Department of Computer Science

BSc (Hons) Computer Science (Artificial Intelligence)

Academic Year 2022 – 2023

Investigating and Optimising Machine Learning Algorithms for Clinical Application

Tanveer Dalal

2010357

A report submitted in partial fulfilment of the requirements for the degree of Bachelor of Science

Brunel University London
Department of Computer Science
Uxbridge
Middlesex
UB8 3PH
United Kingdom
T: +44 1895 203397
F: +44 (0) 1895 251686

Abstract

Heart Disease is a leading cause of death worldwide, with risk factors including chest pain, breathing difficulty, chest congestion, high cholesterol, anxiety, and many more (Donovan, 2020).

This research was conducted to address a gap in previous studies, and proposes an algorithm that measures the importance of significant features that contribute to predicting heart disease. The study focuses on predicting heart disease based on the scores of these significant features using various machine learning algorithms.

This experiment was performed on the Kaggle dataset, widely used by data scientists and this research yielded the highest confidence of 98% in predicting heart disease.

This study managed to provide a significant contribution in computing the strength scores with significant predictors in heart disease prediction. From the evaluation results, this study achieved the highest confidence score by using confusion matrices, f1 score, precision, recall and accuracy.

Moreover, the web page is designed using the Shiny.R package which is a user-friendly application that can be easily used by healthcare professionals for predicting the heart disease of their patients.

Acknowledgements

First, I express my gratitude to Lord Vishnu for His Continuous Blessings and Guidance.

My grandfather, Sube Singh Sangwan, is my role model. During the toughest moments of my life, he stood by me. His unwavering belief in me and constant encouragement has motivated me to give my best.

The unwavering support, affection, and motivation provided by my family throughout my life has been a source of inspiration for me to make positive contributions to the world. Their tireless efforts and dedication continue to inspire me to strive for excellence.

I am extremely grateful to my supervisor, Dr Giuseppe Destefanis, for his invaluable guidance and support throughout my project.

I certify that the work presented in the dissertation is my own unless referenced.

Signature Tanveer

Date 31/03/2023

Total Words: 11438

Table of Contents

List of Figure

List of Tables

Chapter 1 - Introduction

This chapter will provide an overview of the problem concerning heart disease from which thousands of people worldwide die in a minute. With the advancements in machine learning and artificial intelligence, it can identify patterns and risk factors of heart disease among people. Having a well-defined project aim and set of objectives, we can ensure that our efforts are focussed for developing an effective heart disease prediction that can improve early detection.

1.1 Introducing the Problem

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United Kingdom. On the basis of statistics, more than 160,000 deaths occur each year with an average of 460 deaths each day or one every three minutes in the UK (British Heart Foundation, 2023). Heart disease refers to many conditions like coronary heart disease, heart arrhythmias, cerebrovascular disease, and other conditions. In the UK, hospitals are receiving more than 100,000 admissions each year due to heart attack. The most important behavioural risk factors include a poor diet, lack of exercise, obesity and smoking (Mayo Clinic, 2023). Heart diseases are becoming more and more dangerous for our society and predicting the heart disease before it could lead towards death can save millions of lives. Collecting the patient's data can help us to understand the cause of heart disease and can be used to predict the future. Whereas, machine learning algorithms can help in predicting the risk of heart disease by examining patient data and identifying patterns that can indicate potential health issues. Moreover, this project will predict the outcome of the patients who are at high risk of developing heart disease and can be prevented by reducing those risks.

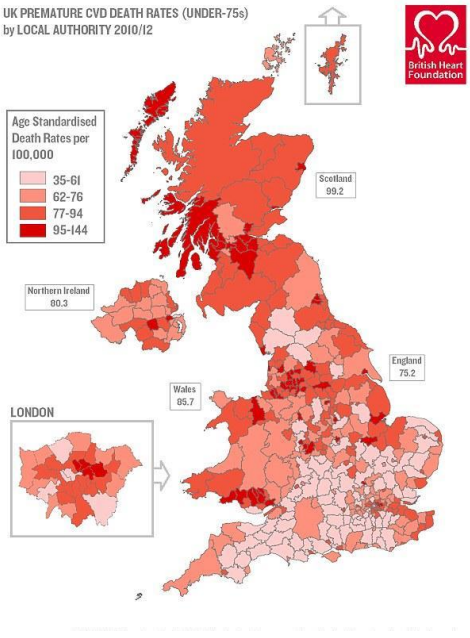


Figure 1: UK Map

During research, there are large volumes of data with patient records available on different platforms like UCL and Kaggle. It gives an opportunity to access and develop a technology that can predict an outcome through analysing. Through these data, various kinds of research can be carried out to diagnose the patients behaviour correctly and using these research with computer technologies can prevent heart disease. For the healthcare industry, machine learning is particularly valuable because it can analyse large amounts of data and identify patterns within electronic health records which is nearly impossible to find manually. Machine learning has the potential to learn by analysing the records of patients and is capable enough to predict the future outcome (Barth, 2022).

1.2 Aims and Objectives

The aim of this project is to identify the heart disease using multiple different machine learning algorithms that can accurately predict the risk of heart disease of patients based on their factors (Age, Gender, Chest Pain, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Electrocardiographic, heart rate, and many more). Based on the accuracy of these models, this project will compare the better result of their efficiency while predicting the patient's output and the best model will be displayed on the Shiny Web application where healthcare professionals can add the details of their patients to get results on heart disease.

For achieving our ultimate aim, each objective has been designed to build upon the previous one: -

1. Conducting background research to understand the various approaches and techniques used for heart disease prediction in the healthcare industry, as well as the types of data that can meet the requirements of such predictions, is crucial for developing a robust and effective heart disease prediction.
2. Choosing the correct methodology is essential before starting a heart disease prediction project. There are different methodologies that can be chosen for this project like CRISP-DM, KDD and SEMMA. The methodology is dependent on factors such as the size and complexity of the project.
3. Collect relevant dataset that can be used for predicting the outcome which is based on the factors (chest pain, blood pressure, thal, slope, and many more) and prepare the data to build multiple models.
4. Evaluating and selecting an appropriate multiple different machine learning algorithms. The choice of the algorithms will depend on the factors such as the nature of the data, the required performance metrics and the complexity of the problem. It is necessary to go through multiple algorithms to determine the most suitable approach.
5. Developing a user-friendly interface using Shiny.R package which is a crucial step for being used by healthcare professionals. The interface should be easy to navigate, with clear instruction for inputting patient's data and receiving a predicted outcome.
6. Evaluate all the models using various metrics such as confusion matrices, F1 score, precision, recall, and accuracy. Comparing these models and selecting the most appropriate model for heart disease prediction.

By achieving these objectives, we aim to develop a heart disease prediction which can accurately predict the future of heart risks and can assist healthcare professionals by identifying the patients that are at risk of developing heart disease.

1.3 Project Approach

This section will justify how the project is undertaken by going through with different methods. The heart disease prediction project will involve the following steps which is shown in **Figure 2**.

- **Background Research -** In this research, heart disease prediction will be discussed in detail. The importance of predicting heart disease accurately is paramount, as it can potentially save lives and reduce healthcare costs. Additionally, previous research and studies related to heart disease need to be discussed, and their strengths and limitations will be analysed.
- **Methodology -** The methodology chosen for this project is the Cross-Industry Standard Process for Data Mining (CRISP-DM). This is considered an industry-standard approach due to its six-phase structure, which accurately describes the data science life cycle. Various methodologies are compared before deciding on the CRISP-DM approach as the most appropriate for the project.
- **Design -** In this stage, the research findings from the background research and methodology will be merged. The aim is to conduct further research to make a decision on RStudio and select the relevant open-source libraries to develop multiple machine learning algorithms and gather heart disease data.
- **Implementation -** The implementation of all the machine learning models will be developed. The developed models will then be integrated into the application to make predictions based on user input.
- **Testing -** The functionality of the application will be tested using black box testing and snapshot-based tests to ensure that all the requirements are met. The best trained model will be integrated into the application for predicting heart disease.
- **Evaluation -** To assess the performance of the models, various metrics such as confusion matrices, f1 score, precision, recall, and accuracy will be used. These metrics will help in comparing and selecting the best model for the heart disease prediction. It will provide detailed analysis and interpretation of the results obtained from the models, highlighting their strengths and weaknesses. It will also discuss the impact of hyperparameter tuning and cross-validation techniques on the models accuracy.

1.4 Dissertation Outline

The structure of the dissertation is given below:

- **Chapter 1: Introduction**
  - This chapter will provide an overview of the problem concerning heart disease from which thousands of people worldwide die in a minute. A set of objectives are listed and provided with a project approach.
- **Chapter 2: Background**
  - This chapter will provide an importance of heart disease prediction and previous research will be shown.
- **Chapter 3: Methodology**
  - This chapter will explain why the CRISP-DM methodology is selected and how this methodology is suitable to complete the objectives of this project.
- **Chapter 4: Design**
  - This chapter will introduce the platform in which this project is designed. The tools and packages are listed with the images of design and diagram.
- **Chapter 5: Implementation**
  - This chapter will walk through the implementation of all machine learning models and display the heart disease prediction in the deployment stage.
- **Chapter 6: Testing**
  - This chapter will carry out the various tests on the heart disease prediction such as integration testing, and unit testing.
- **Chapter 7: Evaluation**
  - This chapter will evaluate all the models using confusion matrices, f1 score, precision, recall, and accuracy and compare their values to find the best model.
- **Chapter 8: Conclusions**
  - This chapter will reflect the whole project and summarise the work for future.

Chapter 2 - Background

In this chapter, the Heart Disease Prediction project will be explained in detail. The significance of preventing heart disease will be highlighted, and the importance of addressing this issue will be discussed. The chapter will also compare and contrast previous research and resources.

2.1 What Exactly is Heart Disease?

Heart disease is a major problem in the UK and globally, with heart attacks being the third leading cause of death in the UK. Many people are dying every minute due to heart disease, and it is the leading cause of death worldwide. The number of new cases of heart disease is also on the rise, and it has a significant impact on the quality of life of those affected. If these trends continue, a large percentage of men and women will develop heart disease and it's estimated that two out of three men and one out of two women will develop heart disease in their life-times (Whyte & McGraw, 2023).

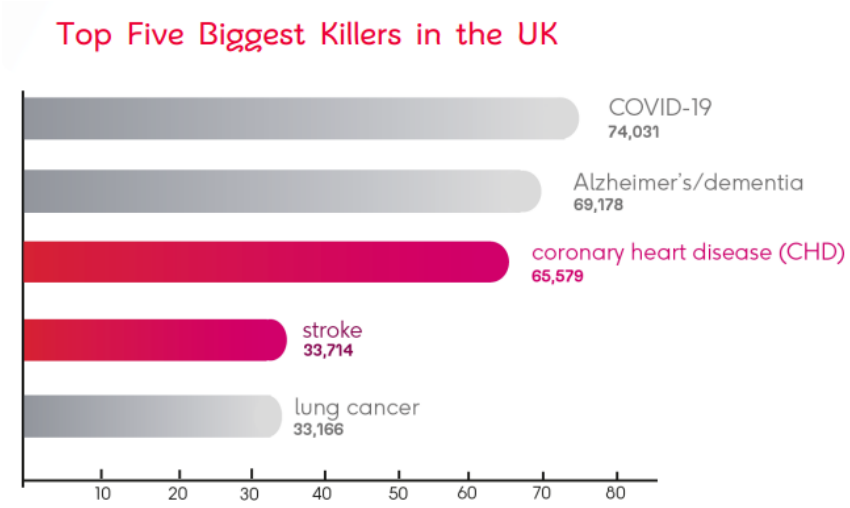


Figure 2: Rate of Heart Disease in UK

It may seem like heart attacks occur randomly or out of nowhere. We all have heard of someone seemingly in perfect health who died from a heart attack (Whyte & McGraw, 2023). For most people, though, a heart attack is not random. Heart damage typically occurs over many years, a result of a combination of genetics and lifestyle (Whyte & McGraw, 2023). Recent research suggests that for most people, genetic factors make up less than 20 percent of the risk of heart disease. The rest, around 80 percent, is caused by lifestyle which indicates the quality and quantity of our food, the amount of daily physical activity we engage in, the quality of our sleep, and how much chronic stress we experience (Whyte & McGraw, 2023).

2.1.1 How Does the Heart Function?

The heart has four chambers. They're similar on the left and right. The two on the top are the atria, and two on the bottom are ventricles (Katzenstein & Pinã Ileana, 2007). One of the reasons the heart is so important for life is the role it plays in delivering oxygen. Oxygen-rich blood from your lungs flows to the left atrium, then to the left ventricle, which pumps it out to your body. Blood returns to the right atrium, then the right ventricle, which sends it back to your lungs for more oxygen. It's an intricate, fine-tuned circuit that doesn't like disruptions (Katzenstein & Pinã Ileana, 2007).

