effectiveness of deep neural networks (DNN) was evaluated by comparing it to other classifiers as part of the validation process. The study used several classifiers, including support vector machines (SVM), NB, KNN and deep neural networks (DNN) to evaluate the performance of these models, and the results were: SVM (86.2%), naive Bayes (83.97%), KNN (81.43%), and DNN (81.9%)

In a study conducted by Ayon et al. (2020), various computational intelligence techniques were evaluated for their ability to predict coronary artery heart disease. Specifically, seven techniques, namely Logistic Regression (LR), Support Vector Machine (SVM), Deep Neural Network (DNN), DT, NB, Random Forest (RF), and KNN were employed, and a comprehensive comparison was ultimately performed. The performance of each technique was assessed using the Statlog and Cleveland heart disease datasets, sourced from the UCI machine learning repository database, and a variety of other evaluation techniques. As per the research findings, the deep neural network attained the maximum precision of 98.15%, with corresponding sensitivity and precision values of 98.67% and 98.01%. The researchers acknowledged that these outcomes surpassed the results of earlier cutting-edge investigations regarding heart disease prognosis.

Premsmith et al. (2021) proposed a model that had a total 303 instances and 75 attributes to predict the presence of cardiovascular diseases by using data mining algorithms. The researchers used Python programming language with the Scikit-Learn library for scientific machine learning in their study. They used a confusion matrix table to evaluate their models, which included measures such as accuracy, precision, recall, and F-Measure. The results confirmed that the Logistic Regression

model performed better than the Neural Network model, with a precision of 95.45% and an accuracy of 91.65%.

Absar et al. (2022) conducted a subsequent study wherein machine learning algorithms were employed to rapidly identify heart disease at a reduced cost. The researchers utilised four different machine learning models, namely Random Forest, Decision Tree, AdaBoost, and K-nearest neighbour, to identify heart disease. They also devised a universal algorithm to scrutinize the elements that played a role in predicting heart disease. The accuracy of the models was calculated using several datasets from Kaggle, including Cleveland, Hungary, Switzerland, and Long Beach, with accuracy rates of 99.03%, 96.10%, 100%, and 100%, respectively. The study also incorporated Streamlit to design a computer-assisted intelligent system for heart disease prediction, which could potentially provide a practical tool for the diagnosis of cardiovascular disease. In conclusion, the study significantly contributed to an understanding of the predictive capabilities of various factors in determining the prognosis of heart disease.

In summary, a review of related research suggests that while various data mining techniques could aid medical professionals in predicting and diagnosing heart disease, more work is needed to improve their performance. Due to the serious nature of heart disease, relying on a single technique is not recommended, as accurate and timely diagnosis is crucial for proper diagnosis and subsequent treatment.

CHAPTER THREE: RESEARCH METHODOLOGY AND DESIGN

3.1 Chapter Overview

In this chapter, the research methodology and philosophy employed in this study are presented. The chapter follows on from the literature review presented in Chapter 2, which connected the research to antecedent studies and highlighted the existing gaps in knowledge. This chapter focuses on the methodology selected to address the research questions formulated in the initial chapter. A robust research methodology ensures that the research questions are effectively addressed, and the research objectives are achieved. Therefore, this chapter outlines the research methods and techniques that were utilised in this study to collect and analyse data.

3.2 Research Methodology

In an academic setting, research typically refers to rigorous and systematic inquiry or investigation of a particular area, with the aim of discovering or revising facts, theories, applications, and even methodological approaches in the field. The goal of research is to generate and disseminate new knowledge, typically through publications in academic journals or conference proceedings.

In Computer Science and Information Systems, there are various research methods that can be employed, such as experimental research, observational research, case studies, surveys, and simulations. Each of these has its own strengths and weaknesses, and researchers must carefully choose the most appropriate approach based on the research question, available resources, and other factors.

In this research, constructivist research methodology is used. According to Oyegoke (2011), constructivist research is a problem-solving method that enhances the performance of a system, thereby contributing to the existing knowledge. This approach involves creating a practical or theoretical artifact that addresses a specific problem in a particular field, resulting in new knowledge about how to solve or understand the problem. In essence, the constructivist research methodology involves building an artefact that solves a problem and generates knowledge about how the problem could be explained. The phases of constructivist research include:

(a) Finding a practically relevant problem.

(b) Obtaining an understanding of the topic and the problem.

(c) Constructing a solution idea.

(d) Demonstrating that the solution works.

(e) Establishing theoretical connections and research contribution.

(f) Examining the scope of applicability

It is crucial to note that the steps involved in constructivist research are typically not carried out in a sequential manner. Although there is a typical order in which they are executed, it is not uncommon for the process to be iterative or for steps to be revisited. In fact, the process can be recursive, meaning that researchers go back and forth between steps as they generate more information and refine their understanding of the problem. Therefore, the process of constructivist research may not always follow a linear, straightforward sequence.

To identify a problem using constructivist research, there are various methods such as conducting a literature review, seeking input from colleagues or relying on a researcher's previous experience. The problem chosen for investigation should be practically relevant. In this research, the problem is to propose a cardiovascular prediction system that assists medical practitioners to detect cardio related diseases efficiently, based on data not intuition.

To gain a comprehensive understanding of the problem, the researchers engaged in an extensive review of relevant literature. Based on this review, they identified the problem and then formulated a solution idea. This solution is designed to effectively address the identified problem and is supported by a clear theoretical framework. To achieve this goal, the researchers utilize data mining techniques and develop a tool that serves as a practical implementation of their proposed solution.

In signifying that the solution works, a validation, or testing exercise is performed. In this research data is deployed, trained, and tested for prediction as a way of validating the algorithms used.

3.3 Research Design and Methods

3.3.1 Design

This section illustrates the framework design. A flow chart is shown in Figure 3.1, explaining the stages involved in entire network design. Initially the data is gathered by a download from UCI website. The downloaded data is then pre-processed to eliminate errors and missing values and discretized data. The network is then built, trained, and tested. The flow diagram in Figure 3.1 shows the framework design and the Figure 3.3 shows the classification algorithm flow chart.

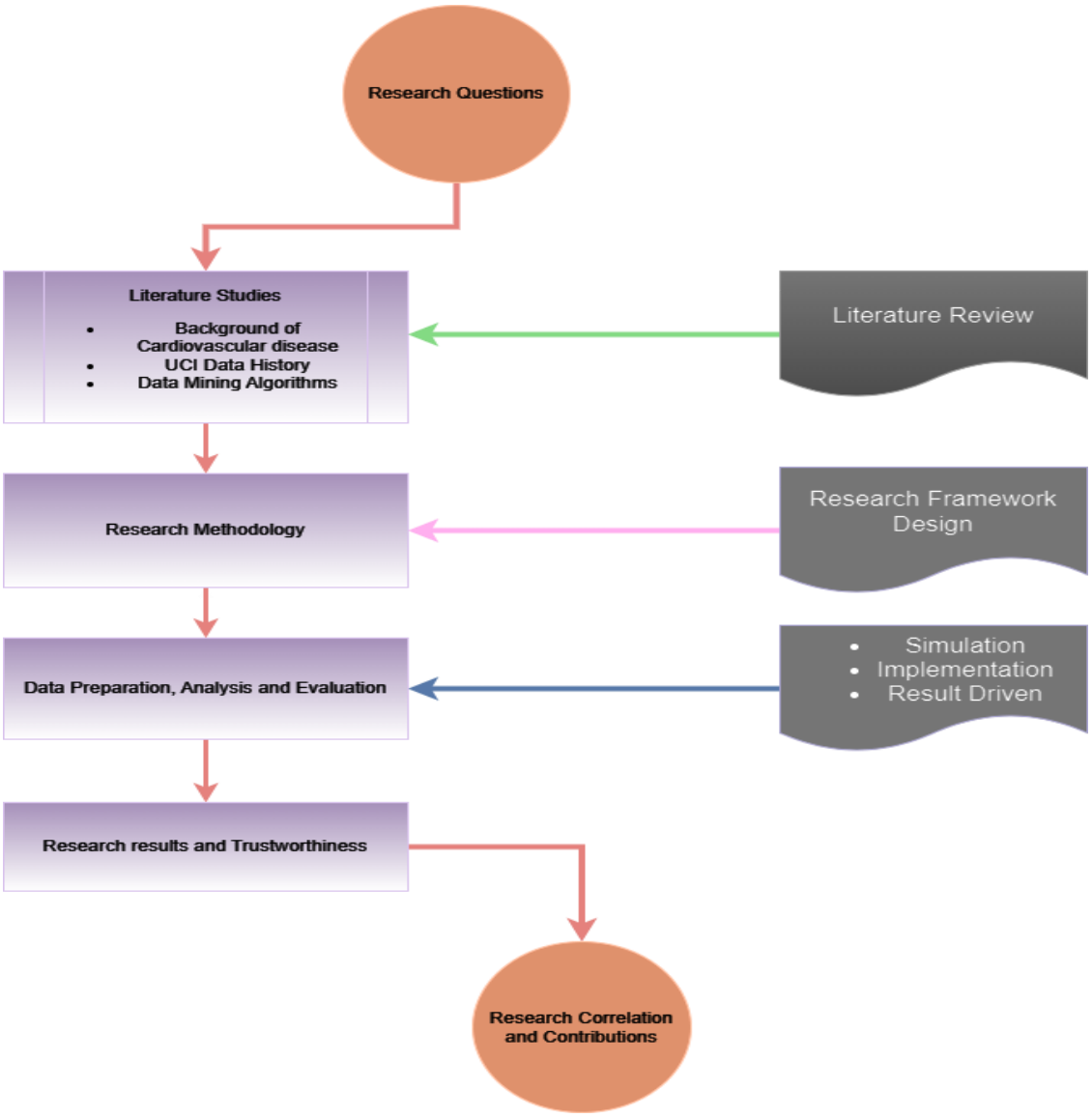


Figure 3. 1: Research Framework Design

The process of constructing a cardiovascular disease prediction model involves utilising DT, Neural Network, K-NN, and Bayesian Classifier techniques. Once the models are designed, their individual performances are assessed and compared. This section provides a thorough discussion of the algorithms employed to construct the models and the matrices employed for evaluating their effectiveness within the heart disease detection system.

3.3.2 Methods

Qualitative and quantitative approaches are methods of analysis employed to achieve solution in the research. The terms qualitative and quantitative are typically considered opposite in meaning, but in this study, they are seen as a spectrum, with varying degrees of overlap, or as distinct concepts with well-defined boundaries. Moriarty (2011) suggests that these often-contrasting perspectives can both be useful in understanding the relationship between qualitative and quantitative approaches. Qualitative is used for literature review and the framework design. Quantitative is then subsequently used for experiment with the dataset collected using ML algorithms.

3.3.2.1 Qualitative Method

Qualitative research investigates phenomena by focusing on understanding their intrinsic nature, diverse manifestations, and the contextual or subjective perspectives through which they could be interpreted. However, it does not encompass the examination of their range, frequency, or cause-and-effect relationships (Busetto, Wick, & Gumbinger, 2020). Qualitative research takes on different methods such focus groups, record keeping, and case study research.

Cardiovascular diseases are better understood by deep delving into and thoroughly understanding the research topic through consulting previous publications and findings. Case study research is used for explaining the field of study and its entity. Despite its frequent application in discrete fields, qualitative research remains relatively underrepresented in the realm of healthcare research.

3.3.2.2 Quantitative Method

Quantitative research involves the collection and analysis of numerical data. It primarily uses data gathered through qualitative methods to make observations and understand phenomena. The notable advantage of employing quantitative research in healthcare lies in its capability to

transform data into numerical form, including the capacity for replication of studies to confirm and validate findings (Everest, 2014).

However, quantitative data only provides a limited understanding of research questions and objectives and may not fully clarify the complexity of cardiovascular disease (Mahoney & Goertz, 2006). This research is based on experiments and aims to establish and confirm deeper insights through secondary data analysis.

3.4 Data collection and analysis

3.4.1 Data collection

Patient data, comprising personal details, medical history, and laboratory test results, is typically generated and stored during medical decision-making. Its purpose is to enhance decision-making regarding additional testing or treatment options and to provide future access when required.

3.4.1.1 Data description

The dataset used for this study is open and publicly available database from UCI (University of California, Irvine, CA) Machine Learning Repository and is available via http://archive.ics.uci.edu/ml/datasets/Heart+Disease (Aha, 1988). The datasets utilized in this study share a uniform instance format and characteristics. They consist of 76 initial attributes, with only 14 of them deemed significant (Aha,1988). The Cleveland Clinic Foundation dataset includes 303 patient records, while the Hungarian Institute of Cardiology dataset includes 294 patient records. Both datasets were combined, resulting in a total of 597 instances in the dataset analysed in this research.

3.4.1.2 Data Features

Table 3. 1 Selected Cleveland Heart Disease Dataset Attributes

S/N	Attribute	Description	Values
1	Age	Age in years	Continuous
2	Gender	Male or Female	1=Male, 0=Female
3	Cp	Chest Pain Type	1=Typical angina 2=Atypical angina 3=Non-anginal pain 4= Asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg)	Continuous
5	Chole	Serum Cholesterol (in mg/dl)	Continuous
6	FBS	Fasting Blood Sugar	1 >= 120 mg/dl 0 <= 120 mg/dl
7	Restecg	Resting electrocardiographic results	0=Normal 1=Having ST_T wave abnormality 2=Left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous
9	Exang	Exercise induced angina	1 = Yes 0 = No
10	Old peak	ST depression induced by exercise relative to rest	Continuous
11	Slope	The slope of the peak exercise segment	1 = Up sloping 2 = Flat 3 = Down sloping

12	Ca	Number of major vessels coloured by fluoroscopy	(1 – 4)
13	Thal	Thallium scan	3= Normal 6 = Fixed defect 7 = Reversible defect

3.4.2 Data Analysis

This project adopts predictive analysis to analyse the results. Predictive data analysis can be used to enhance the results for predicting cardiovascular by incorporating various techniques such as machine learning, statistical modelling, and data mining. Machine learning algorithms like Decision Tree, K-Nearest Neighbor, Naïve Bayes, and Artificial Neural Network can be trained on historical data to identify patterns and predict the likelihood of a patient developing cardiovascular complications. Statistical modelling can also be used to analyse risk factors such as age, gender, and lifestyle habits, and to identify potential interactions between these factors. By combining these techniques, predictive data analysis could provide more accurate and reliable predictions of heart disease, helping medical practitioners to make informed decisions about patient care.

To analyse the relationships between features and the target variable, a correlation matrix with heatmap and boxplot outliers was implemented. The correlation matrix offers a visual depiction of the extent to which features are interconnected with one another or with the target value. The heatmap within the matrix emphasizes the features that exhibit the strongest correlations. It should be noted that correlation can be either positive or negative. Additionally, the outlier detection model was employed to identify any anomalous data entries within the dataset. This approach enables the identification of outliers, both individually and in relation to other numerical and categorical data points.

3.5 Research Tools and Technologies

This research study employs a variety of tools, all of which are free and open source. These tools include.

1. Google Collaboratory

2. Jupyter Notebook
3. Python
4. Excel
5. Pgmpy
6. Pandas
7. Numpy
8. Matplotlib
9. Seaborn

Google Collaboratory, also known as "Colab," is a tool created by Google Research which enables users to run any Python code via a web browser. It is specifically designed for data analysis and education. On the other hand, Jupyter Notebook is a freely available web-based program that enables interactive data science and scientific computing.

Python, a widely used and versatile programming language, was selected for this project due to its open-source nature and extensive documentation and support community. Excel, a software product of Microsoft, was employed for its cost-effectiveness and widespread availability on various computer platforms and architectures.

Pgmpy is a Python module designed to facilitate working with probabilistic graphical models (PGMs). Its capabilities include the ability to learn the structure of a PGM from data, as well as perform inference and parameter learning tasks.

Pandas is a popular open-source Python library utilized for data manipulation and analysis. The primary components of this approach encompass data structures and tools specifically designed for the management and manipulation of numerical tables and time series data.

Numpy is an essential module for scientific computation in Python that offers robust capabilities for managing extensive, multi-dimensional arrays and matrices of numerical data. Additionally, it features an extensive library of mathematical functions that can operate on these arrays.

Matplotlib is a Python library used for creating visualizations and charts, and it is built on top of NumPy, a numerical mathematics extension for Python. It offers an object-oriented Application Programming Interface (API) for integrating plots into applications that use general-purpose Graphical User Interface (GUI) toolkits like Tkinter, wxPython, Qt, or GTK.

Seaborn is a Python library that enhances data visualization capabilities and is built as an extension to matplotlib. It offers a user-friendly interface for generating visually appealing and informative statistical graphics. Seaborn is particularly useful for creating visualizations of large datasets and statistical models.

3.6 Selected machine learning

There are different techniques in data mining, but the techniques commonly used are classification, prediction, clustering, and association rule (Ramageri, 2010). In this research, classification technique are adopted; namely four classification algorithms are deployed to predict and diagnose cardio-vascular disease, the performance of these algorithms are subsequently compared and analysed. The considered algorithms are NB, DT, K-NN, and ANN.

3.7 Performance Evaluation Measures

Evaluation metrics are employed to quantify the efficacy of data mining algorithms on a specific dataset. These metrics provide quantitative measures of the algorithms' performance and allow for a systematic assessment of their efficacy.

3.7.1 Confusion Matrix

The confusion matrix serves as a mechanism to assess the performance of a classification system by providing comprehensive details regarding the real and predicted classifications. One important metric in the confusion matrix is TP, which represents the number of true positive instances - these are cases where heart disease is accurately classified as heart disease by the algorithm. A false negative (FN) occurs when the algorithm incorrectly predicts that a patient does not have heart disease, even though they actually do. A true negative (TN) occurs when the algorithm correctly predicts that a patient does not have heart disease. In this case, the algorithm's classification is correct. On the other hand, a false positive (FP) occurs when the algorithm mistakenly predicts that a healthy patient has heart disease (Damtew, 2011).

