

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

\_\_\_\_\_\_\_\_\_\_

**Jason Jay Dookarun**

**URN**: 6494278

**Supervisor**: Dr Tom Thorne

**Department of Computer Science**

**University of Surrey**

**Declaration**

I, Jason Jay Dookarun, of the Department of Computer Science, University of Surrey, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

Jason Jay Dookarun

2023

**Abstract**

Cardiovascular diseases (CVDs) are a significant global health concern, demanding the development of effective risk prediction systems for early identification and intervention. This thesis focuses on the development of a cardiovascular disease risk prediction system using machine learning algorithms and patient health data.

The study begins with a comprehensive review of relevant literature, exploring existing risk assessment models and identifying gaps in current approaches. A thorough analysis of available patient health data is conducted, encompassing clinical measurements, demographic information, and lifestyle indicators, to capture a comprehensive understanding of the factors contributing to cardiovascular disease risk.

Various statistical models, including logistic regression, decision trees, random forests, and support vector machines, are examined and compared for their predictive performance. The models are trained and evaluated using a large-scale dataset containing anonymized patient records, allowing for the identification of patterns and risk factors associated with cardiovascular diseases.

The outcome of this research is a robust cardiovascular disease risk prediction system that enables healthcare professionals to assess an individual's propensity for cardiovascular disease based on their health data. The developed system provides a personalized risk score for each patient, empowering clinicians to make informed decisions regarding preventive interventions and patient management strategies.

The evaluation of the developed system includes performance metrics such as accuracy, precision and recall, ensuring its reliability and effectiveness. Additionally, potential limitations and areas for future research are discussed to guide further advancements in cardiovascular disease risk prediction using statistical modelling techniques.

Overall, this thesis contributes to the field of cardiovascular disease prevention by integrating statistical models and patient health data to develop an accurate and personalized risk prediction system. The findings have significant implications for early intervention and tailored healthcare strategies, ultimately reducing the burden of cardiovascular diseases on individuals and healthcare systems.

**Acknowledgements**

**Table of Contents**

[List of Figures 7](#_Toc137554945)

[List of Tables 8](#_Toc137554946)

[List of Abbreviations 9](#_Toc137554947)

[1 Introduction 10](#_Toc137554948)

[1.1 Background 10](#_Toc137554949)

[1.2 Problem Statement 10](#_Toc137554950)

[1.3 Aims and Objectives 10](#_Toc137554951)

[1.4 Approach to Solution 11](#_Toc137554952)

[1.5 Organisation of Document 12](#_Toc137554953)

[2 Literature Review 14](#_Toc137554954)

[2.1 Background 14](#_Toc137554955)

[2.2 Existing Systems 14](#_Toc137554956)

[2.3 Selection of Datasets 19](#_Toc137554957)

[2.4 Machine Learning Modelling Principles (focus) 20](#_Toc137554958)

[2.5 Supervised vs Unsupervised 20](#_Toc137554959)

[2.6 Models Considered 21](#_Toc137554960)

[Decision Tree 21](#_Toc137554961)

[Bayesian Learning 21](#_Toc137554962)

[SVM 21](#_Toc137554963)

[Clustering 21](#_Toc137554964)

[Stacking 21](#_Toc137554965)

[2.8 Evaluation Metrics 21](#_Toc137554966)

[3 Methodology 22](#_Toc137554967)

[3.1 Data Exploration and Extraction 23](#_Toc137554968)

[3.2 Data Preparation 23](#_Toc137554969)

[3.3 Data Comprehension 23](#_Toc137554970)

[3.4 Data Modelling 23](#_Toc137554971)

[3.5 Data Evaluation 23](#_Toc137554972)

[4 Results + Testing 24](#_Toc137554973)

[5 Discussion 25](#_Toc137554974)

[6 Conclusion & Future Work 26](#_Toc137554975)

[Reflections 27](#_Toc137554976)

[References 28](#_Toc137554977)

[Appendices 30](#_Toc137554978)

[Appendix A: Multiple Regression Analysis to Predict Log-Transformed Reynolds Risk Score 30](#_Toc137554979)

# List of Figures

[**Figure 1**: Reynolds Risk Score System (Ridker, 2018) 12](#_Toc136954679)