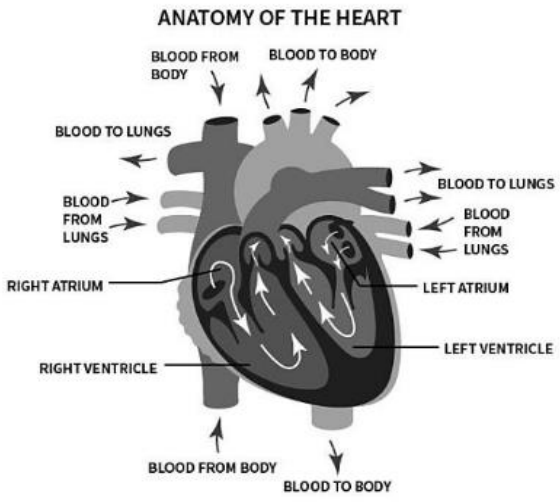


Figure 3: Function of the Heart

Although the heart may appear large, it is actually about the size of a list and weighs between ten and twelve ounces (Katzenstein & Pinã Ileana, 2007). Despite its relatively small size, the heart is a vital organ with immense power. It beats about one hundred thousand times per day, pumping more than two thousand gallons of blood daily, at a rate of five to six quarts per minute (Katzenstein & Pinã Ileana, 2007). Additionally, the heart plays a crucial role in the cardiovascular system.

2.2 Importance of Heart Disease Prediction project

The heart disease prediction project has the potential to predict an individual's risk of developing heart disease in the future. This project has the potential to save countless lives by enabling doctors to identify individuals at risk of heart disease and doctors can provide medications and precautions according to it.

2.2.1 Detecting early stages of heart disease

The heart disease prediction project uses various factors such as cholesterol levels, high blood pressure, chest pain, and many more to detect the early stages of heart disease. By utilising machine learning models, this project can predict the risk of an individual developing heart disease in the future and identify those who are at high risk. Early detection can prevent heart disease or delay the onset of heart disease.

2.2.2 Preventing heart disease before it lead to risk

There are different types of heart conditions that can lead to suffering and instant death. Detecting the early stages of risk factors is crucial to managing the condition effectively. Healthcare professionals can provide appropriate precautions to prevent the progression of the disease and reduce the risk of complications. Patients can also play a crucial role in preventing heart disease by making lifestyle changes such as improving diet and increasing their physical activity levels. By understanding their conditions and taking appropriate measures, patients can prevent further damage to their heart before it leads to significant risk (Katzenstein & Pinã Ileana, 2007).

Figure 4: Prevent Heart Disease

2.2.3 Prevent people from unnecessary medications or surgery

This project can prevent patients from undergoing unnecessary medications and surgery. By detecting the early stages of heart disease, healthcare professionals can provide appropriate treatment plans and health exercises, which may reduce the need for medications or invasive procedures such as surgery. By preventing unnecessary medications or surgery, patients can avoid potential side effects of these treatments and improve their overall quality of life.

2.2.4 Health Impact on Public

It can promote healthy lifestyle modifications and encourage individuals to take part in regular health check-ups. By detecting the early stages of heart disease, healthcare professionals can provide personalised recommendations for diet, exercise, and medications. This project can have a significant impact on healthcare costs by reducing the number of patients who require surgeries and medications. By promoting this project, it can improve public health outcomes, reduce healthcare costs, and improve the overall quality of care (Katzenstein & Pinã Ileana, 2007).

2.2.5 Educating people through heart disease prediction project

This project can be a crucial tool for educating younger generations about the risk factors related to heart disease. By understanding the factors, individuals can take steps to reduce their risk and improve their overall health. Health professionals can stay up-to-date with the latest advances in heart disease prediction. Overall, this project can be a valuable tool for raising awareness and encouraging individuals to take proactive steps to reduce their risk of heart disease.

2.3 Previous studies and methods to prevent heart disease

Several individuals have conducted experiments by creating various machine learning models. These models were then analysed, and the one with the higher accuracy was selected. However, these researchers did not use any techniques to improve the accuracy of the model's performance, nor did they deploy any applications that could be utilised by healthcare professionals.

First team of researchers collaborated to develop a new method for predicting heart disease. Their objective was to evaluate the effectiveness of their algorithm by utilising important features that are associated with heart disease prediction. They conducted a study using Associative Rule Mining to predict heart disease based on the scores of these significant features (Yazdani, Varathan, Chiam, Malik & Wan Ahmad, 2021). They used the UCL dataset in their experiments and achieved a very high confidence score of 98% in predicting heart disease (Yazdani, Varathan, Chiam, Malik & Wan Ahmad, 2021).

Second team of researchers conducted a study on various machine learning models, including Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, K-nearest neighbour, Artificial neural network and Random Forest (Uddin, Khan, Hossain & Moni, 2019). Among these models, the Random Forest algorithm showed superior accuracy comparatively which is 53% (Uddin, Khan, Hossain & Moni, 2019). To evaluate the performance of their model, they used Receiver Operating Characteristic (ROC) curves for diagnostic test evaluation. This technique involves plotting the true positive rate against the false positive rate at various threshold settings (Uddin, Khan, Hossain & Moni, 2019). It is commonly used in medical diagnosis to assess the accuracy of a diagnostic test. By using ROC curves, the researchers were able to analyse the performance of their model and determine its effectiveness in predicting heart disease.

Rawat is another researcher who used various machine learning models like SVM, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, LightGBM, and XGBoost (Rawat, 2021). All of his project code was implemented in Python. Unlike the previous researchers, Rawat used different visualisation methods and plotted all the factors against the target variable. The highest accuracy for the test set was achieved using Logistic Regression and SVM, which were both equal to 80.32% (Rawat, 2021). It's important to note that all the algorithms were implemented with default parameters only. Moreover, he used the confusion matrix which allows to evaluate the performance of the model. By analysing the confusion matrix, one can gain insights into the model's strengths and weaknesses and identify areas of improvement.

All the researchers conducted studies on heart disease prediction using different approaches and machine learning algorithms. Although the studies conducted by the researchers provide valuable insights into the use of machine learning for heart disease prediction.

2.4 Related Research on this Project

This section aims to present a study that identifies gaps in previous research on heart disease prediction and proposes solutions to address these gaps. By building on the work of previous researchers and incorporating various techniques, this study aims to develop more effective and reliable tools for predicting heart disease. The research will investigate previous projects and propose a way to achieve the project's objectives, which include improving the accuracy of heart disease prediction using machine learning algorithms and developing user-friendly applications for healthcare-professionals.

2.4.1 Gaps in previous research

The research mentioned in section 2.3 indicates that various researchers have worked on heart disease prediction by either comparing the accuracy of different machine learning algorithms or by using a single algorithm and enhancing its accuracy through techniques such as cross-validation. However, hyperparameter tuning is an important step in improving the performance of machine learning algorithms. By optimising these techniques, the performance of the model can be challenging for healthcare professionals to integrate these models into their practice. Furthermore, many studies have developed accurate machine learning models for heart disease prediction, they often lack practicality for healthcare professionals. This is because these models are typically designed for research purposes and require a high level of technical expertise to use. Without a user-friendly application, it can be challenging for healthcare professionals to integrate these models into their practice.

2.4.2 Solution to cover gaps

Previous research on heart disease prediction projects has often lacked a brief comparison between machine learning algorithms. While individual studies have focused on the accuracy of specific algorithms, there has been little effort to compare the performance of multiple algorithms in a standardised way. However, such a comparison could be highly informative for researchers and clinicians seeking to improve heart disease prediction. By examining the accuracy of different algorithms, we can gain a better understanding of which algorithms may be most effective in identifying patients at risk for heart disease. Additionally, we can investigate the factors that influence algorithm performance, such as the size of the training dataset and complexity of the model.

Another factor that can enhance the performance of the model is by using hyperparameter tuning and cross-validation techniques together. In this case, this study will perform this technique and the final results will be compared among models. The best model will be used in the project which can predict heart disease.

Finally, a user-friendly application will be designed to enable healthcare professionals to easily identify patients with heart disease. The application will be designed with a user interface that is easy to navigate, and it will provide accurate and reliable results for predicting heart disease in patients.

a) Machine Learning Technique

Heart disease prediction requires a machine learning model to be trained on a dataset that is labelled to indicate which patients have heart disease and which ones do not. This approach, known as supervised learning, involves training the model on labelled data where the target variable is already known and provided during training (Mehta, 2022). This trained model can then be used to make predictions on new, unseen data. In contrast, unsupervised learning involves training a model on an unlabelled dataset where the target variable is unknown, and the model attempts to find patterns or structure in the data without prior knowledge of what the output should be (Mehta, 2022). However, since we already have labelled data for heart disease prediction, supervised learning is the appropriate approach for this task.

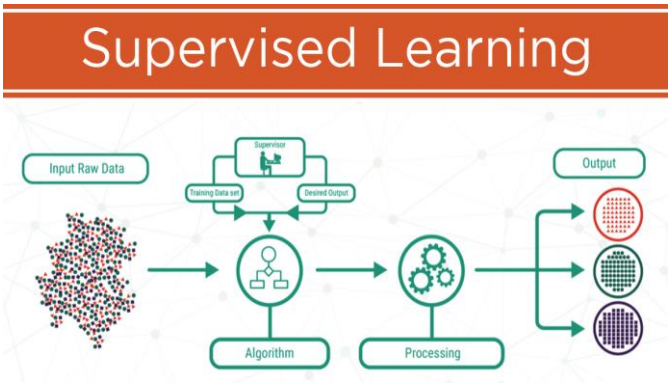


Figure 5: Supervised Learning

b) Supervised Learning

This study aims to predict the occurrence of heart disease using supervised machine learning techniques. The input features include age, gender, cholesterol levels, blood pressure, chest pain, and other factors. The two types of supervised learning techniques, regression and classification, will be utilised in this study (Mehta, 2022). Linear regression will be used for regression, and decision tree, support vector machine, naive bayes, and XGBoost will be used for classification (Mehta, 2022). Hyperparameter tuning and cross-validation will be applied to these algorithms to improve their performance. The analysis of these techniques and their results will be discussed. In addition, the performance of each algorithm will be evaluated by generating a confusion matrix, which shows the number of true positive, true negative, false positive, and false negative results (Suresh, 2021). Using these matrices, the precision, recall, f1-score, and accuracy of each algorithm will be calculated and analysed (Suresh, 2021).

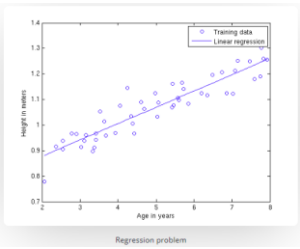


Figure 6: Regression

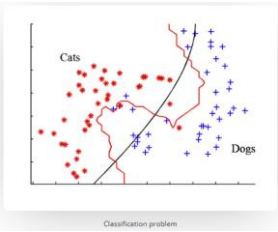


Figure 7: Classification

c) Application Development using Shiny R

Shiny is an R package developed by RStudio that provides a framework for building interactive web applications using R (Shiny, 2012). In the context of heart disease prediction, healthcare professionals can input patient data (age, gender, cholesterol levels, blood pressure, chest pain, etc.) and get a prediction on whether the patient is at risk of heart disease. The app is designed to use any of the machine learning models discussed earlier that have been trained on a heart disease dataset. Overall, a Shiny app can provide a user-friendly interface for healthcare professionals to predict heart disease risk.

Chapter 3 - Methodology

This chapter will explain the data science methodology will be choosed to achieve the aim and objectives listed in Chapter 1. In previous research, no efforts have been made to improve the model's accuracy or develop a practical application for healthcare professionals. Therefore, this study will select an appropriate data science methodology by exploring various methodologies and identifying the one that best suits the research requirements.

3.1 Data Science Methodologies

Data science methodology is typically used for projects that involve the analysis of data to gain insights and make predictions, while software development is used for building software systems (Sharma, 2022). In case of this research, the focus is on analysing and predicting heart disease using machine learning techniques, which aligns more closely with data science methodology. The goal is to build machine learning algorithms that can be used by healthcare professionals to identify patients at risk of heart disease. Therefore, data science methodology is a more appropriate choice for this project.

With the growing use of data science and machine learning in various industries, including healthcare, finance, and e-commerce, organisations are increasingly deploying data science models into larger software systems (IntelliPaat, 2023). This is because data science models can provide valuable insights and predictions that can be integrated into software systems to improve their functionality and decision-making capabilities.

3.1.1 What is CRISP DM?

CRISP DM (Cross-Industry Standard Process for Data Mining) is a widely used data mining process that provides a structured approach to planning, building, implementing, and evaluating data mining solutions (Hotz, 2023).