To evaluate the efficiency of this proposed system, precision, recall, and accuracy assessments were employed, each involving the utilisation of TP, TN, FN, and FP. The computations for these metrics are outlined below, following the recommendations of Damtew (2011).

Table 3. 2 Confusion Matrix

ACTUAL VALUE		**PREDICTED VALUE**	
		+(1)	-(0)
+(1)		True Positive	False Negative
-(0)		False Positive	True Negative

3.7.2 Accuracy

Accuracy is characterised as the ratio of correctly classified instances to the total number of instances. It is calculated by adding the number of true positive (TP) and true negative (TN) instances and dividing the result by the total number of instances.

$$\text{Accuracy (Acc)} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%.$$

3.7.3 Precision

Precision is defined as the proportion of the true positive instances which are classified as positive. It measures the accuracy of positive predictions, indicating how closely predicted values align with actual values.

$$\text{Precision} = \frac{TP}{TP+FP}$$

3.7.4 Recall

Recall, in the context of classification, is defined as the proportion of positive instances that are correctly identified as positive by the classifier.

Recall is also often called sensitivity.

$$\text{Recall} = \frac{TP}{TP+FN}$$

3.7.5 F-Measure

F-measure is conveys the balance between recall and precision.

$$\text{F-Measure} = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.8 Receiver Operating Characteristics (ROC) Area

A Receiver Operating Characteristics (ROC) curve is a method used to assess and compare the performance of classifiers. It involves plotting the true positive rate (TPR) against the false positive rate (FPR) for various threshold settings. TPR represents the proportion of positive instances correctly identified by the classifier, whereas FPR represents the proportion of negative instances inaccurately identified as positive by the classifier (Olson, 2008).

The ROC curve allows for the visualisation, organisation, and selection of classifiers based on their performance. The plot depicts TPR on the y-axis and FPR on the x-axis, and an optimal classifier exhibits a curve that closely aligns with the upper-left corner of the chart, indicating high TPR and low FPR. The area under the curve (AUC) is also used as a metric for comparing classifiers, with a higher AUC indicating better performance (Olson, 2008).

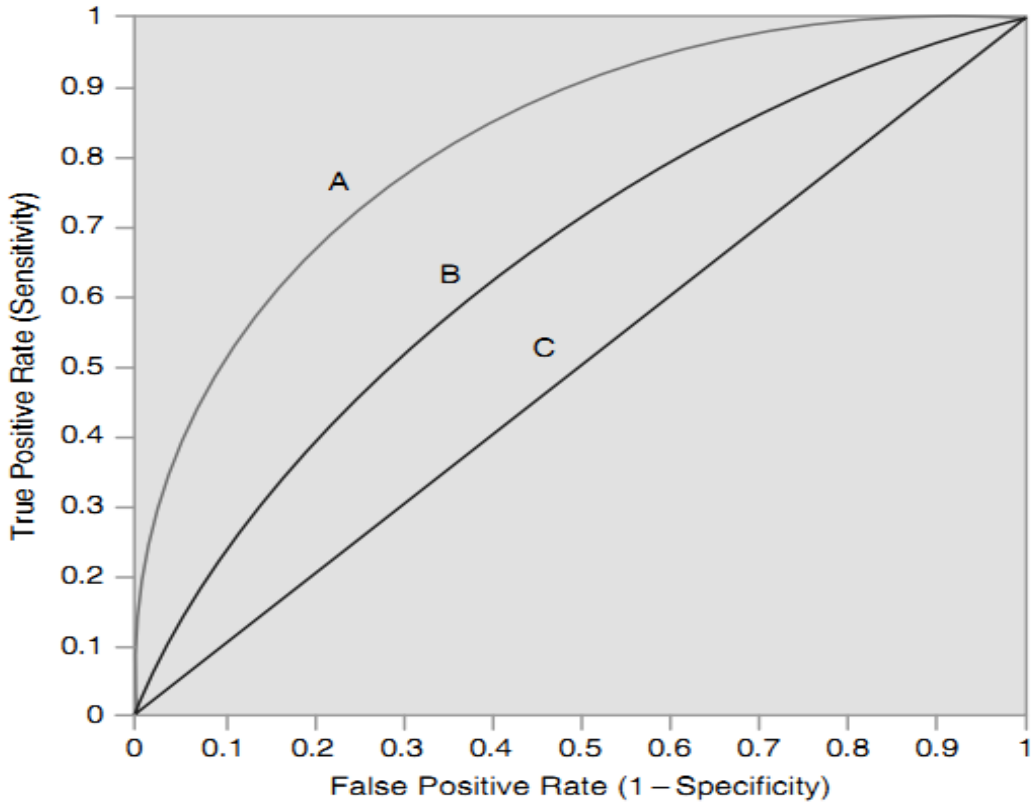


Figure 3. 2: A sample ROC (Damtew, 2011)

3.9 Research Process

Fayyad (1996) proposed a nine-step life cycle for data mining projects. It is important to note that the steps involved in this methodology are iterative and interactive, with users required to make several decisions along the way. Additionally, the process is iterative at each stage, meaning that it may be necessary to revisit previous steps. The methodology commences by identifying the goals of KDD and concludes with the implementation of the acquired knowledge.

The rationale for selecting the KDD methodology for this study can be attributed to three key reasons. Firstly, it is regarded as the most appropriate methodology for academic research. Secondly, using the KDD methodology reduces the level of expertise required for knowledge discovery. Thirdly, the KDD methodology is tool-agnostic, allowing researchers to employ any desired technique throughout the study. As a result, this data mining project was structured on the following nine steps, in line with the KDD methodology (Fayyad, 1996).

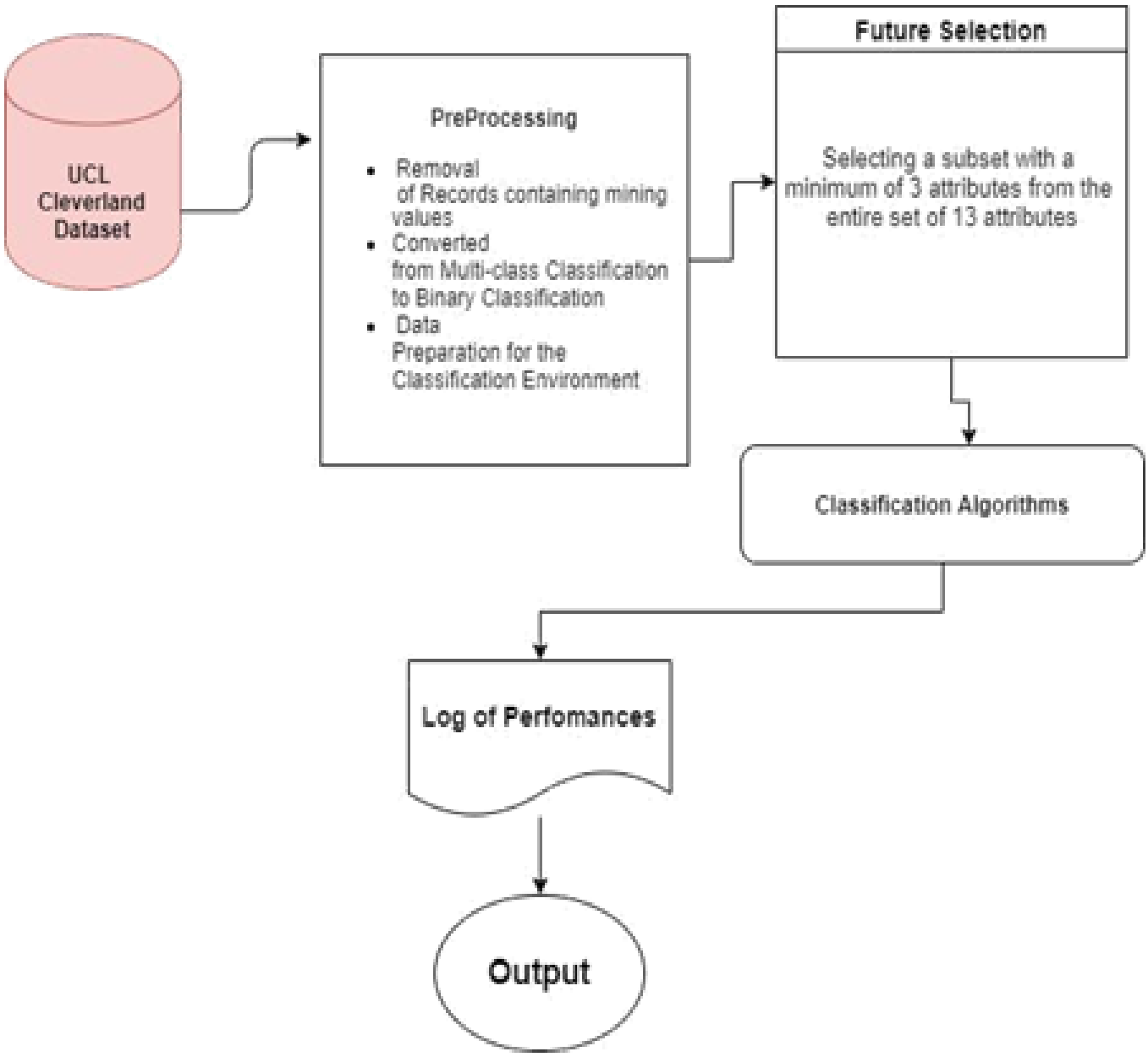


Figure 3. 3: Research Process

3.9.1 Selecting and Creating the Target Dataset

The initial step in KDD methodology is designed to identify and select the dataset used for this experimental project. The UCI heart disease dataset was selected and publicly available.

3.9.2 Pre-processing and Cleaning

The data that was chosen underwent an examination to identify any noise, inconsistency, or missing values using distribution frequency. Additionally, box plots were utilized to detect

outliers. Any noises or anomalies found within the dataset were rectified through manual means, while missing values were substituted with the most probable value calculated using regression. Similarly, outliers were replaced with the average value of the attributes. All the cleaning of data was carried out after resolving the problems and specifications for the pre-processing phase. The entire process and methods used during the pre-processing step are elaborated further in Chapter 4.

3.9.3 Data Transformation

Several modifications were made to the dataset to increase its compatibility with data mining algorithms. To limit the number of features that the classification algorithm had to evaluate and to minimize errors arising from irrelevant attributes, attribute selection was necessary. An attribute selection method was employed to select the most appropriate traits by assigning scores to each element based on its relevance to the problem. Attributes with high scores were retained while those with low scores were eliminated, resulting in a reduced dataset with only the most pertinent attributes. This enabled the classification algorithm to focus solely on the significant factors in making accurate predictions or classifications.

3.9.4 Choosing the Appropriate Data Mining Task

The main objective in this stage was to determine the appropriate data modelling method. Choosing a suitable modelling approach is reliant on the research objective, and since the aim of this study is to identify heart disease through data mining methods, a classification approach was employed to construct predictive models.

3.10 Choosing the Data Mining Algorithm

When selecting the most suitable algorithms, various factors were considered, including the algorithms' performance, the data mining objectives established during the initial phase, the structure of the available data, and the algorithm's applicability in the medical domain.

Following the creation of the models, their performance was assessed and compared against one another. This section delves into the algorithms employed for building the models, as well as the matrices utilised for gauging and contrasting their effectiveness, in the development of the heart disease detection system.

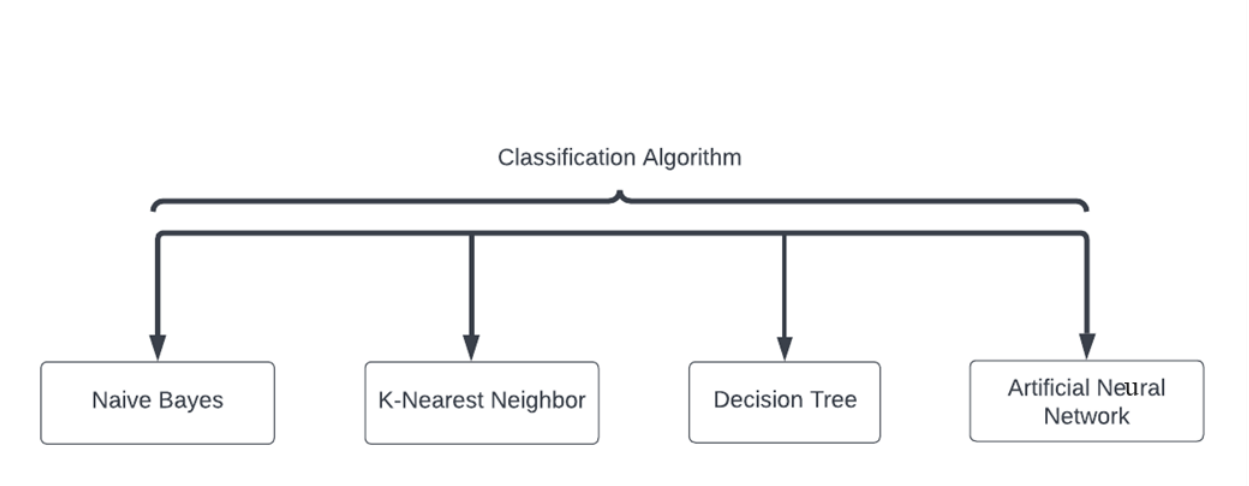


Figure 3. 4: Classification Algorithms

3.10.1 Naïve Bayes

Naive Bayes is a data mining classification technique that can be used to diagnose heart disease in patients (Fayyad et al. 1996; Khatibi & Montazer, 2010). This approach relies on probability theory and seeks to identify the classifications that are most probable (Razali, 2009; Kim, 2004). Although some research shows that Naive Bayes performs similarly to decision tree and neural network classifiers in specific domains, it is essential to bear in mind that in practical applications, Naive Bayes may not always achieve the theoretical minimum error rate. This is because its assumptions, such as class conditional independence and the absence of probability data, can lead to inaccuracies (Han, 2006). As outlined by Han and Kamber (2006), the Naive Bayesian classifier, also known as the simple Bayesian classifier, operates as follows:

We start with a training dataset D consisting of tuples and their corresponding class labels. An n-dimensional attribute vector, X = (x1, x2, . . . , xn), is used to represent each tuple, where each attribute, A1, A2, . . . , A*n*, represents *n* measurements made on the tuple.

Suppose there are m classes, C1, C2, . . . , C*m*. When given a tuple X, the classifier predicts that X belongs to the class with the highest posterior probability, which is conditioned on X. The Naïve Bayesian classifier predicts that tuple X belongs to class C*i* if and only if

$$P(Ci|X) > P(Cj|X)\ for\ 1 \leq j \leq m, j \neq i$$

Thus, we maximize $P(Ci|X)$ where the class Ci with the maximum posteriori probability is selected. Bayes' theorem is used to compute $P(Ci|X)$ and we estimate the class prior probabilities by

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

As P(X) is constant for all classes, only $P(X|Ci)\ P(Ci)$ needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C1) = P(C2) = \ldots = P(Cm)$, and we would therefore maximize $P(X \mid Ci)$. Otherwise, we maximize $P(P|Ci)\ P(Ci)$. Note that the class prior probabilities may be estimated by $P(Ci) = |Ci, D|/|D|$ is the number of training tuples of class Ci in D.

If we have datasets with many attributes, it would be computationally expensive to compute $P(X|Ci)$. To simplify the calculation of $P(X|Ci)$, the naive assumption of class conditional independence is adopted. This assumption presumes that the values of the attributes are conditionally independent of each other, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X|Ci) = \prod_{K=1}^{n} P(Xk|Ci$$

$$P(X|Ci) = P(X1|Ci) * P(X2|Ci) * ..* P(Xn|Ci)$$

The probability can be easily estimated by $P(X1|Ci), P(X2|Ci), \ldots, P(Xn|Ci)$ from the training tuples. Xk refers to the value of attribute Ak for tuple X. For each attribute, we look at whether the attribute is categorical or continuous. For instance, to compute $P(X|Ci)$, we consider the following:

If Ak is categorical, then $P(Xk|Ci)$ is the number of tuples of class Ci in D having the value xk for Ak, divided by $|Ci, D|$, the number of tuples of class Ci in D.

If Ak is continuous, then we need to do more work, but the calculation is straightforward. A continuous-value attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

So that

$P(Xk|Ci) = g(Xk, \mu ci, \sigma ci)$

To compute μci and i , which are the mean (i.e., average) and standard deviation, respectively, of the values of attribute Ak for training tuples of class Ci, we then plug these two quantities into the equation stated at section B, together with xk, to estimate $P(Xk\,|Ci\,)$.

To predict the class label of $X, P(XjCi)P(Ci)$ is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

$$P(X\,|Ci\,)P\,(Ci\,) > P(X\,|Cj\,)P\,(Cj\,) for\ 1 \leq j \leq m, j \neq i$$

In other words, the predicted class label is the class C*i* for which $P(X|Ci)P(Ci)$ is the maximum.

3.10.2 K- Nearest Neighbor

The k-nearest neighbor algorithm (k-NN) is a non-parametric technique utilised for identifying patterns in data, specifically for classification and regression. This method involves finding the group of training data that is most similar to the test data. The algorithm operates by determining the shortest distance between the new test object and the existing training data. The result of the k-NN algorithm is dependent on whether it is used for classification or regression. Both cases involve using the nearest examples of training data in the feature space (Muhammad, 2023). The k-NN algorithm is often referred to as a memory-based method and is known for its fast convergence speed and simplicity. This algorithm is preferred over other classification techniques. Figure 3.5 illustrates the process of k-nearest neighbor classification.

