University of Surrey

Department of Computer Science

Development of a Cardiovascular Disease Risk Prediction System using Machine Learning and Patient Health Data

Jason Jay Dookarun

URN: 6494278

Supervisor: Dr Tom Thorne

A dissertation submitted in partial fulfilment of the
requirements for the award of
Master of Science in Data Science

Declaration

I, Jason Jay Dookarun, hereby certify that this dissertation is my own work and that the figures, tables, formulas, and code segments in this report are original, with the exception that any work of others is acknowledged, quoted, and referenced.

I give consent to a copy of my report being shared with future students as an exemplar, to members of the University of Surrey and the general public with an interest in teaching, learning, and research.

Jason Jay Dookarun
August 31, 2023

Abstract

Cardiovascular diseases (CVDs) are a major global health concern, necessitating the development of effective risk prediction systems for early detection and prevention. The goal of this project is to create a comprehensive cardiovascular disease risk prediction system by using a combination of traditional statistical models and advanced machine learning methods and algorithms.

The study begins with a thorough review of the existing literature, dissecting established risk assessment frameworks while identifying gaps and limitations in current methodologies. An extensive dataset of anonymised patient records with clinical, demographic, and lifestyle variables is analysed in depth to gain a nuanced understanding of factors influencing cardiovascular disease risk.

Building on this foundation, we use a variety of predictive models for training and evaluation, ranging from logistic regression and support vector machines to more advanced techniques like ensemble methods. The performance of each model is rigorously evaluated using metrics such as accuracy, precision, and recall, establishing the dependability and effectiveness of our approach.

This research results in a system that can effectively predict if an individual is at high risk of CVD. This system, in effect, utilises machine learning models to effectively understand how a model has performed and present findings to the user. These scores can be used by healthcare professionals to make data-driven decisions on preventive interventions and patient management strategies, thereby improving the quality of cardiovascular care.

The research concludes with an assessment of the system's performance metrics, confirming its suitability for practical, clinical applications. We also discuss potential limitations and future research directions, laying the groundwork for future advances in the rapidly evolving field of cardiovascular disease risk prediction using machine learning techniques.

Keywords: Cardiovascular Diseases (CVD), CRISP-DM Model, Data Science, Data Modelling, Ensemble Techniques, Data Analysis, Bootstrap Aggregation

Total Word Count: 17246

Acknowledgements

There are several individuals to whom I would like to express my gratitude. Firstly, I extend my heartfelt thanks to Dr Tom Thorne for his invaluable guidance and support throughout the course of this research project. I greatly appreciate his acceptance of my work ethic, his unwavering support, and his enthusiasm that motivated me to pursue this project in a field that I find deeply fascinating.

I would also like to offer my sincere thanks to my friends for their consistent support and encouragement. They have been instrumental in my educational journey, often providing critical academic support when needed.

Lastly, my deepest appreciation goes to my siblings, Ryan and Laura, as well as my parents, Hemlatah and Prithiviraj Dookarun. Their ceaseless motivation and encouragement have inspired me to strive to become the best version of myself. The completion of this project, and indeed my entire educational journey at the University of Surrey, would not have been possible without their unwavering support and love. I dedicate all my hard work and efforts to them.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
ASCVDRE	ASCVD Risk Estimator
BP	Blood Pressure
BMI	Body Mass Index
CAD	Coronary Artery Disease
CSV	Comma-Separated Values File
CV	Cross Validation
CVD	Cardiovascular Diseases
DL	Deep Learning
DS	Data Science
DT	Decision Tree
EDA	Exploratory Data Analysis
ECG	Electrocardiogram
ETL	Extraction, Loading, Transformation
FRS	Framingham Risk Score
HDL	High-Density Lipoprotein
KNN	K-Nearest Neighbors
LDL	Low-Density Lipoprotein
LR	Logistic Regression
ML	Machine Learning
MSE	Mean Squared Error
NB	Naïve Bayes
NN	Neural Networks
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristics
RRS	Reynolds Risk Score

Chapter 1

Introduction

1.1 Background

Cardiovascular diseases, also known as CVDs, have emerged as one of the leading contributors to mortality over the past decades, accounting for a significant proportion of deaths worldwide. In the UK alone, approximately 66,000 deaths are attributed to coronary heart disease, equating to an average of 180 people per day [Foundation, 2021]. Coronary heart diseases encompass a range of conditions affecting the heart and blood vessels, including coronary artery disease, myocardial infarction (also known as heart attacks), and strokes. Additionally, CVDs also represent other conditions such as cerebrovascular disease and rheumatic heart disease (RHD). According to the World Health Organisation, cardiovascular diseases are the "leading cause of death globally, responsible for approximately 17.9 million deaths each year" [Foundation, 2021], a statistic published prior to the COVID-19 pandemic of 2019.

Several factors influence the prevalence of cardiovascular diseases, including age, genetic predisposition, lifestyle choices, and underlying health conditions. Physical inactivity, smoking, poor diet, and excessive alcohol consumption are lifestyle choices that contribute to the development of risk factors such as hypertension, obesity, high cholesterol levels, and diabetes. These risk factors, in turn, increase the likelihood of developing cardiovascular diseases.

1.2 Problem statement

Even though cardiovascular diseases (CVDs) have significant adverse effects on public health, there is an urgent need for precise and trustworthy prediction methods. The complexity of the disease may not be fully captured by current methods of CVD prediction, which might result in subpar accuracy and predictive performance. Current methods for CVD prediction frequently rely on conventional risk factors and clinical assessments. As a result, there is a void in the ability to accurately identify those at risk of CVDs and execute preventative actions.

Developing a CVD prediction system using machine learning (ML) techniques offers a promising solution to this problem. This project aims to design and implement a machine learning-based CVD prediction system to improve the accuracy and efficacy of CVD risk assessment. The system will include a number of variables, such as medical history and key metrics, allowing for a comprehensive analysis that goes beyond traditional risk factors.

1.2.1 Use Cases and User Stories

To effectively build upon a successful product, it is important to consider use cases and user stories. Use cases, as defined by Matz and Germanakos in "The Increasing the Quality of Use Case Definition Through a Design Thinking Collaborative Method and an Alternative Hybrid Documentation Style", illustrate "the aim and subsequent objections of a system and the assigned user role (named actors), by expressing a list of steps and interactions among them towards a common goal" [Germanakos and Matz, 2016]. By introducing and utilising use cases and this principle, one can then comprehend the end goal of a product based on its functionality with a stakeholder, namely the user.

For this project, both use cases and user stories are to be employed to maximise the scope of this project and to elaborate and explore all avenues that may be attained as part of this project, albeit with different perspectives and users.

User stories are brief, casual narratives that portray desired functionality from the viewpoint of the end-user. They frequently stick to a straightforward format: "As a [user role], I want [goal] for the reason that [reason]." User experiences highlight the value or benefit that the user seeks from the product by focusing on their needs, wants, and motivations. They are frequently employed in Agile methodologies. Use cases, on the other hand, give a more organised and in-depth account of how users and the system interact. A use case typically consists of a series of scenarios or steps that show how a user (or actor) interacts with the system to complete a particular task. They provide a more formal definition of the actions, inputs, and anticipated outcomes.

The support with the explanation of the key differences between the user story and use case, the following table (see Table 1), has been presented below, developed by the team from Azoft.

Table 1.1: Comparison between User Stories and Use Cases

User Stories	**Use Cases**
Focus is on the value	The focus is on the behaviour
Less vital details are overlooked	The complete scenario is described
Easy for businesses to understand	Created to be understood by the team

There are numerous benefits associated with use cases and why they are beneficial. The data that was analysed for this study, as outlined in "The Use of Effectiveness of User Stories in Practise" by Garm Lucassen, Fabiano Dalpiaz, Jan Martijn E.M. van der Werf, and Sjaak Brinkkemper, outlines that by using user stories, a template, "productivity and the quality of their work deliverables" were improved [Lucassen et al., 2016]. Furthermore, use cases present other benefits, such as the ability to apply these points to the development of a product. I, as the solution's developer, can effectively align with a user's requirements and expectations by using use cases.

To elaborate further, a set of user stories has been listed below.

Table 1.2: Use Case Descriptions and Stakeholders

Use Case #	Use Case Description	Stakeholders Involved
UC1	As a healthcare provider, I want to know which factors may contribute to CVD	Healthcare Provider
UC2	As a researcher, I want to understand what the most relevant features from the dataset for an accurate CVD risk prediction system are.	Data Scientist, Supervisors and Thesis Researchers
UC3	As a researcher, I want to compare how different ML models perform, and which is the most effective towards the final product.	Data Scientists, Framingham Dataset Group

The project can be viewed from a broader perspective by utilising the user stories and use case process. As previously stated, using cases has numerous advantages, such as the ability to preview a project and outline expectations. Individuals such as developers and stakeholders can view a project's projection, goals, and objectives while adjusting before deployment. In effect, the project intends to effectively contribute to the advancement of the CVD risk prediction system by following and utilising these use cases as a foundation.

1.3 Aims and objectives

The project's goal is to create a predictive tool using a variety of resources such as case studies, user stories, and data. In effect, the tool would give the user the tools they need to understand the impact certain factors have on a nation, including health-related factors. To that end, the primary goals of the project are as follows:

- **Risk Prediction:** This project aims to develop a risk prediction system. To achieve this, the model should be able to identify patterns and use this knowledge to predict if an individual is at risk.
- **Decision Support:** By implementing this style of system, we intend to support the decision process made by individuals or healthcare professionals. As outlined, this project aims to further bolster the decision-making system, and, in effect, by using this tool, healthcare professionals have further backed to support their decision-making.
- **Integration and Accessibility:** As a user, and a member of the public, it can often be difficult to retrieve or use a system that can be accessible by all. By developing this application, we aim to allow for ease of use and access when using this product.
- **Validation and Performance Assessment:** To assess the accuracy and reliability of the developed tool, compare its predictions against real-world data and known outcomes. To ensure the tool's effectiveness in risk prediction, conduct thorough performance evaluations that include metrics such as accuracy, sensitivity, recall and precision.

By pursuing these goals and objectives, the project hopes to provide healthcare professionals with a reliable tool that can be used to accurately predict cardiovascular disease risks. This will allow for earlier detection, more personalised interventions, and better patient outcomes in the prevention and management of cardiovascular disease.

1.4 Approach to Solution

A comprehensive solution approach has been established to ensure that the aims and objectives are met and successfully achieved. The first step is to identify relevant data sources or warehouses that will allow for a comprehensive understanding of the available data and optimise the retrieval of detailed insights.

The following steps will be taken to begin the process:

1. **Data Source Identification and Exploration:** Determine potential data sources or data warehouses that contain the information required for the project at hand. This could entail analysing and comprehending the chosen dataset through explanatory data analysis, data cleansing, transformation, and label encoding.
2. **Data Modelling:** Apply appropriate modelling techniques to analyse the data and make predictions or classifications based on the project's objectives. This step involves selecting suitable modelling approaches, feature engineering, and model training using the prepared data.
3. **Model Evaluation:** Assess the performance of the developed models by evaluating relevant metrics. This step helps in determining the effectiveness and reliability of the models and identifying areas for improvement.
4. **Data Visualisation:** Present the analysed data and model results using visualization tools and techniques. This step involves creating insightful charts, graphs, dashboards, or interactive visualisations to effectively communicate findings and patterns in the data. Visualisation aids in understanding complex relationships and conveying information to stakeholders.

It is critical to iterate and refine the process based on the outcomes and feedback received throughout these steps. This iterative approach allows for continuous improvement while also ensuring that the goals and objectives are met. This comprehensive approach allows for informed decision-making and increases the project's success in meeting its objectives.

1.5 Stakeholders

As part of this project, it is critical to consider any stakeholders who may be involved in the project, whether from a short-term to a long-term perspective. A stakeholder, as defined by McGrath and Whitty in "Stakeholder, defined," is "a member who has some control over the activity" [McGrath and Whitty, 2017]. Among the key stakeholders involved in this project are:

- **Jason Jay Dookarun,** author, and developer of this new CVD Development System. As part of this project, the term "we" will be used for documentation, however no additional support or input has been provided for this project.

- **Dr. Tom Thorne,** supervisor of this Master's Project.
- **The National Heart, Lung, and Blood Institute,** providing access to the public Framingham Heart Study Group dataset.

Regarding the deployment of the project, we believe that this project could be beneficial to countless users. This would include not only medical personnel but would also include those who wish to further query their risks. By developing a tool that also takes advantage of Machine Learning algorithms and methodologies, one can then utilise the technology to enhance the probability of a correct prediction and thus support the future stakeholder in decision-making.

1.6 Organisation of the report

This project has been structured comprehensively as part of the documentation process to effectively present and elaborate on various components associated with the topic. The following outline highlights the report's key sections:

- **Literature Review:** This section provides a review of relevant literature, including previous studies, research papers, and scholarly articles on the subject. Its goal is to lay a durable foundation of knowledge and understanding in the field.
- **Models Research:** In this section, various models, algorithms, and methodologies relevant to the research topic are explored. The models will also consist of understanding and analysing how they function.
- **Exploration of Alternative Modelling Algorithms, Including Stacking:** This section investigates alternative modelling methods that may improve the performance and robustness of the models. The concept of "stacking" is specifically investigated, and its potential benefits are assessed.
- **Data Understanding:** Building on the EDA, this section focuses on gaining a deeper understanding of the dataset's characteristics, such as its structure, variables, and relationships. This understanding serves as the foundation for future modelling efforts.
- **Using Information to Rebuild Models Through Stacking:** Using the knowledge gained in the preceding steps, this section focuses on rebuilding and improving the models using the stacking technique. To improve predictive accuracy and generalisation, multiple models are combined.
- **Testing:** Extensive testing is performed to evaluate the robustness and reliability of the models developed. To validate the performance and ensure the models can deliver accurate predictions, various validation techniques and metrics are used.
- **Conclusion:** Finally, the report concludes by summarising the key findings, discussing the research's implications, and emphasising any recommendations for future work. This section summarises the entire study, emphasising its significance and contributions to the field.

Chapter 2

Literature Review

The following section will provide a detailed explanation of the literature aspects pertaining to this project. It encompasses various components, including an overview of the existing background in this field, an examination of the current solutions that might already exist, and an analysis of the advantages they offer. Moreover, this section will delve into research, elaborating on important aspects such as dataset selection, considerations in modelling, and relevant research in the field of data science, as applicable. To support this discussion, a thorough exploration of different models will be conducted to ascertain the most suitable one for the selected dataset. Additionally, ensemble methods will be examined, exploring their principles and potential applications. Finally, from a theoretical standpoint, this section will provide a concise summary of the proposed approach, followed by the practical implementation of data exploratory analysis.

2.1 Background

Cardiovascular diseases, also known as CVD, refers to a category of conditions that are linked and related to the cardiovascular system. CVD is a broader term utilised to classify conditions and varying disorders that are related to the CVD system, affecting the heart, arteries, veins, and capillaries. The disease, also often referred to as heart disease can be divided into four key "entities", namely, "coronary artery disease (CAD) which is also referred to as coronary heart disease (CHD), cerebrovascular disease, peripheral artery disease (PAD), and aortic atherosclerosis" [Lopez et al., 2022]. CVD in general, from a perspective of statistics, leads to an estimated 17.9 million deaths a year, with more than 4/5 deaths due to heart attacks or strokes, and 1/3 of deaths recorded in people under the age bracket of 70 years old [Organization, 2022].

CVD diseases can occur for a variety of reasons, which may include genetic, environmental, and lifestyle factors. From the perspective of lifestyle factors, these can encompass behaviours such as smoking or exposure to second-hand smoke, obesity, an unhealthy diet, and physical inactivity [Ding et al., 2020]. While certain factors, like genetics, family history, age, and sex, cannot be changed, many other elements, particularly those related to lifestyle, can be modified to reduce the risk of developing CVD.

One of the primary mechanisms underlying the development of cardiovascular disease (CVD) is atherosclerosis [Björkegren and Lusis, 2022]. Atherosclerosis significantly contributes to the progression of CVD. It occurs when fatty deposits, cholesterol, calcium, and other substances

gradually build up within the walls of arteries, leading to plaque formation. Over time, these plaques can harden, causing the arteries to narrow and thereby obstructing blood flow to vital organs such as the heart, brain, and extremities.

High blood pressure, also known as hypertension [Opa, 2019], significantly increases the risk of developing cardiovascular disease (CVD). The constant strain that high blood pressure places on the delicate inner lining of arterial walls can result in damage and inflammation. This damaged state creates an ideal environment for the onset and progression of atherosclerosis, thereby narrowing the arteries further and escalating the likelihood of cardiovascular events such as heart attacks and strokes. Thus, maintaining healthy blood pressure levels is crucial for preserving cardiovascular health and mitigating the risk of CVD-related complications.

The arteries that supply blood to the heart muscle are particularly susceptible to a condition named coronary artery disease (CAD), which is a specific type of cardiovascular disease (CVD). CAD occurs when the coronary arteries become narrowed or blocked due to the presence of atherosclerosis [Shahjehan and Bhutta, 2023]. In cases where a blood clot completely obstructs an artery, the reduced blood supply to the heart can lead to the development of angina (chest pain) or resulting in a heart attack. It is crucial to address risk factors and manage CAD effectively to prevent these potentially life-threatening cardiovascular events.

It is important to recognise that cardiovascular disease (CVD) comprises a wide range of problems in addition to those that have already been described, such as peripheral artery disease, heart valve issues, and distinct types of strokes. Each condition has unique origins and mechanisms, but they are all characterised by a reduction in the cardiovascular system's ability to operate properly. Whether it is the narrowing of peripheral arteries, the malfunctioning of heart valves, or disruptions in blood flow to the brain, these various manifestations of CVD collectively contribute to the overall burden on the cardiovascular system. Understanding the diverse nature of CVD empowers healthcare professionals to develop comprehensive strategies for prevention, diagnosis, and management, aiming to address the unique challenges posed by each condition within the broader spectrum of cardiovascular health.

Once an individual is diagnosed with CVD, one can experience a series of symptoms, as outlined by the British Heart Foundation. The BHT [Ano, n.d.] outlines that dependent on an individual's condition, symptoms of heart disease can include:

- Chest pain
- Pain, weakness, or numb legs and/or arms
- Breathlessness
- Extremely fast or slow heart beats, or palpitations
- Feeling dizzy, lightheaded, or faint
- Fatigue
- Swollen limbs

In addition, those who have experienced a heart attack or stroke may experience long-term consequences including movement difficulties, or speech abnormalities. Furthermore, some

CVD treatments (both short-term and long-term), such as prescription drugs or surgical procedures like stent surgery or bypass surgery [Ric, 2022], may have unfavourable side effects, such as bleeding or infection. It is also worth outlining that stent surgery patients are still prone to CVD diseases such as artery blockages, leading to further discomfort and angina.