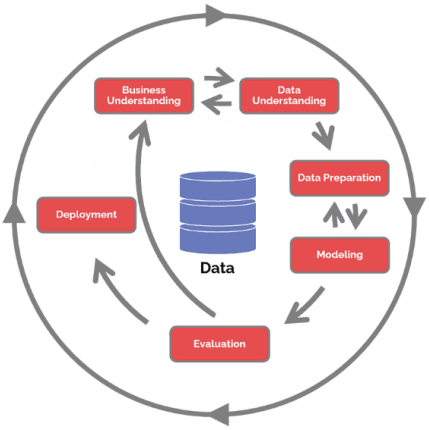


Figure 8: CRISP-DM Workflow

It consists of six phases that naturally describe the data science life cycle:

1. Business Understanding- What the project aims to achieve.
2. Data Understanding- Understanding the available or required data, and to ensure that it is clean and reliable
3. Data Preparation- Preparing the data for modelling, which includes selecting relevant features and cleaning the data
4. Modelling- The appropriate modelling techniques are selected and applied to the prepared data

5. Evaluation- The performance of the models is evaluated to determine which model best meets the business objectives
6. Deployment- Deploying the chosen model and ensuring that stakeholders can access the results

3.1.2 What is KDD?

KDD stands for Knowledge Discovery in Databases. It is the process of discovering useful and unknown knowledge from large volumes of data stored in databases or other digital repositories (Hotz, 2023). KDD involves a combination of data mining, machine learning, and statistical techniques to extract patterns and knowledge from data. The main goal of KDD is to turn raw data into actionable knowledge that can be used to make informed decisions (Hotz, 2023).

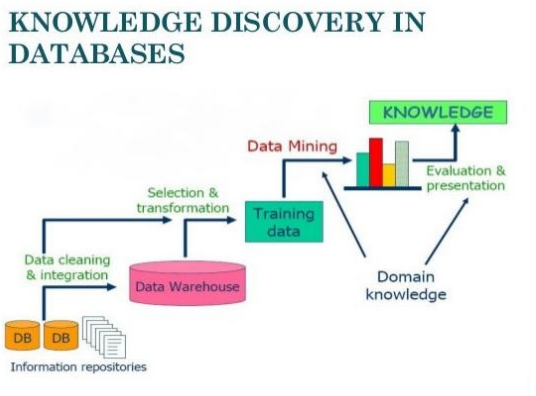


Figure 9: KDD Workflow

The following are the steps involved in the KDD process:

1. Selection- Selecting the relevant data that needs to be analysed. The data may be collected from multiple sources and can be in various formats such as structured or unstructured.
2. Pre-Processing- Data needs to be cleaned, integrated, and transformed into a suitable format for further analysis. This involves tasks such as removing duplicates, handling missing values, normalisation, and data aggregation.
3. Transformation- The pre-processed data is transformed into a suitable format that can be used for analysis.
4. Data Mining- This step involves applying various data mining techniques such as clustering, classification, association rule mining, and outlier detection to identify patterns, and trends in the data.
5. Evaluation- Visualising the data and identifying the insights that can be used to make informed decisions.

3.1.3 What is SEMMA?

SEMMA is a data mining methodology developed by SAS (Statistical Analysis System) Institute, which stands for Sample, Explore, Modify, Model, and Assess. It provides a structural approach for conducting data analysis and predictive modelling projects (Hotz, 2023).

1. Sample- In this step, a smaller and representative subset of the dataset is selected for building the model. The aim is to identify the relevant independent and dependent variables that can affect the outcome. The collected data is then categorised into two parts: preparation and validation (Hotz, 2023).
2. Explore- Data exploration and visualisation to identify patterns, trends, and outliers in the data

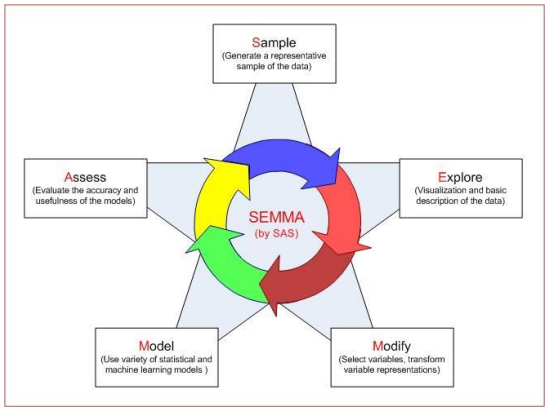


Figure 10: SEMMA Workflow

3. Modify- Data pre-processing techniques are applied to clean, transform, and prepare the data for modelling
4. Model- Selecting an appropriate modelling technique and building predictive models using the prepared data.
5. Assess- The performance of the models is evaluated using appropriate evaluation measures, and the best model is selected for deployment

3.2 The Chosen Methodology

The CRISP-DM approach was selected for this research project because it is highly recognised and has become more popular than KDD and SEMMA for various reasons. The CRISP-DM methodology is more flexible and iterative, enabling modifications and adjustments throughout the process, it also provides a well-structured and defined framework of steps to guid the project from initiation to completion, which is advantageous for new or less experienced data scientists (Hotz, 2023). Additionally, CRISP-DM is supported by a community of users, providing access to resources and best practices that can ensure the success of the project.

3.2.1 Phases and Tasks

The following are the steps that can be used for heart disease prediction using CRISP-DM:

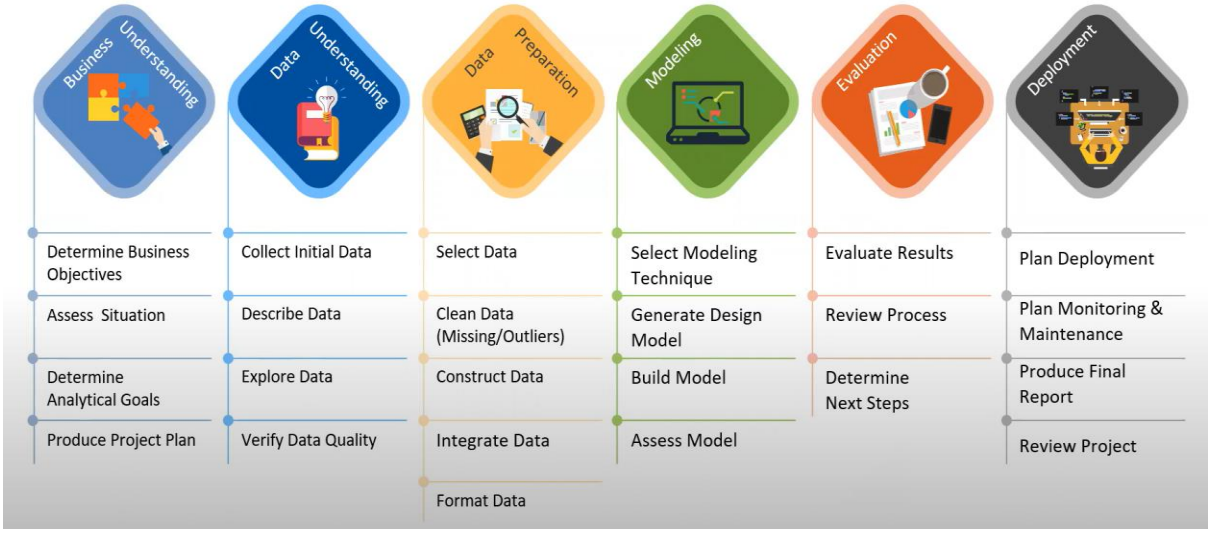


Figure 11: CRISP-DM Phases and Task

a) Business Understanding

In this stage, the focus is on identifying the specific requirements and objectives of the healthcare professionals and organisations. These objectives and aims have been clearly defined in Chapter 1.2, which helps to establish the scope of the project and ensure that the model will be both effective and relevant for the health organisation. Additionally, the project plan has been outlined in Chapter 4.

b) Data Understanding

The data understanding stage involves gathering and assessing the relevant data that will be used in the heart disease prediction. This includes identifying the sources of data, understanding the structure of the data, assessing the quality of the data, and identifying any data gaps or inconsistencies. The goal is to ensure that the data is complete, accurate, and relevant to the problem being assessed (Hotz, 2023). In Chapter 4.3.2 the steps are explained in detail.

c) Data Preparation

In this process, all the raw data will be converted into a well-organised and clean format that is suitable for analysis. It involves tasks such as data cleaning, integration, transformation , and reduction which is explained in Chapter 5. The primary goal of data preparation is to ensure that the data is consistent, accurate, and ready to be used for modelling. The quality of data used for modelling has a direct impact on the accuracy and effectiveness of the modelling.

d) Modelling

In this stage, machine learning techniques are applied to the prepared data to create a predictive model. The objective of modelling is to identify the relationships between the input variables and target variables, and to use these relationships to predict future outcomes. All the different models are built using the prepared data, and it is iteratively refined and tested until the best possible performance is achieved. Once the models are developed, it can be used to predict outcomes for new data points that were not part of the original dataset. These practices are outlined in Chapter 5.3.

e) Evaluation

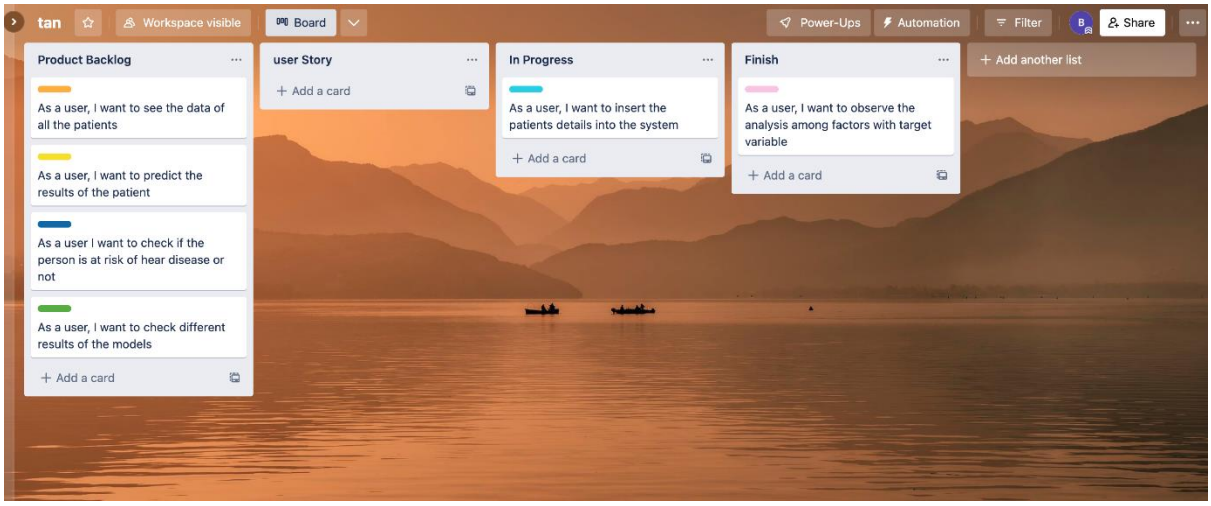
The evaluation stage assesses the models using techniques of cross-validation and hyperparameter tuning, and performance metrics including accuracy, precision, recall, and F1 score. The main goal is to determine which model best meets the project's objectives and requirements, and to ensure that it is effective and dependable for its intended use by healthcare professionals.

f) Deployment

The deployment stage involves implementing the final model onto the cloud so that healthcare professionals can use it to predict the likelihood of heart disease in patients and take appropriate action. Chapter 5.4 of the project report demonstrates the deployment of the model, while the final report and project handover are presented in Chapter 7 and 8, respectively.

3.3 Project Management Tool

Trello is a web-based project management tool that provides and easy and visual way to manage and track projects. Trello is used to create boards for each stage of the CRISP-DM methodology. Each board contain lists of tasks that need to be completed, with cards representing individual tasks.



Chapter 4 - Design

This section of the project introduces the machine learning platform that was utilised to create various models along with a list of the tools and libraries used. A workflow is then created based on the knowledge and insights gained from the previous chapters.

4.1 RStudio

RStudio is a popular integrated development environment (IDE) for the R programming language, which is commonly used in statistical computing and data analysis (Wikipedia, 2023). RStudio provides a user-friendly interface and a wide range of tools and libraries for machine learning (Wikipedia, 2023), making it a suitable choice for building and evaluating algorithms for heart disease prediction.