[**Figure 2**: QRisk Calculator (Nottingham, 2018) 12](#_Toc136954680)

# List of Tables

# List of Abbreviations

AI Artificial Intelligence

CSV Comma-Separated Values File

COVID-19 Coronavirus Disease

DS Data Science

ETL Extraction, Loading, Transformation

EDA Exploratory Data Analysis

CVD Cardiovascular Diseases

RRS Reynolds Risk Score

FRS Framingham Risk Score

CV Cross Validation

DL Deep Learning

ML Machine Learning

AUC Area Under the Curve

NN Neural Networks

MSE Mean Squared Error

CAD Conorary Artery Disease

ECG Electrocardiogram

DT Decision Tree

HDL High-Density Lipoprotein

LDL Low-Density Lipoprotein

BP Blood Pressure

BMI Body Mass Index

LR Logistic Regression

ROC Reciever Operating Characteristics

# 1 Introduction

## 1.1 Background

Cardiovascular diseases also known as CVDs have emerged to be one of the greatest contributors to causes of death over the past decades, accounting to a substantial proportion of mortality worldwide. Around 66,000 deaths in the UK, an average of 180 people die from coronary heart disease (British Heart Foundation, 2023). Coronary heart diseases range from a multitude of conditions that affect the heart. These can range from conditions that affect heart and blood vessels, including coronary artery disease, myocardial infraction, also known as heart attacks and strokes. Similarly, CVDs represent additional diseases such as cerebrovascular disease, rheumatic heart disease and other conditions. According to the World Health Organisation, cardiovascular diseases (CVDs) are the “leading cause of death globally, responsible for approximately 17.9 million deaths each year” (World Health Organisation , 2021), with this statistic being published pre-COVID 2019 pandemic.

Age, genetic predisposition, lifestyle choices, and underlying health conditions all influence the prevalence of cardiovascular diseases. Smoking, physical inactivity, poor diet, and excessive alcohol consumption all contribute to the development of risk factors such as hypertension, obesity, high cholesterol levels, and diabetes. These risk factors, in turn, raise the chances of developing cardiovascular disease.

## 1.2 Problem Statement

// insert problem statement

## 1.3 Aims and Objectives

With consideration to the existing use cases and user stories, the aim of this project is to develop a tool whereby a user is able to determine if a patient is prone to cardiovascular diseases. In order to achieve this, the project has outlined the following goals:

* Risk Prediction:
* Early Detection:
* Decision Support:
* Integration and Accessibility:
* 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, specificity, and AUC-ROC.
* Continuous Improvement:

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 data 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.

The project team can effectively identify relevant data sources, explore the data obtained from these sources, perform modelling, evaluate the models, and present the findings through visualisations by following this solution approach. This comprehensive approach allows for informed decision-making and increases the project's success in meeting its objectives.

## 1.5 Organisation of Document

This report 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 comprehensive review of relevant literature, including previous studies, research papers, and scholarly articles on the subject. Its goal is to lay a strong foundation of knowledge and understanding in the field.
* Models Research: In this section, various models and methodologies relevant to the research topic are explored and analysed. The emphasis is on comprehending existing approaches and their relevance to the problem at hand.
* Exploration of Alternative Modelling Methods, 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.

# 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 my own 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, the concept of "stacking" will be examined, exploring its 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

// talk about existing systems that are present as part of

## 2.2 Existing Systems

Based on a set of parameters provided by the user, multiple approaches can be used to determine if an individual is susceptible to cardiovascular diseases. These approaches can be used in a variety of settings, including medical facilities and private residences. To facilitate this, the system can be implemented via a website or a solution that allows for information input and output. The system can be deployed through a user-friendly website or a dedicated solution to ensure accessibility and convenience, allowing for seamless entry and retrieval of relevant data. Individuals can easily input their information and receive accurate assessments of their susceptibility to cardiovascular diseases by using such platforms.

An extensive review and examination of several existing systems was carried out in order to gain a comprehensive understanding of this topic. The goal is to understand the various elements associated with each of these systems, such as the factors that contribute to their success, the formulation or creation processes used, the specific models used to determine their output, 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 (Ridker, 2018), developed by Dr Paul M Ridker is a system that focuses on taking an 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 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 in the patient's parents before the age of 60, and SBP. This score system, in particular, 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.