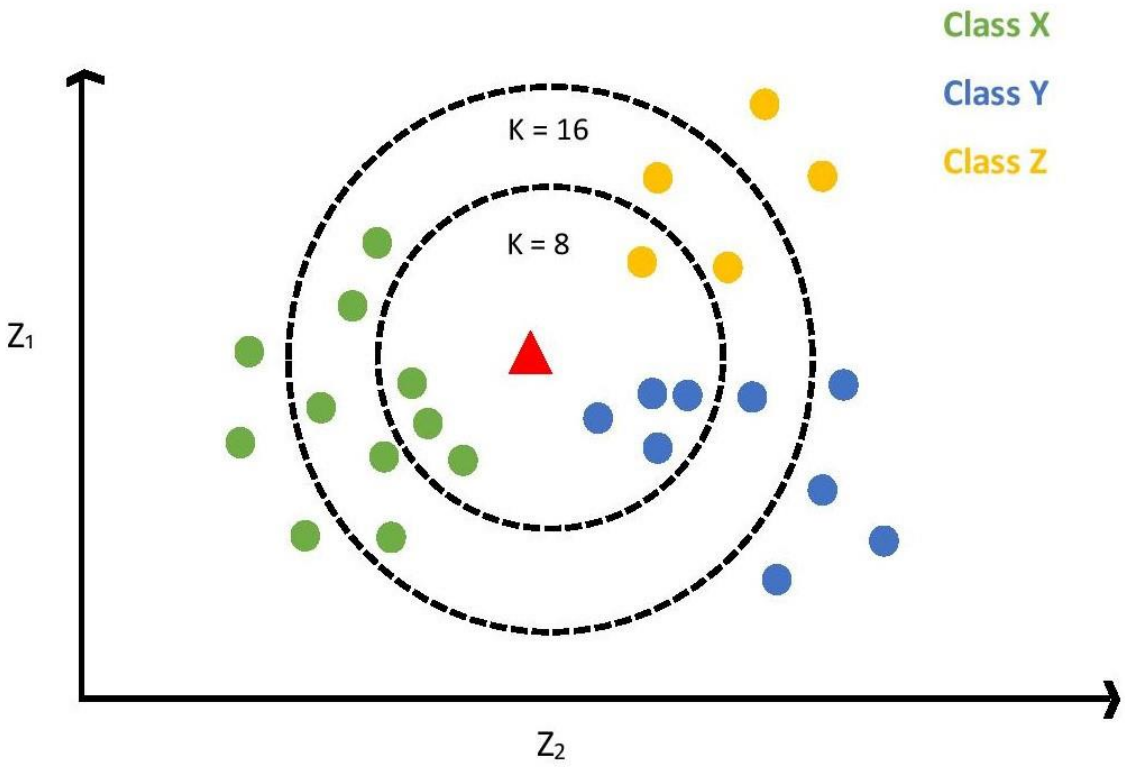


Figure 3. 5: KNN Classification Algorithm

If ***a*** is the first sample (training data) denoted by (*a1*, *a2*,. . . , *an*), and **b** is the second sample (testing data) denoted by (*b1*, *b2*,. . . , *bn*), the distance between them is calculated by relation in the Euclidean equation below.

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 \ldots\ldots\ldots + (a_n - b_n)^2}$$

One significant challenge associated with using the Euclidean distance formula is that it tends to give smaller weights to high-frequency terms. To solve this problem, continuous attributes are standardised in such a way that they have the equal influence on distance measurement between instances (Liao, 2002).

To calculate the K-Nearest Neighbor Algorithm, the following steps are followed:

a. Determine the number of nearest neighbors to consider, denoted by **K**.
b. Compute the Euclidean distance between the query instance and each object in the sample data provided, by squaring the difference in each dimension and summing them up.
c. Group the objects based on their Euclidean distance in ascending order.
d. Select the **Y** categories that are the nearest neighbors (referred to as Nearest Neighbor Classification).
e. Predict the value of the query instance based on the category that appears most frequently among the nearest neighbors.

The K-NN algorithm is simple to implement because it does not need the creation of a model or any other assumptions. This algorithm is adaptable and used for classification, regression, and search. Despite being the most straightforward algorithm, its accuracy is hampered by noise and irrelevant features (Shah, 2020).

3.10.3 Decision Tree

A decision tree (DT) can be utilized to separate a large collection of data into smaller sets by using a series of simple decision rules, as stated by Berry and Linoff (2004). The groups formed by this process become increasingly similar to each other with each subsequent division. There are several decision tree algorithms available, with J48 being the most commonly used. J48 utilizes a pruning technique to create an efficient decision tree, which aims to eliminate irrelevant data and improve prediction by removing overfitting.

The J48 algorithm continuously categorises data until the divisions are as accurate as possible. High-dimensional data can be managed through DTs, and their tree-like representation of learned information makes it easy to understand. The process of learning and classification in decision tree induction is straightforward and efficient. DT classifiers often have acceptable accuracy for linearly separable issues (Hand et al., 2001). The goal is to construct a tree that balances flexibility with accuracy (Dangare & Apte, 2012).

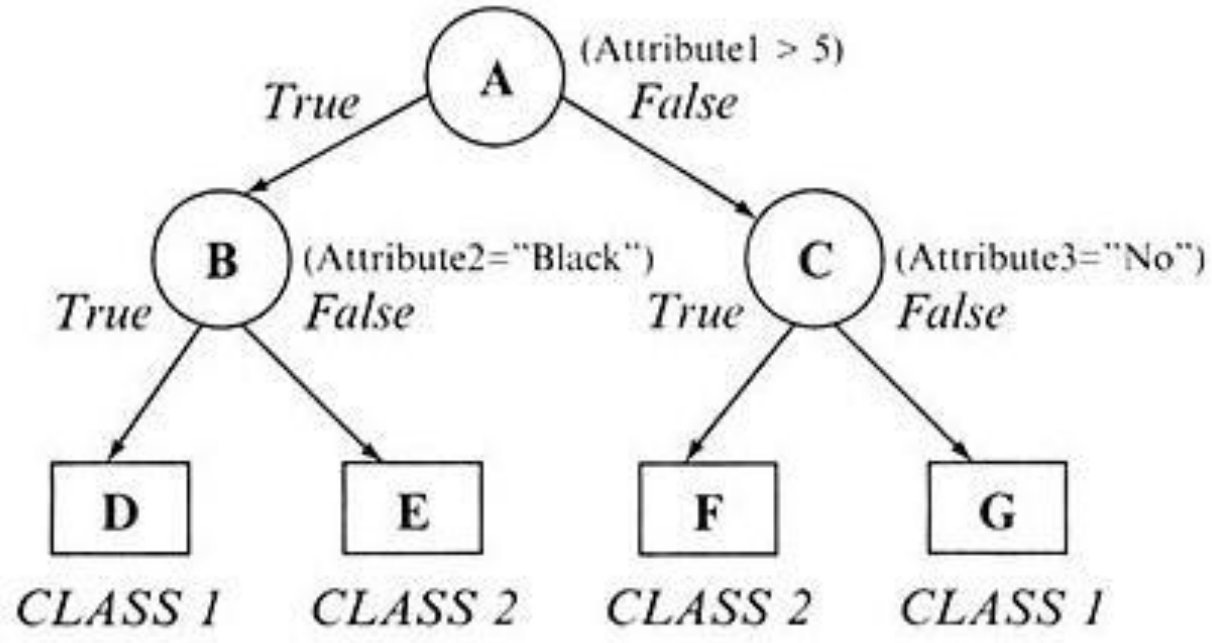


Figure 3. 6: A Simple Decision Tree (Merghadi, 2020)

Many application areas have adopted the DT algorithm for classification such as medicine, manufacturing and production and astronomy. The representative of a DT is depicted in the Figure 3.6.

To construct a decision tree (DT), a training dataset is necessary as it allows the algorithm to identify the relevant variables that may have predictive potential in the observations. The training dataset includes observations that represent the same variables of interest as those in the validation set and is used by the algorithm to establish decision rules that are then incorporated into the DT. The algorithm aims to generalise and identify patterns by selecting the best query that splits the instances into distinct classes, giving rise to the hierarchical structure of the DT (Merghadi, 2020).

Steps to calculate the Decision Tree Algorithm method:

a. Collect and prepare the data set for analysis.
b. Select the attribute (or feature) used as the root node of the tree.
c. Divide the data set into subsets based on the selected attribute.
d. Repeat steps (b) and (c) for each subset, using the remaining attributes until all the data is classified or a stopping criterion is reached.

e. Assign a class label to each leaf node based on the majority class of the instances reaching that leaf.
f. Prune the tree to remove unnecessary branches and improve the accuracy of the model.
g. Test the accuracy of the decision tree using a separate data set or by computing the prediction error rate.

The interpretation of decision trees is relatively straightforward, as the path from the root node to a leaf node clearly illustrates the decision-making process. However, this simplicity is contingent upon the tree being relatively shallow. As the DT's depth increases, it becomes progressively challenging to comprehend the underlying decision rules.

3.10.4 Artificial Neural Networks

The architecture of Artificial Neural Networks (ANN) involves the arrangement of neurons and the connections between them, referred to as weights. The Multilayer Perceptron Neural Network (MLPNN) with Backpropagation (BP) algorithm is a commonly used technique in developing ANN systems. Typically, an ANN architecture comprises three layers, with the first layer serving as the input layer, receiving data or input from external sources. There are various types of ANN architectures, and the input layer is responsible for receiving signals from external nodes.

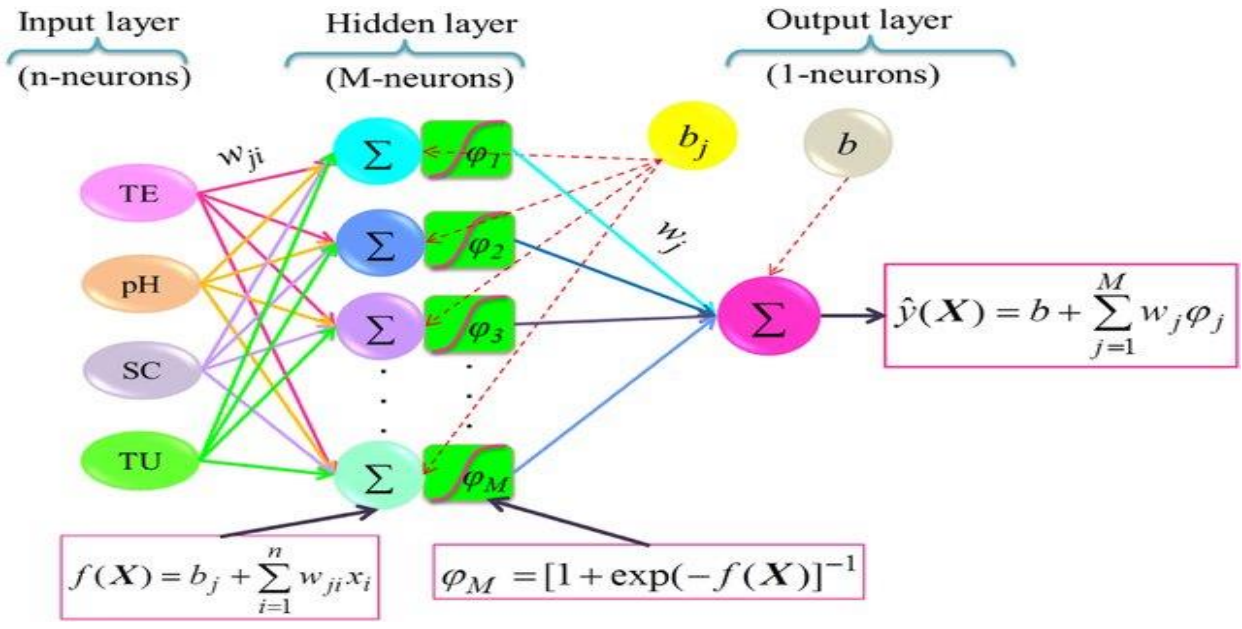


Figure 3. 7: Schematic view of MLPNN (Keshtegar and Heddam, 2019)

The next layer is a hidden layer with neurons that are more capable of extracting data or electrical signals than the previous input layer. One or more neurons may be present in the hidden layer, where the data or electrical signal is processed utilising the functions like arithmetic and mathematics that are available. Data processing results from the hidden layer are then directed to the output layer, where they play a part in assessing the validity of data that are evaluated based on the presence of limits in the activation function. Weighted links are utilised to transmit results after computations are completed. The output layer receives information from the hidden layer, which is then processed and produces a final output.

Multilayer Perceptron Neural Network is summarised and carried out in the following steps.

1. The input layer receives input data for processing, resulting in a predicted output.
2. The predicted output is compared to the actual output, and the difference is calculated as an error value.
3. The network applies the Backpropagation algorithm to adjust its weights based on the error value.
4. Weight adjustment starts from the output layer nodes to the last hidden layer nodes and then works backwards through the network.
5. After the Backpropagation process is complete, the network resumes the forwarding process.
6. This cycle is repeated until the error between the predicted and actual output is minimized.

The fact that NN are extremely resistant against noisy data is one of its benefits. The neural network can adapt and overcome the presence of unhelpful or inaccurate data points within the dataset. This is possible due to the network's large number of artificial neurons and the weights assigned to each connection between them, allowing it to learn and make accurate predictions despite the presence of such data. (Hand, 2001). However, decision trees provide rules that are comprehensible to individuals without a technical background, but neural networks are more difficult for humans to read. Additionally, training neural networks typically takes more time than training decision trees, sometimes up to several hours (Merghadi, 2020)

In light of the numerous DM algorithms available, the objective of this study is to select and utilize an algorithm that is open-source and easily accessible. This decision is based on the desire for wide applicability and to avoid any limitations imposed by licensing and financial constraints. After

conducting an extensive review of available algorithms, it was determined that the Google Colab software, which includes pre-defined DT, ANN, K-NN and Bayesian Classifier, is a suitable candidate for the current study.

3.11 Data Mining Experiments

Four experiments were designed to employ the selected classification algorithms, and these were conducted using a complete training dataset comprising 303 instances. Two scenarios were considered in all experiments, one containing training dataset and the other containing test dataset. Google Colab software was used for these purposes.

Although there are many data mining algorithm software available, our aim is to select and use the one that is open source and easily available. This is for wide applicability and to mitigate license related constraints. After thoroughly checking the available algorithms, Google Colab software has predefined DT, ANN, K-NNN and Bayesian Classifier which made it a candidate choice for this study.

3.12 Evaluation

In this stage, the models are evaluated to determine their effectiveness in achieving the data mining objectives established in the initial step. The evaluation process involves interpreting the results, determining the novelty and relevance of the new information, evaluating the medical significance of the results, and assessing the impact on and alignment with the project goals. The assessment of the algorithms is based on their classification accuracy and the confusion matrix. The models that pass the evaluation are retained and subject to further examination to identify any errors or flawed steps. The models are also compared in terms of performance and the model that demonstrates the best performance is ultimately selected as the optimal classifier for the dataset.

3.13 Using the knowledge established

This final step of the knowledge discovery process is crucial to the success of the entire methodology. Upon completion of all preceding steps, including target data selection, data preparation, application of data mining techniques, and evaluation of model performance, a comprehensive report is compiled to summarize the research findings and outcomes of the study.

3.14 Conclusion

When conducting research, a significant portion of time is dedicated to the methodology, as it is the foundation for obtaining reliable and valid results. To summarise, four algorithms, namely NB, K-NN, DT, and ANN were employed for model building and prediction. The evaluation of classification and prediction methods involved utilizing metrics such as predictive accuracy, TP Rate, TN Rate, and Precision. The predictive accuracy, TP Rate, TN Rate are employed to compare the performance of the models. The subsequent chapter focuses on data preparation and mining.