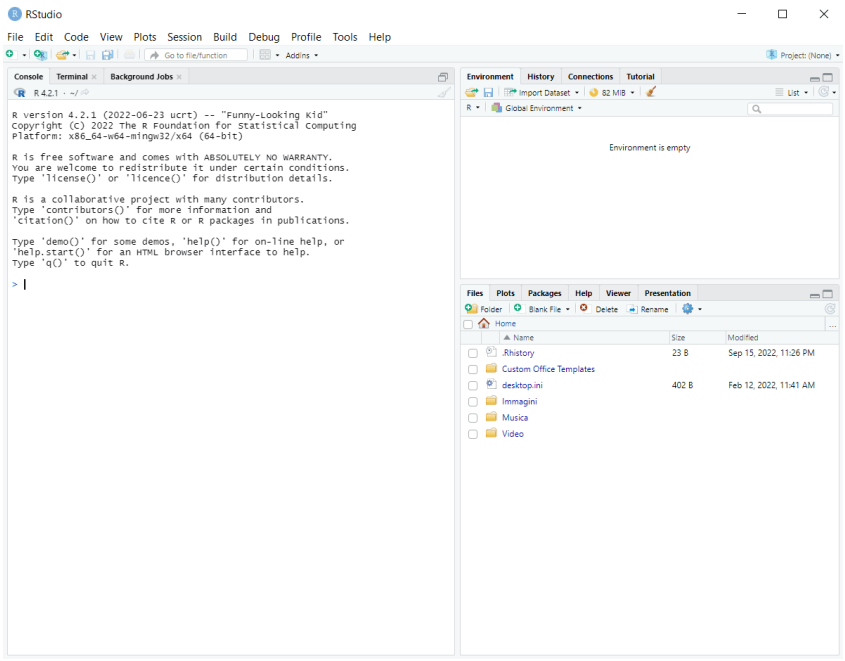


Figure 12: RStudio

One of the reasons to choose RStudio for this project is its flexibility and versatility. RStudio provides a wide range of statistical and graphical techniques, allowing data scientists to explore and analyse data in depth, which is essential for building effective machine learning algorithms (Kosourova, 2022). Additionally, RStudio allows for easy integration with other tools and platforms, which can further enhance the workflow of the project. Moreover, it is easy to access a wide range of libraries, tools, and resources that can be used to build and evaluate machine learning algorithms for heart disease prediction.

4.2 Essential Libraries

This section is dedicated to listing and explaining the various libraries that are required for the project before building any machine learning models in RStudio. The installation of these libraries is done using the “install.packages()” command in RStudio to ensure that all the necessary libraries are properly installed.

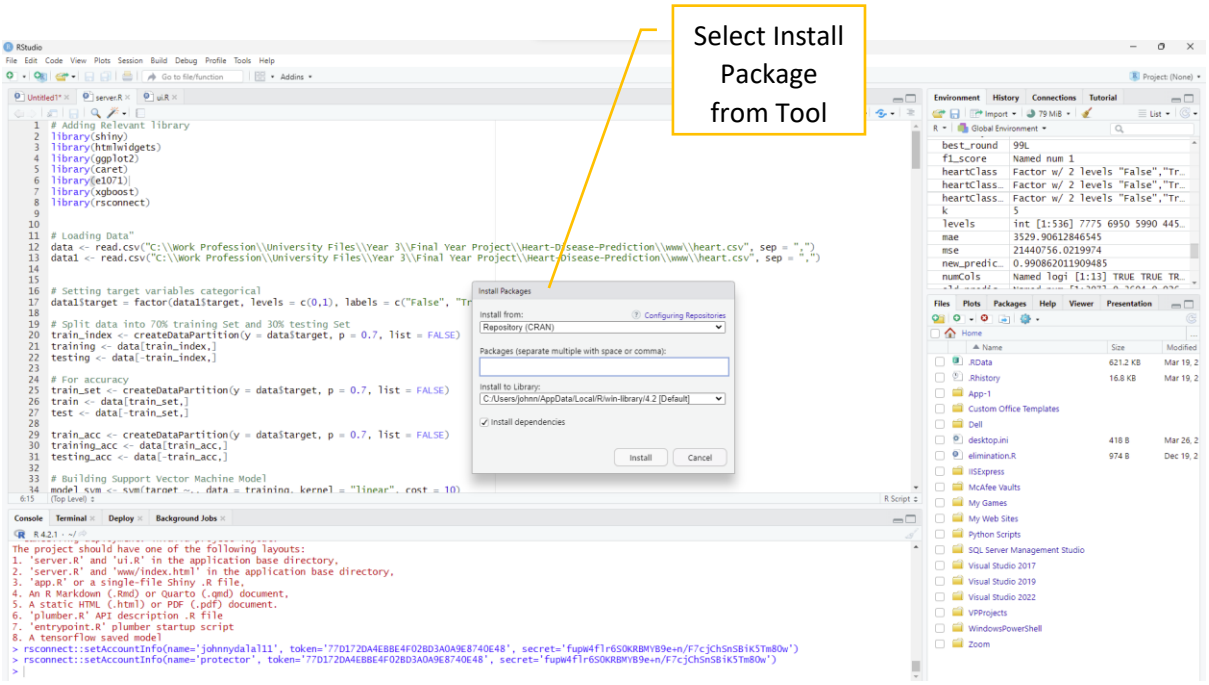


Figure 13: install.packages

4.2.1 library(shiny)

Shiny is an R package that allows developers to build interactive web applications with R. The architecture of Shiny is based on a client-server model (Shiny, 2012). The server runs the R code that generates the application’s output, while the client is the user’s web browser that interacts with the server. This tool will be used to build interactive web-based heart disease prediction where healthcare professionals can predict the heart disease of patients.

4.2.2 library(htmlwidgets)

The htmlwidgets library enables the creation of interactive HTML widgets using R (cran.r, 2023). In the context of heart disease prediction, these widgets can be used to develop a web-based interface where healthcare professionals can input patient details and obtain the predicted outcome of their disease in a user-friendly and interactive manner.

4.2.3 library(ggplot2)

This package is used for data visualisation and creating graphs. This library is widely used in data science and machine learning projects for exploratory data analysis and communication. Moreover, it provides a wide range of features for creating customised and complex plots with simple and easy-to-understand.

4.2.4 library(rsconnect)

This library is used to deploy Shiny applications to various hosting platforms such as shinyapps.io and RStudio Connect (Shiny, 2012).

4.2.5 library(caret)

This package used to provide over 200 different models as well as tools for data pre-processing, model tuning and model evaluation. This library is used to build and test multiple models, allowing data scientists to identify the most relevant approach for a problem.

4.2.6 library(e1071)

This library is used in support vector machines and naive bayes to predict heart disease. It also provides functions for classification and regression.

4.2.7 library(xgboost)

This library is used to build machine learning models on large datasets which can be used for both regression and classification problems. XGBoost is known for its high accuracy and speed.

4.2.8 library(rpart)

This package provides functions for building decision trees which can be used for both regression and classification tasks. The decision tree is used to make predictions about new data on heart disease.

4.3 Work Flow of Shiny Architecture

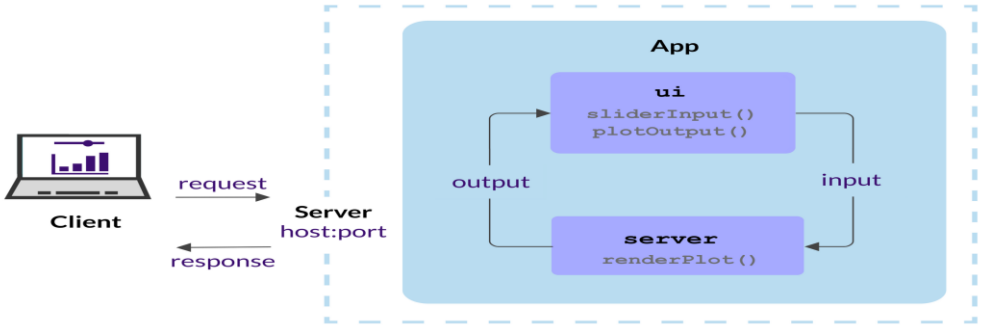


Figure 14: Shiny Architecture

4.3.1 Ui Interface

The web application is designed for predicting heart disease in a file called ui.R. The UI is defined using the "fluidPage()" function from the Shiny package which consists of a title panel, a navbar, and a tab panel for displaying the inputs and outputs for the prediction model.

The navbar consists of a three tab panel named "Home", "Analysis", "Prediction". In the "Prediction" tab, the healthcare professionals can input the data of their patients for predicting health disease if they carry or not. The inputs are defined using various input functions from the Shiny package, such as "sliderInput()", "radioButtons()", "selectInput()", and "numericInput()". The UI also includes various "verbatimTextOutput()" functions for displaying the inputs in real-time as the user interacts with the app.

Overall, the code can provide a user-friendly interface for collecting the necessary data for predicting heart disease.

4.3.2 Server File

This is a file which is connected from the Ui file. The code within it starts by loading the data of heart disease and setting the target variable as categorical. It then splits the data into a 70% training set and a 30% testing set. Next, it builds various kinds of machine learning models like Linear Regression, Support Vector Machine, XGBoost, Decision tree, and Naive Bayes.

a) Heart Disease Data

The dataset is available on Kaggle for heart disease prediction which contains various factors that are commonly associated with the risk of heart disease. These factors include age, gender, cholesterol levels, blood pressure, chest pain, and many more. The dataset also contains information about the presence or absence of heart disease in the patients.

Using this data, a study can be carried out to predict the risk of heart disease in patients based on their health-related factors.

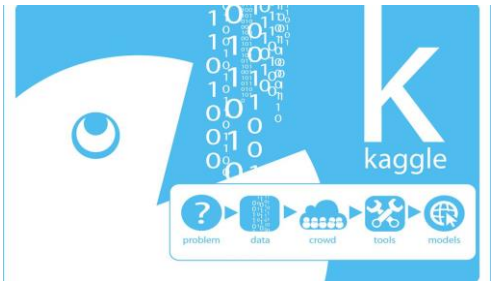


Figure 15: Kaggle

b) Data Preparation

It is a crucial step to prepare the data in order to get an accurate result of the predictive model. The first step is to clean the data to remove any incomplete entries and address null values. This can be achieved by using various techniques such as imputing missing values or removing rows with missing data.

Once the data has been cleaned, it is important to balance the dataset with respect to the target variable, which can be performed using oversampling technique. After balancing the dataset, the next step is to

check and remove outliers, which can be identified using boxplot then the target variable will be factored to ensure that the data is in a suitable format for machine learning algorithms.

Finally, the data will then undergo a train-test split, where the train dataset will be used to train various machine learning algorithms and the test dataset will be used as a validation metric for the model.

c) Machine Learning Algorithms design

Below, different machine learning techniques will be discussed and the reasons weather the chosen algorithm is useful or not:

1. **Naive Bayes**

   The Naive Bayes model is based on Bayes' theorem, which is a mathematical formula for calculating the probability. The "naive" part of the name comes when the input data are conditionally independent of each other, which simplifies the calculation of probabilities (Wikipedia, 2023).

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Figure 16: Naive Bayes Formula

   It is a probabilistic algorithm that can handle missing data. In the case of heart disease prediction, missing or incomplete data is a common problem that can impact the accuracy of the model. It can work with incomplete data by estimating the probability distribution of the missing values based on the available data.

2. **Linear Regression**

   Linear regression is a statistical modelling technique that is used to describe the linear relationship between a dependent variable and one or more independent variables. The relationship between the independent variables and the dependent variable is represented by a straight line (Wikipedia, 2023).

In a multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

Figure 17: Linear Regression Formula

   According to the study, it may not be the best choice for heart disease prediction because it assumes a linear relationship between the dependent variable and the independent variables.

However, it can be useful for predicting certain risk factors such as blood pressure and cholesterol levels, which have a linear relationship with the risk of developing heart disease.

3. **Support Vector Machine**

SVM is commonly used for both classification and regression. It works by finding the optimal hyperplane that separates the data into different classes based on their attributes (Wikipedia, 2023).

It has the ability to handle both linear and non-linear relationships between variables. It can use a kernel function that can further transform the data into higher-dimensional space where it can be linearly separated, even when the relationship between variables is non-linear. This can improve the accuracy of heart disease prediction.

4. **Decision Tree**

Decision tree is commonly used for both classification and regression. It works by recursively splitting the data into subsets based on the values of the input features, and creating a tree-like structure that represents a set of rules for predicting the output variable (Wikipedia, 2023).

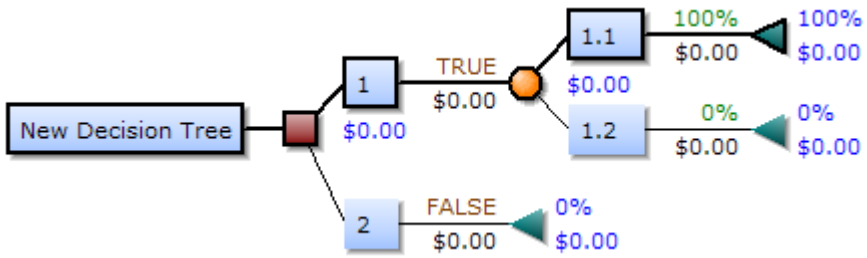


Figure 18: Decision Tree Example

This model can provide insight into the factors that can contribute to the risk of heart disease. The tree structure of the model can be easily visualised and understood, which can help in identifying the most important risk factors.