Regarding current methods of identification and treatments in place for CVD, the methods involve a comprehensive approach aimed at managing and altering certain factors within the lifestyle factor bracket of an individual's life. This can include the improvement of quality of life, as a result contributing to a lower risk of CVD. Similarly, there are varying treatments that are currently present within the medical space, from adopting healthy diets to cholesterol-lowering drugs. Studies show that statins, also recognised within the medical space as cholesterol-lowering drugs reduce health risks related to atherosclerosis. By having such drugs, a step inhibited by the liver to synthesise cholesterol is employed, and as a result, applied [Ziaeian and Fonarow, 2017]. Other medical substances such as blood pressure medications, and antiplatelet agents can be utilised to control heart rhythm abnormalities. In more severe cases, interventions such as bypass surgery or stent implants are utilised, as these are necessary to allow for effective restoration of blood flow. Alternative and advanced treatments can also be used in acute cases, including implantable devices such as pacemakers or defibrillators, to optimise heart function.

A variety of techniques are used to evaluate the existence and seriousness of cardiovascular problems when it comes to the identification of CVD. Measurements of blood pressure, cholesterol, and blood sugar may be part of routine screens to look for CVD risk factors or early warning symptoms. Heart function, anatomy, and blood flow are all revealed through diagnostic procedures including electrocardiograms (ECGs or EKGs), stress tests, echocardiograms, and cardiac imaging methods (such CT scans or MRI).

It is significant to highlight that, as medical science develops, approaches for treatment and detection continue to change. To enhance results and give more accurate evaluations of CVD, new medicines, minimally invasive techniques, and innovative diagnostic technologies are continually being developed. To be informed about the most recent treatment choices and to ensure the early identification and management of CVD, regular follow-up appointments with medical specialists are necessary. Additionally, systems that allow people to input information to determine their risk of CVD are already in place and will be covered later.

Addressing this significant global health issue necessitates a comprehensive understanding of the fundamental causes and underlying processes of cardiovascular disease (CVD). By recognising the risk factors associated with CVD, promoting healthy lifestyles, and implementing targeted therapeutic approaches, we can effectively reduce the burden of CVD and strive towards improved cardiovascular health worldwide.

As a result, my intention as part of this project is to develop a system, such that a user can then effectively detect if they are at risk of CVD, with consideration to a multitude of factors, namely lifestyle and genetic factors. To successfully understand and analyse how to build the correct system to apply such infrastructures, existing systems are to be examined. By examining existing systems, we can understand the existing market, and what is working well, what is not working well, and what could be effectively adopted, yet altered for a better solution for the stakeholders involved, and future clientele.

2.2 Existing Systems

Based on a set of parameters specified by the user, various methods can be employed to assess an individual's susceptibility to cardiovascular diseases. These methods are adaptable to diverse settings, including healthcare facilities and private homes. To enable this, the system could be hosted on a user-friendly website or implemented through a dedicated solution, both of which would allow for straightforward data input and output. Such platforms make it simple for individuals to input their information and receive precise assessments of their vulnerability to cardiovascular diseases."

A thorough review and assessment of several existing systems were conducted to acquire a comprehensive understanding of this subject matter. The objective is to grasp the diverse elements associated with each system, including factors contributing to their success, the methods employed in their formulation or creation, the specific models used for generating outcomes, and the benefits identified within each system.

To commence, the first product/system that was reviewed as part of this literature review focused on analysing the functionality of the Reynolds Risk Score. The Reynold Risk Score [Rid, 2018], developed by Dr. Paul M Ridker is a system that focuses on taking input from the user to then predict whether one may be deemed to be prone to cardiovascular diseases in the next 10 years. The existing system, as shown in Figure 2.1, allows the user to enter metrics in the form of parameters. The Reynolds Risk Score, according to Klisi [Klisić, 2018] in Acta Clinica Croatica, incorporates several factors, including age, gender, total cholesterol, HDL-d, smoking status, hs-CRP level, family history of heart attack before the age of 60, and SBP. This score system was used in conjunction with the Framingham system. The Reynolds Risk Score System, which was developed primarily for women, uses traditional risk factors such as CRP levels and parental history, as discussed by Lloyd-Jones [Lloyd-Jones, 2011]. The goal is to improve the modelling process and the Bayes information criterion.

The ability to reclassify women in the United States is a significant advantage of using this method rather than a traditional Framingham calculator. According to Lloyd-Jones in Concepts of Screening for Cardiovascular Risk Factors and Diseases, "the Reynolds Risk Score revealed that when applied, 5.8 percent of 8149 women were reclassified compared to FRS, with approximately an equal number reclassified upward and downward" [Lloyd-Jones, 2011].

A variety of renditions of the Reynolds Risk Score are available for personal use and can be obtained. We will use the official Reynolds Risk Score to evaluate its functioning and performance in this case.

Figure 2.1: Reynolds Risk Score System [Rid, 2018]

When examining the success of the RRS (Reynolds Risk Score) system and its performance in comparison to its competitors, particularly the FRS (Framingham Risk Score), several pieces of research are available for reference. This includes an article authored by Dr Tomoyuki Kawada [Kawada, 2023], in which Kawada outlines various components and elements by which RRS outperforms FRS, as well as specific factors contributing to RRS's shortcomings. As described in the article, RRS outperforms the FRS system in certain areas, notably the incorporation of family history, which has a significant impact on CVD. Furthermore, creating such a tool is advantageous as it can be 'useful for recommending risk-lowering actions for each individual, such as smoking cessation, regular exercise habits, and dietary regulation of calorie and salt intake.' However, comparative investigations, such as the one cited in Dr Kawada's study and published in the Journal of the American College of Cardiology [De Filippis et al., 2011], suggest that RRS may be linked to metabolic syndrome, as several components overlap.

Furthermore, some limitations relating to ethnicity and age are presented in the RSS, as cited by Dr Kawada [Kawada, 2023]. These are important factors because different ethnicities, such as black Africans, African Caribbeans, and South Asians in the UK, are more likely to develop either high blood pressure or type 2 diabetes than white Europeans [Foundation, 2021].

The QRisk Calculator system was the next example/existing system examined as part of this study. The QRisk calculator was created by the QRisk Research Team to estimate one's risk of developing CVD over a 10-year period [Hippisley-Cox et al., 2017]. It is mostly used in the United Kingdom and takes into account a variety of risk factors to provide a personalised risk assessment.

Age, gender, ethnicity, smoking status, systolic blood pressure, body mass index (BMI), diabetes status, cholesterol levels (total cholesterol/HDL cholesterol ratio), family history of CVD, and use of antihypertensive medication are all risk factors considered by the QRisk Calculator. These variables are used to calculate a person's CVD risk score.

In terms of performance, QRisk has been shown to perform well in predicting CVD risk in the UK population. Several studies have demonstrated its accuracy and superiority over other risk assessment tools, such as the Framingham Risk Score (FRS) and ASCVD Risk Estimator, in predicting CVD events in the UK population [Hippisley-Cox and Coupland, 2012]. Moreover, the QRisk calculator functions and presents the results in the form of a percentage.

An example of this can be seen and found below, in Figure 2.2. Like Figure 2.1, the QRisk calculator focuses on collecting a range of parameters to produce the most accurate model. When analysing the content presented both in Figure 2.2 and 2.2 , one can immediately identify differences between the requested data.

Figure 2.2: QRisk Calculator [Not, 2018]

Upon analysis, numerous changes can be identified, which may be factors and reasons why QRisk is a more popular choice in the UK and one that may provide higher levels of accuracy, albeit while incorporating features used in the Framingham calculator. The incorporation of ethnicity was one of the notable features considered for the purposes of QRisk. According to the 2021 UK consensus [GOV, 2022], more than 18% of the UK's population is black, Asian, mixed, or other ethnic group. This feature is critical because ethnicity can influence one's risk of developing potential CVD risks. According to Chaturvedi, "migrants of South Asian descent worldwide have elevated risks of morbid and mortal events because of ischaemic heart disease (IHD)"[Chaturvedi, 2003] and that "in the UK, mortality from IHD in both South Asian men and women is 1.5 times that of the general population (see Appendix D), and South Asians have not benefited to the same extent from the general decline in deaths caused by IHD over the last few decades."

The final example that was examined as part of this literature review focused on the ASCVD Risk Estimator. The ASCVD Risk Estimator is a tool developed by the American College of Cardiology (ACC) and the American Heart Association (AHA) to "estimate an individual's risk of developing atherosclerotic cardiovascular disease (heart attack and stroke) over the next 10 years" [of Cardiology, 2021]. It is used to assess the risk of heart attack and stroke in adults.

The risk estimator takes into account a number of risk factors that have been linked to the development of atherosclerotic cardiovascular disease. Age, gender, race, total cholesterol, high-density lipoprotein (HDL) cholesterol, systolic blood pressure, blood pressure-lowering medication use, diabetes status, and smoking status are all risk factors [of Cardiology, 2016].

When reviewing and comprehending all three existing systems that are in place to review one's

CVD risk, all three existing systems provide unique functionality. Each share additional factors that contribute to greater accuracy, namely factors such as ethnicity, and diabetic status. This can also be outlined via the ASCVDRE system, whereby the system implements similar features.

2.3 Selection of Dataset

Numerous datasets can be used from the datasets and databases that are available for use in this project. These can be obtained from the owner's source, such as the Framingham Dataset, Kaggle, and so on. Several datasets have been identified using Kaggle as a result of research and investigation. The Heart Disease UCI dataset, CVD Dataset, and CVD Risk Prediction Dataset were among them.

The Framingham Dataset is the first set of data being considered. The information is derived from the Framingham Heart Study (see Appendix B), a long-running cardiovascular study that began in 1948 and includes citizens of the town of Framingham, Massachusetts, in the United States. In addition to clinical parameters like blood pressure, cholesterol levels, and body mass index (BMI), these attributes also include demographic data such as age, gender, and educational attainment. The dataset also takes into consideration aspects of behaviour and lifestyle choices including drinking, smoking, and physical activity levels.

There are several constraints to consider when using this dataset. To begin with, the dataset's primary focus on Framingham residents may limit its applicability to other communities or racial or cultural groups. Furthermore, because the data was compiled over a long period of time, it may not include recent medical developments or specific emergent risk factors.

The National Health and Nutrition Examination Survey (NHANES) dataset [Dinh et al., 2019] is the following source being considered. The NHANES dataset is a highly renowned and commonly used dataset in the field of public health and nutrition research in the United States, which is viewed from the perspective of what the dataset contributes to the project. Like the Framingham Study (FS) Dataset, this dataset comprises a broad range of variables and features related to health and nutrition. It contains data on demographics including age, gender, race, and socioeconomic status. The NHANES dataset also includes several clinical parameters, including waist circumference, blood pressure, cholesterol, and glucose levels, as well as physical characteristics like height, weight, and age.

It offers a plethora of knowledge that makes it possible to investigate different medical diseases, dietary deficiencies, and the influence of lifestyle factors on population health outcomes. Researchers can better understand public health issues and create interventions and policies that are based on scientific evidence by studying this dataset to find trends, patterns, and connections between various variables. With a representative sample of the US population, the survey attempts to enable extrapolation of results to a larger population, and in effect, presents another advantage of using the NHANES dataset.

The NHANES dataset does, however, have some restrictions, just like any other dataset. First, it relies on self-reported data, which could contain errors or recall bias. In addition, because the survey is repeated, there are gaps in the data-gathering cycles. This delay can make it more difficult to detect recent changes or emerging health patterns.

"The Heart Disease UCI" [Sony, 2020] dataset is a collection of medical records from patients evaluated for heart disease at the Cleveland Clinic Foundation. It includes numerous clinical measurements, demographic data, and characteristics that can help predict the presence of heart disease. Age, gender, blood pressure, cholesterol levels, and electrocardiogram (ECG) readings are a few examples of these characteristics. The dataset's goal is to make it easier to create predictive models and algorithms that can correctly identify and forecast cardiac disease.

These datasets give researchers the chance to explore the fields of risk assessment, disease diagnosis, and cardiovascular health. Researchers can find patterns, correlations, and predictors of heart disease and cardiovascular risk by analysing these datasets. These discoveries can help in the creation of risk prediction models, individualised interventions, and more precise diagnostic tools for cardiovascular health.

Additionally, they share relevance to the real world is one of their advantages. They are derived from actual patient data and give information about the traits and risk factors that are common in populations that are assessed for heart disease. The ability to apply research findings in clinical practice is increased thanks to this authenticity, which also makes it easier to convert research findings into better patient care.

It is critical to realise and acknowledge that particular datasets have drawbacks and may not be a suitable match for this solution. As with any dataset, absent or omitted entries may have an impact on studies and predictions. Furthermore, these datasets only reflect a small percentage of the population and may not be completely representative of other geographic areas or the total population. As a result, when extending results to broader populations, caution is advised.

Among the options, the Framingham Dataset is the most suitable option for numerous reasons. Firstly, the data originates from the renowned Framingham Heart Study, a continually evolving study that has accumulated an extensive amount of knowledge over many years. It contains a wealth of demographic data, including age, gender, and educational attainment, which enables researchers to evaluate how these factors affect the onset and development of cardiovascular conditions. As a result, based on the above information and elaboration, we believe that to appropriately develop a solution, the Framingham Dataset would be the most appropriate, for the reasons outlined.

2.4 Machine Learning Modelling Principles

2.4.1 Supervised vs Unsupervised Models

It is critical to understand the underlying differences between supervised and unsupervised learning when addressing modelling frameworks used to ML or Data Scientific principles in general. Supervised learning and unsupervised learning are two distinct techniques that play critical roles in a variety of machine learning applications.

To begin, supervised learning entails training a model with tagged data. The supervised learning process focuses on using a user's inputs to generate a collection of outputs. These inputs are frequently in many formats, ranging from picture to text. According to IBM, "it

is defined by its use of labelled datasets to train algorithms that accurately classify data or predict outcomes" citepIBM. The supervised learning method focuses on not only forecasting the future value but also analysing the amount of accuracy generated. To understand the accuracy, consider the values from the original data source, also known as features, as inputs that are used to generate an output.

The fundamental approach of supervised learning can be divided into two categorial types, namely classification and regression. Classification as a concept and approach focuses on the process by which the aim and objective of the model is to predict a distinct and discrete class for a said input.

Classification tasks in the supervised modelling space focus on assigning input data points to pre-defined classes, and in effect, using this information to comprehend if a value belongs to a class or sample set. An example of this in the application can be spam detection. With the concept of spam detection, the algorithm and model can be effectively trained to understand if something such as an email or text belongs to a category or class named "spam," or if it should be marked as "non-spam." This can be achieved by using data samples, provided in the form of datasets [Dada et al., 2019]. By using this method, the model can successfully learn from features such as words, subject lines, sender and effectively output whether an input is in a class label of "spam" or "not spam."

Alternatively, supervised learning can also be classified as a regression model [Gup, 2021]). When a model is described to follow the protocols of a regression model, the objective is to predict a continuous output variable or an output in the form of a value based on inputs provided by the user. In effect, by doing so, the input data is studied to gain knowledge to effectively make a prediction.

An example of this in effect could include a task whereby a user wishes to predict the price of a house [Vol, 2019] based on a set of parameters, including the number of rooms if the property has a garage, the neighbourhood, date since construction, nearby stations, and additional factors. As this case focuses on predicting the value of a property, and as this task would involve handling discrete class labels, a regression model would be suited best for this task, as regression deals with continuous numerical values, aiming to predict a quantity or measurement based on input features.

Multiple models can be used for modelling and comprehending how a model may perform. From the context of decision trees, Song and Lu outline in their paper several common usages for using decision trees. These include variable selection, assessing the relative importance of variables, handling missing values, prediction, and data manipulation [Song and Lu, 2015]. Additionally, multiple Machine Learning models are recommended for supervised learning, both from the perspective of classification and regression. From the perspective and viewpoint of classification, models such as KNN (K-Nearest Neighbours), Random Forest, AdaBoost/XGBoost and Decision Trees can be used, as outlined in "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance" [Strecht et al., 2015]. Moreover, for regression, as outlined in the conference paper, examples include OLS (Ordinary Least Squares), SVM, CART, KNN, Random Forest, and AdaBoost.R2, a variation on the existing AdaBoost.

On the contrary, unsupervised learning is another machine learning technique that can be

utilised. "Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a dataset with no pre-existing labels and with a minimum of human supervision," as defined by Saman Siadati [Siadati, 2018]. One of the key differences with unsupervised learning is that the data provided is unlabelled and therefore not classified in advance. In the case of unsupervised learning, the aim and objective of the model are to discover underlying patterns and/or relationships between features without using pre-labelled data. Examples of unsupervised learning include clustering, PCA, and GANs. Clustering algorithms such as K-means focus on understanding how data points relate to one another; this is achieved by identifying any patterns or similarities that may be present. An example of this application can be demonstrated by Afolabi, who applies K-Means clustering to a FIFA dataset [Afolabi, 2023].

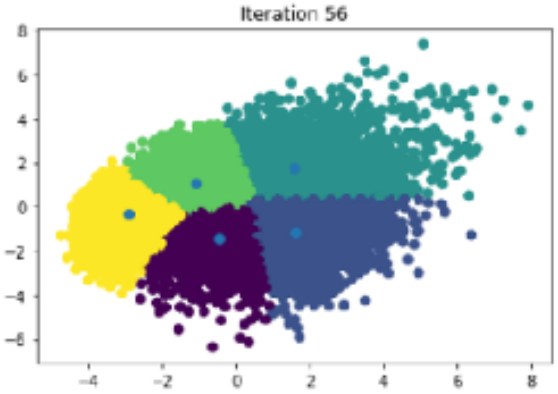

Figure 2.3: Cluster Formation After Iteration from Implementing K-Means Clustering. Each data and cluster is represented in a distinct colour with the cluster centroid being positioned and indicated by a blue marker.

The results present a variation of colours, presenting a series of clusters/categorised groups of data points that show relation to one another. Figure 2.3 uses the player dataset from the video game franchise FIFA, whereby features consist of player details, and statistics in dissimilar categories such as passing, shooting etc. Each cluster represents an average value for players ranked on FIFA, using data such as overall, potential, wage, value, and age as shown by Figure 5 in the technical report [Afolabi, 2023].

Consequently, other models that utilise unsupervised learning principles include PCA. PCA (Principal Component Analysis) is a technique and method used for feature extraction and dimensionality reduction. An example of this application, along with examples outlining how PCA functions, can be found in "The Tutorial on PCA and Appropriate PCA and Approximate Kernel PCA" by Sanparith Marukatat [Marukatat, 2022]. As mentioned by Marukatat, "PCA relies on a linear combination of these features to construct a principal subspace; this is the main subspace on which most of the feature vectors lie. PCA uses variance as a measure of the information content of the subspace." To understand the steps undertaken to complete PCA, a set of steps is outlined below, depicting each one.

Algorithm 1 Principal Component Analysis [OSu, 2020]

Step 1: Standardise the Data via standard data cleansing processes.
Step 2: Compute the Covariance Matrix of the Feature from the Dataset.
Step 3: Perform the eigen-decomposition of the covariance matrix.
Step 4: Order the eigenvectors in decreasing order based on the magnitude of corresponding eigenvalues.
Step 5: Determine the value k, representing the number of top principal components.
Step 6: Construct the new projection matrix using the top principal components selected.