CHAPTER FOUR: DATA PREPARATION AND MINING

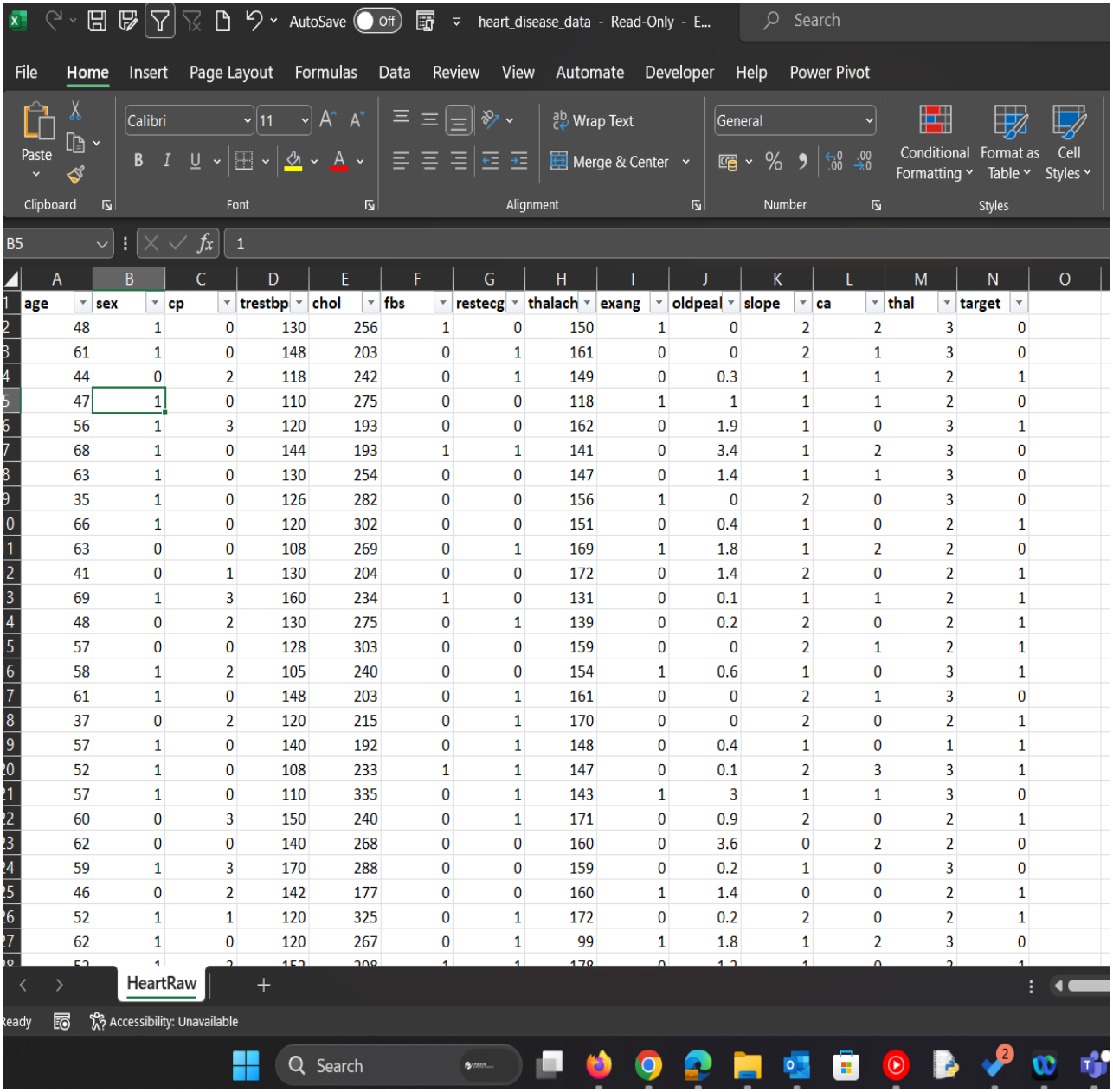
4.1 Introduction

In this chapter, data preparation and data mining transformations are presented. Data preparation or pre-processing is a technique employed in data mining to gather, combine, structure and organize data so it can be used in business intelligence (BI), analytics and data visualization application. Usually, real-world data is imperfect, with inconsistencies, missing information, and may not exhibit immediate patterns or trends, and it is ridden with inaccuracies. Data pre-processing therefore becomes an integral part of machine learning and prepares data for further processing that establishes patterns and trends.

4.2 Data Retrieval

The initial step in the process of creating a model was to collate the data from the UCI website. We downloaded the repository data and renamed it to HeartRaw.csv. We added attribute names as outlined in the documentation in Table 1. We chose the CSV format as it is easy to handle for data pre-processing with the use of Pandas. Figure 4.1 illustrates the raw Cleveland data before any processing.

The data underwent necessary cleaning and addressed null values, outliers, and duplicate entries as part of the pre-processing stage. Figure 4.1 displays the processed version of HeartRaw, renamed to Heart.csv, which went through the stages of pre-processing as shown in Figure 4.2.



age	sex	cp	trestbp	chol	fbs	restecg	thalach	exang	oldpeal	slope	ca	thal	target
48	1	0	130	256	1	0	150	1	0	2	2	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
44	0	2	118	242	0	1	149	0	0.3	1	1	2	1
47	1	0	110	275	0	0	118	1	1	1	1	2	0
56	1	3	120	193	0	0	162	0	1.9	1	0	3	1
68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
63	1	0	130	254	0	0	147	0	1.4	1	1	3	0
35	1	0	126	282	0	0	156	1	0	2	0	3	0
66	1	0	120	302	0	0	151	0	0.4	1	0	2	1
63	0	0	108	269	0	1	169	1	1.8	1	2	2	0
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
69	1	3	160	234	1	0	131	0	0.1	1	1	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
57	0	0	128	303	0	0	159	0	0	2	1	2	1
58	1	2	105	240	0	0	154	1	0.6	1	0	3	1
61	1	0	148	203	0	1	161	0	0	2	1	3	0
37	0	2	120	215	0	1	170	0	0	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
52	1	0	108	233	1	1	147	0	0.1	2	3	3	1
57	1	0	110	335	0	1	143	1	3	1	1	3	0
60	0	3	150	240	0	1	171	0	0.9	2	0	2	1
62	0	0	140	268	0	0	160	0	3.6	0	2	2	0
59	1	3	170	288	0	0	159	0	0.2	1	0	3	0
46	0	2	142	177	0	0	160	1	1.4	0	0	2	1
52	1	1	120	325	0	1	172	0	0.2	2	0	2	1
62	1	0	120	267	0	1	99	1	1.8	1	2	3	0

Figure 4. 1: A Sample of Pre-processed Cleveland data

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
10	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1

Figure 4. 2: A Sample of Processed Cleveland Dataset

4.3 Exploratory data analysis results

Harper and Flammia (2020) propose that a correlation matrix plot serves as a valuable tool to evaluate the extent of linear correlation among variables. The correlation matrix plot effectively illustrates both the magnitude and direction of the correlation between two variables, with possible values ranging from -1 to +1. It displays the correlation between the coefficients and can be represented as a heat map to visualize the relationship between features. The study found a positive correlation between chest pain (cp) and heart disease, with higher levels of chest pain indicating a higher likelihood of heart disease. Cp is an ordinal feature with 4 categories: typical angina, atypical angina, non-anginal pain, and asymptomatic pain. Furthermore, the study found a negative correlation between old peak, exercise-induced angina, and heart disease. This may be due to narrowed arteries slowing the blood flow during exercise.

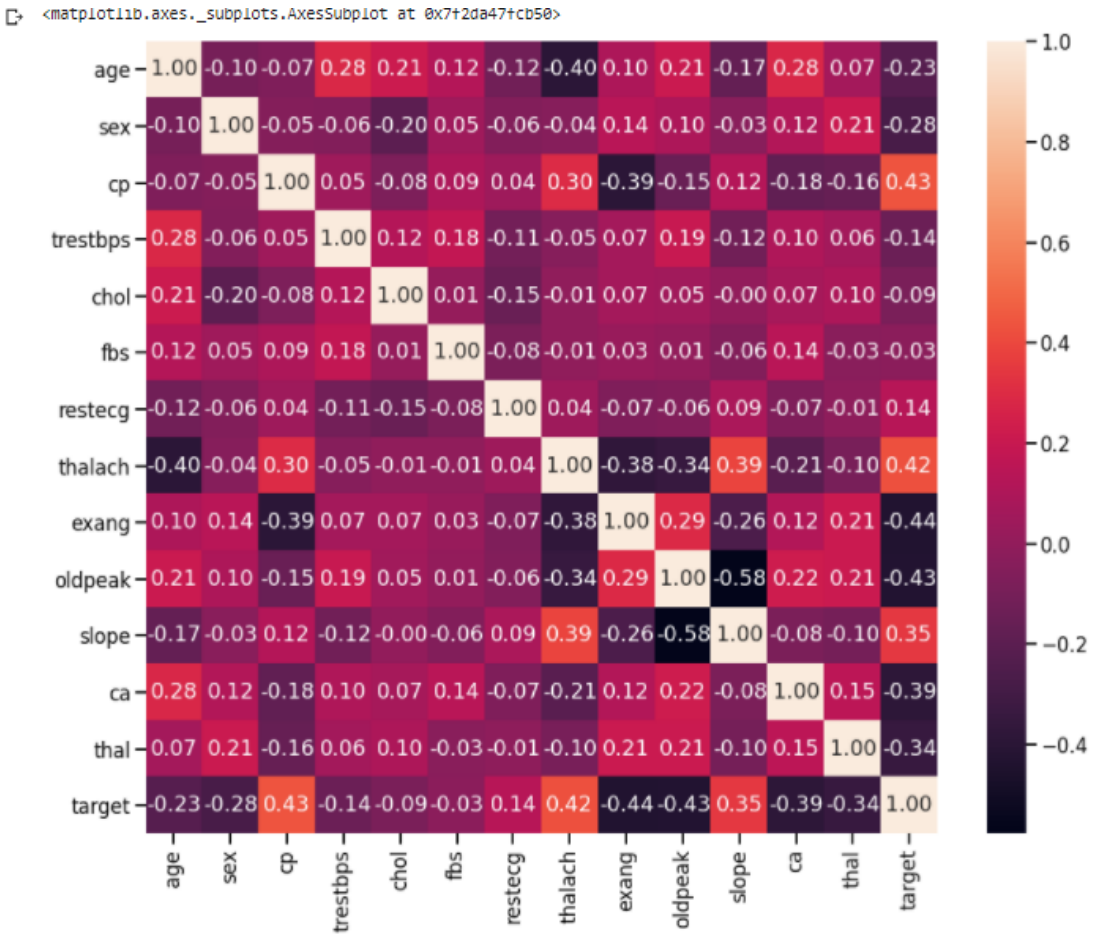


Figure 4. 3: Correlation coefficient matrix for UCI's heart disease dataset

Figure 4.3 illustrates the relationship between heart disease and other variables, particularly showing the likelihood of heart disease based on age. The Figure 4.3 also indicates that individuals with heart disease have a higher likelihood of an abnormal reading on the thallium test and more likely to experience induced angina. Additionally, the study found that people with blood vessels 0 occupy most of the data, and a greater number of heart attacks were observed when CA=0.

The following observations were recorded from Figure 4.4

- *Sex* : The number of males are way more than the number of females in our data (male=1, female = 0).
- *CP*: People with type 0 chest pain (typical angina) are far more in number than the other groups. Type 3 cp (asymptomatic angina) are the least in number.
- *FBS*: People with fasting blood sugar <120 are greater in number than people with blood sugar levels>120.

- *RESECG*: 0 (normal) and 1(having ST-T wave abnormality) are almost equal in number. This is useful for predicting heart attack. Type 2 is almost negligible.
- *EXANG*: People without exang (0) are almost double the number of people with exang.
- *CA*: People with blood vessels 0 occupy most amount of our data. More number of heart attacks were observed when CA=0 (previous analysis) .
- *THAL*: People with thal 2 are more in number. No information was given about this (may not include this in predictions).

```
df.hist(figsize=(15,15))
plt.show()
```

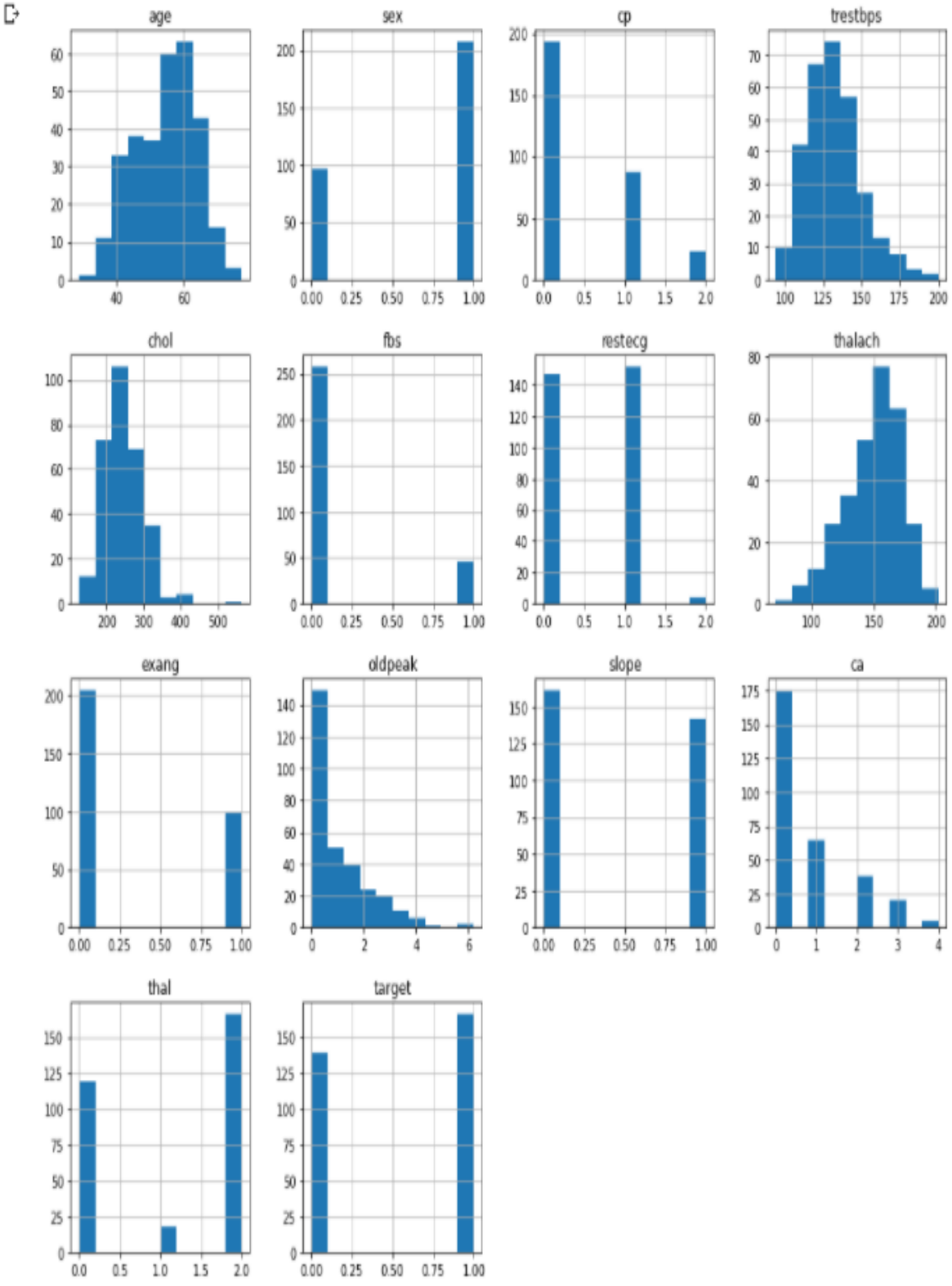


Figure 4. 4: Target versus variables

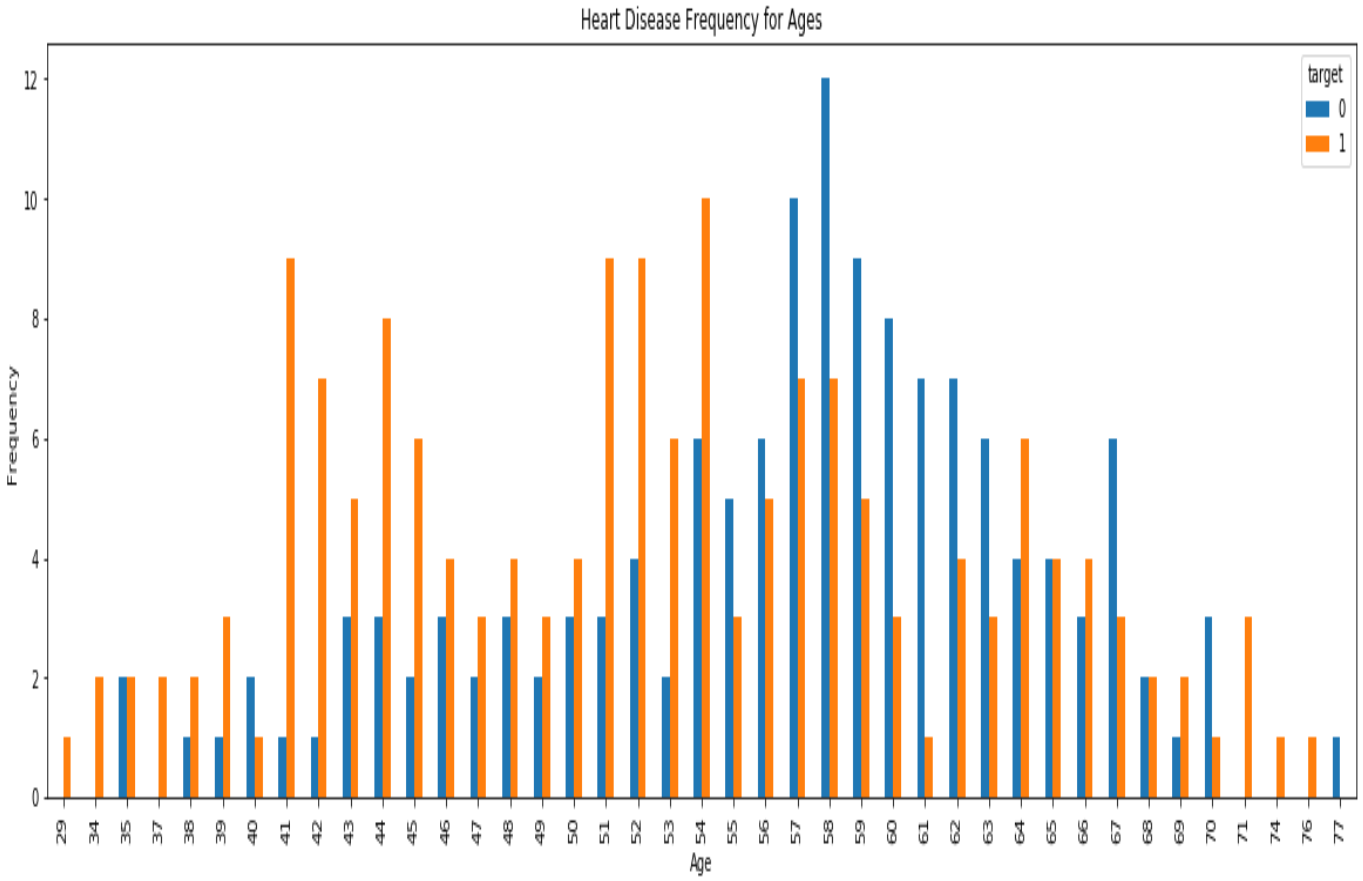


Figure 4. 5: Heart Disease Frequency for Ages

The Figure 4.5 demonstrates the potential occurrence of heart diseases relative to age. The high occurrence of cardiovascular disease (CVD) in this population, as shown in Figure 4.5, has been linked to various factors such as increased oxidative stress, inflammation, apoptosis, and overall deterioration and degeneration of the myocardium. It is clear from the figure that older adults are at a higher risk of developing CVD. As a person ages, they become more vulnerable to functional and electrical abnormalities that could cause cardiovascular disease (CVD) due to an increase in oxidative stress. The American Heart Association (AHA) reported that from 2013 to 2017, hypertension or high blood pressure was diagnosed in 77.8% of women and 70.8% of men aged 65 to 74 (Benjamin et al., 2019). The diagnosed hypertension rates further increased to 85.6% for women and 80.0% for males over 75 years old (Benjamin et al., 2019).