5. **XGBoost**

XGBoost is used for both classification and regression. It is an optimised implementation of gradient boosting, a popular learning method that can combine multiple weak models to create a strong predictor (Wikipedia, 2023).

It is capable of handling both categorical and continuous input variables, which is important in the case of heart disease prediction as it involves several risk factors. The algorithm can provide a clear boundary between different classes of heart disease and can identify the most important risk factors for heart disease.

d) Hyperparameter and cross-validation test

When constructing models for tasks like classification or regression, it is important to enhance their performance. Hyperparameters are factors that dictate the model's training and complexity. By adjusting these hyperparameters, the model may achieve increased accuracy and better generalisation to unfamiliar data.

Another widely used technique to assess and enhance model performance is cross-validation. This method divides the data into multiple subsets or folds, training the model on a part of the data, and testing it on another part. This procedure is carried out for each fold, providing a more dependable estimate of the model's performance than a single train-test split.

To evaluate the effectiveness of hyperparameter tuning and cross-validation, we will compare the performance of the models using metrics such as F1 score, precision, recall, and accuracy. The F1 score is a combined measure of precision and recall computes the ratio of true positives among all actual positives.

By comparing the outcomes obtained through hyperparameter tuning and cross-validation, we can identify which method is more effective in enhancing the model's accuracy and generalisation capabilities.

e) Deployment of the models

Based on the test results which can be found in Chapter 7, the best model will be deployed to predict heart disease. Whenever healthcare professionals input the data of their patients then they will receive the results of whether the patient has heart disease or not.

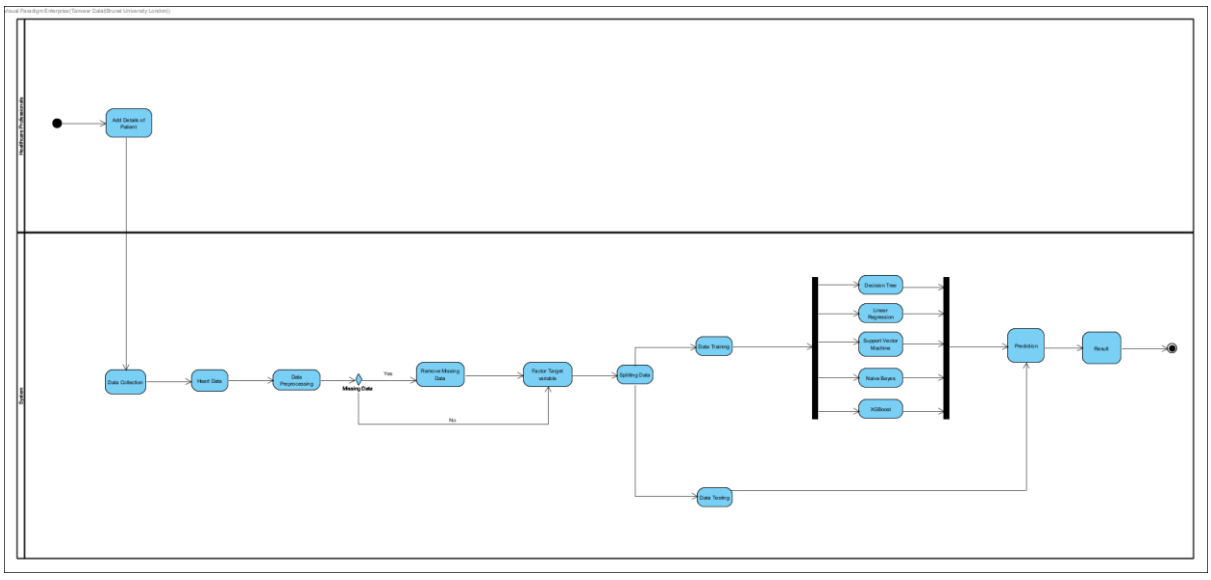


Figure 19: Activity Diagram

Chapter 5 - Implementation

This chapter focuses on the implementation of the heart disease prediction model using multiple machine learning algorithms and displaying the predicted results to the patient. The workflow of the Shiny architecture created in the previous chapter will be utilised to ensure a successful completion of this chapter. This chapter will provide details on the training of the machine learning algorithms and the process of evaluating the predictive performance of the model. Additionally, the chapter will describe how the results will be displayed to the users, including the user interface and the functionality of the Shiny application.

5.1 Understanding of Heart Disease Dataset

To understand the data some tests have occurred like generating the plots of factors distribution based on the target variable. It is important to visualise because it shows how much these factors are related to the target variable, which is the presence or absence of heart disease. By examining these plots, healthcare professionals can gain insights into the potential relationship between all the factors with heart disease. The given code specifies the data used in the plot and the mapping of the cp variable on x-axis and the target variable on the colour of the plotted lines.

```
output$cp_plot <- renderPlot({
  ggplot(data1, aes(cp, colour = target)) +
    geom_freqpoly(binwidth = 1) +
    labs(title = "Chest Pain Distribution by Target")
})
```

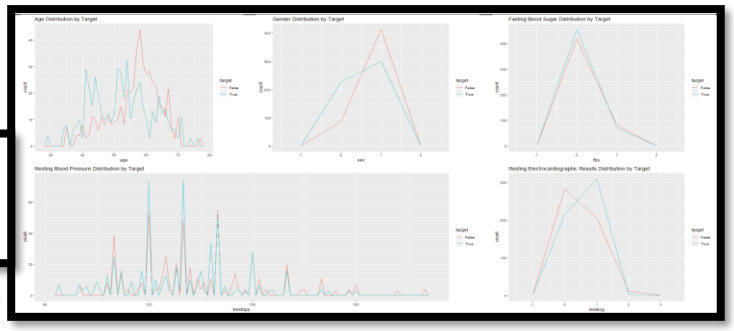


Figure 20: Analysis

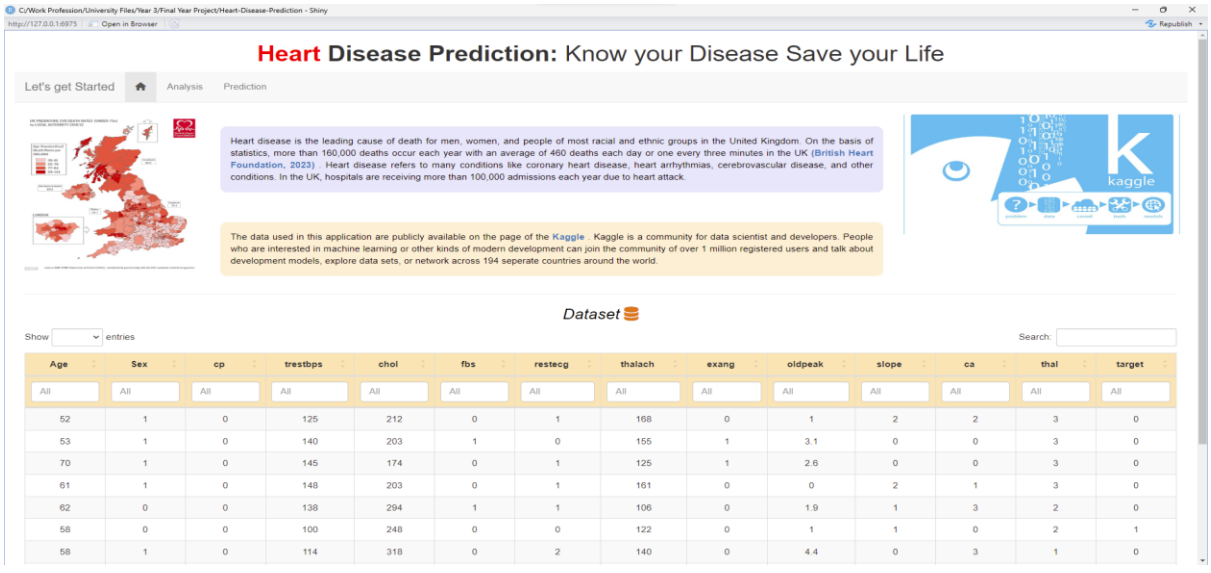


Figure 21: Kaggle Dataset

5.2 Data Preparation

This stage will follow the steps to clean the data and prepare it to use in the machine learning model.

First, the given code will remove all the duplicate rows from the original "data" dataframe. This is done by using the unique() function in R, which returns only the unique rows in a dataframe, effectively removing any duplicate rows.

```
#checking for duplicate values
duplicate_rows <- duplicated(data)
print(duplicate_rows)
delete_dupli <- unique(data)
```

Second, the line of code assesses the target variable column from the dataframe. It specifies the target variable as a factor with two levels (0 and 1), where 0 represents the absence of heart disease, and 1 represents the presence of heart disease.

```
# Setting target variables categorical
data$target = factor(data$target, levels = c(0,1), labels = c("False", "True"))
```

Finally, splitting the data into a training set and a testing set using the 'createDataPartition()' function from the 'caret' package. Where the data is divided into 70% training set and the remaining 30% for testing set.

5.3 Building Models with Hyperparameter and Cross-Validation

This section focuses on constructing and compiling several models using hyperparameter tuning and cross-validation techniques. These techniques involve optimising the performance of the machine learning models by selecting the best set of hyperparameters and using cross-validation to evaluate the model's performance on the data.

5.3.1 Naive Bayes

Hyperparameter tuning is defined using the 'expand.grid()' function. In this case, the 'tune_grid' object defines a grid of hyperparameters that includes the 'fl' parameter, which is the threshold frequency.

A 10-fold cross-validation object is created using the 'trainControl()' function to evaluate performance.

The 'train()' function is then called to perform hyperparameter tuning and cross-validation on a Naive Bayes model using the training data. The 'method' argument specifies the machine learning algorithm to use which is Naive Bayes, the 'trControl' argument specifies the cross-validation object to use, and the 'tuneGrid' argument specifies the hyperparameters to tune over.

Once the model is trained, the 'predict()' function is used to make predictions on the testing set or it can predict the new data as well.

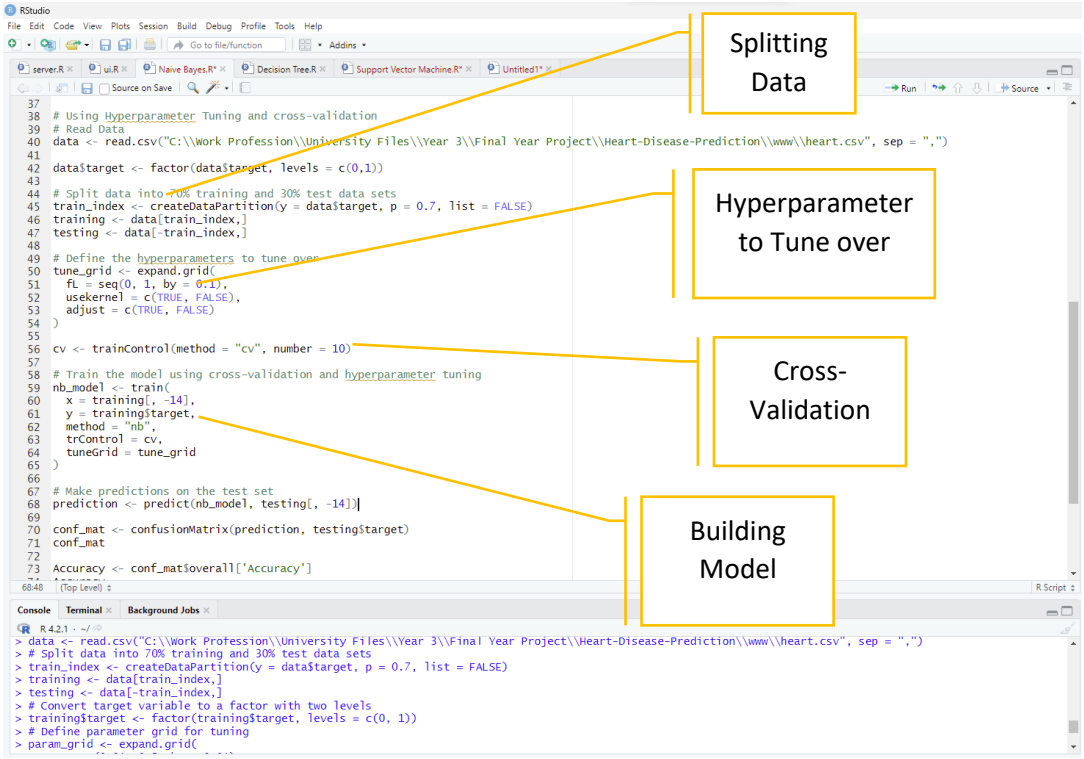


Figure 22: Naive Bayes Model

5.3.2 Decision Tree

The 'target' variable in the training set is converted to a factor with two levels using the 'factor()' function.