Step 7: Compute the new k-dimensional feature space. $=0$

By using this method, numerous advantages can be drawn upon, including the ability to reduce any form of noise within the data and the capacity to produce uncorrelated features. Subsequently, key advantages of using PCA also include low noise sensitivity [Karamizadeh et al., 2013]. Owing to its low noise sensitivity, Principal Component Analysis is more robust to small fluctuations or random errors in the dataset. However, despite these advantages, PCA also has some drawbacks, such as information loss. In effect, when applying PCA to a dataset, the total number of features is effectively minimised. This consequently results in lower-ranked components being lost, along with any information they may have contained.

2.5 Selection of Models

2.5.1 Decision Trees

To begin, one of the models that will be explored as part of this project is based on a decision tree. The decision tree concept re-aligns with the concept of supervised learning, allowing it to be applied to both classification and regression models. The most common form of decision trees revolves around classification.

Decision tree algorithms, whether that is via application through classification or regression involve the partitioning of data based on a selected feature. The aim and objective of a decision tree is to increase the information gain, whilst ensuring that impurities are minimised. An example of this can be seen and presented below, illustrating the split of weather conditions, presented by Yadav [Yadav, 2018].

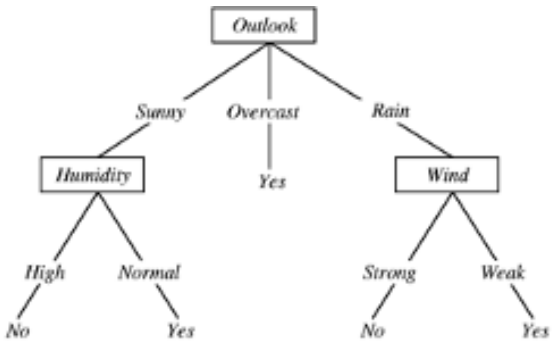

Figure 2.4: Decision Tree Principles[Yadav, 2018]

There are a series of metrics that are utilised as part of decision trees to comprehend and understand how accurately or how well a decision tree performs. This includes metrics such as the Gini index or entropy score[Kohavi and Quinlan, 2022]. A Gini score, as a metric and measurement, is utilised to measure impurity, employed to distinguish and identify how "well a split separates classes or categories within datasets" [Brownlee, 2020a]. The objective of applying a Gini index is to ensure that impurities are not present within the decision tree starting from the root node. To calculate the Gini index of a decision tree, the following formula is used (See Equation 2.1 below):

$$\text{Gini Index} = 1 - \sum_{i=1}^{j} p_i^2 \tag{2.1}$$

Figure 2.5: Gini Index [Tangirala, 2020]

An example of a decision tree in application could involve the concept of assessing one's mortgage loan eligibility. According to existing requirements for applying for a mortgage, certain factors are taken into consideration before an application is accepted. Factors commonly reviewed include annual/monthly income, age, existing credit score, and employment status [Exp, 2023]. These factors can be considered and utilised as features for input, while the output focuses on modelling and visualising certain factors to support the decision-making process and predict risk categories. The figure below illustrates an example of this in application.

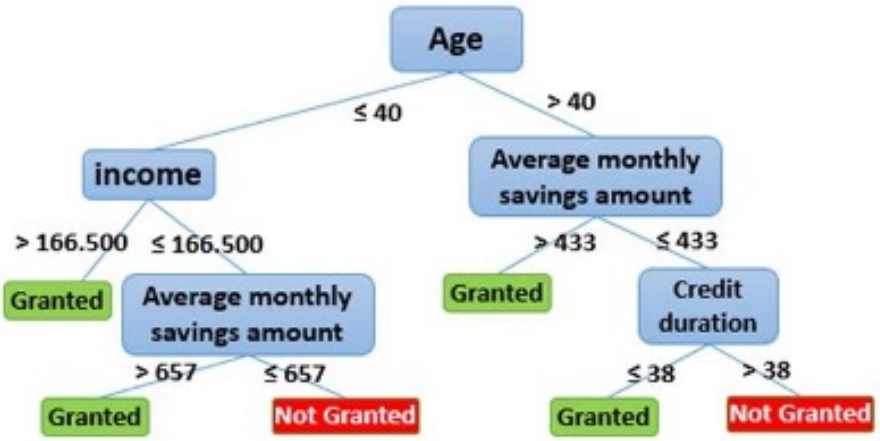

Figure 2.6: Decision Tree Visualisation[Youssef, 2018]

There are numerous advantages and disadvantages associated with decision trees. One of the key advantages associated with decision trees is interpretability, an example being Figure 2.6. As shown, the results can be effectively understood, from understanding the root nodes down to child nodes, splits associated, reasoning between the splits (based on values such as greater or equal to 657) etc. Successively, another advantage associated with a decision tree includes the ability for the tree to function without the requirement of assumptions and the ability to handle missing values. By having a model that does not need to presume information from certain features i.e., Feature X, decision trees can effectively distribute data/feature relations.

However, despite the advantages present in this model, some disadvantages need to be taken into consideration, namely the potential for overfitting, instability, and bias towards certain features. By having features within a dataset, with the presence of multiple levels, it must be noted that issues may arise such as being biased towards a certain feature. Similarly, minor adjustments to the training data such as an imbalance between the training and testing data percentage can often lead to instability, effectively impacting the outcome and output of the product.

To further bolster decision trees, hyperparameters can be exploited. Hyperparameters play a crucial role, as they are adjustable parameters that can control various aspects of the process when modelling. In the context of decision trees, these parameters allow for tuning and adjustments during the building process, allowing developers to fine-tune the model to produce an output with greater accuracy. The mentioned hyperparameters for a decision tree can be found below, with a cross-reference to the journal "An Empirical Study on Hyperparameter Tuning of Decision Tree" by Rafael Gomes Mantovani et al and EDUCBA [EDU, 2023].

Table 2.1: Table Listing Hyperparameters for Decision Trees [Mantovani et al., 2018]

Parameter Name	**Explanation**
Maximum Tree Depth	Defines how deep a tree can grow. This can often be further extended depending on the complexity of the model structure and if further investigation is required from the perspective of understanding training errors.
Minimum Sample for Split	Defines the minimum number of samples before a node can be split to host child nodes. By default, the value is set to 2, however, if the number is less than the internal node, that node will, in effect become a leaf node by default.
Minimum Samples per Leaf	This, similar to minimum sample for split focuses on the minimum number of samples needed/required before a node can become a leaf node. Therefore, with both this parameter and the previous, a split can only take place if both conditions are met. Default $= 1$
Max Features	Defines the number of features that are to be used within the decision tree and used when identifying the correct split. If a dataset or data frame were to consist of 20 features, the max features parameter allows for a distinguished number to be set so that only a set number of features are used.

2.5.2 Bayesian Learning

Bayesian learning is another modelling approach that can be utilised in machine learning that focuses on estimating probabilities and predictions. The framework focuses on incorporating information such as prior knowledge and making informed decisions.

Bayesian learning primarily incorporates the concept of Bayes' Theorem, a fundamental principle used in the Probability Theorem. The theory utilises a mathematical formula utilised in statistics to speculate and take into consideration multiple combinations of probabilities of an event or set of events taking place. A set of conditions must be adhered to, however, before

using Bayes' Theorem, as outlined below:

1. All events in the event set A must be mutually exclusive.
2. All events in the event set A must be exhaustive.
3. The probability of event set B is not 0.

To represent this from a visual standpoint, the following diagram has been developed by Khan Academy, representing the pre-requisites.

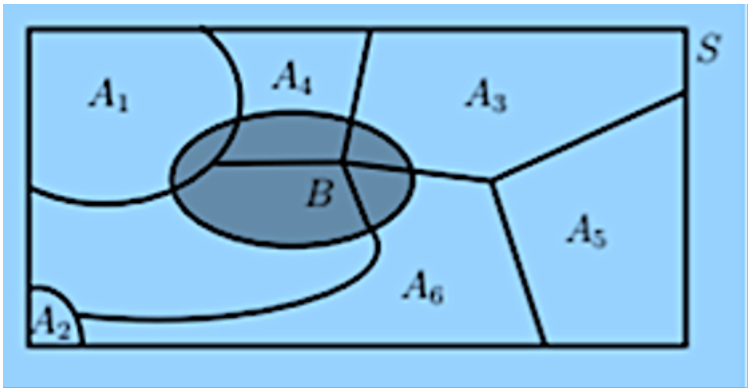

Figure 2.7: Pre-Requisite Spaces in Bayes Theorem [Academy, n.d.]

As shown by the diagram in Figure 2.7, numerous spaces are present, ranking from set A_1 to A_6. Each space represents the probability of an event taking place. As outlined above, a set of prerequisites must be adhered to before applying Bayes' Theorem. The events of B interlinking with A_1 for instance are mutually exclusive, and therefore, no other events can occur if one already takes place.

Bayes' Theorem is represented in the form of an equation. The equation states that the posterior probability of an event, given observed evidence, is proportional to the product of the prior probability of the event and the likelihood of the observed evidence given the event. In other words, Bayes' Theorem enables us to revise probabilities by combining prior knowledge with observed data.

Bayes' Theorem Equation [Joyce, 2021]

$$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}, \tag{2.2}$$

whereby:

$P(h) =$ prior probability of hypothesis h,
$P(D) =$ prior probability of training data D,
$P(h|D) =$ probability of h given D,
$P(D|h) =$ probability of training data D given hypothesis h.

As shown by the equation above, Bayes' Theorem consists of multiple components, whilst considering the probability of an event taking place, considering the alternative event. To elaborate and with cross-reference to a journal by M Webb and Sidebotham [Cornfield, 1967]. Each constraint involved within the equation is outlined in Eq 2.2. An example that can be used to support this, is the analogy of coin tosses. Using the analogy of coin tosses, two events must be taken into consideration, namely head or tails. The likelihood of an event taking place from a coin is at 50%, heads or tails.

There are several types of Bayesian Learning techniques, each with its characteristics and applications. One widely used example is Naive Bayes, a simple yet effective algorithm for classification tasks. Naive Bayes assumes independence between features given the class label, making it computationally efficient and well-suited for large-scale datasets.

The prior probabilities and class-conditional probabilities are computed from the training data in Naive Bayes. These estimations are then used to compute the classes' posterior probabilities given the observed features. The predicted class is the one with the highest posterior probability. When interpretability is critical, Naive Bayes is a popular alternative. It provides a straightforward and interpretable approach to categorisation.

A number of fundamental components are linked to and associated with the NB concept. Elements such as random variables and probability distribution citepWebb2010 are included. To begin, a random variable can represent a variety of values, most commonly class labels in the context of the Nave Bayes model. A probability distribution is a function that represents the possibility of an event occurring. This should be the sum of all probability and so must equal 1.

As part of Naïve Bayes, and NB modelling, fewer hyperparameters are offered and available to use in comparison to other models such as SVM, and decision trees. This is due to several factors, including:

1. No need for regularisation as the model is naturally regularised based upon its approach.
2. Different distributions in the form of Gaussian, Multinomial etc.
3. NB makes assumptions that all features are independent and therefore as this is the case, there is no requirement or need for hyperparameter tuning.

As an alternative, automated hyperparameter optimisation can be applied. As outlined by Joel Ostblom in their Jupyter Notebook report and ML lectures [Ost, 2021], there are numerous advantages to using automated HP (hyperparameter) optimisation/tuning, such as:

1. Less error prone
2. Reduces human error.
3. Data-driven may be more effective
4. Overfitting on the validation set.

When using AHO (Automated Hyperparameter Optimisation), there are two major ideas to consider: grid search or randomised AHO search. The principle focuses on "performing hyperparameter tuning in order to determine the optimal value for a given model" [Gre, 2023] when using Grid Search CV (Cross Validation). In the case of GridSearchCV, in contrast to

the lack of parameters while using base NB, additional parameters can be used. These are listed in the table below.

Table 2.2: GridSearchCV Parameters [Kopal, 2021]

Parameter Name	**Explanation**
Estimator	Model of interest, in this case Gaussian Naïve Bayes would be applied.
Param_Grid	Dictionary with parameter names as a string, with a list of parameter settings to try as values.
Verbose	Represents the verbosity, presenting detailed processing information, marked as 0 or 1, whereby 1 = True.
CV	Represents the cross-validation value often defaulted to 10-fold cross-validation.
N_Jobs	Represents the max number of concurrently running workers, when set to -1, all CPUs are used.

Similarly, RandomizedSearchCV functions similarly to GridSearch whereby parameters are optimised for the best performance. The only difference is that randomized search has focused on the parameters used [Scikit-learn.org, 2019]. Similar to Grid Search, Randomized Search utilises `n_jobs`, `estimator`, and `param_grid` (known as `param_distributions` in Randomized Search). In addition, `n_iter` is also used, referencing the number of parameter settings sampled.

After fine-tuning the parameters, and discovering the optimal combination, a test is conducted on the testing data to identify how the hyperparameters have performed, with consideration to minimising predefined loss. To identify this, evaluation metrics are used, reviewing values such as accuracy score, precision, recall and F1 Score.

In conclusion, Bayesian learning offers a moral framework for incorporating prior convictions, measuring uncertainty, and developing reliable predictions. We can change our opinions based on observed facts and openly indicate uncertainty by using Bayes' theorem. One of the well-known Bayesian Learning algorithms, Naive Bayes, provides a straightforward yet efficient method for categorising objects. We will examine the ideas, benefits, and applications of Bayesian learning throughout this course, with a concentration on Naive Bayes.

2.5.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) are a strong and popular class of supervised machine learning algorithms that are mostly utilised for classification and regression applications. Vapnik and Cortes introduced SVM in the 1990s, and since then it has attracted a lot of attention for its outstanding capacity to handle complex and high-dimensional data with exceptional accuracy.

SVM functions by creating a decision boundary also known as a hyperplane to separate classes. The concept is extremely similar to K-Means Clustering whereby nodes are positioned to the closest cluster based on their distance to the centroid node. However, in the case of SVM, support vectors (nodes) are positioned on alternate sides of the margin, supporting the clas-

sification of the node in class X or Y [DataCamp, 2019].

One of the elements of understanding the functionality of SVM is understanding the mathematical elements attached to the method. As previously mentioned, SVM functions very similarly to K-Means, and therefore relies on the calculation and length of vectors, and the dot product. The dot product is a representation of how vectors are related in the form of a scalar quantity. To calculate this, the following formula is used:

Dot Product in SVM [MLMath.io, 2019]

$$u \cdot v = \|u\|\|v\| \cos(\theta) \tag{2.3}$$

Sebastian Raschka demonstrates an example of SVM in action by showing how SVM kernels function using the Iris dataset [Raschka, 2023]. The example includes two images: the left-hand-side image presents two features using linear SVM, while the right-hand-side image showcases the application of SVM with kernels.

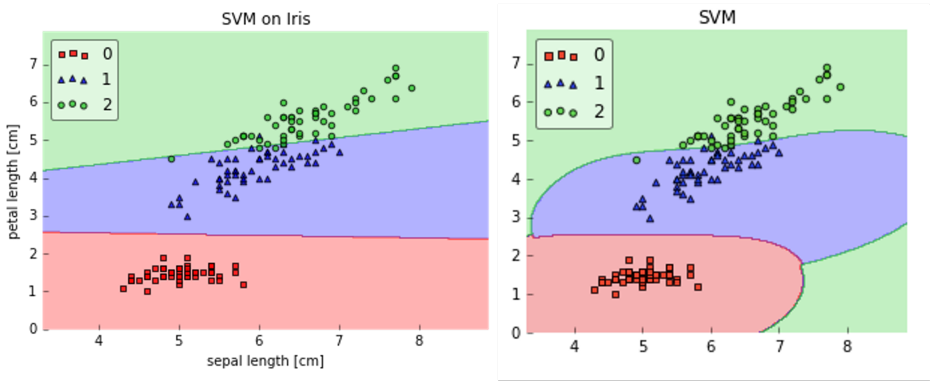

Figure 2.8: SVM with Linear Boundary vs SVM with Kernel

By using kernels, the complexity can effectively grow like the training set. A key difference between linear SVM and kernel-based SVM lies in how the kernel successfully adapts to the data. This adaptation is achieved because the kernel function transforms the training set so that a non-linear decision surface becomes a linear equation in a higher-dimensional space, thereby allowing the calculation of the inner product between two vectors. To calculate the kernel, the following equation is used:

Standard Kernel Function GeeksforGeeks [2020a]

$$K(\bar{x}) = \begin{cases} 1, & \text{if } \|\bar{x}\| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

With any hyperparameter, different aspects of the model are altered, and, as a result, impact the outcome. The accuracy of the SVM's predictions can be greatly improved by carefully tweaking these parameters in line with the data's characteristics and the problem being addressed. The essential parameters include:

- **Kernel:** focuses on mapping data to a higher-dimensional space whereby the linear decision boundary can be found. Types: linear, polynomial, sigmoid.

- **C:** Also known as the regularisation parameter focuses on balancing the margin and minimising the training error within the model.
- **Gamma:** Also known as the Kernel coefficient, affects the shape of the decision boundary.

2.5.4 K Nearest Neighbours

Clustering is an important technique in both supervised and unsupervised machine learning that is essential for data exploration, pattern recognition, and understanding the underlying structure of datasets. In contrast to supervised learning, which employs labelled data for training, clustering works with unlabeled data, aiming to combine similar data points based on their intrinsic qualities. The fundamental goal of clustering is to organise the data into clusters, with each cluster comprising data points that are more comparable to one another than to those in nearby clusters. This method allows for the detection of organic clusters, the discovery of cryptic patterns, and the acquisition of meaningful knowledge about how data is dispersed. By minimising the within-cluster variance, which is calculated as the sum of the squared distances between each data point and its matching cluster centroid, the algorithm iteratively refines the clusters.

KNN and K-Means clustering follows similar patterns to one another, with minuscule differentiating factors [Zhao et al., 2021]. K-Means clustering focuses on the discovery of unlabelled data groups and identifying what groups or categories the said data points belong to. This, in effect, is an unsupervised learning algorithm. On the other hand, KNN, known as K-Nearest Neighbours focuses on predicting what class/category the next point or value should belong to.

One of the key aspects attached to KNN focuses on the measurement of distance from one node to another. To accomplish this, a method defined as the Euclidean distance is utilised[Hu et al., 2016]. It is important to know that this method can be altered in the form of a parameter to an alternative, namely, Manhattan distance. To calculate the distance between nodes via Euclidean distance, the following equation can be used, where p and q are the subjects with n characteristics [Zhang, 2016].

Euclidean Distance

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \tag{2.5}$$

Hyperparameters, similar to a decision tree, can be used to improve the accuracy and value of evaluation metrics for clustering. The quality of final clusters can be greatly improved by employing appropriate hyper-parameters. To help further, the table below contains a list of hyperparameters that can be utilised on a basic K-Nearest Neighbours model, along with what the parameter represents and what purpose it serves during the modelling process.