4.4 Data Cleaning

Data cleaning encompasses a variety of tasks aimed at enhancing the quality of data. These tasks include imputing missing values, reducing noise in the data, identifying, and removing outliers,

and resolving inconsistencies within the dataset. This process is necessary for an efficient data mining process, but it is also time consuming and labour-intensive procedure. By removing the duplicates and data conflicts features, the entire process improves the data quality.

4.4.1 Handling Noisy Data

In the context of data, noise refers to the presence of random errors or fluctuations in a measured variable. This noise can originate from various sources, such as faulty data collection instruments, errors made during data entry, or technological limitations. It is worth noting that a significant portion of the errors found in the UCI dataset can be attributed to human error or misrepresentation.

4.4.2 Handling Missing Values

Handling missing values in heart disease prediction datasets is a crucial step in the data pre-processing stage. Multiple strategies exist for addressing missing values, including the following approaches:

1. Removal or deletion: This method involves deleting the rows or columns with missing values: This is the simplest method; however, caution should be exercised to ensure that valuable information is not lost in the process of removal or deletion.
2. Mean/median imputation: In this approach, missing values are replaced with the mean or median value of the respective variable. This method assumes that the missing values are missing at random and does not consider the relationships between variables.
3. Regression imputation: This technique utilises regression models to predict missing values based on other variables. It estimates the missing values by considering the relationships among the various variables.
4. Using ML algorithms: Some ML algorithms such as Random Forest, K-NN and Multiple Imputation can be used to impute missing values.
5. Multiple imputation: Multiple imputation involves creating multiple plausible imputations for the missing values, thereby accounting for the uncertainty associated with imputation. Statistical techniques are employed to generate the missing values.

The missing data was addressed by removing the rows that contained them. As the presence of missing values can impede accurate prediction, the pre-processing stage was performed to fill these missing values. Each missing value was substituted with the mean of the other values in the

attribute. The dataset included a considerable amount of irrelevant or undesirable data, including numerous missing values. This process is known as "data cleaning," which involves the elimination of unnecessary or undesired data. The mean value was calculated and utilised to replace the missing values (Hong, 2020). The code implementation to identify missing values is presented in Figure 4.6.

```
#Count the empty values in each column
df.isna().sum()
age       0
sex       0
cp        0
trestbps  0
chol      0
fbs       0
restecg   0
thalach   0
exang     0
oldpeak   0
slope     0
ca        0
thal      0
target    0
dtype: int64
#To view missing values
df.thal.value_counts()
df.ca.value_counts
#This code is to check if there are missing values
df.isnull().values.any()
False
```

Figure 4. 6: Python code for handling missing values

4.4.3 Handling Outliers

Outliers are extreme values that are close to data range limits or that deviate from the data set's trend. Identifying outliers is critical, according to Larose (2014), since they may indicate mistakes in the data entry.

To identify outliers in the datasets, boxplots were employed. Figure 4.7 shows the resulting boxplot plot after defining the outliers. Once the outliers were identified, the corresponding attributes were removed and replaced with the mean value.

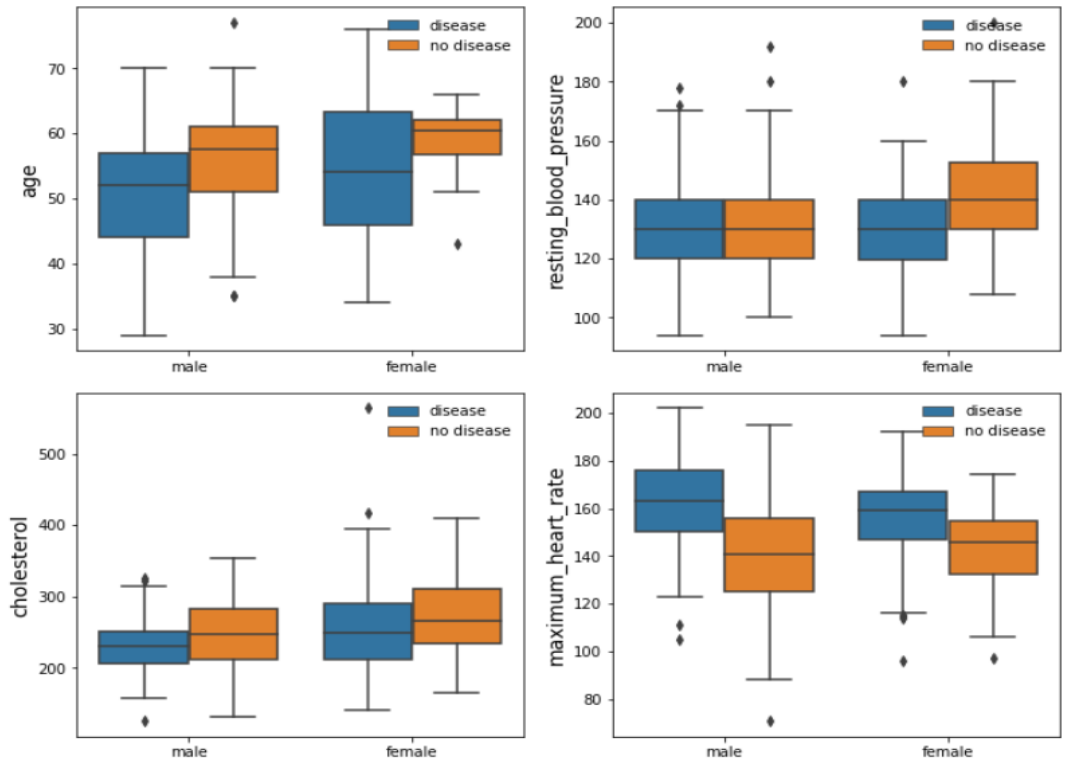


Figure 4. 7: Boxplots of Attributes vs Target

From Figure 4.7, it is clear that age, resting blood pressure and cholesterol are not factors in determining the heart health of a person. The likes of maximum heart rate and ST depression are good attributes in determining the health of the heart. The heart rate is directly proportional to the presence of a cardiovascular disease, meaning that the higher the heart rate, the higher the chances of having cardiovascular disease. There exists an inverse relationship between the ST depression level and the likelihood of cardiovascular disease, such that a lower ST depression value is associated with a higher incidence of cardiovascular disease.

4.4.4 Handling Inconsistent Data

In this research, UCI datasets were utilised that had been entered by various individuals, resulting in the same data being represented in different ways. It is crucial to consider and correct these differences to ensure that the outcome of the research is accurate and reliable. An instance of this is apparent in the Sex category, where two separate values (male and female) are present, yet they

are documented in distinct ways. Occasionally, the abbreviation (M and F) is used to record the patient's sex, while in other cases, the complete terms (Male and Female) are employed. To eliminate any discrepancies, the results documented using abbreviations were adjusted to align with the way the cases were reported using the full form of the terms. This ensures that the data is consistent, and any conclusions drawn from it are accurate.

4.4.5 Label Encoding

This dataset contains three categorical variables Cp, Slope and Thal that are not well represented. Machine Learning algorithm produces errors if the variables are not encoded. We converted the Cp values of (1,2,3,4) to (0,1,2,3), Slope values of (1,2, 3) to (0, 1, 2) and the Thal values of (3,6,7) to (0,1,2).

```
#To label encode
df['cp'].replace([1,2,3,4] , [0,1,2,3], inplace=True)
df['slope'].replace([1,2,3] , [0,1,2], inplace=True)
df['thal'].replace([3,6,7] ,[0,1,2], inplace=True)
```

Figure 4. 8: Label Encoding

4.5 Data Discretization

Discretization is the process of transforming continuous variables to discrete values by using a restricted number of labels to describe the original variables. In a continuous spectrum, discrete values can have a finite number of intervals, whereas continuous values can have an unlimited number of intervals (Damtew, 2011). Variables in classification models are discrete in nature. In this dataset, 5 out the 14 attributes are continuous, therefore we need to make this data discrete for possible classification. In this dataset, age, trestsbp, chol, thalach and oldpeak are continuous.

Oslon et al., (2008) outlined some reasons why many scientists prefer discrete values to continuous values in working with predictive models. Olson suggests that discrete values are more closely aligned with a knowledge-level representation, and that reducing and simplifying data into discrete categories benefits both users and experts (Olson et al., 2008).

4.5.1 Age attribute discretization

To discretize the age attribute, the code below in Figure 4.9 was used, and Table 4.1 shows the results of the discretising Age.

```
#To Discretize Age attribute
bins = [28,45,64,79]
names = [0,1,2]
df['AgeC'] = pd.cut(df['age'], bins, labels=names)
```

Figure 4. 9: Python code to discretize Age

Table 4. 1: Results to discretize Age attribute

Age	Class	Variable
29 – 45	**Young**	**0**
46 – 64	**Middle**	**1**
64– 77	**Old**	**2**

4.5.2 Tresbps attribute discretization

To discretize the Tresbps attribute, the code in Figure 4.10 below was used, and Table 4.2 shows the results of the discretising Tresbps.

Tresbps attribute ranges from 94mmHg to 200mmHg; the code below categorises Tresbps, respectively.

```
#To Discretize tresbps attribute
bins = [90,120,140,201]
names = [0,1,2]
df['trestbpsC'] = pd.cut(df['trestbps'], bins, labels=names)
```

Figure 4. 10: Python code to Discretize Tresbps attribute

Table 4. 2: Results to Discretize Tresbps attribute

Tresbps(mmHg)	Class	Variable
90 – 120	**Ideal**	**0**
120 – 140	**Pre-high**	**1**
140 – 200	**High**	**2**

4.5.3 Chol attribute discretization

To discretize chol. research finds that chol ranges from 126mg/dl to 564mg/dl. Figure 4.11 below discretize chol respectively and Table 4.3 shows the results of discretized chol attribute:

```
#To Discretize chol attribute
bins = [125,200,240,565]
names = [0,1,2]
df['cholC'] = pd.cut(df['chol'], bins, labels=names)
```

Figure 4. 11: Python code to Discretize Chol attribute

Table 4. 3: Results to Discretize Chol attribute

Chol(mg/dl)	Class	Variable
125 – 200	**Normal**	**0**
200 – 239	**Mid**	**1**
240 – 565	**High**	**2**

4.5.4 Thalach attribute discretization

To discretize thalach, the science expects that the maximum heart rate is 220 minus the age. Initially, we create the column that holds values(200-age). This column is then later compared with thalach attribute. When the thalch is less (200-age), we categorize it as 0 or else as 1. Figure 4.12 shows the Python code to discretize thalach attribute.

```
#To Discretize thalach attribute
df['220 - age'] = 220 - df['age']
df['thalachC'] =df['200-age'] < df['thalach']
df['thalach'].replace([True,False],[1,0],inplace=True)
```

Figure 4. 12: Python code to discretize Thalach

4.5.5 Oldpeak Attribute discretization

The degree of ST depression induced by exercise compared to rest can be classified as 0 for a range of 0-2, indicating irreversible ischemia, and 1 for a range of 2-6.2, indicating reversible ischemia. The Figure 4.13 below shows the implementation.

```
#To Discretize oldpeak attribute
bins = [-1,2,6.3]
names = [0,1]
df['oldpeakC'] = pd.cut(df['oldpeak'], bins, labels=names)
```

Figure 4. 13: Python code to discretize Oldpeak

4.6 Data Mining

4.6.1 Training the data

After the data was meticulously cleaned, converted into discrete values, and deemed suitable for modelling, the next step was the modelling or data mining phase. During this phase, various models are constructed and implemented to predict the value of the goal variable and identify which independent variable displays the most optimal performance using a set of predictors or independent variables. Various categorisations are utilised to evaluate a collection of observations for the purpose of identifying the most precise model that can predict individuals prone to experiencing cardiac arrests. To arrive at the logical conclusions, a variety of algorithms utilising different methods of data input and output computation are employed to analyse the problem at various levels of complexity. The implementation of the code, written in Python, was used to divide the dataset into train and test datasets as shown in Figure 4.14

```
#Splitting the data again, into 75% training data and 25% testing data set

from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25, random_state = 1)

#Feature scaling, Scaling the data values in the data to the value between 0 and 1
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)
```

Figure 4. 14: Python Code to train and test dataset

4.6.2 Testing the data

The model has been fitted to the data successfully, and it is now time to test the model. The test dataset is used to make predictions, and the code implementation for this process is shown in Figure 4.15.

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
def print_score(clf, X_train, y_train, X_test, y_test, train=True):
    if train:
        pred = clf.predict(X_train)
        clf_report = pd.DataFrame(classification_report(y_train, pred, output_dict=True))
        print("The Train result:\n)
        print(f"The algorithm accuracy Score: {accuracy_score(y_train, pred) * 100:.2f}%")
        print("")
        print(f"Classifcation Algorithm report:\n{clf_report}")
        print("")
        print(f"The Confusion Matrix: \n {confusion_matrix(y_train, pred)}\n")
    elif train==False:
        pred = clf.predict(X_test)
        clf_report = pd.DataFrame(classification_report(y_test, pred, output_dict=True))
        print("The Test result:\n)
        print(f"The algorithm accuracy Score: {accuracy_score(y_test, pred) * 100:.2f}%")
        print("")
        print(f"Classifcation Algorithm report:\n{clf_report}")
        print("")
        print(f"The Confusion Matrix: \n {confusion_matrix(y_test, pred)}\n")
```

```
from sklearn.model_selection import train_test_split
X = df.drop('target', axis=1)
y = df.target
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42
```

Figure 4. 15: Python code to test the trained dataset

This chapter detailed the processes necessary in making the data set ready for data mining. The data was cleaned, and inconsistency was removed, and the process went to split and train the dataset. The following chapter presents and explains the results in each data mining algorithm implemented on the UCI heart disease dataset.

CHAPTER FIVE: RESULTS AND DISCUSSION

This presents and discusses the research findings, illustrating how the research questions were addressed and the research objectives explored. An overview of the research findings is presented in the following sections of this chapter.

5.1 Experimental Setup

Four classification algorithms were used in this experiment. It was conducted on a full trained dataset. The dataset was cleaned, and inconsistencies were removed.

To prepare for data mining, the researchers loaded the entire dataset saved in a Microsoft Excel csv file into the Google Collaboratory modelling tool. Google Collaboratory allows for automatic conversion of csv files to .ipynb format, which is necessary as Google Collaboratory can only work with test files saved in the .ipynb format.

Testing was conducted separately for each classification model and performance parameters computed separately. The same training and test were used for all the models. In this research, the effectiveness of the algorithms was assessed using standard performance metrics such as accuracy, precision, recall, F Score, and False Negative. These metrics were calculated by employing the Confusion Matrix. The conventional metrics of accuracy, precision, recall, F Score, and False Negative, which were computed using the Confusion Matrix, were utilised to assess the effectiveness of the algorithms in this study.

The procedures taken to resample the data are described in the algorithm below.

Algorithm 5.1: Classification of dataset

Step 1. Start

Step 2. Open Google Collaboratory modelling software

Step 3. Load the libraries

Step 4. Import File (Load .csv)

Step 5. Load Visualisation

Step 6. Split Data into Train and Test

Step 7. Load Classification model namely: K Nearest Neighbour, Naïve Bayes, Decision Tree, Artificial Neural Network

Step 8. Test Accuracy on all Classification Algorithms

Step 9. Stop

With dataset loaded in the environment, the mining process begins by making use of the model built for testing of the hidden test data.

5.2 Model Building

5.2.1 Model Building Using K-Nearest Neighbour (KNN)

The initial model constructed in this study using the Google Collaboratory modelling tool is the KNN model. Algorithm was performed under the same conditions as mentioned above in 5.1. Table 5.2 below shows precision, recall, and F-Measure results of KNN without applying neighbour parameters in the algorithm and Table 5.1 depicts the confusion matrix of the predicted results of the experiment.

Table 5. 1: K- Nearest Neighbour Confusion Matrix

Confusion matrix		
Actual	**Yes (Predicted)**	**No (Predicted)**
YES	26 (87)	15 (50)
NO	13 (43)	37 (123)

Out of the total of 303 patients, the K-NN algorithm accurately identified 210 patients with cardiovascular disease, while 93 patients were not correctly identified. With an accuracy rate of 69.23%, this algorithm proved to be an ineffective choice. Specifically, the algorithm correctly predicted that 87 out of the 303 patients with cardiovascular disease had the disease, while incorrectly classifying 50 patients as disease-free despite them having the disease. The algorithm also demonstrates proficiency in identifying negative cases, accurately diagnosing 123 patients out of the 166 patients who did not have heart disease, while misclassifying the remaining 43 patients as having the disease when they did not. Overall, the algorithm obtained an average precision of 66.67%, indicating that the algorithm effectively generated significant values for all categories.

Table 5. 2: Summary of test dataset results obtained from KNN classifier model

Metrics	**Train Value**	**Test Value**
Precision	74.44%	66.67%
Recall	69.07%	63.41%
F-Measure	71.66%	65.00%

The accuracy that was achieved before applying k-neighbour is 100% for training dataset and 52.46% for testing dataset respectively. The next step was to evaluate if KNN would perform differently or better if various parameters were used as inputs, such as n-neighbour, where n is a number of a nearest neighbour. With an F-Measure score of 65%, this indicates the model's Precision and Recall are significantly balanced, implying that they are weighted averages of the two metrics.

Table 5. 3: Summary test results obtained from K(n) Neighbour

K(n)	**Train Accuracy (%)**	**Test Accuracy (%)**
1	100	52.46
2	79.75	59.02
3	78.10	63.93
4	76.03	63.93
5	78.10	63.93
6	74.38	65.57
7	72.31	67.21
8	75.00	69.23
9	73.14	67.21

An overview of the outcomes from using various values of k in the KNN algorithm is shown in Table 5.3. From the results, we see that value of K-neighbours (8) are optimal with the highest accuracy score of 69.23% of predicting the probability of cardio-vascular disease and Figure 5.1 illustrates the performance of the model in Google Collaboratory.