A parameter grid for tuning is defined using the 'expand.grid()' function. In this case, the 'param_grid' object specifies a grid of values for the 'cp' parameter, which is the complexity parameter that controls the amount of pruning in the decision tree.

The 'train()' function is then called to perform hyperparameter tuning and cross-validation on a Decision Tree model using the training data. The 'method' argument specifies the machine learning algorithm to use which is Decision Tree, the 'trControl' argument specifies the cross-validation object to use, and the 'tuneGrid' argument specifies the hyperparameters to tune over.

Once the model is trained, the 'predict()' function is used to make predictions on the testing set. The 'type' argument is set to 'raw' to ensure that the predicted values are returned as probabilities.

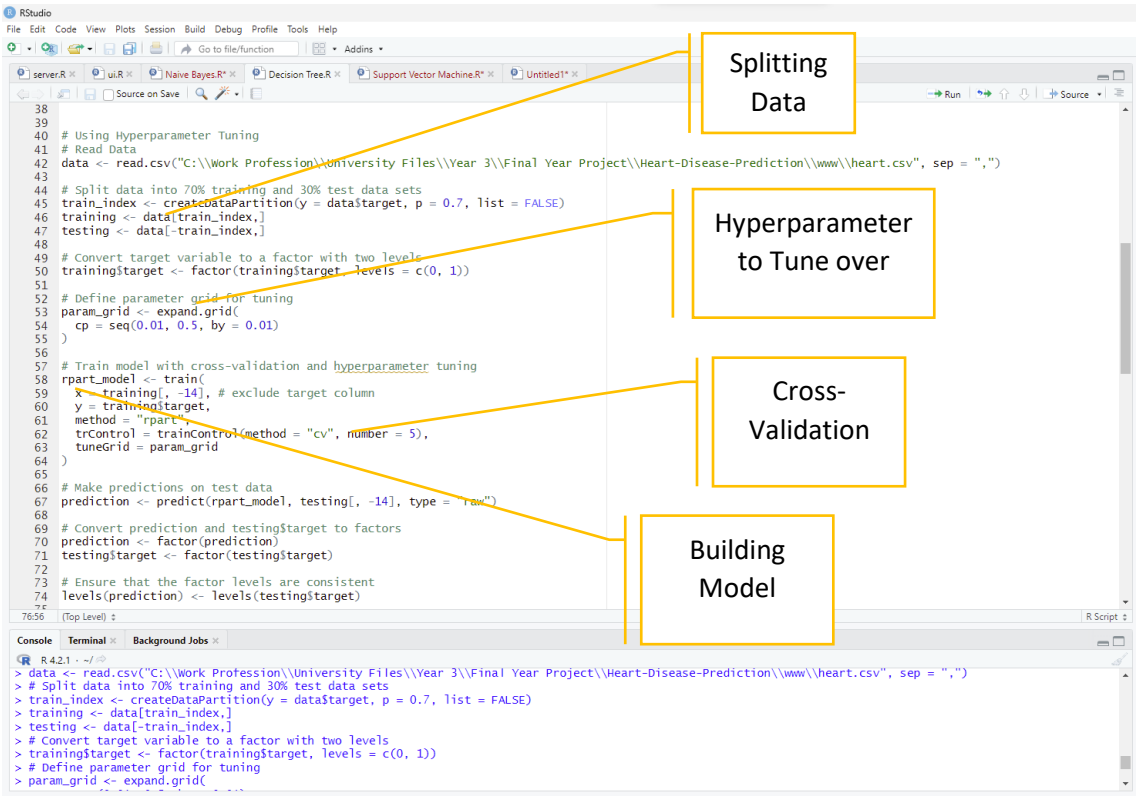


Figure 23: Decision Tree Model

5.3.3 Support Vector Machine

A parameter grid for tuning is defined using the 'expand.grid()' function. In this case, the 'param_grid' object specifies a grid of values for the 'C' parameter, which controls the trade-off between maximising the margin and minimising the classification error, and the 'gamma' parameter controls the shape of the radial basis function kernel

The 'tune.control()' function is then used to specify the parameters for cross-validation during the tuning process. In this case, 'sampling' is set to 'cross' which is cross-validation.

The 'tune()' function is then called to perform hyperparameter tuning on an SVM model using the training data. The 'svm()' function is used as the first argument, the 'target ~.' formula is is used to specify all the predictors to predict the target variable, the 'kernel' argument is set to 'radial' to specify a radial basis function kernel, the 'ranges' argument specifies the hyperparameters to tune over, and the 'tunecontrol' argument specifies the tuning parameters

Once the tuning process is complete, the 'best.model' object is extracted from the 'tune_result' object, which stores the trained and tuned SVM model.

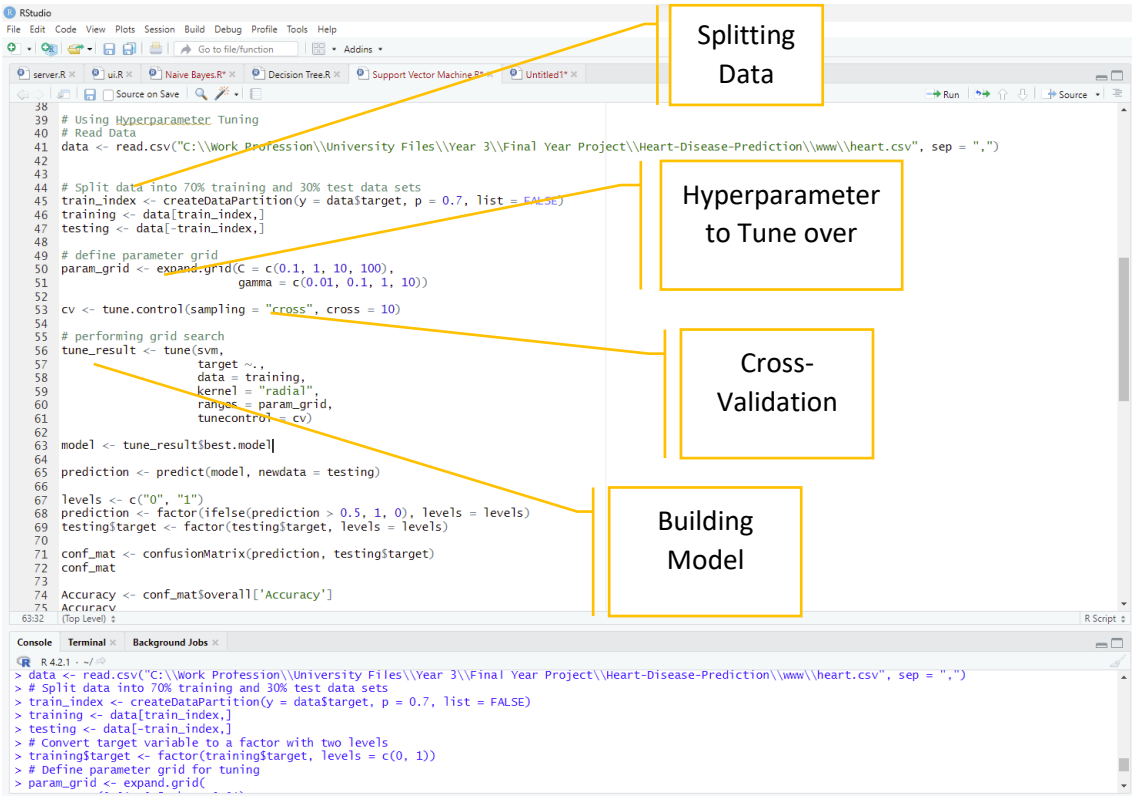


Figure 24Support Vector Machine Model

5.3.4 XGBoost

A parameter grid for hyperparameter tuning is defined using the 'expand.grid()' function. In this case, the 'param_grid' object specifies a range of values for several hyperparameters, such as 'nrounds', 'max_depth', 'eta', 'gamma', 'colsample_bytree', 'min_child_weight', and 'subsample'.

A 10-fold cross-validation object is created using the 'trainControl()' function to evaluate performance.

The 'train()' function is then called to perform hyperparameter tuning on an XGBoost model using the training data. The 'as.matrix()' function is used to convert the training data to a matrix format. The 'tuneGrid' argument specifies the hyperparameters to tune over , and the 'objective' argument specifies the objective function to optimise.

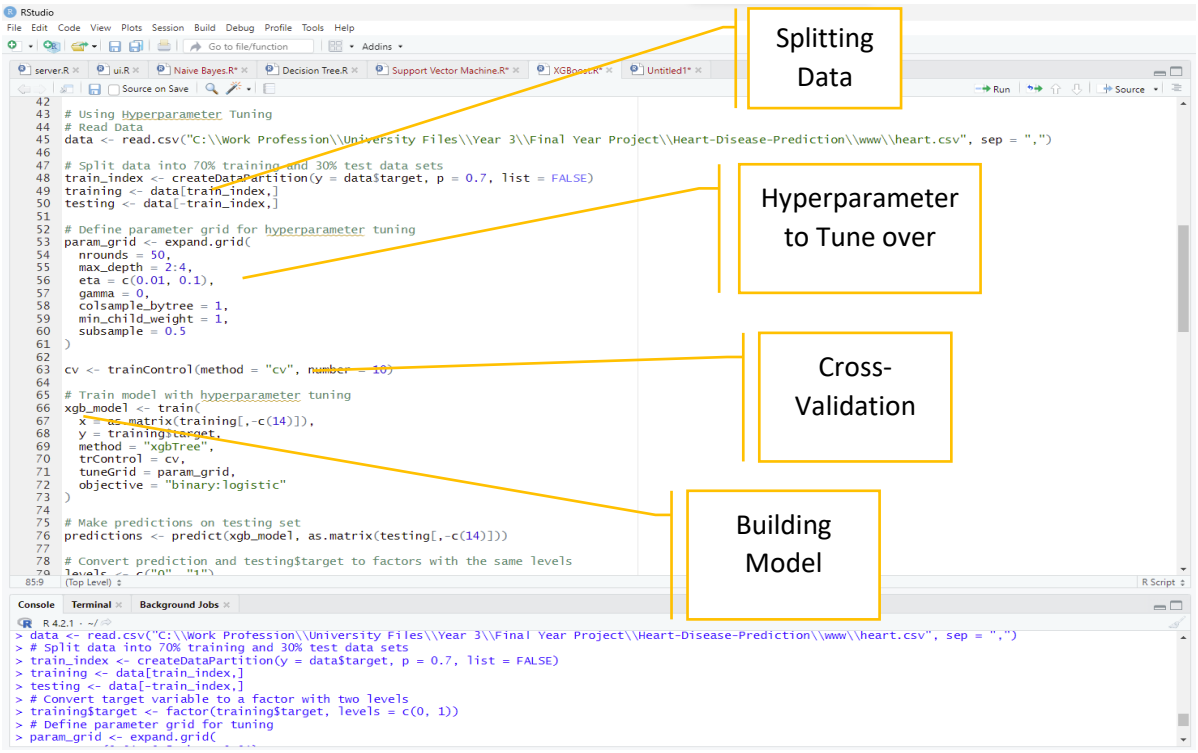


Figure 25: XGBoost

5.3.5 Linear Regression

The 'createFolds()' function is used to create cross-validation folds. The 'y' argument specifies the target variable to use for stratified sampling, the 'k' argument specifies the number of folds to create, and the 'list' and 'returnTrain' arguments are used to create a list of the training indices for each fold.

The 'train()' function is then called to perform cross-validation on a linear regression model using the training data. The 'method' argument is set to 'lm' to use a linear regression model. The 'trControl' argument is used to specify the parameters for cross-validation during the tuning process. In this case, 'method' is set to 'cv' to use cross-validation, and 'index' is set to the 'folds' object.

Once the cross-validation is complete, the 'predict()' function is used to make predictions on the testing set. the predicted values are stored in the 'prediction' object.