2.5.5 Ensemble Methods

In the Machine Learning field, ensemble methods are a technique used to aggregate a large number of models to get a final product/prediction [Dietterich, 2000]. Using this strategy, the precision developed is frequently greater than that obtained with a single model. In this way, a model's weakness can be compensated for by attributing to the strength of another

Table 2.3: KNN Hyperparameters [scikit-learn developers, 2019]

Parameter Name	**Explanation**
N_Neighbours	Number of neighbours/categorical groups set by default.
Weights	The weight function is used during prediction, whereby a uniform/default weight is present, and a distance, represents the weight points in the inverse of their distance.
Metric	The metric represents the method utilised. By default, it is set to $p = 2$, whereby the Euclidean distance is represented.

model utilising ensemble methods and aggregating a series of strengths.

There are several advantages that can be taken advantage of by using this method, such as improved accuracy, robustness and reduced overfitting [Brownlee, 2020b]. By combining multiple models and key parameters, the level of accuracy can, in effect be bolstered. Additionally, since multiple models are combined, ensemble methods become more robust. By balancing the trade-off between bias and variance and by employing alternative subsets and properties of the data, they can also lower the risk of overfitting and underfitting. There are several types of ensemble methods, including:

- Random Forest
- Boosting
- Bootstrap Aggregating (Bagging)
- Stacking

By optimising the benefits of diverse models while avoiding their drawbacks, ensemble techniques can be a useful addition to the machine learning toolbox. They may be worth considering for many applications, even if they are more difficult to design and grasp because to the potential advantages in predictive performance. Each of these ensemble approaches will be explored in terms of the theoretical functionality they provide, the parameters they provide for tuning, and the benefits received from applying these methods.

2.5.6 Random Forest

Random forest is an ensemble method that encompasses the structure utilised by decision trees[Breiman, 2001]. The model focuses on building a multitude of decision trees during the training period, using the training data allocated during the initial training testing split. Once achieved, the result is returned to the user, with the output resulting in the mode of classes. By building multiple decision trees, Random Forests can often achieve better results from the perspective of evaluation metrics, in comparison to a single decision tree.

The node significance equation based on Gini importance can be used to implement and comprehend how the Random Forest algorithm works. It is expected that just two child nodes are active, as presented by Stacey Ronaghan.

Node Importance using Gini Importance Formula

$$n_{i;j} = W_j C_j - W_{\text{left}(j)} C_{\text{left}(j)} - W_{\text{right}(j)} C_{\text{right}(j)} \tag{2.6}$$

where $n_{i;j}$: the importance of node j,
W_j : weighted number of samples reaching node j,
C_j : the impurity value of node j,
left(j) : child node from left split on node j,
right(j) : child node from right split on node j.

When considering what hyperparameters can be utilised, similar parameters offered in decision tree modelling can be utilised, however, additional hyperparameters can be altered. This includes:

- **N_estimators**: Number of trees processed and developed. By having more trees, the performance can be improved, whilst increasing computational cost.
- **Criterion**: Function utilised to measure the quality of a split. Options that can be used include Gini and entropy.

Using subsamples of the dataset, numerous decision trees are used to build Random Forests, which incorporate the predictions from the individual decision trees. This aids in lowering the model's variance and enhancing generalisation to new data.

2.5.7 Boosting

Boosting is another prominent algorithm used in machine learning to reduce the number of errors identified or suspected in data analytics. Boosting has various advantages, including increased predictive accuracy, lower model variance, and the capacity to handle a combination of categorical and numerical information. Boosting techniques with built-in feature selection, such as AdaBoost, Gradient Boosting, and XGBoost, are frequently resistant to overfitting, especially when the data set is vast and diverse. This approach is applicable to a variety of models, including random forests and decision trees. The list below depicts a series of procedures taken to use boosting appropriately, as stated by Amazon AWS [AWS, n.d.].

Algorithm 2 Boosting Algorithm

Step 1: Assign equal weight to each data sample, by feeding data from the initial ML model applied, ranging from decision tree to random forest. A level of data separation (training and testing) must have already been applied.
Step 2: The boosting algorithm assesses the model predictions and increases the sampling weight accordingly with a greater error. Weights are also assigned based on the model's performance. The model then outputs a prediction with a high amount of influence following the boosting procedure.
Step 3: The algorithm passes the weighed data to the next decision tree.
Step 4: Steps 2-3 are repeated until instances of the training errors are below a set threshold. $=0$

There are different forms of boosting that can be applied to any form of machine learning models, namely Adapting Boosting, Gradient Boosting and XG (Extreme Gradient) Boosting.

2.5.8 Adaptive Boosting

Adaptive boosting, also known as AdaBoosting, was one of the first boosting models to be created and deployed. The ML model focuses on self-correction during each iteration of boosting, which is quite similar to the beginning steps of boosting. AdaBoosting is applied by giving both the same weight. Once this is accomplished, the weighting is evenly distributed after each iteration and decision tree. Corrections are made when needed by using the following. By doing so repeatedly, the difference between the actual and anticipated values, known as the residual error, is reduced to an acceptable level.

This technique can be used in a range of scenarios and manners including predictors and classification. An example where this ensemble method could be used can be in a series of projects such as binary classification and multi-class classification.

2.5.9 Gradient Boosting

Similarly, Gradient Boosting, like AdaBoosting, is an ensembled technique that may be used to improve performance after an initial ML model has been deployed [GeeksforGeeks, 2020b]. AdaBoost and Gradient Boosting (GB) differ in how they handle misclassified items. In contrast to AdaBoost, Gradient Boosting does not give extra weight to incorrectly detected objects. Instead, GB constructs a series of base learners, each better than the previous one, to steadily improve the loss function. It varies from AdaBoost in that it focuses on achieving precision from the start rather than constantly correcting errors, which may make GB a more exact tool. Gradient boosting is quite adaptable, as it works well for classification and regression issues.

2.5.10 Extreme Gradient Boosting

Finally, the third type of boosting discussed as part of this investigation is Extreme Gradient Boosting, often known as XGBoost. Extreme Gradient Boosting, or XGBoost, extends the gradient boosting concept [Ibrahem Ahmed Osman et al., 2021] to meet the demands of modern computational needs. XGBoost employs a large number of CPU cores to support parallel learning during the training phase and to accelerate computation. The approach is used in big data applications due to its capacity to handle enormous datasets.

The primary features that distinguish XGBoost from other boosting techniques include its support for distributed computing across multiple machines, its ability to run calculations in parallel, and its efficient use of cache memory. These attributes collectively make XGBoost a potent tool that offers speed and scalability, making it a desirable option for those handling large-scale, complex data processing projects. When using XGBoost over alternative methods, additional parameters can be utilised for parameter tuning. All available parameters are listed below:

2.5.11 Bootstrap Aggregating: Bagging

Bagging, short for Bootstrap Aggregating is an ensemble technique utilised within the data science scope and machine learning to further improve and bolster the accuracy of a model's

Parameters	AdaBoost	Gradient Boosting	XGBoost
max_depth	◆	◆	◆
learning_rate	◆	◆	◆
n_estimators	◆	◆	◆
subsample	◆	◆	◆
max_features/colsample_bylevel	◆	◆	◆
colsample_bytree			◆
reg_alpha			◆
reg_lambda			◆

Figure 2.9: AdaBoost vs Gradient Boosting vs XGBoost

performance [Resti et al., 2023]. This technique often is used to resolve overfitting for classification or regression matters. The concept of bagging works through the following steps, as outlined by Leo Breiman [Breiman, 2001]:

1. **Bootstrapping:** The initial step undertaken focuses on creating diverse techniques. This consists of generating different subsets of the training dataset by selecting data points at random and with replacements.

2. **Parallel Training:** Each sample is then trained independently and in parallel with each other using base learners from the initial run when the ML model was applied or by using weak learners.

3. **Aggregation:** Based on the expected task, an average of the prediction is taken to compute an estimate.

Using the Bagging (Bootstrap Aggregating) concept has several benefits, including reduced overfitting. By training on different subsets, bagging'smooths out' any irregularities in the dataset or models. Another advantage is parallelisation, which allows individual models to be trained concurrently, increasing computational efficiency. When applied to larger datasets, bagging is particularly effective.

Rather than using the original sample from the initial decision tree model, an alternative approach involves integrating random forests with bootstrap samples. This is used in 'The Random Forest Algorithm for Statistical Learning' by Matthias Schonlau and Rosie Yuyan Zou [Schonlau and Zou, 2020]. Following specific underlying principles, bootstrap aggregation is applied to random forests in this work to mitigate overfitting.

```
for i ← 1 to B do
    Draw a bootstrap sample of size N from the training data;
    while node size != minimum node size do
        randomly select a subset of m predictor variables from total p;
        for j ← 1 to m do
            if jth predictor optimizes splitting criterion then
                split internal node into two child nodes;
                break;
            end
        end
    end
end
return the ensemble tree of all B subtrees generated in the outer for loop;
```

Figure 2.10: Random Forest Algorithm [Schonlau and Zou, 2020]

Bagging is a reliable technique that is frequently employed in a variety of machine-learning applications because it can improve a model's prediction performance by lowering its variance and boosting its stability.

2.5.12 Stacking

Stacking as a concept and a model is a combination of previous theoretical aspects, outlined in bagging, random forest and boosting whereby the principles of stacking focus on providing the best insights. An example of this can be found and is shown by Barton and Lennox upon their experimentation, attempting to improve prediction and variable importance robustness for soft sensor development[Barton and Lennox, 2022]. The concept focuses on these key theoretical functionalities:

1. **Training Multiple Base Models:** Stacking begins by training multiple base models on the provided dataset, often of various types. These base models can be any machine learning algorithm, including but not limited to decision trees, SVMs, and linear regression.
2. **Prediction-Making for "Meta" Dataset:** Once the training phase is complete, the next step focuses on making predictions using the validation set. These predictions serve as features and are collated into a new "meta-dataset". In doing so, all predictions and outcomes from each base model are recorded as features within this new dataset.
3. **Training Using New Dataset:** Using this new "meta" dataset, a subsequent round of training is applied. Stacking shows its strength at this stage by focusing on combining the initial predictions from the base models. It seeks the best methods for improving results by enhancing evaluation metrics.
4. **Final Prediction/Output:** Following the conclusion of the training phase, the predictions are fed into the meta-model, which uses these as inputs to generate the final prediction.

Stacking has the advantage such as improving prediction accuracy over individual models by combining the capabilities of several types of models, perhaps providing a model with higher overall performance [Soni, 2023]. Stacking also provides flexibility because it may be utilised with a variety of models, making it adaptive. However, there are some drawbacks to consider. Training numerous models, as well as a meta-model, can be computationally and time-consuming. Stacking can lead to overfitting if not implemented appropriately, especially if the basis models are highly complicated or there are insufficient validation data points for training the meta-model. When using stacking in a machine learning assignment, this balance of benefits and drawbacks must be evaluated.

2.5.13 Evaluation Metrics

Evaluation metrics play a crucial role in assessing the performance and effectiveness of models in data science. They offer impartial evaluations of a model's predictions' precision, recall, accuracy, and other crucial properties. Data scientists can compare various models, make informed decisions, and improve the performance of their models thanks to these metrics.

Models are trained in the field of data science to make predictions or categorisations based on input data. However, merely being able to predict outcomes is insufficient to assess a model's quality. Metrics for evaluation offer a quantitative evaluation of how well a model does on tasks. They act as a common method for gauging and contrasting the effectiveness of various models, or iterations of the same model.

The type of problem being solved determines the evaluation metrics that should be used. For instance, metrics like accuracy, precision, recall, and the F1 score are frequently employed in classification tasks to assess the model's capacity to correctly classify instances into various categories. The accuracy of numerical predictions is, however, frequently assessed in regression tasks using metrics like mean absolute error (MAE), mean squared error (MSE) [JJ, 2016], and root mean square error (RMSE). Beyond model evaluation, evaluation metrics have a wider impact. They have effects on the creation, verification, and application of models.

Evaluation metrics offer information on potential areas for improvement by quantifying a model's performance. The metrics can be examined by data scientists to find patterns, comprehend model flaws, and come to wise feature engineering, model selection, or hyperparameter tuning decisions. Metrics can be split into two key categorical sections, namely accuracy metrics and error-based metrics. Both serve unique purposes, which, in effect can support the user comprehending how a model can be performing based on the certain constraints or parameters applied.

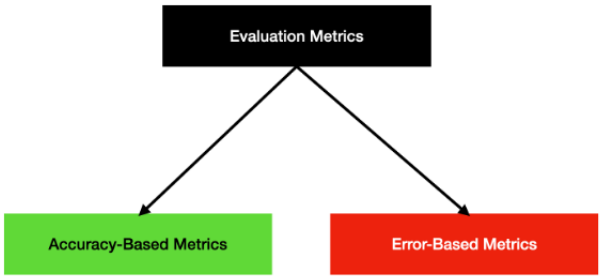

Figure 2.11: Hierarchy of Evaluation Metrics [Ramzai, 2020]

In terms of accuracy-based metrics, their relevance often varies from model to model, and some may carry more weight than others. These metrics are employed to gauge the correctness and reliability of a prediction, depending on the specific model used. As a result, they aid in understanding whether certain hyperparameters are functioning as intended, or if the model requires further tuning. Examples of such metrics include F1 score, accuracy, precision, and recall. Importantly, all these metrics use similar variables for their calculations, which will be further elaborated upon. The table below outlines the metrics discussed.

Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp+tn}{tp+fp+tn+fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp+fn}{tp+fp+tn+fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp+fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn+fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp+fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp+tn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2*p*r}{p+r}$	This metric represents the harmonic mean between recall and precision values
Geometric-mean (GM)	$\sqrt{tp*tn}$	This metric is used to maximize the *tp* rate and *tn* rate, and simultaneously keeping both rates relatively balanced
Averaged Accuracy	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+}}{l}$	The average effectiveness of all classes
Averaged Error Rate	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+}}{l}$	The average error rate of all classes
Averaged Precision	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	The average of per-class precision
Averaged Recall	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	The average of per-class recall
Averaged F-Measure	$\frac{2*p_M*r_M}{p_M+r_M}$	The average of per-class F-measure

Figure 2.12: Threshold Metrics for Classification Evaluations [Hossin and Sulaiman, 2015]

To conclude, numerous accuracy-based metrics such as the F1 score, accuracy, precision, and recall are critical tools for evaluating the performance of different machine learning models. These metrics provide useful insights into the validity of the model's predictions by employing

similar variables to compute their respective scores. The comparative table offered in this part encompasses various evaluation measures, assisting in understanding the impact of specific hyperparameters and leading towards prospective model optimisations. By methodically analysing these indicators, researchers and practitioners can make educated judgements on model selection and tweaking, matching the model's performance with the unique requirements of the work at hand.

Chapter 3

Methodology

The methodology section of this study aims to provide a comprehensive understanding of the selected dataset through Exploratory Data Analysis (EDA). The study adheres to the CRISP-DM (Cross-Industry Standard Process for Data Mining) cycle, as illustrated by Rüdiger Wirth and Jochen Hipp of the Università di Bologna, which is shown below. In addition to understanding the dataset, our objectives also include applying modeling techniques and evaluating the performance of the developed models.

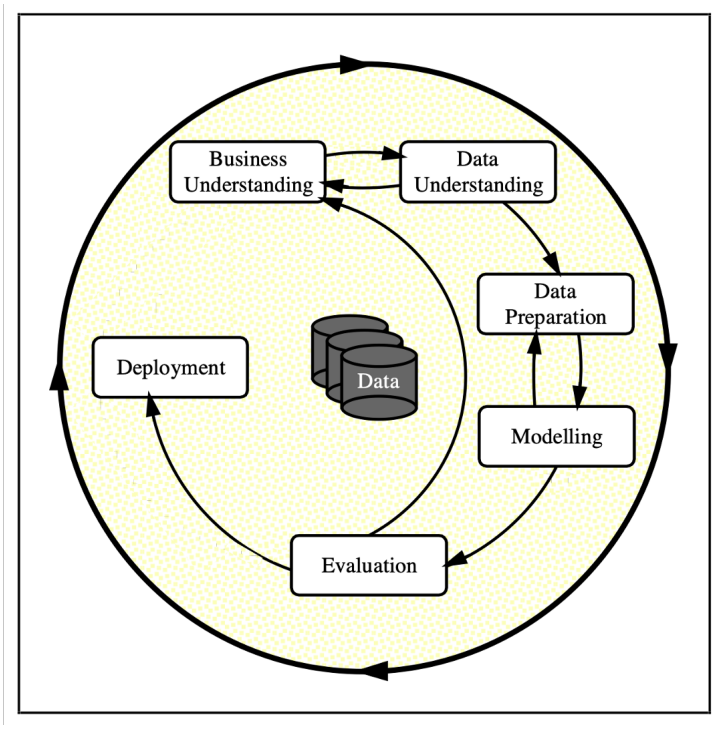

Figure 3.1: CRISP-DM Cycle [Wirth and Hipp, n.d.]

Following the CRISP-DM cycle, we will employ various phases to ensure a systematic and rigorous approach to data analysis. Although the primary focus of this section is on the Data Understanding phase, we will also touch upon other relevant phases, such as Data Preparation, Modelling, and Evaluation.

The Data Understanding phase involves gaining insights into the dataset's characteristics,

identifying data quality issues, and discovering initial patterns or trends. Through EDA, we will employ statistical and visual techniques to explore the dataset's features, distribution, and relationships. This process will help us uncover missing values, outliers, inconsistencies, and understand the central tendencies and spread of the variables. Furthermore, correlation analysis and visualisation will assist us in identifying potential associations between variables.

Once we have a solid understanding of the dataset, we will proceed to the Data Preparation phase. This phase involves transforming and cleaning the data to ensure its suitability for modelling. We will address missing values, handle outliers, and apply feature engineering techniques to enhance the predictive power of the variables.

Subsequently, in the Modelling phase, we will develop and apply appropriate models to address the research questions and objectives of the study. We will select suitable algorithms based on the nature of the problem, the available dataset, and the desired outcome. Our goal will be to build models that effectively capture the underlying patterns and relationships present in the data.

Following model development, we will move on to the Evaluation phase. Here, we will assess the performance of the models by employing appropriate evaluation metrics and techniques. This evaluation will enable us to determine the effectiveness and accuracy of the models in predicting the desired outcomes. If necessary, we may iterate through the Modelling and Evaluation phases to refine and improve the models' performance.

By comprehensively exploring the dataset, preparing the data for modelling, and utilising visualisations, this section sets the foundation for effective machine learning modelling. Comprehensive examinations of each subject, including the specific techniques used for data preparation, understanding, modelling, and evaluation, will be provided in the following sections.