```
The Test result:
**********************************************************************
The algorithm accuracy Score: 69.23%

----------------------------------------------------------------------
Classifcation Algorithm report:
                    0          1  accuracy  macro avg  weighted avg
precision    0.666667   0.711538  0.692308   0.689103      0.691321
recall       0.634146   0.740000  0.692308   0.687073      0.692308
f1-score     0.650000   0.725490  0.692308   0.687745      0.691478
support     41.000000  50.000000  0.692308  91.000000     91.000000

----------------------------------------------------------------------
The Confusion Matrix:
 [[26 15]
 [13 37]]
```

Figure 5. 1: The KNN classifier model showing the performance of the training and test dataset

The AUC for this model was 69.00 and ROC of the model is presented in Figure 5.2.

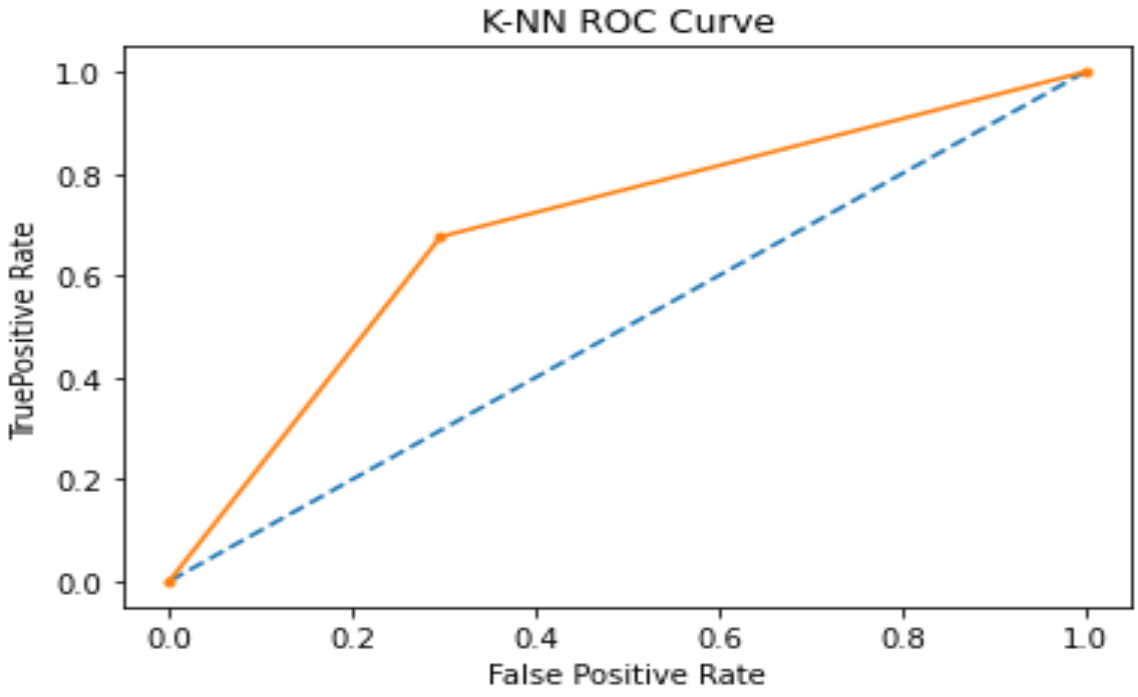


Figure 5. 2: ROC curve showing the performance output of the KNN Classifier Model

5.2.2 Model Building using Naïve Bayes

The second algorithm was developed to assess the ability of the Naïve Bayes classifier in predicting cardiovascular disease. This classifier was constructed under the same conditions as mentioned above in 5.1 where all 14 attributes and 303 instances are selected. Table 5.5 illustrates the precision, recall, and F-Measure outcomes obtained from the implementation of the Naïve Bayes algorithm. Table 5.4 displays the confusion matrix, showcasing the predicted results of the experiment.

Table 5. 4: Naïve Bayes Confusion Matrix

Confusion matrix		
Actual	**Yes (Predicted)**	**No (Predicted)**
YES	21 (104)	6 (30)
NO	3 (15)	31 (154)

The Naive Bayes classifier, which was constructed using all attributes, correctly predicted 258 (85.25%) instances as having cardiovascular disease, while incorrectly identifying 45 (14.75%) instances as not having the disease. The Confusion matrix in Table 5.4 demonstrates that the classifier accurately predicted that 104 patients (True Positive) had cardiovascular disease and incorrectly classified 30 patients (False Negative) as not having the disease. The model performed well in identifying negative cases, correctly diagnosing 154 patients out of the 169 patients who did not have heart disease and misclassifying the other 15 patients as having the disease when they did not have.

The algorithm obtained an average precision of 85.25%, indicating that the model is successful in producing correct values for each class. The F-Measure score, which is 87.32%, indicates that the model's Precision and Recall are evenly matched, meaning that it is a weighted average of both metrics. The Naive Bayes algorithm exhibited superior performance compared to the initial algorithm in predicting the presence of cardiovascular disease.

Table 5. 5: Summary of test dataset results obtained from Naïve Bayes classifier model

Metrics	**Value**
Precision	83.78%
Recall	91.17%
F-Measure	87.32%

The accuracy score achieved using Naive Bayes is 85.25 % which is a moderately successful overall accuracy.

Table 5-4 shows how Naïve Bayes was able to obtain an accuracy of 85.25%. Moreover, the achieved AUC of 0.872 shown in Figure 5.3 explains how the naïve Bayes model performed in terms of avoiding false classifications.

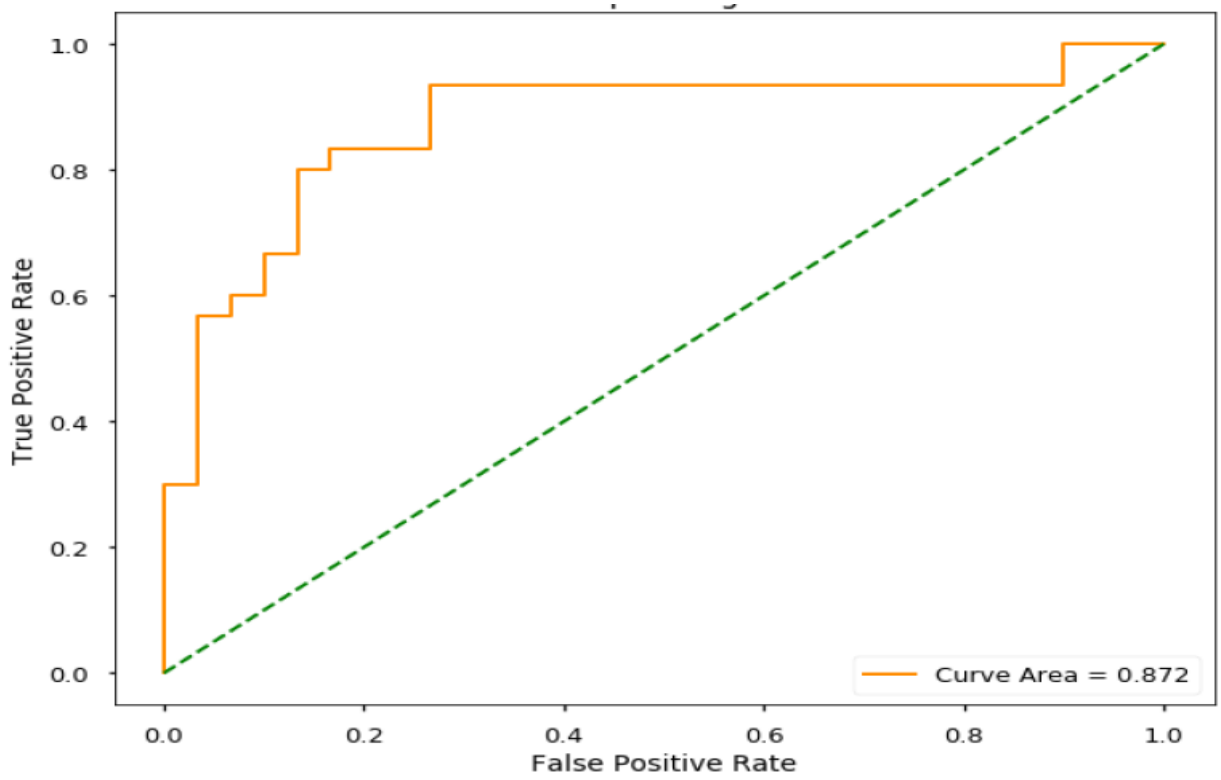


Figure 5. 3: AUC of ROC for Naïve Bayes.

5.2.3 Model Building using Decision Tree

The third model was developed to assess the performance of the Decision Tree classifier in predicting cardio-vascular disease. The classifier builds a tree to demonstrate its classification and it is one of the most popular methods used in Data Mining and decision making.

This classifier was constructed under the same conditions as mentioned above in 5.1 where all 14 attributes and 303 instances are selected. Table 5.6 shows the confusion matrix of the predicted results of the experiment and Table 5.7 displays precision, recall, and F-Measure results of Naïve Bayes algorithm.

Table 5. 6: Decision Tree Confusion Matrix

Confusion matrix		
Actual	**Yes (Predicted)**	**No (Predicted)**
YES	30 (100)	11 (36)
NO	15 (50)	35 (117)

The Decision Tree classifier that was built on all attributes accurately predicted 217 (71.43%) instances as having the disease whereas the other 86 (28.57%) patients did not have a disease. The accuracy of this model is significantly high.

From the Confusion matrix in Table 5.6, the classifier accurately predicted that 100 patients (True Positive) had cardiovascular disease and the 36 patients (False Negative) were falsely classified as not having cardio-vascular disease when they had the disease. The algorithm performed perfectly in identifying negative cases, correctly diagnosing 117 patients of the 167 patients without cardiovascular disease and misclassifying the other 50 patients as having the disease while they did not.

The algorithm obtained an average precision of 76.08%, allowing the conclusion that the model is successful in predicting pertinent values of each class. With F-Measure value of 72.91%, it can be concluded that the Precision and the Recall of the model are significantly balanced, meaning that

it is weighted average of both metrics. Decision Tree algorithm performed better than the first built algorithm in predicting the existence of cardio-vascular disease and Figure 5.4 depicts the performance of the Decision Tree model in the programming environment of python.

Table 5. 7: The summary of test dataset results obtained from Decision Tree model.

Metrics	**Value**
Precision	66.67%
Recall	73.17%
F-Measure	69.77%

```
The Train result:
****************************************************************
The algorithm accuracy Score: 100.00%

_____________________________________________________________________
Classifcation Algorithm report:
              0      1  accuracy  macro avg  weighted avg
precision   1.0    1.0       1.0        1.0           1.0
recall      1.0    1.0       1.0        1.0           1.0
f1-score    1.0    1.0       1.0        1.0           1.0
support    97.0  115.0       1.0      212.0         212.0
_____________________________________________________________________
The Confusion Matrix:
 [[ 97   0]
 [  0 115]]

The Test result:
*********************************************************************
The algorithm accuracy Score: 71.43%