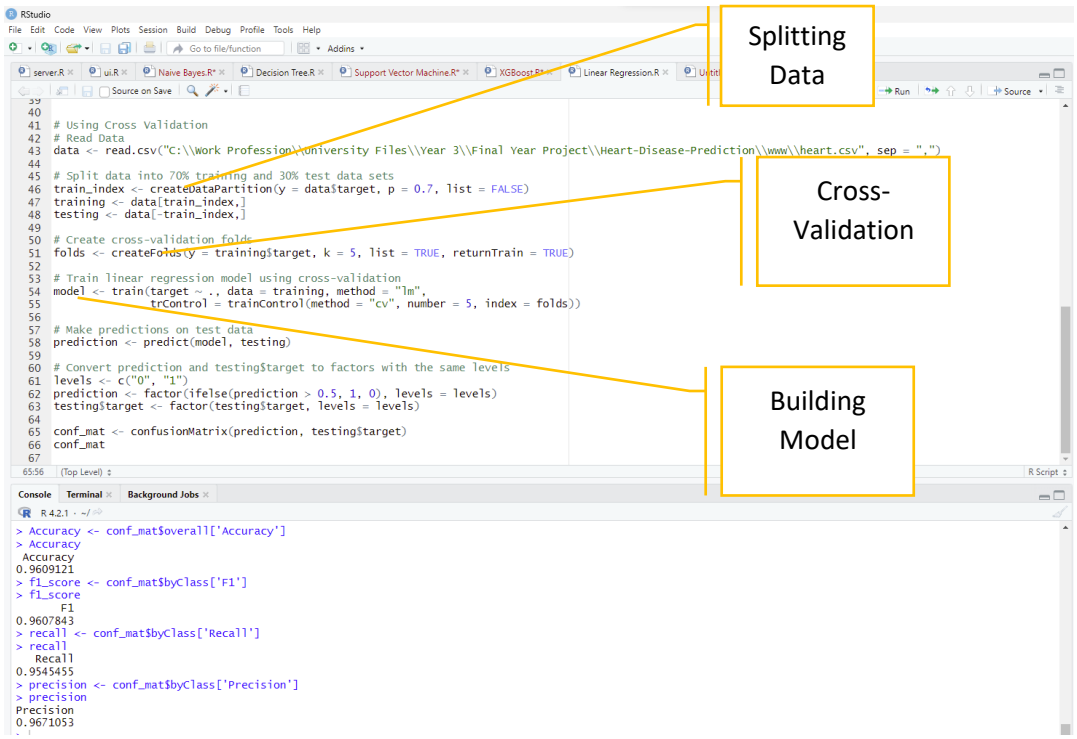


Figure 26: Linear Regression Model

5.4 Deployment of Shiny Web Application

Shinnyapps.io is a cloud-based platform that enables users to host and deploy Shiny applications, including the heart disease prediction model.

Once the application is ready, the user can connect to Shinyapps.io and create a new account. After creating an account, the user can deploy the application by uploading the code to Shinyapps.io. The platform will take care of the rest, and the application will set up the web server.

After deploying the application, the user can test it by accessing the URL provided by Shinyapps.io.

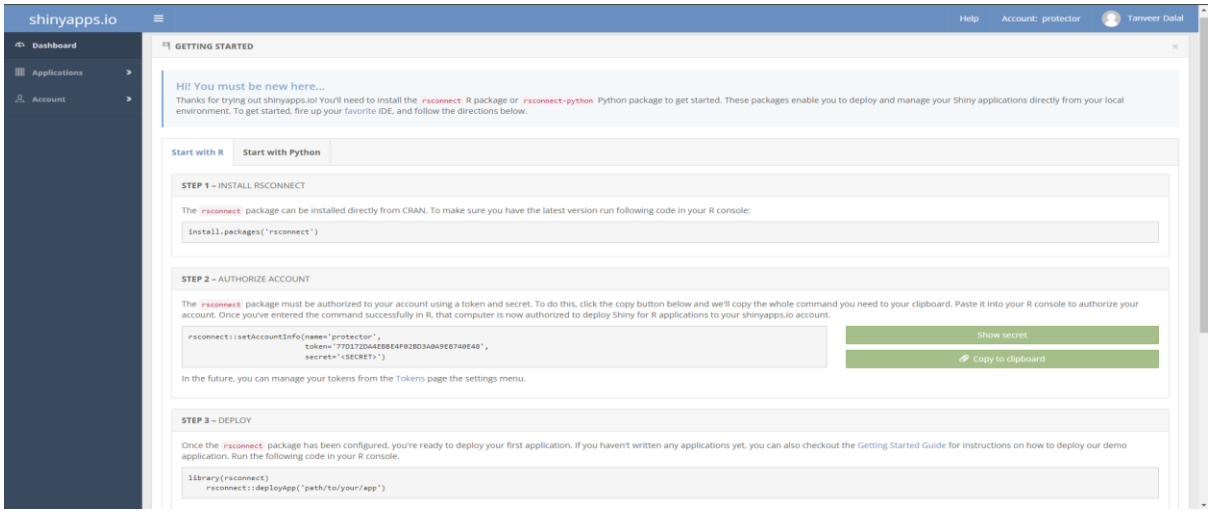


Figure 27: Shinyapps.io

Chapter 6 - Testing

Testing is a crucial part of developing software applications, including a Shiny web application and Black-box testing.

6.1 Snapshot-based tests

This test involves creating an interactive web interface where users are allowed to input patient information, and then use predictive models to estimate the risk of heart disease.

6.1.1 User Interface

UI is a part of an application where users can provide patient data, such as age, gender, chest pain type, blood pressure, cholesterol levels, and many more. In addition, the UI has a button to trigger the prediction and an area to display the prediction result.

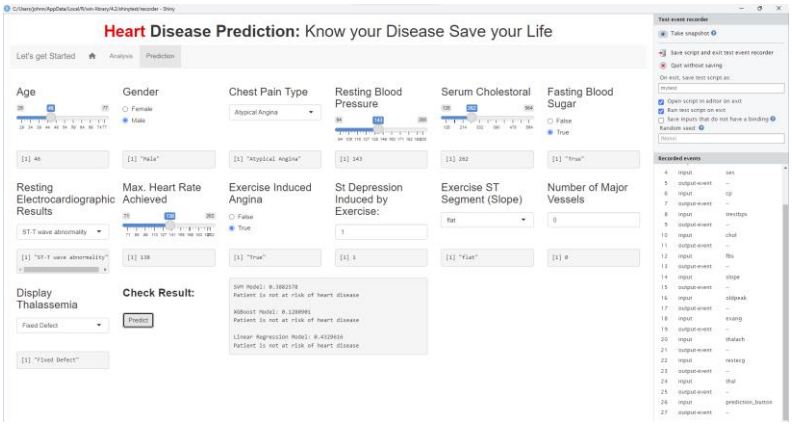


Figure 28: Snapshot-based test

6.1.2 Server-side logic

This part of the application is used to process the inputs, where the data is prepared for predictive models and can generate the risk.

6.1.3 Testing

To ensure the application works as expected, certain tests have been performed on both the UI and the server-side logic. The report on this test is an ok report. Which suggests that the working of this application is perfectly good and all the models within it are working correctly.

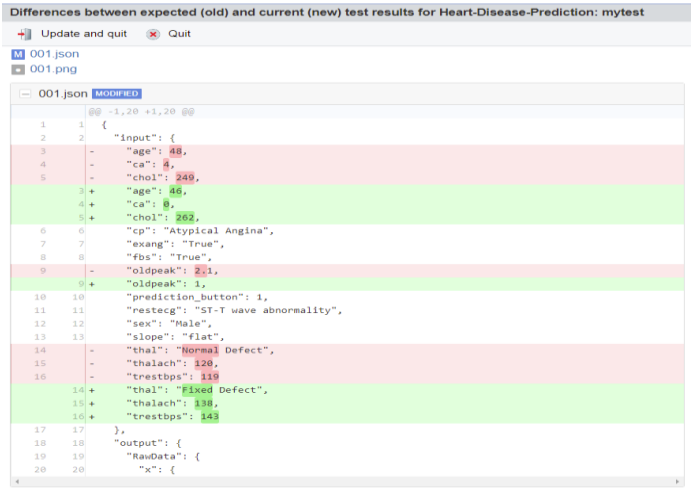


Figure 29: Tests Result

6.2 Black-Box testing

Black-box testing is a method of testing an application without knowing its internal code or structure. It focuses on the functionality of the application, ensuring that it works as expected and meets the requirements specified by the user (Imperva, 2020).

Black-box testing is an appropriate choice for this heart disease prediction because it deals with the focus of ensuring that the application meets the user's requirements and functions as expected. It is also useful when testing the application's usability, user interface, and performance by focusing on the application's functionality, and can be verified that it is easy to use, and responds quickly to user input.

Overall, black-box testing is a good way to verify that the heart disease prediction application meets the user's expectations.

Table 1: Black-Box Testing

Test Case	Input	Expected Output	Actual Output	Pass/Fail
Input Validation	Chest Pain Type: Atypical Angina	Chest Pain Type: Atypical Angina	Chest Pain Type: Atypical Angina	Pass
Output Validation	Predict Heart Disease for a patient	Result: "Low Risk"	Result: "Low Risk"	Pass
Output Validation	Explain Results	Explanation: "Based on the given details patient is not at risk of heart disease"	Explanation: "Based on the given details patient is not at risk of heart disease"	Pass
Performance Testing	Running Application	All requests processed within 5-10 seconds	All requests processed within 5-10 seconds	Pass
User Interface Testing	Navigate through the application	Easy to understand and use	Easy to understand and use	Pass

Chapter 7 - Evaluation

This section of the research will assess all the machine learning algorithms used in this study. A discussion will be presented on which models are considered to be the most appropriate to use. A comparison between the older models and the model constructed using both hyperparameter tuning and cross-validation will be provided.

7.1 Machine Learning Algorithms

The study utilised different machine learning algorithms to build predictive models for heart disease prediction. These algorithms included Linear Regression, Support Vector Machine, Naive Bayes, Decision Tree, and XGBoost. However, these algorithms were not used in their default form, but rather were enhanced using hyperparameter tuning and cross-validation techniques to improve their performance.

To evaluate the performance of these models, various metrics were used, such as confusion matrices, F1 score, precision, recall, and accuracy. Confusion matrices help to evaluate the performance of a classification model by providing the number of true positive, false positive, true negative, and false negative predictions. F1 score, precision, and recall are metrics commonly used for evaluating the performance of classification models. Accuracy is also a widely used metric that measures the percentage of correctly predicted instances out of all instances.

Finally, a comparison of these models is presented in a table that provides an easy reference for their performance. This comparison helps in selecting the most appropriate model for heart disease prediction.

Table 2: Model Evaluation

Models	F1 score	Precision	Recall	Accuracy
SVM	0.8431373	0.8958333	0.7962963	0.8436482
Naive Bayes	0.7765568	0.8548387	0.7114094	0.8006536
Decision Tree	0.8837209	0.8926174	0.875	0.8859935
XGBoost	0.9803922	0.9615385	1	0.980456
Linear Regression	0.8028169	0.7916667	0.8142857	0.8175896

Table 3: Model Evaluation using Hyperparameter Tuning

Models	F1 score	Precision	Recall	Accuracy
SVM	0.9865772	1	0.9735099	0.9869707
Naive Bayes	0.85	0.9083969	0.7986577	0.8627451
Decision Tree	0.8983607	0.883871	0.9133333	0.8990228
XGBoost	0.9345794	0.9375	0.931677	0.9315961
Linear Regression	0.8633094	0.9090909	0.8219178	0.8762215

7.2 Deploying The Best Model

Based on the evaluation metrics of F1 score, precision, recall, and accuracy, the SVM model has performed the best among all the other models used in this project. The SVM model was created using hyperparameter tuning and cross-validation techniques, which improved its accuracy by 14% compared to its previous version. Therefore, the SVM model can be considered the most suitable model for this project.