3.1 Data Extraction

Following data selection and approval from the author of the dataset, the first stage of this project focuses on the second stage of the CRISP-DM cycle, namely the Data Understanding and Data Preparation cycle. At this process, the focus is on ensuring the data is correctly imported and extracted in a manner whereby the data is viewable. To achieve a few techniques and methods are utilised for the purposes of importing the data. As part of data analytics, a Jupyter Notebook must be utilised.

A Jupyter notebook file is a file format and one used within the Data Science space, as it is a format the incorporates a multitude of components and formats, such as code, visualisations, and other multimedia components. Jupyter notebook as a format is compatible with three programming languages [Lorena et al., 2019]. This, as a result allows for flexibility within the file, presenting a solution appropriately.

Another advantage that Jupyter notebook has focuses on the ability to work with an interactive and reproducible environment. Jupyter notebook as a file and as a standalone software platform allows for one to execute code blocks and decode in real-time whilst gaining information. By having such an infrastructure of a system, users can effectively share code and files without the need of re-running the entire file to comprehend what is occurring.

A key benefit of Jupyter notebooks is their capacity to integrate text and code into a single, cohesive document. This makes it simple to share insights, document analysis steps, and provide thorough justifications alongside the corresponding code.

To effectively establish that the dataset has been correctly synchronised with the Jupyter Notebook, a few libraries have had to be imported, namely the Pandas library [PyD, 2023]. The Pandas library is a library that can be imported into the Jupyter notebook space, allowing for users to modify and alter data. One of the core structures of the Pandas library involves the Dataframe element whereby the Pandas library provides a 2D table-like structure, like a spreadsheet. This can be used for the purposes of manipulation and data transformation. For the purposes of this section, the data loading and storage methods are to be used to verify the import was successful. This can be found below.

```
df = pd.read_csv("frmgham2.csv")
```

Listing 3.1: Reading CSV file

Once this has been achieved, the next step that is often employed within the data science space is the application of using the head() method. The head() function focuses on returning the first 5 rows from the dataframe. In this instance, the dataframe, assigned the value df investigates the top 5 rows within the dataframe. This is an effective way on comprehending if the data source has been correctly read, and whether it has been read in the appropriate format. This can be seen in the Listing below, and Figure below, whereby the code has been applied, with the results providing the top 5 columns. As we can see, based off the information provided from the Framingham documentation (see Appendix B and B.1), the columns have been appropriately loaded. Likewise, as no form of transformation has been applied at this stage, all the values within the columns have been retained. It is also worth noting the prior to importing the dataset file, no form of file conversion was applied, and therefore any form of data loss can be voided.

```
df.head()
```

Listing 3.2: Application of Head() Function

	RANDID	SEX	TOTCHOL	AGE	SYSBP	DIABP	CURSMOKE	CIGPDAY	BMI	DIABETES	...	CVD	HYPERTEN	TIMEAP	TIMEMI	TIMEMIFC	TIMECHD
0	2448	1	195.0	39	106.0	70.0	0	0.0	26.97	0	...	1	0	8766	6438	6438	6438
1	2448	1	209.0	52	121.0	66.0	0	0.0	NaN	0	...	1	0	8766	6438	6438	6438
2	6238	2	250.0	46	121.0	81.0	0	0.0	28.73	0	...	0	0	8766	8766	8766	8766
3	6238	2	260.0	52	105.0	69.5	0	0.0	29.43	0	...	0	0	8766	8766	8766	8766
4	6238	2	237.0	58	108.0	66.0	0	0.0	28.50	0	...	0	0	8766	8766	8766	8766

5 rows × 39 columns

Figure 3.2: df.head() Output

3.2 Data Preparation

After successfully completing data extraction, the next step in collaboration with the CRISP-DM model focuses on data preparation. This process is combined with EDA, also known as Exploratory Data Analysis, as part of this step. Data preparation, on the other hand, focuses

on ensuring that the data status is at a point where it may be processed for further understanding. "Data preparation is typically an iterative process of manipulating raw data, which is often unstructured and messy, into a more structured and useful form that is ready for further analysis" [Abdallah et al., 2017]. One of the most important stages is to ensure that there is no duplication or null value. As a result, one of the earliest phases, which focused on filling in any missing numbers with a mean value, was completed, achieved by Listing 3.3.

```
df = df.fillna(df.mean())
```

Listing 3.3: Filling Missing Values with Column Names

The next stage that needed to be finished was removing any rows that would be considered useless in order to make sure the data was properly cleaned and prepared for processing via modelling. This was accomplished by using EDA.

3.3 Exploratory Data Analysis

First, after data preparation and extraction, the data needed to be reviewed and cleaned to remove any redundant data to fully understand the dataset, with the application of Exploratory Data Analysis. EDA, also known as Exploratory Data Analysis, is a fundamental step undertaken during the process of data comprehension and data exploration that delves further into comprehending the dataset's structure, patterns, and relationship between variables. As stated by C Chatfield, "the ingredients of EDA are discussed, and two main objectives delineated, namely data description and model-formation" [Chatfield, 1986].

Figure 3.3 shows the heatmap that was used in this regard. The heatmap was a useful tool for displaying correlations and locating outliers or missing values in the dataset [Kumar, 2022].

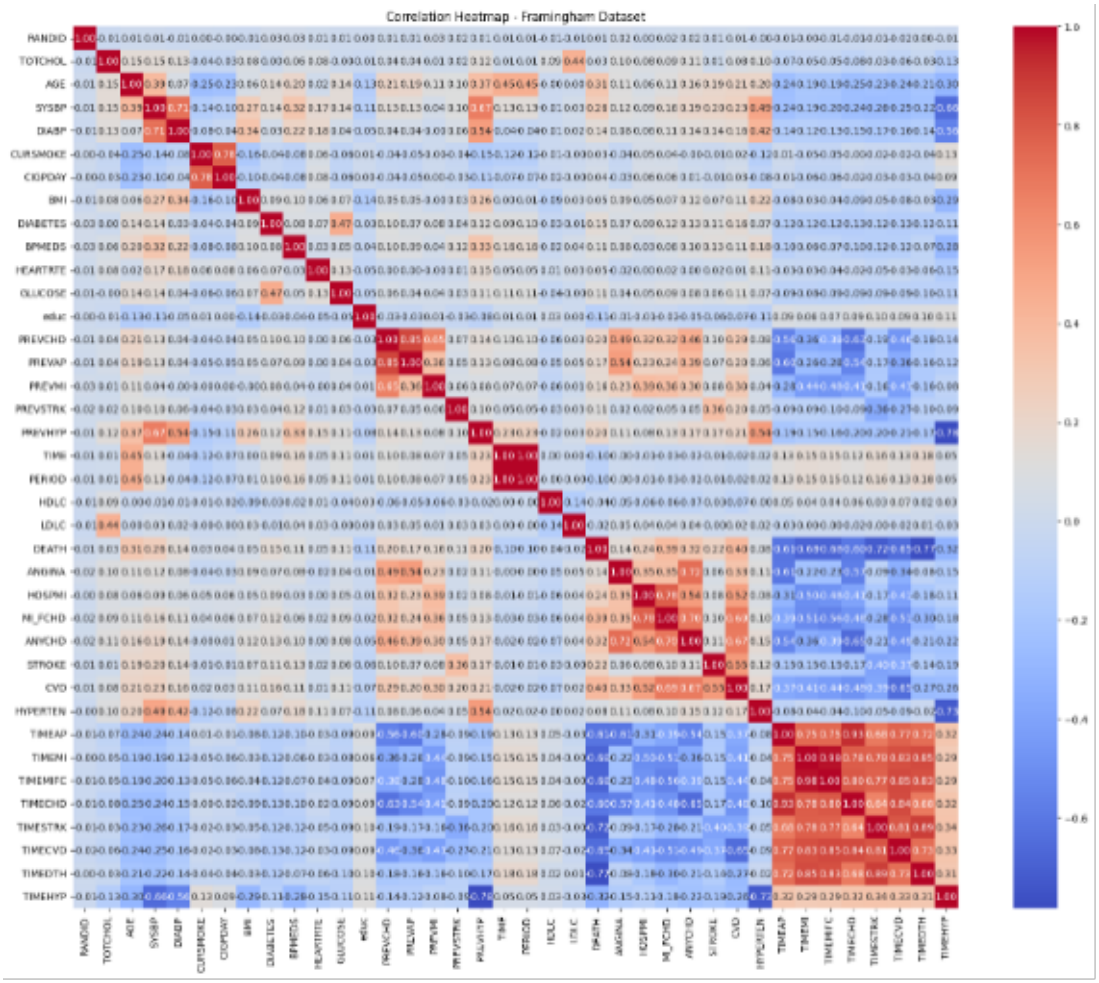

Figure 3.3: Correlation Heatmap: Framingham Dataset (Pre-Filtration)

After running the following heatmap, as illustrated via Figure 3.3, a large variation of detail was presented, whilst taking into consideration every feature that was initially present within the dataset. Any feature that contained data whereby it contained any null values accordingly as previously defined.

Based on the initial presented heatmap, the data suggests that features from TIMEAP to TIMEHYP on the Y Axis of the heatmap contribute each other. This was further examined accordingly to further evaluate the data present and what this could mean and why this may be present within the data. As shown in 3.4, numerous features were outlined as correlated to one another.

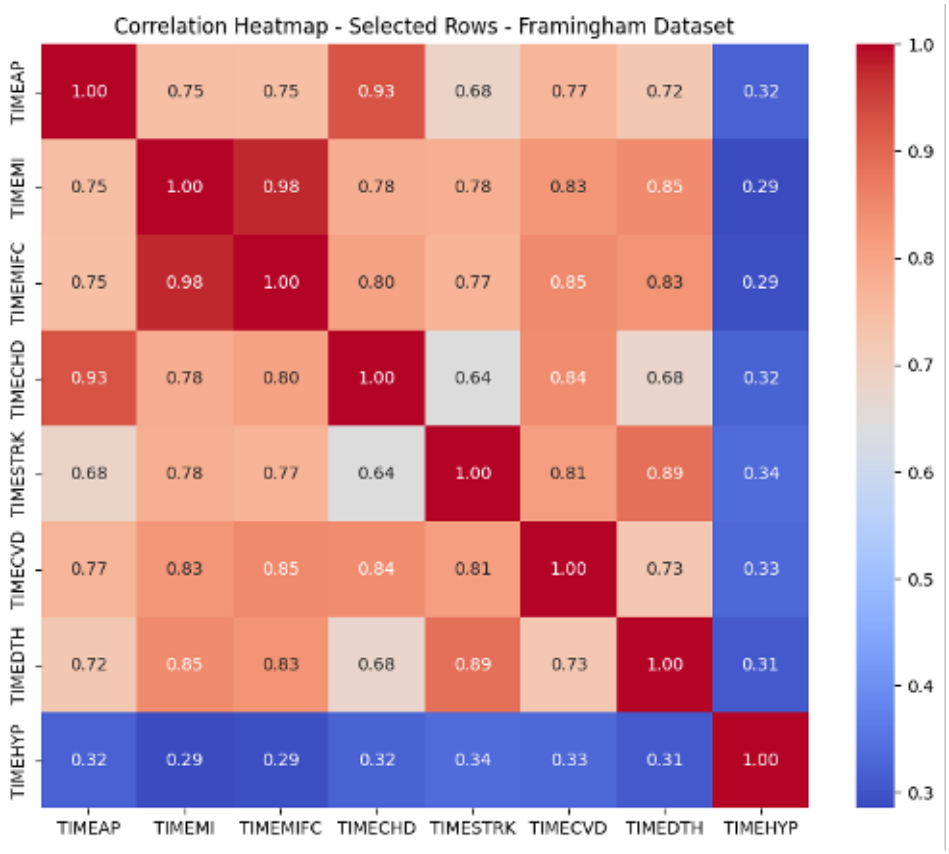

Figure 3.4: Heatmap Correlation for Columns 31-38

Upon interpretation, the data presented in Figure 3.4 suggests a multitude of information. To commence, based on the heatmap generated, the feature names `TIMEMIFC` and `TIMEMI` illustrate that 98% of the time, the values match. To comprehend what these features mean and represent, a cross-reference was conducted between Appendix B and Appendix C.

The appendix and data suggest that the feature `TIMEMIFC` is an indicator of whether the event of a Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease has been recorded. Based upon the data definition that can be found in Appendix B, the feature `TIMEMIFC` is a clone of the value of `MI_FCHD`, which can immediately contribute to `TIMEMIFC`.

Additionally, as previously mentioned, the values of `TIMEMIFC` and `TIMEMI` exhibited a match rate of 98%. This is logical considering the definition of both features. As outlined in the data definition document and as shown in Figure 3.5, both `HOSPMI` (clone of `TIMEMI`) and `MI_FCHD` (clone of `TIMEMIFC`) include reference to Hospitalized Myocardial Infraction.

HOSPMI	Hospitalized Myocardial Infarction
MI_FCHD	Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease

Figure 3.5: Framingham Data Definition for HOSPMI and `MI_FCHD`

As per principles of OR clauses, the values of `MI_FCHD` are likely to be 1 if one of the conditions are met, i.e. A = 1 and B = 0, the output of the clause results in 1, and therefore TRUE [Kaur, 2017]. Since this logic, output of 98% is consistent based upon the data being replicated. As a result, and based on the values presenting repetitive values, both features of TIMEMI and TIMEMIFC were dropped. Further examination was conducted upon the data, by which, based on AppendixB and C, features commencing with TIME were repeated values. This as a result was deemed unnecessary due to redundant data, and therefore was effectively removed.

To further investigate the values that were present as part of the dataset, the next step we wanted to undertake as part of this project involved examining and understanding the correlation between certain features. This included: number of cigarettes a day, systolic blood pressure, age, diabetes, if they were a current smoker , total cholesterol levels, HDLC and LDLC, BMI, Glucose and Diabetes (Boolean). The correlation can be outlined below, as presented in Figure 3.6.

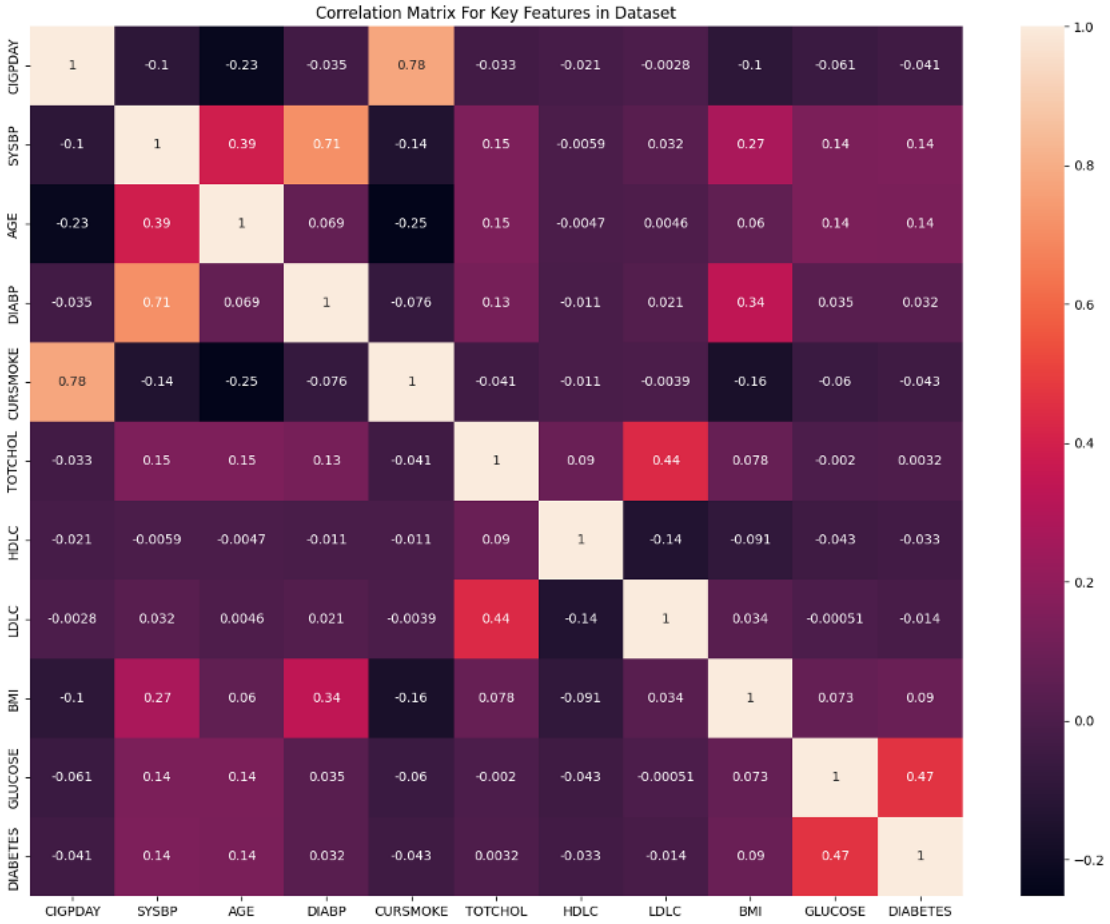

Figure 3.6: Correlation Matrix Heatmap between Key Features Showing Relation

Further analysis was undertaken as part of EDA but focusing into the correlation of various aspects to further expand on the knowledge obtained as part of this dataset. This phase is critical because it allows one to correctly identify functions that may be of use and support, which is critical and important during the development and modelling stages, by offering features that may be used. In order to comprehend the relationship between certain features,

the following insights were obtained. This includes insights such as the relationship between smoking statuses between defined genders, the association between cholesterol and age, and ultimately the Systolic Blood Pressure of individuals exhibited throughout time, displayed in a box and plot graph.

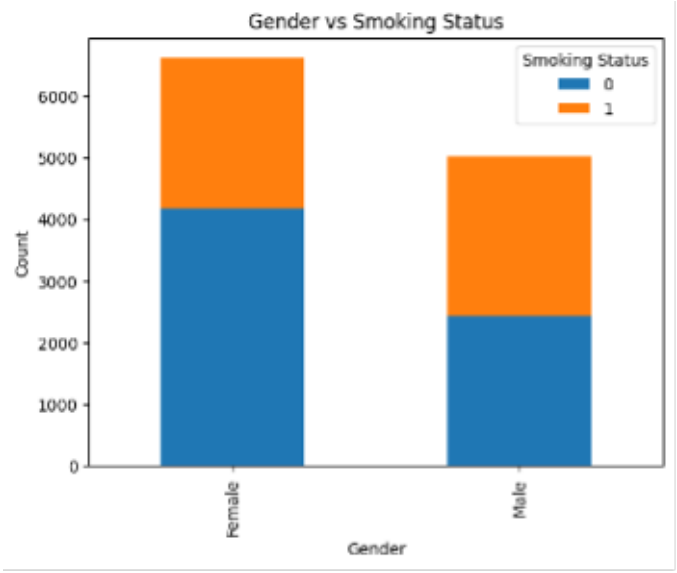

Figure 3.7: Illustration showing the results and correlation between Gender and Smoking Statues

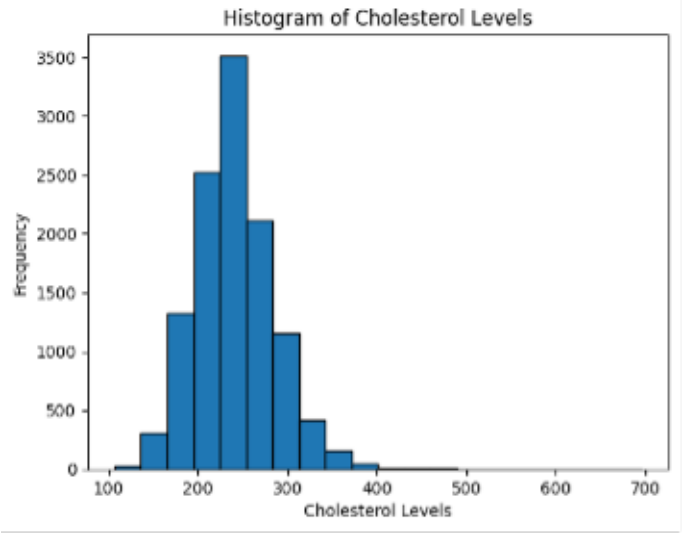

Figure 3.8: Illustration showing a histogram of cholesterol levels.