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

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

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 1**: Reynolds Risk Score System (Ridker, 2018)

When examining the success of the RRS (Reynolds Risk Score) system and how it performed in relation to its competitors, particularly the FRS (Framingham Risk Score), several research are available to refer to. This also includes an article authored by Dr Tomoyuki Kawada (Kawada, 2023), in which Kawada outlines various components and elements by which RRS outperforms FRS and specific factors that lead to RRS's demise. As described in the article, RSS outperforms the FRS system in certain areas, notably the use of family history, which has a significant impact on CVD. Furthermore, the creation of such a tool is advantageous in that it can be "useful for recommending risk-lowering actions for each individual, such as smoking cessation, regular exercise habit, 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 RSS may be linked to the metabolic syndrome since several components overlap.

Additionally, some limitations are presented as cited by Dr Kawada (Kawada, 2023). These limitations include the factors related to ethnicity, socioeconomic status and age in the RSS. These are key factors as different ethnicities are more likely to be prone to CVD, such as black Africans, African Caribbeans and South Asians in the UK, with the likelihood to develop either high blood pressure or type 2 diabetes in comparison to white Europeans (British Heart Foundation, 2021).

The next example/existing system that was analysed as part of this study focused on the QRisk Calculator system.

*The QRisk Calculator is a tool developed by the QRisk research team to estimate an individual's risk of developing cardiovascular disease (CVD) over a 10-year period. It is primarily used in the United Kingdom and takes into account various risk factors to provide a personalized risk assessment.*

*The QRisk Calculator incorporates risk factors such as age, sex, 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. These factors are used to calculate an individual's CVD risk score.*

*The calculator provides an estimated percentage representing the risk of experiencing a cardiovascular event (such as heart attack or stroke) within the next 10 years.*

*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 (FSS) and ASCVD Risk Estimator, in predicting CVD events in the UK population.*

*Bullet points summarizing the QRisk Calculator:*

*- The QRisk Calculator estimates an individual's 10-year risk of developing cardiovascular disease.*

*- It considers factors like age, sex, ethnicity, smoking status, blood pressure, BMI, diabetes, cholesterol levels, family history, and antihypertensive medication use.*

*- The calculator provides a personalized CVD risk score as a percentage.*

*- QRisk has shown good performance in predicting CVD risk in the UK population.*

*- Studies suggest QRisk may outperform the FSS and ASCVD Risk Estimator in predicting CVD events in the UK population.*

*References:*

*1. Hippisley-Cox, J., Coupland, C., & Brindle, P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ, 357, j2099.*

*2. Hippisley-Cox, J., Coupland, C., & Brindle, P. (2014). Derivation and validation of QFractureScores for prediction of hip fracture in patients aged 30-85 years in primary care: a prospective open cohort study. BMJ, 348, g342.*

*3. Collins, G. S., Altman, D. G., & Moons, K. G. (2012). Reporting of risk prediction models: TRIPOD statement. BMJ, 344, e3312.*

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 2**: QRisk Calculator (Nottingham, 2018)

Example 3: 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. It is used to assess the risk of heart attack and stroke in adults.*

*- The risk estimator takes into account several risk factors that have been shown to contribute to the development of atherosclerotic cardiovascular disease. These risk factors include age, sex, race, total cholesterol, high-density lipoprotein (HDL) cholesterol, systolic blood pressure, blood pressure-lowering medication use, diabetes status, and smoking status.*

*- To use the ASCVD Risk Estimator, the following information is required:*

*1. Age: The age of the individual in years.*

*2. Sex: Male or female.*

*3. Race: White, African American, or other.*

*4. Total cholesterol: The total level of cholesterol in millimoles per litre (mmol/L).*

*5. HDL cholesterol: The level of high-density lipoprotein cholesterol in mmol/L.*

*6. Systolic blood pressure: The systolic blood pressure reading in millimetres of mercury (mmHg).*

*7. Blood pressure-lowering medication use: Whether or not the individual is currently taking medication to lower blood pressure.*