Old Model of SVM

```
Confusion Matrix and Statistics


prediction   0   1
         0 129  15
         1  33 130

               Accuracy : 0.8436
                 95% CI : (0.7981, 0.8824)
    No Information Rate : 0.5277
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6884

 Mcnemar's Test P-Value : 0.01414

            Sensitivity : 0.7963
            Specificity : 0.8966
         Pos Pred Value : 0.8958
         Neg Pred Value : 0.7975
             Prevalence : 0.5277
         Detection Rate : 0.4202
   Detection Prevalence : 0.4691
      Balanced Accuracy : 0.8464

       'Positive' Class : 0
```

New Model of SVM

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 147   0
         1   4 156

               Accuracy : 0.987
                 95% CI : (0.967, 0.9964)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9739

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 0.9735
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9750
             Prevalence : 0.4919
         Detection Rate : 0.4788
   Detection Prevalence : 0.4788
      Balanced Accuracy : 0.9868

       'Positive' Class : 0
```

Chapter 8 - Conclusion

This chapter reflects an overview of the research conducted in each chapter. The study's contributions and limitations are examined, and directions for future research arising from this project are also discussed.

8.1 Dissertation Summary

The study aimed to fill a gap in previous heart disease prediction research by providing a standardised comparison of multiple machine learning algorithms and investigating the factors that influence algorithm performance.

Below will be the dissertation summary for how each chapter has meet all the 6 objectives:

1. Chapter 1 of the dissertation provides an introduction to the research problem by highlighting the need for improved heart disease prediction. The chapter outlines the key requirements for heart disease prediction, which inform the aims, objectives, and approach of the research project.

2. Chapter 2 of the dissertation provides a comprehensive background research on heart disease, and the need for accurate prediction methods. This chapter also reviews the existing methods that are used for predicting heart disease and their limitations. This research is important because it provides valuable insights into the current state of heart disease prediction.

For achieving the project aim, **objective 1** was met in this chapter.

3. Chapter 3 of the dissertation provides a comparison of various data science methodologies and their management tools. This chapter discusses the benefits and limitations of each methodology, and ultimately, the preferred methodology of CRISP-DM is used to fit the project's needs. This chapter provides an overview of the different stages of the CRISP-DM methodology, including data preparation, modelling, evaluation, and deployment.

Objective 2 was met in this chapter.

4. Chapter 4 of the dissertation describes the platform used for building the heart disease prediction model and the Shiny application. This chapter discusses the required libraries for constructing multiple machine learning algorithms and building the user interface for the application. This chapter serves as a guide for readers to understand the technical details of the heart disease prediction model and the Shiny application.

5. Chapter 5 of the dissertation presents the implementation details of the data preparation techniques and the construction of multiple machine learning models for heart disease prediction. This chapter describes the different algorithms used in the study, the process of model building. Additionally, this

chapter describes the deployment process of the Shiny application on the Shinyapps.io platform, a cloud-based service that enables the deployment of Shiny applications on the web.

Objective 3, **objective 4,** and **objective 5** were met in this chapter.

6. Chapter 6 of the dissertation describes the testing and deployment of the heart disease prediction model and the Shiny application. This chapter provides details on the two significant tests conducted to evaluate the performance of the model and the application. This chapter serves as a guide to understand the testing of the heart disease prediction model and the Shiny application.

7. Chapter 7 of the dissertation will cover the evaluation of several machine learning algorithms, where their performance will be assessed based on various metrics such as confusion matrices, f1 score, precision, recall, and accuracy. The chapter provides an overview of the best performing model, which will be utilised for the shiny web application.

Objective 6 was met in this chapter.

8. Chapter 8 highlights the research, including the development of a multiple machine learning model to predict heart disease and the creation of a user-friendly Shiny application for healthcare professionals. This chapter also acknowledges the limitations of the study and suggests future research to address these limitations. This chapter provides a comprehensive overview of the research project, its contributions, and future implications for heart disease prediction research.

8.2 Future Work

In the future, researchers may consider integrating additional data sources to enhance the accuracy of the heart disease prediction model. Although this study compared various machine learning algorithms, further research could examine the use of hybrid models to achieve improved prediction accuracy. It is also suggested that the Shiny application developed in this study can be evaluated for a longer period to improve its efficiency and receive feedback from healthcare professionals. The development of this application can be further enhanced by adding other medical conditions, such as diabetes or cancer.

References

Donovan, R. (2020, February 27). *Heart disease: Risk factors, prevention, and more*. Available at: https://www.healthline.com/health/heart-disease#heart-disease-symptoms-in-women (Accessed 10 February 2023)

British Heart Foundation. (2023, January). *Facts and figures*. British Heart Foundation. Available at: https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures (Accessed 11 February 2023)

British Heart Foundation. (2023, January). *Can I still work if I have a heart condition?*. BHF. Available at: https://www.bhf.org.uk/informationsupport/heart-matters-magazine/wellbeing/working-and-heart-problems (Accessed 11 February 2023)

Mayo Foundation for Medical Education and Research. (2022, August 25). *Heart disease*. Mayo Clinic. Available at: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118 (Accessed 11 February 2023)

Barth, S. (2022, November 15). *Machine learning in healthcare - benefits & use cases*. ForeSee Medical. Available at: https://www.foreseemed.com/blog/machine-learning-in-healthcare#:~:text=Machine%20learning%20in%20healthcare%20is%20becoming%20more%20widely,support%20and%20the%20development%20of%20clinical%20care%20guidelines. (Accessed 13 February 2023)

Katzenstein, L., & Pinã Ileana L. (2007). *Living with heart disease: Everything you need to know to safeguard your health and take control of your life*. AARP/Sterling Pub. Co (Accessed 13 February 2023)

Whyte, J. J., & McGraw, P. C. (2023). *Take control of your heart disease risk*. Harper Horizon, an imprint of HarperCollins Focus LLC. (Accessed 14 February 2023)

Rawat, S. (2021, June 28).*Heart disease prediction*. Medium. Available at: https://towardsdatascience.com/heart-disease-prediction-73468d630cfc (Accessed 14 February 2023)

Mordecai, A. (2020, April 28).*Heart disease risk assessment using machine learning*. Medium. Available at: ttps://towardsdatascience.com/heart-disease-risk-assessment-using-machine-learning-83335d077dad (Accessed 15 February 2023)

Yazdani, A., Varathan, K. D., Chiam, Y. K. K., Malik, A. W., & Wan Ahmad, W. A. (2021, June 21). *A novel approach for heart disease prediction using strength scores with significant predictors - BMC Medical Informatics and Decision making*. BioMed Central. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01527-5 (Accessed 15 February 2023)

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019, December 21). *Comparing different supervised machine learning algorithms for Disease Prediction - BMC Medical Informatics and decision making*. BioMed Central. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8 (Accessed 17 February 2023)

Mehta, A. (2022, April 28). *An ultimate guide to understanding supervised learning*. Digital Vidya. Available at: https://www.digitalvidya.com/blog/supervised-learning/ (Accessed 17 February 2023)

Suresh, A. (2021, June 22). *What is a confusion matrix?* Medium. Available at: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 (Accessed 18 February 2023)

Shiny. (2012, November). Shiny. Available at: https://shiny.rstudio.com/ (Accessed 10 January 2023)

Sharma, R. (2022, November 24). *Data scientist vs software developer [Ultimate Comparison Guide]*. upGrad blog. Available at: https://www.upgrad.com/blog/data-scientist-vs-software-developer-comparison-guide/ (Accessed 18 February 2023)

IntelliPaat. (2023, March 21). *Data Science vs Software Engineer: What to choose and why?* Intellipaat Blog. Available at: https://intellipaat.com/blog/data-science-vs-software-engineering/ (Accessed 20 February 2023)

Hotz, N. (2023, January 19). *What is CRISP DM?* Data Science Process Alliance. Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed 20 February 2023)

Hotz, N. (2023, January 19). *What is KDD and Data mining?* Data Science Process Alliance. Available at: https://www.datascience-pm.com/kdd-and-data-mining/ (Accessed 24 February 2023)

Hotz, N. (2023, January 19). *What is SEMMA?* Data Science Process Alliance. Available at: https://www.datascience-pm.com/semma/ (Accessed 24 February 2023)

Wikipedia. (2023, March 3). *RStudio*. Wikipedia. Available at: https://en.wikipedia.org/wiki/RStudio#:~:text=RStudio%20is%20an%20integrated%20development%20environment%20%28IDE%29%20for,allows%20accessing%20RStudio%20using%20a%20web%20browser%20 (Accessed 3 March 2023)

Kosourova, E. (2022, September 21). *RStudio tutorial for Beginners: A complete guide*. DataCamp. Available at: https://www.datacamp.com/tutorial/r-studio-tutorial (Accessed 3 March 2023)

cran.r. (2023, March 17). *Introduction to HTML Widgets*. Introduction to HTML widgets. Available at: https://cran.r-project.org/web/packages/htmlwidgets/vignettes/develop_intro.html (Accessed 5 March 2023)

Wikipedia. (2023, March 25). *Decision Tree*. Wikipedia. Available at: https://en.wikipedia.org/wiki/Decision_tree (Accessed 7 March 2023)

Wikipedia. (2023, March 25). *Naive Bayes classifier*. Wikipedia. Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (Accessed 7 March 2023)

Wikipedia. (2023, March 25). *Linear regression*. Wikipedia. Available at: https://en.wikipedia.org/wiki/Linear_regression (Accessed 7 March 2023)

Wikipedia. (2023, March 25). *Support Vector Machine*. Wikipedia. Available at: https://en.wikipedia.org/wiki/Support_vector_machine (Accessed 7 March 2023)

Shiny. (2012, November). Shiny. Available at: https://shiny.rstudio.com/deploy/ (Accessed 10 March 2023)

Wikipedia. (2023, March 25). *XGBoost*. Wikipedia. Available at: https://en.wikipedia.org/wiki/XGBoost (Accessed 10 March 2023)

Imperva. (2020, September 24). *What is Black Box Testing: Techniques & Examples*. Learning Center. Available at: https://www.imperva.com/learn/application-security/black-box-testing/ (Accessed 12 March 2023)

Appendix A Personal Reflection

A.1 Reflection on Project

The project on heart disease prediction was undertaken with the aim of addressing the global issue of heart disease. To do so, the research began by conducting a thorough analysis of previous studies and research on heart disease prediction.

This research was focused on heart disease and the different machine learning algorithms used to predict it. Certain studies have been conducted to improve the accuracy of these models by using hyperparameter tuning and cross-validation techniques. It is essential to ensure that the research was completed without any test failure.

Overall, for developing accurate prediction models, this study has the potential to contribute to the ongoing efforts to address the global issue of heart disease.

A.2 Personal Reflection

If I had more time, I could have explored additional machine learning algorithms that may have potentially improved the accuracy of the heart disease prediction model. I could have investigated the impact of other medical conditions, such as diabetes or cancer. The application on Shiny can be further enhanced by adding more advanced features and functionalities. However, the study still provides valuable insights and can serve as a foundation for future research on heart disease prediction using machine learning algorithms.

Appendix B Ethics Documentation

B.1 Ethics Confirmation

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

17 January 2023

LETTER OF CONFIRMATION

Applicant: Mr Tanveer Dalal

Project Title: Heart Disease Prediction

Reference: 41705-NER-Jan/2023- 43177-1

Dear Mr Tanveer Dalal

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has confirmed that, according to the information provided in your application, your project does not require ethical review.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research, you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

Figure 30: Ethics Letter

Appendix C Confusion Matrices of Hyperparameter

C.1 Support Vector Machine

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 147   0
         1   4 156

               Accuracy : 0.987
                 95% CI : (0.967, 0.9964)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9739

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 0.9735
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9750
             Prevalence : 0.4919
         Detection Rate : 0.4788
   Detection Prevalence : 0.4788
      Balanced Accuracy : 0.9868

       'Positive' Class : 0
```

C.2 Naive Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 119  12
         1  30 145

               Accuracy : 0.8627
                 95% CI : (0.819, 0.8993)
    No Information Rate : 0.5131
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7245

 Mcnemar's Test P-Value : 0.008712

            Sensitivity : 0.7987
            Specificity : 0.9236
         Pos Pred Value : 0.9084
         Neg Pred Value : 0.8286
             Prevalence : 0.4869
         Detection Rate : 0.3889
   Detection Prevalence : 0.4281
      Balanced Accuracy : 0.8611

       'Positive' Class : 0
```

C.3 Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 137  18
         1  13 139

               Accuracy : 0.899
                 95% CI : (0.8597, 0.9304)
    No Information Rate : 0.5114
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7981

 Mcnemar's Test P-Value : 0.4725

            Sensitivity : 0.9133
            Specificity : 0.8854
         Pos Pred Value : 0.8839
         Neg Pred Value : 0.9145
             Prevalence : 0.4886
         Detection Rate : 0.4463
   Detection Prevalence : 0.5049
      Balanced Accuracy : 0.8993

       'Positive' Class : 0
```

C.4 XGBoost

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 150  10
         1  11 136

               Accuracy : 0.9316
                 95% CI : (0.8973, 0.9572)
    No Information Rate : 0.5244
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8629

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9317
            Specificity : 0.9315
         Pos Pred Value : 0.9375
         Neg Pred Value : 0.9252
             Prevalence : 0.5244
         Detection Rate : 0.4886
   Detection Prevalence : 0.5212
      Balanced Accuracy : 0.9316

       'Positive' Class : 0
```

C.5 Linear Regression

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 120  12
         1  26 149

               Accuracy : 0.8762
                 95% CI : (0.8341, 0.9109)
    No Information Rate : 0.5244
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7507

 Mcnemar's Test P-Value : 0.03496

            Sensitivity : 0.8219
            Specificity : 0.9255
         Pos Pred Value : 0.9091
         Neg Pred Value : 0.8514
             Prevalence : 0.4756
         Detection Rate : 0.3909
   Detection Prevalence : 0.4300
      Balanced Accuracy : 0.8737

       'Positive' Class : 0
```