_____________________________________________________________________
Classifcation Algorithm report:
                   0          1  accuracy  macro avg  weighted avg
precision   0.666667   0.760870  0.714286   0.713768      0.718427
recall      0.731707   0.700000  0.714286   0.715854      0.714286
f1-score    0.697674   0.729167  0.714286   0.713421      0.714978
support    41.000000  50.000000  0.714286  91.000000     91.000000
_____________________________________________________________________
The Confusion Matrix:
 [[30 11]
 [15 35]]
```

Figure 5. 4: The Decision Tree classifier model showing the performance of the training and test dataset

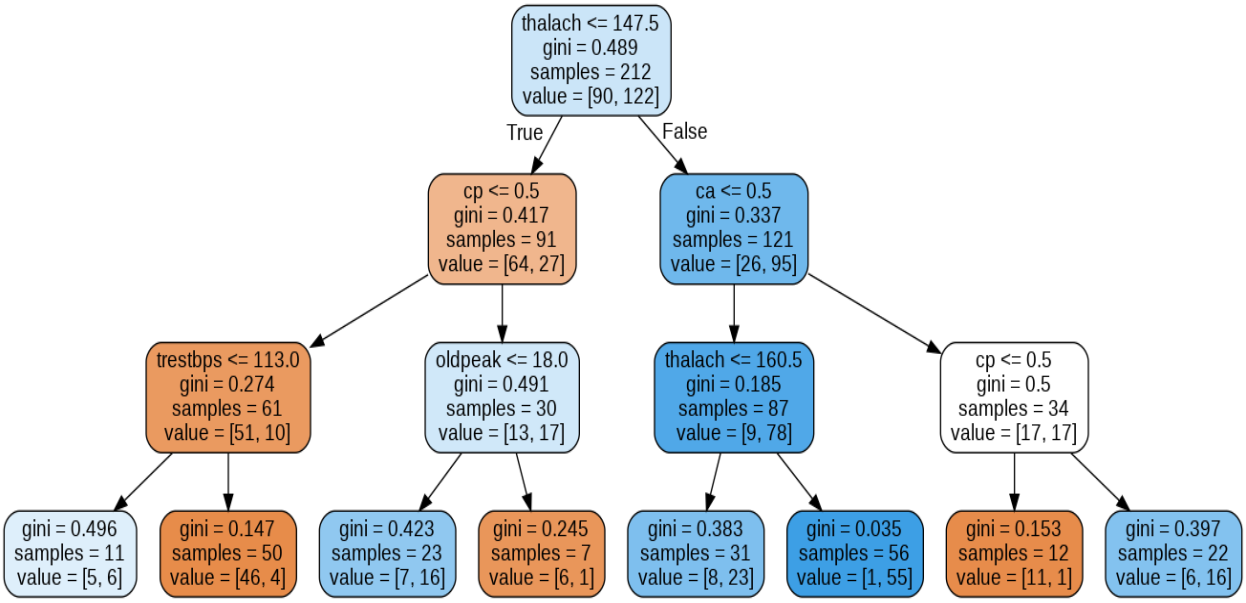


Figure 5. 5: Decision Tree Classifier

Figure 5.5 is a representation of a simple Decision Tree. The model achieved the 71.43% accuracy. As evident in Figure 5.4, the training set's accuracy is 100 %, yet the test set's accuracy is significantly lower. This means the tree is over-adapted or overfitting, and it is not properly generalised to new data. As a result, the tree must be pruned ahead of time. We have set maximum depth to three. Overfitting is reduced by limiting the depth of the tree. The training set's accuracy suffers as a result, but the test set's accuracy improves.

Table 5. 8: Decision Tree Accuracy for Pruned Set

Depth	Training Set (%)	Test Set (%)
0	100	71.43
3	84.3	82.0

After rectifying overfitting on the dataset, the new accuracy of the test set increased to 82% which shows a slight improvement on behaviour to dataset. The 10.57% improvement in the results was obtained using grid search to establish the optimal hyperparameters and the tree in Figure 5.3 was used to classify. The difference in accuracy percentage after the depth of the tree has been applied shows that the tree behaves better when pruned. Figure 5.6 shows the tree after it has been pruned.

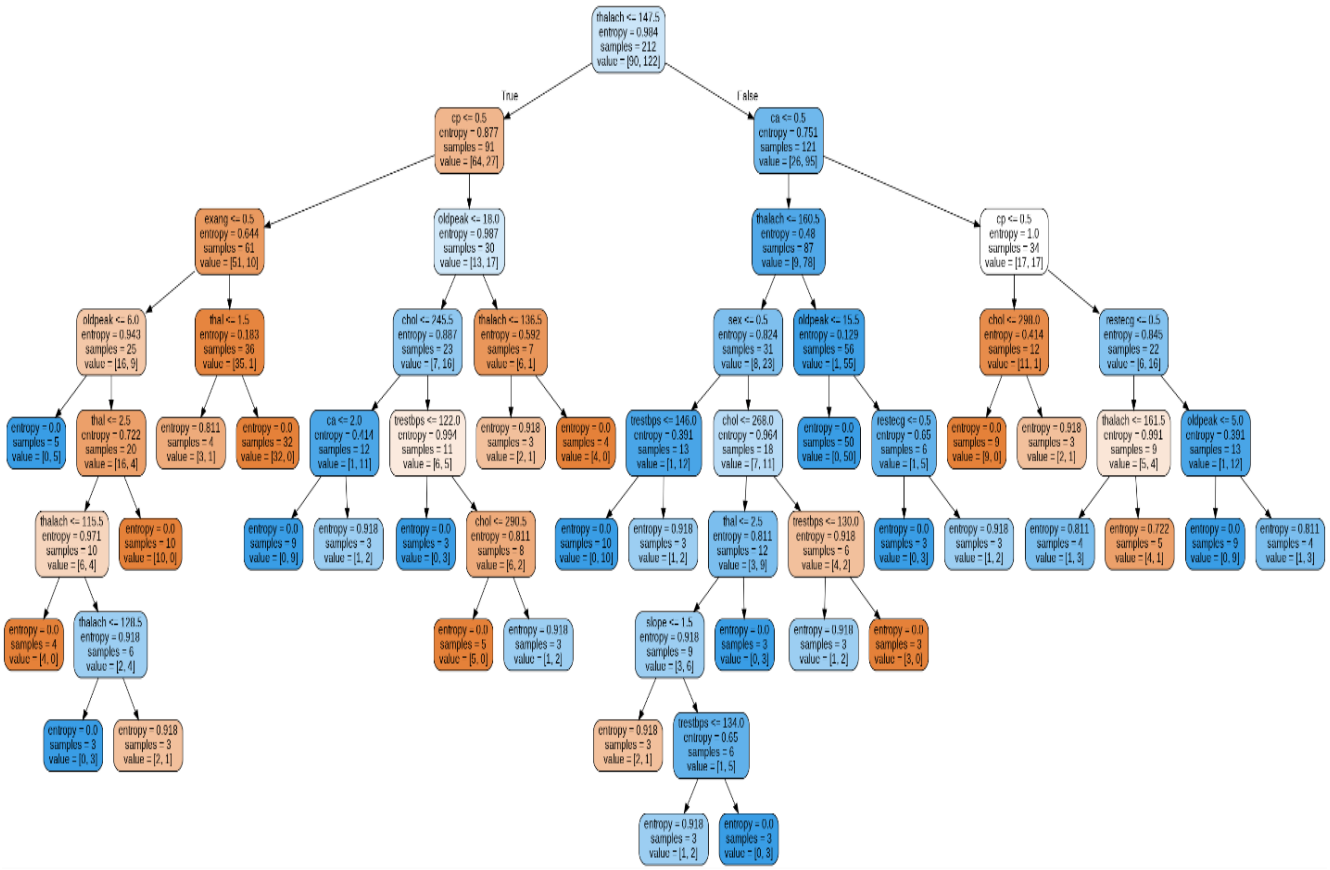


Figure 5. 6: Decision Tree Showing Grid Search

5.2.4 Model Building using Artificial Neural Networks

The purpose of the fourth experiment was to investigate the potential of artificial neural networks (ANN) in predicting the presence of cardiovascular disease. The essence of neural networks lies in the process where input data are available at the input layer, and the network neurons perform computations in the hidden layers until an output value is generated at each output neuron. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables (Two Crows Corporation, 2005). This output serves as an indicator to identify the appropriate class for the input data. In other words, one anticipates a significant output value on the correct class neuron and relatively lower output values on all other neurons.

This classifier was constructed under the same conditions as mentioned above in 5.1 where all 14 attributes and 303 instances are selected. Table 5.9 below displays the confusion matrix of the

predicted outcomes of the experiment and Table 5.10 shows precision, recall, and F-Measure results of Naïve Bayes algorithm.

Table 5. 9: Artificial Neural Network Confusion Matrix

Confusion matrix		
Actual	**Yes (Predicted)**	**No (Predicted)**
YES	24 (119)	3 (15)
NO	2 (10)	32 (159)

It is evident in this case that 278 (91.75%) of the instances were correctly identified using the ANN algorithm, whereas 17 (8.25%) were incorrectly classified. In comparison to the other methods that have been discussed thus far, the model's overall accuracy rate is higher. Out of the 303 individuals who had cardiovascular disease, the algorithm accurately identified 119, but it wrongly classified 15 patients as being disease-free when they had the condition.

The algorithm is better at identifying negative instances, with a True Negative rate of 94.08 percent, correctly recognising 159 of the 169 patients who did not have the condition and the remaining 10 patients who did have the disease when they were free of it.

Table 5. 10: Summary of test dataset results obtained from ANN model

Metrics	**Value**
Precision	91.40%
Recall	94.11%
F-Measure	92.75%

The algorithm used in this study obtained an average precision of 91.40%, indicating that the model is successful in predicting pertinent values for each class. Furthermore, the F-Measure value of 92.75% proposes that the Precision and Recall of the model are well-balanced, meaning that it is a weighted average of both metrics. The results demonstrate that the Artificial Neural Network

(ANN) algorithm demonstrated superior performance compared to other algorithms in predicting the presence of cardiovascular disease.

A key mathematical method for improving the precision of predictions in data mining and machine learning is backpropagation, also known as backward propagation. This method is used to quickly calculate derivatives and is a learning algorithm that ANN employs to compute gradient descent with respect to weights. The system is optimised by fine-tuning the connection weights to minimize the variation between the desired and achieved outputs. The weights are updated in the reverse direction, from output to input, which gives the algorithm its name. The use of backpropagation in this research contributed to the high performance of the ANN algorithm in predicting cardiovascular disease.

5.3 Discussion

The effort of this study was to develop a predictive model that can accurately forecast the presence of cardio-vascular disease being less dependent on the doctor's intuition. The optimal classification model for the data employed in this study was determined by comparing the performances of the four classification methods that were used to model the data. Discussion of the results and performances of the four selected algorithms is elaborated in this section.

5.3.1 Model Comparison

All the experiments were conducted and the next stage was designed to establish which one performed better that the other. All experiments were performed under the same conditions with all the attributes selected.

Table 5. 11: Comparison of the classifiers models performance on the dataset based on correctly and incorrectly classified.

	KNN	**Naïve Bayes**	**Decision Tree**	**ANN**
Correctly Classified	210	258	217	278
Incorrectly Classified	93	45	86	25

The performance of the algorithm models was compared in Table 5.11 based on the correct and incorrect classification of patients with heart disease. The table illustrates that the Artificial Neural Network (ANN) algorithm had the highest performance, correctly identifying 278 patients and incorrectly identifying 25 patients of the dataset. The Naive Bayes algorithm performed second-best, correctly identifying 258 patients and misclassifying 45 patients. The Decision Tree algorithm performed better than the K-Nearest Neighbor algorithm by correctly identifying 7 more patients, with 217 patients correctly classified and 45 patients misclassified. The K-Nearest Neighbor algorithm had the least performance, correctly classifying 210 patients and incorrectly classifying 93 patients.

It is clear from a close examination of Table 5.11 that the ANN classifier performed best at accurately classifying every patient in the dataset. As such, this classifier was chosen as the best for both the training and test datasets. Additionally, several performance metrics, including accuracy, F-Measure, recall, true positive rate, and true negative rate, were used to compare the models. These evaluations further support the conclusion that the ANN algorithm is the most effective for the prediction of cardiovascular disease.

Table 5. 12: Performance Summary of the Algorithms

Model	**Accuracy**	**Precision**		**Recall**	**F-Measure**	**ROC AUC**
K-NN	66.67	66.67		63.41	65.00	69.00
Naïve Bayes	85.25	83.78		91.17	87.32	87.2
Decision Tree	71.43	66.67		73.17	69.77	77.12
ANN	91.75	91.40		94.11	92.75	91.23

As shown in Table 5.12, the study evaluated the performance of the four algorithms in detecting the presence of cardiovascular disease. The highest accuracy achieved was 91.75% while the lowest accuracy was 66.67%. The Artificial Neural Network (ANN) algorithm exhibited the highest accuracy at 91.75%, with a difference of approximately 6% higher than the second-best algorithm, Naive Bayes, which recorded an average accuracy of 85.25%. The K-Nearest Neighbor (KNN) classifier performed the worst among the four algorithms evaluated, despite having the

same attributes as the other algorithms. Additionally, the Decision Tree and KNN classifiers performed below average among all the algorithms, with Decision Tree achieving an average accuracy of 71.43% and KNN recording the lowest prediction accuracy of 66.67%.

The findings of the research are further presented in Figure 5.7, which presents a meta-analysis of the four algorithms chosen for this research. It is evident from the figure that ANN outperformed the other algorithms in all comparison factors. These results demonstrate the effectiveness of ANN in predicting cardiovascular disease and its superiority over other algorithms in this field.

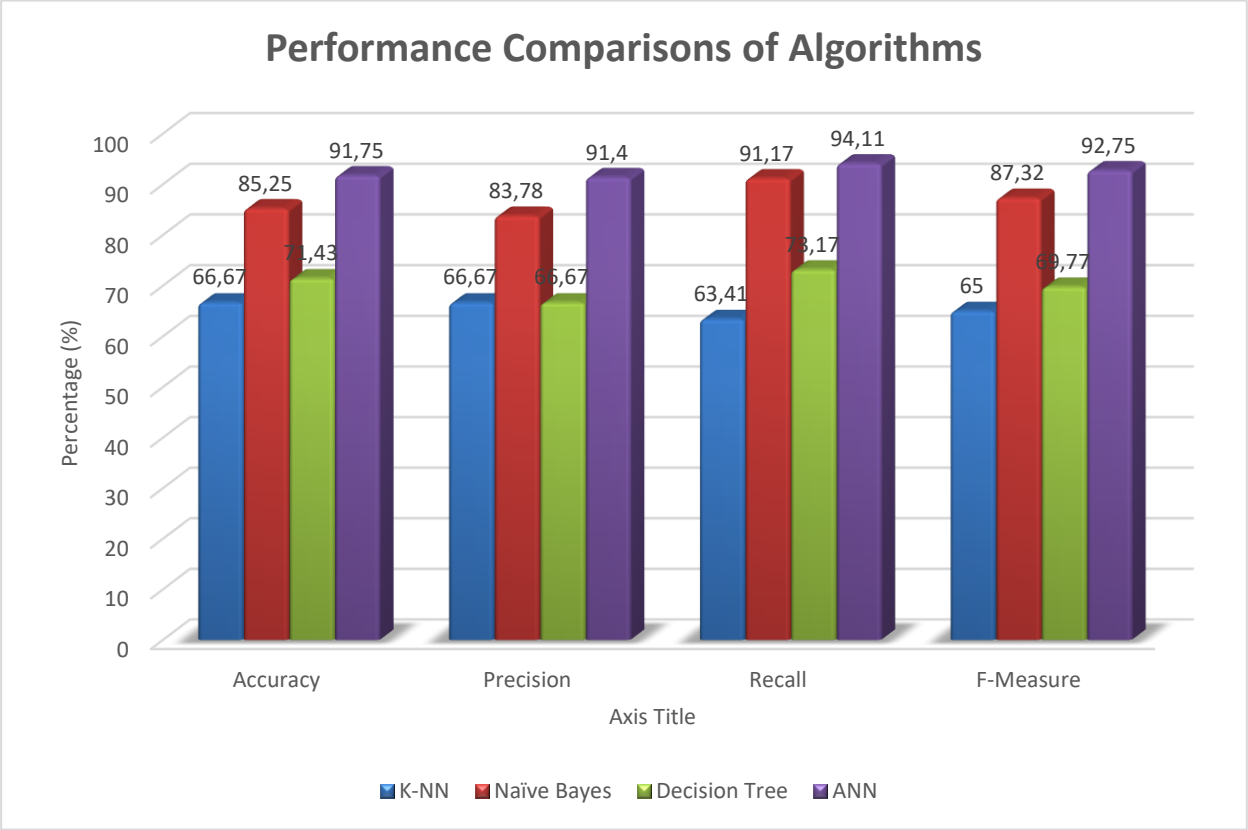


Figure 5. 7: Classification Algorithms Comparisons

5.3.2 ROC AUC of Classification Algorithms.

Table 5.12 of the study shows that the ANN and NB classifiers were the most precise when evaluating the area under the curve (AUC) as a measure of the quality of separation in terms of the Receiver Operating Characteristic (ROC) Area. When compared to data sets from previous studies, it was observed that the neural network classifier applied to specific characteristics achieved an ROC Area that was closer to the "perfect classification" mark. This finding highlights the improved

performance of the ANN classifier in the identification of cardiovascular disease when compared to other classification algorithms, and the usefulness of AUC as a metric to measure the performance of these models.

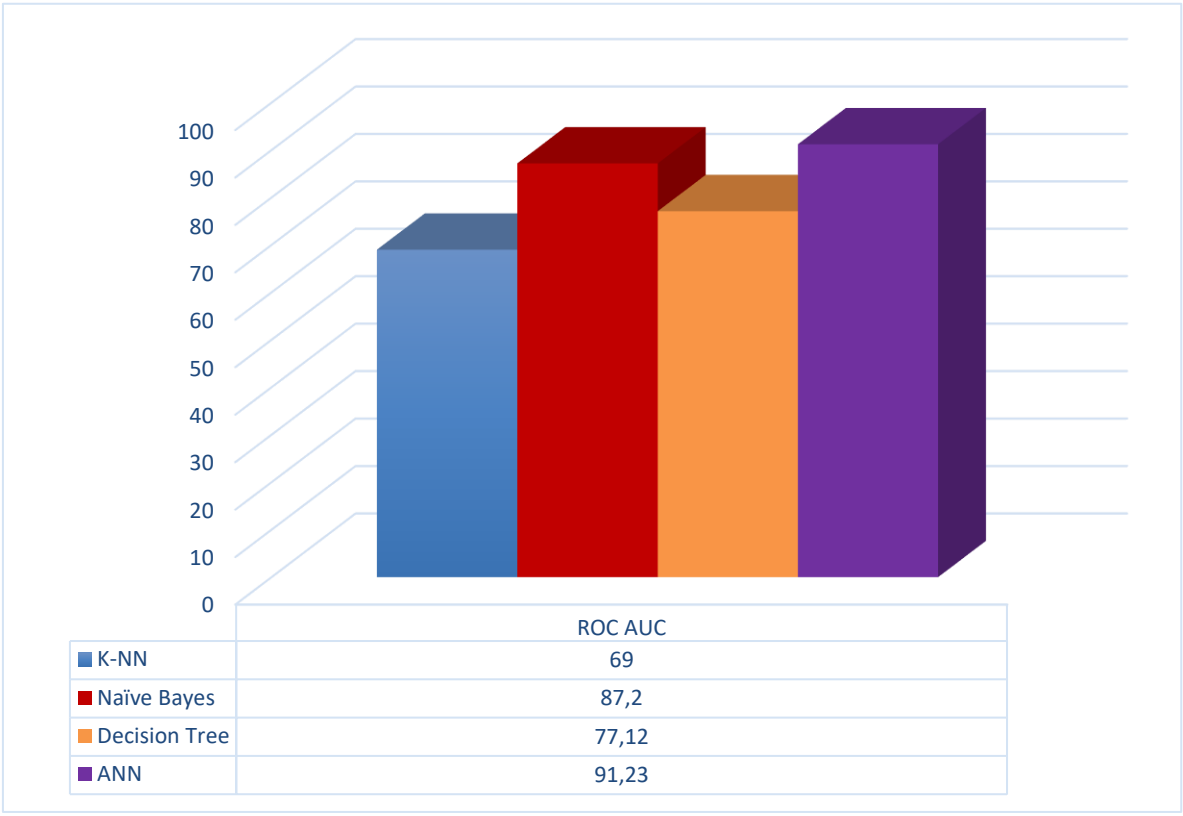


Figure 5. 8: Classification ROC Area performances

Once the comparison of the different models was completed, the subsequent stage involved choosing the optimal model based on the outcomes of the comparison.

5.3.3 Performance of model selection.

The present study aimed to develop a model for the prediction of cardiovascular disease, with a focus on minimizing the rate of false positives, as the detection of this lethal illness is of paramount importance in healthcare. The objective of the study was to determine the true positive rate and to develop a model that could achieve 100% classification accuracy. The accuracy of the prediction depends on the properties or characteristics used throughout the prediction process.

Experimental results of the research verify that the Artificial Neural Network (ANN) classifier outperformed other algorithms such as NB, DT, and K-NN classifiers in the prediction of

cardiovascular disease. Despite this, the highest classification accuracy achieved by the algorithm was 91.75 percent, falling slightly short of the goal of 100% accuracy.

Several observations emerge as to why the algorithms did not provide a perfect accuracy rate. The model's imprecision could be related to the inclusion of missing or incomplete data as well as noise in the dataset utilised in the study, which could have reduced the accuracy of the results. Additionally, the dataset's characteristics, such as excessive dimensionality and unbalanced classes, were found to have an impact on the categorisation accuracy of the models. Furthermore, the True Negative rate was found to be higher than the True Positive rate in all the models, which might have affected the overall performance.

CHAPTER 6: CONCLUSION, RECOMMENDATIONS AND FUTURE WORK

6.1 Conclusion

This chapter examines the main research questions and its sub-questions listed in Chapter One to determine whether they were successfully addressed in the study and to determine whether the study's objective was met.

How can machine learning be used to predict cardiovascular disease to improve diagnosis treatment?

1. What data can be used to identify the features causing cardiovascular disease?
2. How can factors contributing to cardiovascular diseases be designed for the purpose of mining using UML tools?
3. Which algorithm has features which could be applied to identification and diagnostic support for cardiovascular disease?
4. Can the best data mining technique assist in predicting cardio-vascular disease?

6.2 Evaluation of the Research Questions

To answer the research questions, it was necessary to achieve the research objectives and obtain the findings, which ultimately enabled the research goal to be accomplished.

6.2.1 Research Question One

The first research question in this research is: “What data can be used to identify the features causing cardiovascular disease?” The objective designed for the research was:

To identify the features causing cardiovascular diseases from a dataset.

In Section 3.5 of the research, a thorough examination of the cardiovascular disease dataset from the Cleveland Data was conducted, with a specific focus on 13 variables and 303 patients. The results of this examination illustrate that the prevalence of cardiovascular disease is influenced by

a variety of factors, including cholesterol levels, thallium stress test results, exercise-induced angina, resting electrocardiographic results, and others. The findings of this review regarding the causes of cardiovascular disease were instrumental in informing the pre-processing and data mining stages of the study. Through an examination of the data and pinpointing key variables, this study was able to enhance its comprehension of the fundamental factors that impact cardiovascular disease, which in turn guided the research process and ensured that the most relevant data was analysed and utilised.

6.2.2 Research Question Two

The study's second research question was “How can factors contributing to cardiovascular diseases be designed for the purpose of mining using UML tools?” The research aimed to achieve the following objective to address the question:

To identify factors contributing to cardiovascular diseases be designed for the purpose of mining using UML tools.

In Section 3.9 of the research, the research life cycle was described. The data used in this study was stored in an Excel file and was subsequently converted into a machine-readable format to facilitate the research and modelling process. This conversion of data into a machine-readable format allowed for efficient manipulation and analysis of the data, thus enabling the research process to proceed smoothly.

6.2.3 Research Question Three

This study's third research question is “Which algorithm has features which could applied to identification and diagnostic support for cardiovascular disease?” To address this research question, the study strives to accomplish the following objective.

Implementing RO2 using data mining techniques namely Naïve Bayes, Decision Tree, K-Nearest Neighbour and ANN to interpret the results obtained.

The present study focused on assessing the effectiveness of various algorithms in the prediction of cardiovascular disease, with a specific emphasis on the use of the Accuracy, Precision, Recall, and F-Measure metrics. The research utilised both training and test datasets to develop and assess performance models. The findings confirm that among the four algorithms implemented, the Artificial Neural Network (ANN) classifier demonstrated the most superior performance.

After applying various metrics to the dataset and considering the dataset itself, the ANN method was deemed the most appropriate and effective model for predicting cardiovascular disease. The study clearly demonstrated that the ANN algorithm was the best performer among the algorithms implemented, in accordance with the dataset and the metrics applied.

6.2.4 Research Question Four

The last research question in this study is “Can the best data mining technique assist in predicting cardio-vascular disease?” To answer this question, the research aims to meet the objective, designed:

To determine the most efficient algorithm for cardiovascular disease.