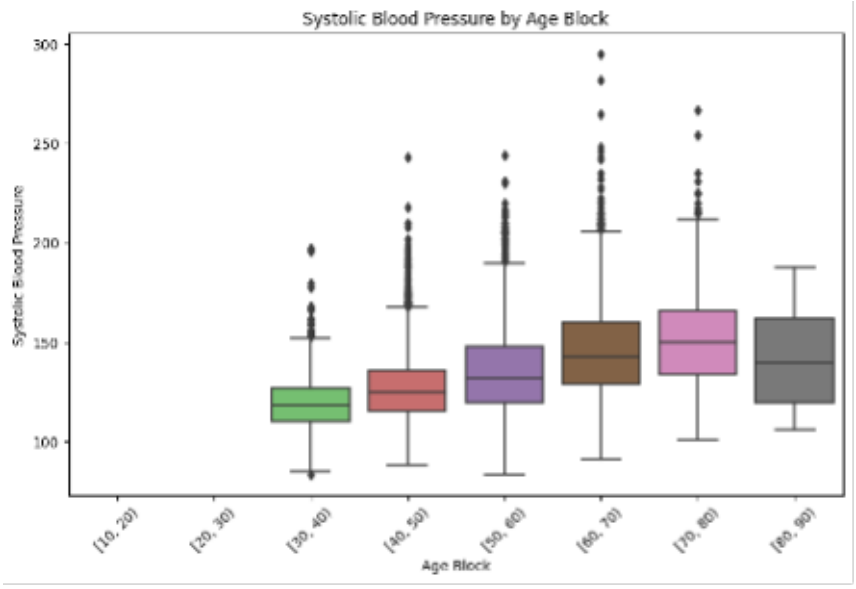

Figure 3.9: Illustration showing Systolic Blood Pressure by Age Block in the form of Box and Whiskers Plot Graphs

3.4 Data Modelling

In the context of the CRISP-DM cycle (see Fig 3.2) and machine learning, data modelling focuses on the construction of a model using the data gathered. Without this process, and without a specific design and approach. As part of this process, numerous elements need to be taken into consideration, including the selection of modelling technique, the model training approach, and model evaluation, including metrics that may be used to comprehend how it performs. This process involves numerous components, including:

- Selection of Modelling Technique: Classification or Regression
- Model Training: Splitting of Data, Training Process
- Hyperparameter Tuning: Tuning techniques and Validation Set
- Model Evaluation: Selection of Evaluation Metrics, Validation, Testing and Comparing Models

As defined in previous sections and as part of the problem statement, the key objective involved with this project is to identify if an individual may be at high risk of developing CVD based on a set of characteristics such as diabetes value, CVD traces in family history, ethnicity and age. As outlined Will Hillier, classification focuses on "predicting or identifying which category or categories an observation belongs to" [Hil, 2022]. Similar, Hillier also mentions that "regression focuses on using input variables to identify a continuous variable", commonly representing weight, height, salary etc. Based upon the task we intend on investigating; clustering would be deemed best, as this project focuses on identifying those who may be at high risk, which would be best presented in the form of categorical classification.

The next step focuses on model training. By default, when training a model, it is recommended that the data should be split at a 70:30 ratio, whereby 70% of the total dataset is used for the purposes of training the model. In effect, the remaining 30% would be used for the purposes of testing the model, and validating if the predictions made by the model are accuracy if not near the true value retained in the initial training dataset. One of the fundamental reasons why the 70:30 or 80:20 split is adopted is due to two key values, namely (1), no overestimation of accuracy is present by adopting this approach, and (2), the more accurate among valid estimates i.e. their overestimation of the approximation error is the smallest possible [Gholamy et al., 2018]. These finds can be found explained in the technical report written by Gholamy, Kreinovich and Kosheleva [Gholamy et al., 2018]. To avoid overfitting during the process of modelling, numerous techniques can be implemented. This includes consideration of cross-validation and the use of validation sets. Validation sets as a concept focuses on reserving a section of the dataset simply for the purposes of validating the performance of models and understanding the results achieved. They can place a crucial role as hyperparameters can then be tuned accordingly based on the validation set, to ensure that there is no trace of overfitting or underfitting from the model. To present this in a visual format, Figure 16 below presents how the validation set is allocated between training and testing data.

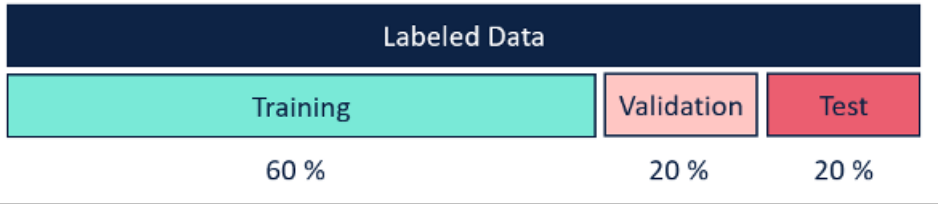

Figure 3.10: Figure showing allocation of validation set for purposes of testing from testing set.

Once data allocation has been achieved, the next step focuses on tuning strategies towards models. GridSearchCV and Bootstrap Aggregation will be utilised for hyperparameter optimisation in order to do this. A range of hyperparameters can be tried with GridSearchCV; hence, by modelling many models with various hyperparameter configurations, the most effective model can be chosen for further modelling.

3.5 Data Evaluation

The final phase of the CRISP-DM cycle prior to deployment entails the evaluation procedure. As defined previously, there are a multitude of ways to evaluate how a model has performed, namely through the use of evaluation metrics. As discussed in Section 2, there are a series of evaluation metrics that can be utilised for the purposes of understanding how a model has performed. Based on the problem statement and the objective of the task, accuracy, and F1/F score are to be utilised.

In addition, to further evaluate the performance of a model, a series of tests are to be conducted outlining:

- Performance with 70:30 split for testing and training.

- Performance with 70:30 split for testing and training with GridSearchCV.
- Performance 70:20:10 split for training, testing and validation.
- Performance 70:20:10 split for training, testing, validation and GridSearchCV and Bootstrap Aggregation.
- Performance using Stacking approach.

As previously indicated, one of the approaches to be investigated is the application of stacking, to assess the impact this ensemble technique may have on performance. Below is an illustration of each model that will be accepted as a feature, as mentioned earlier, to provide a clearer understanding of its application.

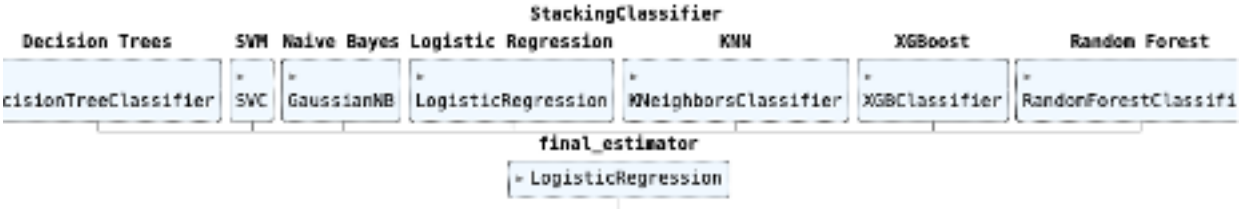

Figure 3.11: Figure showing Stacking Classifier, Visualised

Chapter 4

Results

Following the decisions and application of the CRISP-DM cycle, the subsequent step centres on the deployment of the model and the interpretation of the results. One of the key objectives we sought to achieve during this project was the development of a model that could identify CVD using a specific set of features. Another objective that we believe would be beneficial for future developers is understanding the roles that cross-validation and hyperparameter tuning play. To gain this understanding, numerous tests were carried out. These included a test with a pre-validation split, specifically using a 70:30 ratio, another test with pre-validation split whilst applying grid search to fine-tune the hyperparameters for each model, and finally a 60:20:20 split (training, validation, testing) also utilising grid search.

For the purposes of evaluation, both recall and accuracy score metrics were investigated, owing to the significance they hold. Employing recall as a measure enables successful identification of the number of positive cases in relation to actual positives. In doing so, any missed cases can be appropriately identified. Furthermore, to gauge the success of each model, the macro average was employed, as opposed to the default or weighted average. By using macro averaging, all classes are treated equally, thereby providing a more insightful return value. In the context of medical applications, this approach would be most appropriate, as it enables users to ascertain how a specific metric is performing and whether any classifications are being overlooked.

To commence, the first set of results analysed focuses on reviewing how the model performed prior to hyperparameter tuning and before utilising the 10% validation set. To illustrate this, the corresponding values can be found in the table below. It is important to note that any superior values have been highlighted in bold.

Table 4.1: Table Showing Accuracy and Recall % Post-Runtime (Before 10% Validation Split)

ML Model Used	Accuracy %	Recall % (Macro Average)	Hyperparameters Tuned
Decision Trees	95.9%	98%	None
SVM	92.0%	50%	None
Naïve Bayes	92.0%	50%	None
Logistic Regression	92.0%	50%	None
KNN	91.5%	52%	None

Based on the above results, it is evident that decision trees outperform the alternative models, prior to any form of hyperparameter tuning. This outcome was anticipated due to the inherent properties of decision trees, especially when compared to models like SVM, Naïve Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbours (KNN). Various factors contribute to the superior performance of decision trees (DT), including their ability to handle a mix of categorical and numerical data. In contrast, SVM, NB, and LR are generally more sensitive to mixed data types. Importantly, these observations are also consistent with the recall results. Excluding DT, KNN was the second-highest performing model in terms of recall. Several factors could explain why KNN excelled in the category of recall at the macro average level. These may include its adaptability to data clusters and densities. Additionally, if non-linear decision boundaries exist between the predicted values of PREVCHD, KNN would likely perform better, as the model is designed not to make assumptions about class distributions.

Once this step was completed, the next phase in the data evaluation process focused on applying hyperparameters. As previously outlined, the use of hyperparameters can optimise a model for improved performance. To achieve this, GridSearchCV was utilised to fine-tune the model across a range of scenarios, ultimately selecting the best parameters for application. The results are illustrated below, along with a description of the chosen hyperparameters. As part of this process, ensemble methods and techniques, including XGBoost and Random Forest, were employed to assess whether these approaches could enhance the evaluation metrics accordingly.

Table 4.2: Table Showing Accuracy % Post-Runtime with Hyperparameter Tuning, Including GridSearch and XGBoost (Before 10% Validation Split)

Model Used	Accuracy %	Recall %	Hyperparameters Tuned
Decision Trees	99.6%	99%	Max_depth: 10
SVM	92.1%	50%	C: 0.1, gamma: 1
Naïve Bayes	92.1%	50%	var_smoothing: 1×10^{-9}
Logistic Regression	92.1%	50%	C: 0.1, penalty: l2
KNN	92.2%	51%	metric: manhattan, n_neighbors: 9, weights: uniform
XGBoost	99.7%	98%	colsample_bytree: 0.8, learning_rate: 0.01, max_depth: 3, n_estimators: 50, subsample: 0.8
RandomForest	99.7%	98%	max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 50

This can also be represented as a box and whiskers plot, as illustrated:

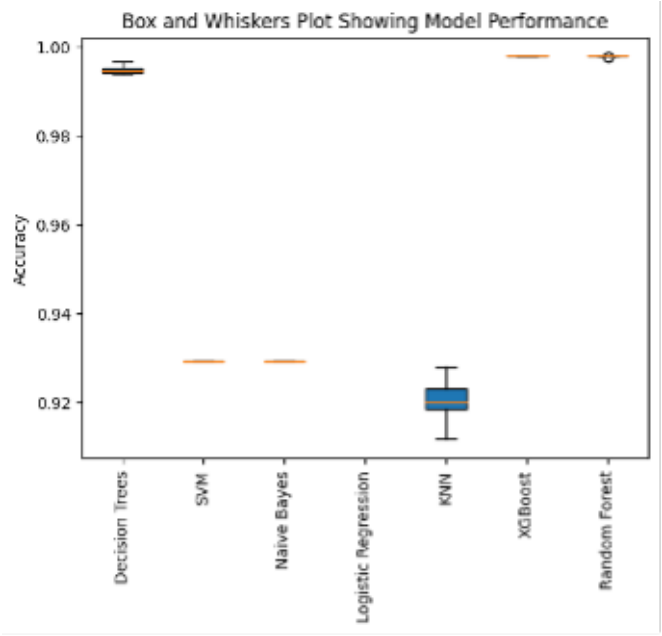

Figure 4.1: Illustration Showing Box and Whisker's Plot (Pre-Validation Set)

Based on the above results, the data suggest that both XGBoost and RandomForest outperformed their counterparts in the test, including DT, SVM, NB, LR, and KNN. This outcome was anticipated, as both XGBoost and RandomForest are ensemble techniques. Consequently, due to their inheritance of various techniques from other models, they were expected to outshine the base models.

There are several reasons for this occurrence. XGBoost and RandomForest are, as previously discussed, ensemble methods that benefit from multiple advantages in comparison to base models such as DT. RandomForest, an ensemble method, is structured based on how a Decision Tree operates. It focuses on creating multiple decision trees, one of its key characteristics, to reduce variance in prediction. This results in a more stable and accurate outcome compared to a singular Decision Tree. Furthermore, ensemble methods like RandomForest are less susceptible to overfitting compared to their counterparts and are also robust to noisy data.

Similarly, XGBoost works by iteratively building weak learners like decision trees and combining them into a stronger model. As a result, errors from preceding models are adapted to, with greater emphasis placed on the new models. Tree pruning and regularisation are also key features that contribute to improved accuracy. Tree pruning in XGBoost and RandomForest adheres to a depth-first approach, thus helping to control tree complexity and avoid overfitting through the use of the "max_depth" hyperparameter. Additionally, the presence of regularisation (both L1 lasso and L2 ridge) in the model minimises overfitting by penalising large coefficients.

Despite the application of hyperparameter tuning to all models via GridSearchCV, the performance metrics improved only by 0.1%.

Once this stage of testing was completed, the next focus was on integrating a validation split as part of the testing, for the purposes of the final results. As discussed, there are numerous

benefits to using a validation split in comparison to using testing data, such as avoiding overfitting. Often, models can learn and, in effect, memorise the patterns associated with testing data. By introducing a validation set, performance is generalised, as the model's performance is tested on an unseen dataset. To evenly distribute the data, it was divided in a 70:10:20 ratio: 70% allocated for training, 20% for testing, and 10% for the validation set.

Table 4.3: Accuracy Results Post-Runtime (With 10% Validation Set, GridSearchCV and Hyperparameter Tuning)

ML Model Used	Accuracy %	Recall % (Macro Average)	Precision % (Macro Average)
Decision Tree	99.6%	99%	99%
KNN	92.1%	50%	71%
SVM	92.1%	50%	46%
Naïve Bayes	92.1%	50%	46%
Logistic Regression	92.1%	50%	46%
Random Forest	99.7%	98%	100%
XGBoost	99.7%	98%	100%

Once again, to represent the statistics generated, a box and whiskers plot has once again been utilised, and can be found presented below:

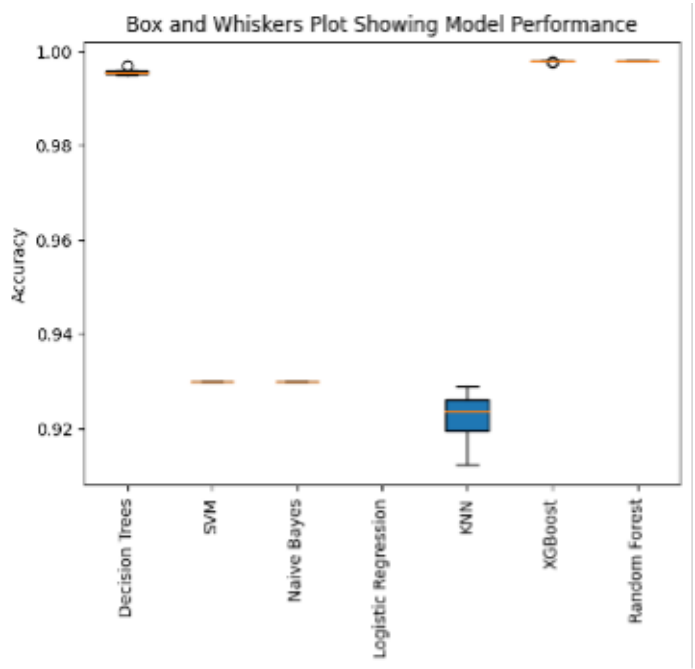

Figure 4.2: Illustration Showing Box and Whisker's Plot (with Validation Set)

Following runtime, the results display a pattern similar to the data presented before the introduction of the validation set. Similar to the previous run, Random Forest (RF) and XGBoost (XGB) outperformed their counterparts, owing to the principles and characteristics they employ. In addition to the accuracy metric, both recall and precision were evaluated as supplementary metrics. Based on the recall values, XGBoost, Random Forest, and Decision

Trees (DT) were the stronger performers, achieving high recall scores.

The reason DT may achieve a high recall score could be attributed to several characteristics inherent to the model. These include the capability to select appropriate features, the depth of the tree, and the utilisation of deterministic rules. Owing to these deterministic rules, which may be strongly indicative of a particular class, the recall score can often be higher. Additionally, a greater tree depth allows the model to capture more nuances in the data.

From the perspective of precision, K-Nearest Neighbours (KNN) also achieved a value above 80%. This could be attributed to the separation in feature space. By employing distance metrics such as the Manhattan or Euclidean methods, the model can effectively capture the similarity between instances of the same class. Furthermore, when features are well-selected, KNN can effectively distinguish between these differences.

Previously, during the research phases, one of the ensemble methods that was discussed centred around the concept of Bootstrap Aggregation, also known as Bagging. To comprehend the implications and impacts that Bagging might present, the models utilised as part of GridSearchCV were also subjected to this technique, generating the following results.