In this research, the best features and the most suitable classification algorithms were identified and used in the design of a system for the prediction of cardiovascular disease, utilising the Cleveland dataset in Google Colaboratory and implemented using the Python programming language. The most effective algorithm, as determined in Section 6.2 of the study, was utilised in the final design of the system. This approach ensured that the system was optimised to effectively predict cardiovascular disease using the Cleveland dataset, and that the best features and algorithms were selected to achieve the highest performance.

6.3 Summary of Conclusions

The present research study aimed to analyse data and apply four different algorithms for the purpose of developing and creating a prediction system capable of detecting the number of patients with cardiovascular disease in the future. To accomplish this, data obtained from Kaggle UCI was utilised, modelling a predictive system to detect cardiovascular related diseases. The data, collected by the University of California, Irvine, comprised 303 instances and 14 attributes were

utilised for input and output to detect the presence of cardiovascular disease. The models were constructed and trained using Google collaborative software, an online platform for hosting and running such models.

The study conducted a thorough examination of the existing literature to obtain an extensive understanding of the subject matter and prior research in the realm of data mining and its utilisation in the healthcare sector, with particular attention given to cardiovascular disease. The practical section involved understanding the problem, visualising the data, and extracting meaningful insights through the implementation of predictive models. These stages were completed using Colab, which allowed for flexibility in working with the data.

Four classification models were built and evaluated, including Naive Bayes, Decision Tree, K-Nearest Neighbor, and Artificial Neural Networks (ANN). The performance of these algorithms was compared using standard metrics such as accuracy, precision, recall, and F-measure. Each of the four models exhibited satisfactory performance in predicting the prevalence of heart disease, with ANN being the most effective with 91.75% computation.

The study successfully met its objectives and demonstrated the potential of using machine learning and artificial intelligence in healthcare to build models that can aid organizations in decision-making and reducing assumptions related to cardiovascular diseases. However, it should be noted that the chosen model did not achieve 100% accuracy, leaving room for further refinement in the form of 8.25% miscomputed cases.

6.4 Limitations and Future work

Due to the availability and volume of data, the project's scope was restricted to the UCI Machine Learning Repository's data; yet, this was adequate to study, comprehend, and implement the prediction of the occurrence of heart attacks.

Further research in this niche could embark on performing different experiments with more classification and regression algorithms and to build the model that could predict exactly the type and state of the cardiovascular disease.

REFERENCES

Absar, N., Das, E.K., Shoma, S.N., Khandaker, M.U., Miraz, M.H., Faruque, M.R.I., Tamam, N., Sulieman, A. and Pathan, R.K., 2022, June. The efficacy of machine-learning-supported smart system for heart disease prediction. In *Healthcare* (Vol. 10, No. 6, p. 1137). MDPI.

Aggarwal, C.C. and Aggarwal, C.C., 2015. *Mining text data* (pp. 429-455). Springer International Publishing.

Ahmed, A. and Hannan, S.A., 2012. Data mining techniques to find out heart diseases: an overview. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *1*(4), pp.18-23.

Ahmad, P., Qamar, S. and Rizvi, S.Q.A., 2015. Techniques of data mining in healthcare: a review. *International Journal of Computer Applications*, *120*(15).

Alkhafaji, M.J.A., Aljuboori, A.F. and Ibrahim, A.A., 2020, June. Clean medical data and predict heart disease. In *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-7). IEEE.

Amin, M.S., Chiam, Y.K. and Varathan, K.D., 2019. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, pp.82-93.

Anbarasi, M., Anupriya, E. and Iyengar, N.C.S.N., 2010. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, *2*(10), pp.5370-5376.

Anitha, S. and Sridevi, N., 2019. Heart disease prediction using data mining techniques. *Journal of analysis and Computation*.

Anooj, P.K., 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), pp.27-40.

Avoiding heart attacks and strokes: don't be a victim-protect yourself. World Health Organization (WHO), World Self Medication Industry (WSMI), World Heart Federation (WHF), and International Stroke Society (ISS), 2005. Retrieved January 11, 2020, from:http://library.health.go.ug/publications/heart-disease/avoiding-heart-attacks-and-

strokesdon% E2%80%99t-be-victim-protect-yourself. Azevedo, A. I. R. L. & Santos, M. F. (2008). KDD, SEMMA

Ayon, S.I., Islam, M.M. and Hossain, M.R., 2020. Coronary artery heart disease prediction: A comparative study of computational intelligence techniques. *IETE Journal of Research*, pp.1-20.

Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A. and Bashir, K., 2019, January. Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619-623). IEEE.

Bell, J. and Waters, S., 2018. *Ebook: doing your research project: a guide for first-time researchers*. McGraw-hill education (UK).

Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R. and Delling, F.N., 2019. Heart disease and stroke statistics—2019 update: a report from the American Heart Association. *Circulation*, *139*(10), pp.e56-e528.

Berry, M.J. and Linoff, G.S., 2004. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Bruxvoort, D., 2009. THE HEALTHY WOMAN: A COMPLETE GUIDE FOR ALL AGES. *Feminist Collections: A Quarterly of Women's Studies Resources*, *30*(3), pp.20-21.

Busetto, L., Wick, W. and Gumbinger, C., 2020. How to use and assess qualitative research methods. Neurological Research and practice, 2, pp.1-10.

Byrne, J., Eksteen, G. and Crickmore, C., 2016. Cardiovascular disease statistics reference document. *The Heart and Stroke Foundation of South Africa*.

Chang, C.L. and Chen, C.H., 2009. Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, *36*(2), pp.4035-4041.

Chaurasia, V. and Pal, S., 2014. Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol*, *2*, pp.56-66.

Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H. and Lin, E.J., 2011, September. HDPS: Heart disease prediction system. In *2011 computing in Cardiology* (pp. 557-560). IEEE.

Chung, K., Yoo, H. and Choe, D.E., 2020. Ambient context-based modeling for health risk assessment using deep neural network. *Journal of Ambient Intelligence and Humanized Computing*, *11*, pp.1387-1395.

Curiac, D.I., Vasile, G., Banias, O., Volosencu, C. and Albu, A., 2009, June. Bayesian network model for diagnosis of psychiatric diseases. In *Proceedings of the ITI 2009 31st international conference on information technology interfaces* (pp. 61-66). IEEE.

D'Souza, A., 2015. Heart disease prediction using data mining techniques. *International Journal of Research in Engineering and Science (IJRES) ISSN (Online)*, pp.2320-9364.

Damtew, A., 2011. Designing a predictive model for heart disease detection using data mining techniques (Doctoral dissertation, Addis Ababa University).

Dangare, C. and Apte, S., 2012. A data mining approach for prediction of heart disease using neural networks. *International Journal of Computer Engineering and Technology (IJCET)*, *3*(3).

Dangare, C.S. and Apte, S.S., 2012. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10), pp.44-48.

Das, R., Turkoglu, I. and Sengur, A., 2009. Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), pp.7675-7680.

David B. and Belcy A. 2018. "Heart disease prediction using data mining techniques." *Journal on Soft Computing*, vol. 9, no. 1, pp. 1824-1830, October 2018

Aha.W.David,"HeartDiseaseDataset,"1988,*Url:http://archive.ics.uci.edu/ml/datasets/Heart+Disease*

Deekshatulu, B.L. and Chandra, P., 2013. Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia technology*, *10*, pp.85-94.

Dehkordi, S.K. and Sajedi, H., 2019. Prediction of disease based on prescription using data mining methods. *Health and Technology*, *9*, pp.37-44.

Everest, T., 2014. Resolving the qualitative-quantitative debate in healthcare research. *Medical Practice and Reviews*, *5*(1), pp.6-15.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), pp.37-37.

Kolling ML, Furstenau LB, Sott MK, Rabaioli B, Ulmi PH, Bragazzi NL, Tedesco LPC. Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development. Int J Environ Res Public Health. 2021 Mar 17;18(6):3099. doi: 10.3390/ijerph18063099. PMID: 33802880; PMCID: PMC8002654.

Hajar R. Risk Factors for Coronary Artery Disease: Historical Perspectives. Heart Views. 2017 Jul-Sep;18(3):109-114.doi:10.4103/HEARTVIEWS.HEARTVIEWS_106_17. PMID: 29184622; PMCID: PMC5686931.

Han, J. and Kamber, M., 2006. *Data mining: Concepts and techniques*. 2nd. University of Illinois at Urbana Champaign: Morgan Kaufmann.

Hand, D. J. Mannila, H. and Smyth, P. 2001. *Principles of data mining* (*adaptive computation and machine learning*). MIT Press, 2001

Harper, R., Flammia, S.T. and Wallman, J.J., 2020. Efficient learning of quantum noise. *Nature Physics*, *16*(12), pp.1184-1188.

Hong, W.C. and Hong, W.C., 2020. Hybridizing Meta-heuristic Algorithms with CMM and QCM for SVR's Parameters Determination. *Hybrid Intelligent Technologies in Energy Demand Forecasting*, pp.69-133.

Khatibi, V. and Montazer, G.A., 2010. A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. *Expert Systems with Applications*, *37*(12), pp.8536-8542.

Kerssens, N., 2019. De-agentializing data practices: The shifting power of metaphor in 1990s discourses on data mining. *Journal of Cultural Analytics*, pp.1-26.

Keshtegar, B., Heddam, S. and Hosseinabadi, H., 2019. The employment of polynomial chaos expansion approach for modeling dissolved oxygen concentration in river. *Environmental Earth Sciences*, *78*, pp.1-18.

Khan, R.U., Hussain, T., Quddus, H., Haider, A., Adnan, A. and Mehmood, Z., 2019, January. An intelligent real-time heart diseases diagnosis algorithm. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-6). IEEE.

Kim, J., Washio, T., Yamagishi, M., Yasumura, Y., Nakatani, S., Hashimura, K., Hanatani, A., Komamura, K., Miyatake, K., Kitamura, S. and Tomoike, H., 2004. A novel data mining approach to the identification of effective drugs or combinations for targeted endpoints—application to chronic heart failure as a new form of evidence-based medicine. *Cardiovascular drugs and therapy*, *18*, pp.483-489.

Krishnan, S. and Geetha, S., 2019, April. Prediction of Heart Disease Using Machine Learning Algorithms. In 2019 1st international conference on innovations in information and communication technology (ICIICT) (pp. 1-5). IEEE

Larose, D.T. and Larose, C.D., 2014. *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.

Liao, S.C., and Lee, I.N., 2002. Appropriate medical data categorization for data mining classification techniques. *Medical informatics and the Internet in medicine*, *27*(1), pp.59-67.

Mahoney, J. and Goertz, G., 2006. A tale of two cultures: Contrasting quantitative and qualitative research. *Political analysis*, *14*(3), pp.227-249.

Massachusetts Institute of Technology School of Management and Brown, S., 2021. *Machine learning, explained*. MIT Sloan management.

Meena, K., Tayal, D.K., Gupta, V. and Fatima, A., 2019. Using classification techniques for statistical analysis of Anemia. *Artificial intelligence in medicine*, *94*, pp.138-152.

Merghadi, A., Yunus, A.P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D.T., Avtar, R. and Abderrahmane, B., 2020. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, *207*, p.103225.

Moriarty, J., 2011. Qualitative methods overview.

Mosca, L., Banka, C.L., Benjamin, E.J., Berra, K., Bushnell, C., Dolor, R.J., Ganiats, T.G., Gomes, A.S., Gornik, H.L., Gracia, C. and Gulati, M., 2007. Evidence-based guidelines for cardiovascular disease prevention in women: 2007 update. Circulation, 115(11), pp.1481-1501

Mozaffarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M., De Ferranti, S., Després, J.P., Fullerton, H.J., Howard, V.J. and Huffman, M.D., 2015. Heart disease and stroke statistics—2015 update: a report from the American Heart Association. circulation, 131(4),

pp.e29-e322.Olson, D.L. and Delen, D., 2008. *Advanced data mining techniques*. Springer Science & Business Media.

Muhammad, G., Naveed, S., Nadeem, L., Mahmood, T., Khan, A.R., Amin, Y. and Bahaj, S.A.O., 2023. Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease using K Nearest Neighbor Algorithm: A Robust Approach. IEEE Access.

Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications* (pp. 108-115). IEEE.

Patil, S.B. and Kumaraswamy, Y.S., 2009. Intelligent and effective heart attack prediction system using data mining and artificial neural network. *European Journal of Scientific Research*, *31*(4), pp.642-656.

Patterson, J. and Gibson, A., 2017. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.".

Premsmith, J. and Ketmaneechairat, H., 2021. A predictive model for heart disease detection using data mining techniques. *Journal of advances in Information Technology*, *12*(1).

Ramageri, B.M., 2010. Data mining techniques and applications. *Indian journal of computer science and engineering*, *1*(4), pp.301-305.

Ramotra, A.K., Mahajan, A., Kumar, R. and Mansotra, V., 2020. Comparative analysis of data mining classification techniques for prediction of heart disease using the weka and SPSS modeler tools. In *Smart Trends in Computing and Communications: Proceedings of SmartCom 2019* (pp. 89-97). Springer Singapore.

Razali, A.M. and Ali, S., 2009. Generating treatment plan in medicine: A data mining approach. *American Journal of Applied Sciences*, *6*(2), p.345.

Saravanakumar, S., and Rinesh S. 2014. “Effective heart disease prediction using frequent feature se-lection method,” International *Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 1, pp. 2767–2774, 2014.

Shafique, U., Majeed, F., Qaiser, H. and Mustafa, I.U., 2015. Data mining in healthcare for heart diseases. *International Journal of Innovation and Applied Studies*, *10*(4), p.1312.

Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6), pp.1-6.

Shouman, M., Turner, T. and Stocker, R., 2012, March. Using data mining techniques in heart disease diagnosis and treatment. In *2012 Japan-Egypt Conference on Electronics, Communications and Computers* (pp. 173-177). IEEE.

Sinaga, L.M. and Suwilo, S., 2020. Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining. In *IOP Conference Series: Materials Science and Engineering* (Vol. 725, No. 1, p. 012106). IOP Publishing.

Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*, *3*(6), pp.2385-2392.

Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), pp.43-48.

Subhadra, K. and Vikas, B., 2019. Neural network based intelligent system for predicting heart disease. *International Journal of Innovative Technology and Exploring Engineering*, 8(5), pp.484-487.

Tarawneh, M. and Embarak, O., 2019. Hybrid approach for heart disease prediction using data mining techniques. In *Advances in Internet, Data and Web Technologies: The 7th International Conference on Emerging Internet, Data and Web Technologies (EIDWT-2019)* (pp. 447-454). Springer International Publishing.

Tayefi, M., Esmaeili, H., Karimian, M.S., Zadeh, A.A., Ebrahimi, M., Safarian, M., Nematy, M., Parizadeh, S.M.R., Ferns, G.A. and Ghayour-Mobarhan, M., 2017. The application of a decision tree to establish the parameters associated with hypertension. *Computer methods and programs in biomedicine*, *139*, pp.83-91.

Tomar, D. and Agarwal, S., 2013. A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), pp.241-266.

Two Crows Corporation (2005). *Introduction to Data Mining and Knowledge Discovery*. Third Edition. Two Crows Corporation. 10500 Falls Road, Potomac, USA

Vijayarani, S. and Muthulakshmi, M., 2013. Comparative analysis of bayes and lazy classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(8), pp.3118-3124

WikimediaCommons,https://commons.wikimedia.org/wiki/File:Diagram_of_the_human_heart.svg Accessed on 04/12/2022

Witten, I.H. and Frank, E., 2002. Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, 31(1), pp.76-77

APPENDIX