Table 4.4: Performance Metrics with Bagging

ML Model Used	Accuracy %	Recall % (Macro Average)	Precision % (Macro Average)
Decision Tree	99.7%	98%	100%
KNN	92.1%	50%	66%
SVM	92.1%	50%	46%
Naïve Bayes	92.1%	50%	46%
Logistic Regression	92.1%	50%	46%
Random Forest	99.7%	98%	100%
XGBoost	99.7%	98%	100%

Additionally a box and whiskers plot has once again been utilised to visualise the outcome, and can be found presented below:

Figure 4.3: Illustration Showing Box and Whisker's Plot (with Validation Set and Bootstrap Aggregation)

Similar to the results presented previously, it can be shown and found from the above results that upon this runtime, Decision Trees performed as well as Random Forest and XGBoost, attaining the same level of accuracy, recall and precision to its counterparts. There are several reasons this may have occurred, including the manner upon which Bagging functions. Firstly, Random Forest and Bagged Decision Trees both leverage a bagging strategy, and XGBoost also employs a form of ensemble learning through boosting.

The implementation of stacking, a crucial ensemble technique in this research, served as the focus for the final testing phase. Stacking is the process of integrating various models into a unified hierarchy with the aim of maximising their advantages and minimising their limitations. To amalgamate the predictions from the carefully selected base models, a meta-model was defined and subsequently fine-tuned. The efficacy of stacking was rigorously evaluated, demonstrating its capacity to outperform individual models and enhance predictive performance. The findings from this phase are presented below.

Table 4.5: Performance Metrics After Applying Stacking Ensemble Method

ML Model Used	Accuracy %	Recall % (Macro Avr)	Precision % (Macro Avr)	F1 Score % (Macro Avr)
Meta-Model	99.74%	98%	100%	99%

The results presented in the table above are consistent with expectations based on stacking as an ensemble method's principles and fundamentals. Stacking, as previously stated, orchestrates the collaboration and interplay of a number of models, including those previously evaluated. As a result, the results are similar to those obtained by XGBoost, RF, and DT. From a characteristics standpoint, this can happen for a variety of reasons. Model similarity

and optimal balance are two of these factors. If the stacking ensemble's base models are similar to the XGBoost model, their predictions may be similar as well. The ability to capture a wide range of patterns by combining different models is often at the heart of stacking's effectiveness. If the ensemble models are closely related, their combined predictive power may not differ significantly from that of a single, powerful model, such as XGBoost. Furthermore, it is possible that the stacking framework's hyperparameter tuning and model selection produced an ensemble with predictive power comparable to that of individual strong performers such as XGBoost, DT, and LR. Finally, because of the nature of stacking, the results are likely to outperform those of single models. The base models are trained and their predictions are used to derive meta-features during both the training and testing phases. As a result, metrics like accuracy from the base models serve as meta-features, helping to improve accuracy and robustness.

To summarise, the comprehensive analysis incorporating the CRISP-DM cycle, various machine learning models, and ensemble techniques provides a nuanced understanding of model efficacy in identifying Cardiovascular Diseases (CVD). Prior to hyperparameter tuning, tree-based models emerged as strong performers, particularly in terms of recall. Subsequently, ensemble methods like XGBoost and Random Forest, optimised through GridSearchCV, proved to be the most effective, while other models such as SVM and KNN showed only marginal improvements through tuning. The inclusion of a validation split further bolstered the evaluation process by providing a more generalised performance metric. Stacking, as an advanced ensemble technique, yielded results comparable to leading individual models like XGBoost and Random Forest, demonstrating the potency of model aggregation. The systematic approach involving data splits, parameter tuning, and the careful selection of evaluation metrics such as accuracy, recall, and precision offers a robust foundation for future developers in the field of medical predictive modelling. Importantly, the use of macro-averaging ensures that the performance metrics are not skewed by class imbalances, providing a more balanced and insightful evaluation. Overall, this study lays a solid groundwork for employing machine learning models, particularly ensemble methods, in the diagnosis and prediction of CVD, thereby highlighting the significant impact these models can have.

Chapter 5

Conclusions and Future Work

5.1 Conclusion of Work

In conclusion, we successfully met the project's aims, objectives, and outlined goals by evaluating and understanding the impact that machine learning models have on the detection of CVD. The project commenced with the careful selection of a dataset appropriate for CVD, which was subsequently processed using various modelling techniques. Adhering to the CRISP-DM methodology, we employed data cleansing techniques and exploratory data analysis (EDA) to gain a thorough understanding prior to proceeding with modelling.

Initially, several models were tested, among which Decision Trees stood out for their effectiveness in terms of recall. These models were further optimised through hyperparameter tuning, employing advanced ensemble methods such as XGBoost and Random Forest, which were refined using GridSearchCV. Based on carefully selected evaluation metrics, including accuracy, recall, and F1-score, this methodical approach yielded positive results and provided a nuanced understanding of the model's impact on the processed data. Additionally, the project explored higher-level ensemble techniques like stacking, thereby demonstrating their superiority over leading individual models.

Overall, the project spanned a wide array of data science and machine learning topics, culminating in a comprehensive solution that successfully predicts scenarios involving the key CVD factor chosen as the target variable.

5.2 Future Modifications

If given the opportunity to conduct this research again, with a similar or even greater time allocation, there are several elements I would change or reconsider to ensure the best possible outcome. One of the key elements I would revisit pertains to the dataset. Although the Framingham dataset was selected due to its rich heritage and detailed information, its age and potential relevance in today's context could be limiting. Given the advent of COVID-19 and the availability of more recent data, I would aim to use a newer dataset to make the data modelling, feature extraction, and correlations more pertinent to current circumstances. However, this could introduce new challenges such as obtaining permissions and navigating GDPR and data protection laws.

Furthermore, if given the chance to develop this project further, I would consider the inclusion of another population, potentially in collaboration with a longitudinal study. The existing research focused solely on the American population and did not account for other diverse communities. Therefore, including additional countries and populations would provide a more comprehensive understanding of the problem. Longitudinal data could also offer deeper insights, as trends and correlations could be observed over time.

Moreover, I would be interested in exploring alternative classification models, such as Inductive Logic Programming (ILP), to evaluate their potential for enhancing model performance. ILP could provide a different perspective on the relationships within the data, thereby potentially offering improved or complementary results to the machine learning techniques used in the initial study.

Chapter 6

Reflection

To summarise and conclude, this project allowed me to apply a variety of techniques, principles, and technologies. I was able to apply what I studied at the University of Surrey, including knowledge from modules such as Principal Business Analytics (PBA), Data Science Principles and Practices, and modelling applications. I gained deeper insights into the operation and utility of various models by applying what I learned in these courses. In addition, I was able to learn new skills, such as building ensemble models, to help with the project. The use of Python was a fundamental aspect of this project, which I was able to expand on to strengthen the project overall.

Several setbacks occurred during the course of the project, including uncertainties related to the selection of an appropriate dataset. The project's primary goal was to investigate cardiovascular diseases (CVD), particularly in the post-COVID-19 context. Despite my lack of medical experience, I aimed to use a reputable dataset that could comprehensively represent CVD-related factors. While there were several datasets available for this purpose, I discovered that the Framingham dataset provided a variety of features that proved useful to the study.

To summarise, this project was a challenge and an educational opportunity. This research project allowed me to further delve into the medical aspects of cardiovascular diseases, and the functionality of machine learning models. This project also allowed me to apply the knowledge developed during my time at the University of Surrey. Additionally, I was able to further develop skills such as time-management skills, documentation skills and follow ML protocols and key principles such as CRISP-DM cycle. These skills allowed for me to appropriately complete this project, and further develop my skill set.

To summarise, this project was both difficult and educational. It enabled me to delve deeper into the medical aspects of cardiovascular diseases as well as investigate the capabilities of machine learning models. Furthermore, the project provided a platform for me to apply and further develop skills I learned at the University of Surrey, such as time management, documentation, and adherence to machine learning protocols and key principles like the CRISP-DM cycle. These abilities were critical to the project's success and the expansion of my overall skill set.

References

Welcome to the qrisk®3-2018 risk calculator, 2018. URL https://qrisk.org. [Accessed 6 June 2013].

Reynolds risk score, 2018. URL http://www.reynoldsriskscore.org. [Accessed 6 June 2023].

Hypertension, 2019. URL https://www.ncbi.nlm.nih.gov/books/NBK470294/. [Accessed 6 June 2023].

Predicting house prices with linear regression, 2019. URL https://shorturl.at/ksF89.

A step-by-step introduction to pca, 2020. URL https://towardsdatascience.com/a-step-by-step-introduction-to-pca-c0d78e26a0dd. [Accessed 20 July 2023].

Regression vs. classification in machine learning: What's the difference?, 2021. URL https://www.springboard.com/blog/data-science/regression-vs-classification/#:~:text=Both%20regression%20and%20classification%20are,along%20with%20correctly%20labeled%20data. [Accessed 28 August 2023].

Bait 509: Business applications of machine learning, 2021. URL https://bait509-ubc.github.io/BAIT509/intro.html. [Accessed 1 August 2023].

Regional ethnic diversity, 2022. URL https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/regional-ethnic-diversity/latest#main-facts-and-figures. [Accessed 20 June 2023].

What is the difference between regression and classification?, 2022. URL https://careerfoundry.com/en/blog/data-analytics/regression-vs-classification/#what-is-classification. [Accessed 21 August 2023].

Stents vs bypass surgery for treating cad, 2022. URL https://www.verywellhealth.com/stents-or-bypass-surgery-1745725. [Accessed 28 August 2023].

Decision tree hyperparameters, 2023. URL https://www.educba.com/decision-tree-hyperparameters/. [Accessed 17 August 2023].

Mortgages in the united kingdom, 2023. URL https://www.expatica.com/uk/housing/buying/your-guide-to-uk-mortgages-747470/#eligibility. [Accessed 30 June 2023].

Hyperparameter tuning with grid search cv, 2023. URL https://www.mygreatlearning.com/blog/gridsearchcv/. [Accessed 15 August 2023].

Pandas, 2023. URL https://pandas.pydata.org. [Accessed 6 July 2023].

British heart foundation: Cardiovascular heart disease, n.d. URL https://www.bhf.org.uk/informationsupport/conditions/cardiovascular-heart-disease. [Accessed 11 July 2023].

Z. S. Abdallah, L. Du, and G. I. Webb. Data preparation. *Encyclopedia of Machine Learning and Data Mining*, pages 318–327, 2017.

Khan Academy. Conditional probability with Bayes' theorem (video). https://www.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/v/bayes-theorem-visualized, n.d.

A. Afolabi. Implementing k-means clustering. *Implementing K-Means Clustering*, 2023.

AWS. What is boosting? guide to boosting in machine learning, n.d. URL https://aws.amazon.com/what-is/boosting/.

Maxwell Barton and Barry Lennox. Model stacking to improve prediction and variable importance robustness for soft sensor development. *Digital Chemical Engineering*, 3: 100034, 2022. ISSN 2772-5081. doi: https://doi.org/10.1016/j.dche.2022.100034. URL https://www.sciencedirect.com/science/article/pii/S2772508122000254.

J. L. Björkegren and A. J. Lusis. Atherosclerosis: Recent developments. *Cell*, 185(10):1630–1645, 2022.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

J. Brownlee. Train-test split for evaluating machine learning algorithms, 2020a. URL https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/. [Accessed 30 June 2023].

J. Brownlee. Why use ensemble learning?, 2020b. URL https://machinelearningmastery.com/why-use-ensemble-learning/. Accessed: insert date here.

C. Chatfield. Exploratory data analysis. *European Journal of Operational Research*, 23(1): 5–13, 1986.

N. Chaturvedi. Ethnic differences in cardiovascular disease. *Heart*, 89(6):681–686, 2003.

J. Cornfield. Bayes theorem. *Review of the International Statistical Institute*, 35(1):34–49, 1967.

Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6): e01802, 2019. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2019.e01802. URL https://www.sciencedirect.com/science/article/pii/S2405844018353404.

DataCamp. Support vector machines with scikit-learn tutorial, 2019. URL https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python. [Accessed 17 August 2023].

A. P. De Filippis et al. Journal of the american college of cardiology. *Journal of the American College of Cardiology*, 58(20):2076–2083, 2011.

T.G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, volume 1857, Berlin, Heidelberg, 2000. Springer. doi: 10.1007/3-540-45014-9_1.

L. Ding, Y. Liang, E.C.K. Tan, et al. Smoking, heavy drinking, physical inactivity, and obesity among middle-aged and older adults in china: cross-sectional findings from the baseline survey of charls 2011–2012. *BMC Public Health*, 20:1062, 2020. doi: 10.1186/s12889-020-08625-5. URL https://doi.org/10.1186/s12889-020-08625-5.

A Dinh, S Miertschin, A Young, and SD Mohanty. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*, 19(1):211, Nov 6 2019. doi: 10.1186/s12911-019-0918-5.

British Heart Foundation. How your ethnic background affects your risk of heart and circulatory diseases, 2021. URL https://www.bhf.org.uk/what-we-do/our-research/research-successes/ethnicity-and-heart-disease. [Accessed 8 June 2023].

GeeksforGeeks. Major kernel functions in support vector machine (svm), July 2020a. URL https://www.geeksforgeeks.org/major-kernel-functions-in-support-vector-machine-svm/.

GeeksforGeeks. Ml - gradient boosting, August 2020b. URL https://www.geeksforgeeks.org/ml-gradient-boosting/.

P. Germanakos and A. Matz. Increasing the quality of use case definition through a design thinking collaborative method and an alternative hybrid documentation style. pages 48–59, 2016.

A. Gholamy, V. Kreinovich, and O. Kosheleva. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Engineering and Applied Computer Science*, 5(2), 2018.

J. Hippisley-Cox and C. Coupland. Derivation and validation of updated qfracture algorithm to predict risk of osteoporotic fracture in primary care in the united kingdom: Prospective open cohort study. *BMJ*, 244, 2012.

J. Hippisley-Cox, C. Coupland, and P. Brindle. Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, 2017.

M. Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process*, 5(2), 2015.

Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 5(1), Aug 2016. doi: https://doi.org/10.1186/s40064-016-2941-7. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4978658/.

Ahmedbahaaaldin Ibrahem Ahmed Osman, Ali Najah Ahmed, Ming Fai Chow, Yuk Feng Huang, and Ahmed El-Shafie. Extreme gradient boosting (xgboost) model to predict the groundwater levels in selangor malaysia. *Ain Shams Engineering Journal*, 12(2):1545–1556, 2021. ISSN 2090-4479. doi: https://doi.org/10.1016/j.asej.2020.11.011. URL https://www.sciencedirect.com/science/article/pii/S2090447921000125.

JJ. Mae and rmse — which metric is better?, March 2016. URL https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d.

James Joyce. Author and citation information for "bayes' theorem", 2021. URL https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=bayes-theorem.

S. Karamizadeh et al. An overview on principal component analysis. *Journal of Signal and Information Processing*, 4:173–175, 2013.

K. Kaur. An illustrative approach to use sql functions: A review. *International Journal of Engineering and Applied Computer Science*, 02(03):114–122, 2017.

T. Kawada. Reynolds risk score as a risk assessment tool for cardiovascular disease after 10 years: Its strong relationship with blood pressure. *The Journal of Clinical Hypertension*, 14 (8):571–572, 2023.

A. Klisić. Cardiovascular risk assessed by reynolds risk score in relation to waist circumference in apparently healthy middle-aged population in montenegro. *Acta Clinica Croatica*, 57(1): 20–30, 2018.

R. Kohavi and J. R. Quinlan. Data mining tasks and methods: Classification: decision-tree discovery. In *Handbook of data mining and knowledge discovery*, pages 267–276. 2022.

Jain Kopal. How to improve naive bayes?, 2021. URL https://medium.com/analytics-vidhya/how-to-improve-naive-bayes-9fa698e14cba. [Accessed 10 August 2023].

A. Kumar. Correlation concepts, matrix & heatmap using seaborn, April 2022. URL https://vitalflux.com/correlation-heatmap-with-seaborn-pandas. Accessed: insert date you accessed this resource.

D. Lloyd-Jones. Concepts of screening for cardiovascular risk factors and disease. pages 433–442, 2011.

E. O. Lopez, B. D. Ballard, and A. Jan. Cardiovascular disease. *StatPearl [Internet]*, 2022.

B. Lorena et al. Teaching and learning with jupyter. 2019. URL https://jupyter4edu.github.io/jupyter-edu-book/index.html. [Accessed 6 July 2023].

Garm Lucassen, Fabiano Dalpiaz, Jan Martijn Van der Werf, and Sjaak Brinkkemper. The use and effectiveness of user stories in practice. pages 205–222, 03 2016. ISBN 978-3-319-30281-2. doi: 10.1007/978-3-319-30282-9_14.

R. G. Mantovani et al. An empirical study on hyperparameter tuning of decision trees. *Machine Learning*, 2, 2018.

S. Marukatat. Tutorial on pca and approximate pca and approximate kernel pca. *Artificial Intelligence Review*, 1(33), 2022.

S. K. McGrath and S. J. Whitty. Stakeholder defined. *International Journal of Managing Projects in Business*, 10(4):721–748, 2017.

MLMath.io. Math behind support vector machine(svm). *Medium*, 2019.

American College of Cardiology. Project risk reduction by therapy, 2016. URL `https://tools.acc.org/ascvd-risk-estimator-plus/#!/calculate/estimate/`. [Accessed 6 June 2023].

American College of Cardiology. Ascvd risk estimator, 2021. URL `https://tools.acc.org/ascvd-risk-estimator-plus/#`.

World Health Organization. Cardiovascular diseases, 2022. URL `https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1`. Accessed: INSERT DATE YOU ACCESSED THE WEBSITE.

Jalal Ramzai. Top 10 model evaluation metrics for classification ml models, May 2020. URL `https://towardsdatascience.com/top-10-model-evaluation-metrics-for-classification-ml-models-a0a0f1d51b9`.

Sebastian Raschka. How do i select svm kernels?, 2023. URL `https://sebastianraschka.com/faq/docs/select_svm_kernels.html`. Accessed: 30 Aug. 2023.

Y. Resti et al. A bootstrap-aggregating in random forest model for classification of corn plant diseases and pests. *Science and Technology Indonesia*, 2023.

M. Schonlau and R. Y. Zou. The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, 20(1), 2020.

scikit-learn developers. sklearn.neighbors.kneighborsclassifier — scikit-learn 0.22.1 documentation, 2019. URL `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`. Accessed: insert date here.

Scikit-learn.org. sklearn.model_selection.RandomizedSearchCV — scikit-learn 0.21.3 documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html`, 2019.

Rai Dilawar Shahjehan and Beenish S. Bhutta. Coronary artery disease. 2023. [Updated 2023 Feb 9]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK564304/`.

Saman Siadati. What is unsupervised learning. 08 2018. doi: $10.13140/\mathrm{RG}.2.2.33325.10720$.

YY Song and Y Lu. Decision tree methods: applications for classification and prediction, Apr 25 2015.

Brijesh Soni. Stacking to improve model performance: A comprehensive guide on ensemble learning in python, May 2023. URL `https://shorturl.at/xGKOV`.

Md. R. K. Sony. Uci heart disease data, 2020. URL `https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data`.

Pedro Strecht, Luís Cruz, Carlos Soares, João Moreira, and Rui Abreu. A comparative study of classification and regression algorithms for modelling students' academic performance, 06 2015.

Suryakanthi Tangirala. Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2), 2020.

S Wild and P McKeigue. Cross-sectional analysis of mortality by country of birth in england and wales, 1970-92. *BMJ*, 314(7082):705, 1997.

Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining, n.d. URL `https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf`.

P. Yadav. Decision tree in machine learning, 2018. URL `https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96`. Available at: `https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96`.

Fenjiro Youssef. Machine learning for banking: Loan approval use case, 2018. URL `https://medium.com/@fenjiro/data-mining-for-banking-loan-approval-use-case-e7c2bc3ece3`.

Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*, 4(11):218, 2016.

Dongdong Zhao, Xiaoyi Hu, Shengwu Xiong, Jing Tian, Jianwen Xiang, Jing Zhou, and Huanhuan Li. k-means clustering and knn classification based on negative databases. *Applied Soft Computing*, 110:107732, 2021. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.107732. URL `https://www.sciencedirect.com/science/article/pii/S1568494621006530`.

Boback Ziaeian and Gregg C. Fonarow. Statins and the prevention of heart disease. *JAMA Cardiology*, 2(4):464, April 2017. doi: 10.1001/jamacardio.2016.4320.

Appendix A

Multiple Regression Analysis to Predict Log-Transformed Reynolds Risk Score

The following table represents the application of multiple regression analysis to predict log-transformed RRS (Reynolds Risk Score) by utilising 7 key components and log-transformed HOMA-IR (Homeostasis Model Assessment for Insulin Resistance). More information on the model can be found in the article developed and written by Tomoyuki Kawada MD, PhD [Kawada, 2023].

Independent Variables	**B (SE)**	**Beta**	***P* Value**
Age	0.036 (<0.001)	0.617	<.001
Systolic blood pressure	0.008 (<0.001)	0.430	<.001
Total cholesterol	0.002 (<0.001)	0.239	<.001
HDL cholesterol	−0.005 (<0.001)	−0.299	<.001
Log10 (hsCRP)	0.100 (0.001)	0.159	<.001
Current smoking	0.173 (0.001)	0.318	<.001
Family history of CVD	0.231 (0.002)	0.211	<.001
Log10 (HOMA-IR)	0.011 (0.002)	0.011	<.001

Figure A.1: Transformed Reynolds Risk Score

Appendix B

Framingham Dataset Documentation

This appendix serves as a concise documentation for the Framingham dataset, providing essential information about its structure, variables, and data sources. The Framingham dataset is widely used in cardiovascular research and contributes to understanding cardiovascular disease risk factors.

Users will find a summary of the dataset's structure, including the number of observations and variables. The documentation highlights the data collection process, sources, and relevant methodologies. It offers a clear understanding of the dataset's origin and reliability.

Including this dataset documentation, the appendix promotes transparency, reproducibility, and reliable research outcomes. It serves as a valuable resource for understanding and utilizing the Framingham dataset effectively in cardiovascular studies.

FHS Teaching Longitudinal Data Documentation 2021a
or scan the QR code below to view the dashboard on any device.

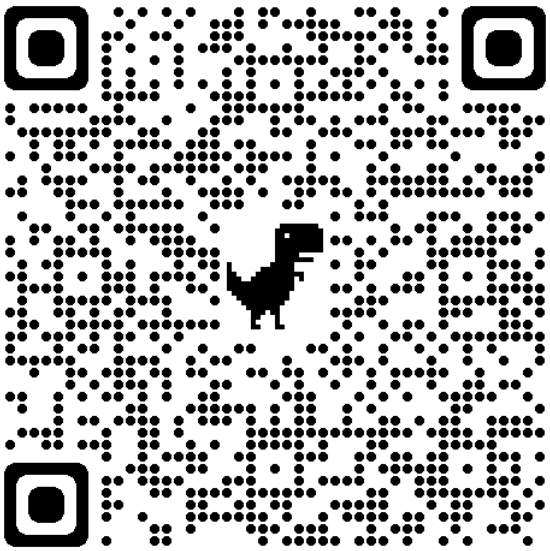

Figure B.1: QR To View Document

Appendix C

Framingham Dataset Feature Definition

This appendix gives a thorough rundown of all the features found in the Framingham dataset. The Framingham dataset is a popular dataset in cardiovascular research and contains useful data for examining risk factors for cardiovascular disease. For accurate data analysis and modelling, it is essential to comprehend the meaning and importance of each feature. Further information on the documentation can be found in Appendix B.

Variable	Description	Units	Range or count
RANDID	Unique identification number for each participant		2448-9999312
SEX	Participant sex	1=Men 2=Women	n=5022 n=6605
PERIOD	Examination Cycle	1=Period 1 2=Period 2 3=Period 3	n=4434 n=3930 n=3263
TIME	Number of days since baseline exam		0-4854
AGE	Age at exam (years)		32-81
SYSBP	Systolic Blood Pressure (mean of last two of three measurements) (mmHg)		83.5-295
DIABP	Diastolic Blood Pressure (mean of last two of three measurements) (mmHg)		30-150
BPMEDS	Use of Anti-hypertensive medication at exam	0=Not currently used 1=Current Use	n=10090 n=944
CURSMOKE	Current cigarette smoking at exam	0=Not current smoker 1=Current smoker	n=6598 n=5029
CIGPDAY	Number of cigarettes smoked each day	0=Not current smoker 1-90 cigarettes per day	
EDUC	Attained Education	1=0-11 years 2=High School Diploma, GED 3=Some College, Vocational School 4=College (BS, BA) degree or more	
TOTCHOL	Serum Total Cholesterol (mg/dL)		107-696
HDLC	High Density Lipoprotein Cholesterol (mg/dL)	available for period 3 only	10-189
LDLC	Low Density Lipoprotein Cholesterol (mg/dL)	available for period 3 only	20-565
BMI	Body Mass Index, weight in kilograms/height meters squared		14.43-56.8
GLUCOSE	Casual serum glucose (mg/dL)		39-478

Figure C.1: Framingham General Data Definition, Page 2

Variable	Description	Units	Range or count
DIABETES	Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more	0=Not a diabetic 1=Diabetic	n=11097 n=530
HEARTRTE	Heart rate (Ventricular rate) in beats/min		37-220
PREVAP	Prevalent Angina Pectoris at exam	0=Free of disease 1=Prevalent disease	n=11000 n=627
PREVCHD	Prevalent Coronary Heart Disease defined as pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina)	0=Free of disease 1=Prevalent disease	n=10785 n=842
PREVMI	Prevalent Myocardial Infarction	0=Free of disease 1=Prevalent disease	n=11253 n=374
PREVSTRK	Prevalent Stroke	0=Free of disease 1=Prevalent disease	n=11475 n=152
PREVHYP	Prevalent Hypertensive. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg	0=Free of disease 1=Prevalent disease	n=6283 n=5344

Figure C.2: Framingham General Data Definition, Page 3

Variable name	Description
ANGINA	Angina Pectoris
HOSPMI	Hospitalized Myocardial Infarction
MI_FCHD	Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease
ANYCHD	Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease
STROKE	Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease
CVD	Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease
HYPERTEN	Hypertensive. Defined as the first exam treated for high blood pressure or second exam in which either Systolic is ≥ 140 mmHg or Diastolic ≥ 90mmHg
DEATH	Death from any cause
TIMEAP	Number of days from Baseline exam to first Angina during the followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup
TIMEMI	Defined as above for the first HOSPMI event during followup
TIMEMIFC	Defined as above for the first MI_FCHD event during followup
TIMECHD	Defined as above for the first ANYCHD event during followup
TIMESTRK	Defined as above for the first STROKE event during followup
TIMECVD	Defined as above for the first CVD event during followup
TIMEHYP	Defined as above for the first HYPERTEN event during followup
TIMEDTH	Number of days from Baseline exam to death if occurring during followup or Number of days from Baseline to censor date. Censor date may be end of followup, or last known contact date if subject is lost to followup

Figure C.3: Framingham General Data Definition, Page 4

Appendix D

Standardised Mortality Ratios for Heart Diseases

The following graph illustrates the standardised mortality ratios (SMR) for heart diseases and stroke that may be present in different ethnicity groups, namely European, South Asian and African Caribbean between the age category of 20-62 between 1989-92. The following has been adapted from Wild and McKeigue [Wild and McKeigue, 1997], and cross-referred to by Chaturvedi[Chaturvedi, 2003] in the Ethnic Differences in Cardiovascular Diseases, written in June 2003.

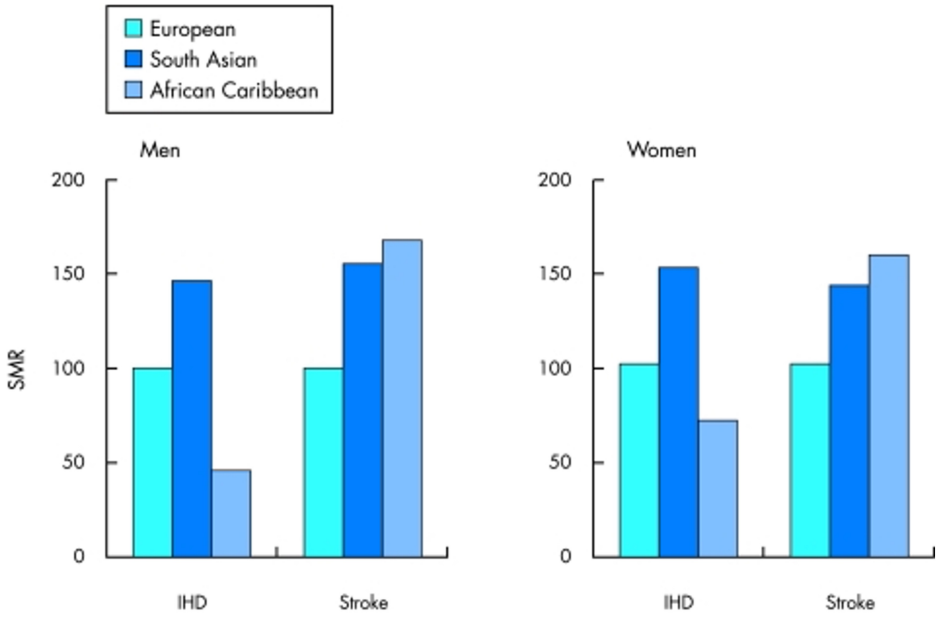

Figure D.1: Standardised Mortality Ratios (SMR) for Heart Diseases Among Different Ethnicity Groups

Appendix E

GitHub Repository

As part of this project, GitHub was used a central repository storage system. As part of this repostitory, users can find a link to all the necessary documentation and files used as part of this project. The repository includes, a copy of the .ipynb file, including data dictionary, and data source.

GitHub Repository

Appendix F

SAGE-HDR Ethics Form

The following appendix attaches the SAGE-HDR (Self-Assessment for Governance and Ethics - Human and Data Research) Ethics form, required for completion as part of this project. The following document includes a self-assessment, validating whether this project requires review by an ethics committee before collection commences. No data was personally collected as part of this project as the data was sourced from a source (National Heart, Blood and Lung Institute).

SAGE-HDR (v3.8 24/04/23)

Response ID	Completion date
1046015-1045997-115244794	23 Aug 2023, 13:13 (BST)

1	**Applicant Name**	JASON JAY DOOKARUN
1.a	**University of Surrey email address**	jd00795@surrey.ac.uk
1.b	**Level of research**	Postgraduate Taught (Masters)
1.b.i	**Please enter your University of Surrey supervisor's name. If you have more than one supervisor, enter the details of the individual who will check this submission.**	DR TOM THORNE
1.b.ii	**Please enter your supervisor's University of Surrey email address. If you have more than one supervisor, enter the details of the supervisor who will check this submission.**	tom.thorne@surrey.ac.uk
1.c	**School or Department**	Computer Science
1.d	**Faculty**	FEPS - Faculty of Engineering and Physical Sciences

2	**Project title**	Development of a Cardiovascular Disease Risk Prediction System using Machine Learning and Patient Health Data

3	**Please enter a brief summary of your project and its methodology in 250 words. Please include information such as your research method/s, sample, where your research will be conducted and an overview of the aims and objectives of your research.**	In this research project, the primary aim is to evaluate the effectiveness of various Machine Learning models in predicting the risk of developing Cardiovascular Disease (CVD). Reaching high evaluation metrics in model performance is the goal, which will help create predictive systems for CVD risk that are more accurate and dependable. Using data from the National Heart, Lung, and Blood Institute, the Framingham Heart dataset is part of the applied methodology. Several variables related to cardiovascular health are included in this dataset, which is used as the project's sample. To ensure the dataset is suitable for modelling, a thorough data preparation phase is conducted, involving multiple data cleansing techniques. The cleaned data is subjected to machine learning models, such as Support Vector Machines, Random Forests, and Decision Trees, among others. Ensemble techniques are also used to improve the performance of individual models. Evaluation of performance metrics, including F1-score, recall, accuracy, and precision, reveals how well the models predict the risk of CVD. The study is carried out under strict adherence to ethical and data protection

		regulations in a controlled setting. This project aims to progress the field of predictive healthcare, specifically in the area of cardiovascular diseases, by achieving the stated goals and objectives.

4	**Are you planning to join on to an existing Standard Study Protocol (SSP)? SSPs are overarching pre-approved protocols that can be used by multiple researchers investigating a similar topic area using identical methodologies. Please note, SSPs are only being used by 3 schools currently and cannot be used by other schools. Using an SSP requires permission and sign-off from the SSP owner**	NO

5	**Are you making an amendment to a project with a current University of Surrey favourable ethical opinion or approval in place?**	NO

6	**Does your research involve any animals, animal data or animal derived tissue, including cell lines?**	NO

8	**Does your project involve human participants (including human data and/or any human tissue*)?**	YES

9	**Will you be accessing any organisations, facilities or areas that may require prior permission? This includes organisations such as schools (Headteacher authorisation), care homes (manager permission), military facilities, closed online forums, private social media pages etc. This also includes using University mailing lists (admin permission). If you are unsure, please contact ethics@surrey.ac.uk.**	NO

10	**Does your project involve any type of human tissue research? This includes Human Tissue Authority (HTA) relevant, or non-relevant tissue (e.g. non-cellular such as plasma or serum), any genetic material, samples that have been previously collected, samples being collected directly from the donor or obtained from another researcher, organisation or commercial source.**	NO

11	**Does your research involve exposure of participants to any hazardous materials e.g. chemicals, pathogens, biological agents or does it involve any activities or locations that may pose a risk of harm to the researcher or participant?**	NO

12	**Will you be importing or exporting any samples (including human, animal, plant or microbial/pathogen samples) to or from the UK?**	NO

13	**Will any participant visits be taking place in the Clinical Research Building (CRB)? (involving clinical procedures; if only visiting the CRB to collect/drop-off equipment or to meet with the research team (i.e. for informed consent/discussion) select 'NO').**	NO

14	**Will you be working with any collaborators or third parties to deliver any aspect of the research project?**	NO

15	**Are you conducting a service evaluation or an audit? Or using data from a service evaluation or audit?**	NO

16	**Does your funder, collaborator or other stakeholder require a mandatory ethics review to take place at the University of Surrey?**	NO

17	**Does your research involve accessing students' results or performance data? For example, accessing SITS data.**	NO

18	**Will ANY research activity take place outside of the UK?**	NO

19	**Are you undertaking security-sensitive research, as defined in the text below?**	NO

20	**Does your project require the processing of special category1 data?**	NO

21	**Have you selected YES to one or more of the above governance risk questions on this page (Q10-Q20)?**	NO

22	**Does your project process personal data2? Processing covers any activity performed with personal data, whether digitally or using other formats, and includes contacting, collecting, recording, organising, viewing, structuring, storing, adapting, transferring, altering, retrieving, consulting, marketing, using, disclosing, transmitting, communicating, disseminating, making available, aligning, analysing, combining, restricting, erasing, archiving, destroying.**	NO

23	**Are you using a platform, system or server external to the University approved platforms (Outside of Microsoft Office programs, Sharepoint, OneDrive Qualtrics, REDCap, JISC online surveys (BOS) and Gorilla)**	NO

24	**Does your research involve any of the above statements? If yes, your study may require external ethical review or regulatory approval**	NO

25	**Does your research involve any of the above? If yes, your study may require external ethical review or regulatory approval**	NO

26	**Does your project require ethics review from another institution? (For example: collaborative research with the NHS REC, the Ministry of Defence, the Ministry of Justice and/or other universities in the UK or abroad)**	NO

27	**Does your research involve any of the following individuals or higher-risk methodologies? Select all that apply or select 'not applicable' if no options apply to your research. Please note: the UEC reviewers may deem the nature of the research of certain high risk projects unsuitable to be undertaken by undergraduate students**	NOT APPLICABLE - none of the above high-risk options apply to my research.

28	**Does your research involve any of the following individuals or medium-risk methodologies? Select all that apply or select 'not applicable' if no options apply to your research.**	NOT APPLICABLE - none of the above medium-risk options apply to my research.

29	**Does your research involve any of the following individuals or lower-risk methodologies? Select all that apply or select 'not applicable' if no options apply to your research.**	NOT APPLICABLE - none of the above lower-risk options apply to my research.

30	**Declarations**	• I confirm that I have read the University's Code on Good Research Practice and ethics policy and all relevant professional and regulatory guidelines applicable to my research and that I will conduct my research in accordance with these. • I confirm that I have provided accurate and complete information regarding my research project • I understand that a false declaration or providing misleading information will be considered potential research misconduct resulting in a formal investigation and subsequent disciplinary proceedings liable for reporting to external bodies • I understand that if my answers to this form have indicated that I must submit an ethics and governance application, that I will NOT commence my research until a Favourable Ethical Opinion is issued and governance checks are cleared. If I do so, this will be considered research misconduct and result in a formal investigation and subsequent disciplinary proceedings liable for reporting to external bodies. • I understand that if I have selected 'YES' on any governance risk questions and/or have selected any options on the higher, medium or lower risk criteria then I MUST submit an ethics and governance application (EGA) for review before conducting any research. If I have NOT selected any governance risks or selected any of the higher, medium or lower ethical risk criteria, I understand I can proceed with my research without review and

		acknowledge that my SAGE answers and research project will be subject to audit and inspection by the RIGO team at a later date to check compliance.

31	**If I am conducting research as a student:**	• I confirm that I have discussed my responses to the questions on this form with my supervisor to ensure they are correct. • I confirm that if I am handling any information that can identify people, such as names, email addresses or audio/video recordings and images, I will adhere to the security requirements set out in the relevant Data Protection Policy