*8. Diabetes status: Whether the individual has diabetes or not.*

*9. Smoking status: Whether the individual is a current smoker or not.*

*- Once these inputs are provided, the ASCVD Risk Estimator calculates the individual's 10-year risk of developing atherosclerotic cardiovascular disease. The risk is presented as a percentage, indicating the likelihood of experiencing a heart attack or stroke within the next 10 years.*

*- It is important to note that the ASCVD Risk Estimator is intended for use in individuals aged 40-79 years who do not have a history of atherosclerotic cardiovascular disease or stroke. The tool provides an estimate based on population-based data and should be used as a starting point for discussions between patients and healthcare providers to guide decisions regarding preventive measures and treatment.*

*- The development and validation of the ASCVD Risk Estimator involved extensive research and data analysis. The details of the methodology and the data sources used are described in the ACC/AHA guidelines on the assessment of cardiovascular risk. These guidelines provide a comprehensive overview of the risk assessment process, including the development and validation of the risk estimator.*

*Reference:*

*Goff Jr., D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., ... & Smith Jr, S. C. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology, 63(25 Part B), 2935-2959.*

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 3**: ASCVD Risk Estimator Plus (American College of Cardiology, 2016)

## 2.3 Selection of Datasets

From the perspective of datasets and databases that are available for the usage of this project, numerous datasets can be utilised. These can be retrieved from the proprietor’s source such as the Framingham Dataset, Kaggle etc. From research and investigation, numerous datasets have been identified through the use of Kaggle. These included the Heart Disease UCI dataset, CVD Dataset, CVD Risk Predication Dataset.

Framingham Dataset

*National Health and Nutrition Examination Survey (US)*

*Cleveleand Clinic Foundation*

*Heart Disease UCI Dataset via Kaggle*

*Cardiovascular Disease Dataset via Kaggle*

*CVD Risk Predication Dataset Kaggle*

*Include other datasets that were considered, and why not used.*

## 2.4 Machine Learning Modelling Principles (focus)

## 2.5 Supervised vs Unsupervised

When discussing modelling structures applied to ML or Data Scientific principles in general, is crucial to comprehend what the fundamental differences between supervised and unsupervised learning. Supervised learning and unsupervised learning are two different approaches that play fundamental and essential roles in various machine learning tasks.

To commence, supervised learning involves training a model using labelled data.

*1. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24. DOI: 10.1016/j.engappai.2006.07.002*

*2. Aha, D. W., & Bankert, R. L. (1997). A comparative evaluation of sequential feature selection algorithms. In Artificial Intelligence in Medicine (Vol. 11, No. 3, pp. 223-242). Elsevier. DOI: 10.1016/S0933-3657(97)00003-2*

*3. Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. In ACM SIGMOD Record (Vol. 25, No. 2, pp. 103-114). ACM. DOI: 10.1145/235968.233324*

*4. Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., ... & Ng, A. Y. (2012). Building high-level features using large-scale unsupervised learning. In Proceedings of the 29th International Conference on Machine Learning (ICML-12) (pp. 91-98). DOI: 10.1145/2247596.2247602*

## 2.6 Models Considered

## Decision Tree

<https://www.analyticsvidhya.com/blog/2020/10/decision-tree-introduction-with-example/>

<https://scikit-learn.org/stable/modules/tree.html>

## Bayesian Learning

<https://www.coursera.org/lecture/bayesian-methods-in-machine-learning>

<https://towardsdatascience.com/a-gentle-introduction-to-bayesian-learning-6d1a9a974b69>

## SVM

<https://scikit-learn.org/stable/modules/svm.html>

<https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python>

## Clustering

<https://scikit-learn.org/stable/modules/clustering.html>

<https://towardsdatascience.com/introduction-to-clustering-algorithms-in-machine-learning-80e18d639b59>

## Stacking

<https://www.kaggle.com/general/18793>

<https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/>

// mathematical representation for formulas

// use Google Scholars

## 2.8 Evaluation Metrics

<https://scikit-learn.org/stable/modules/model_evaluation.html>

<https://towardsdatascience.com/a-gentle-introduction-to-machine-learning-evaluation-metrics-bc88d7226a21>

<https://www.kdnuggets.com/2020/05/evaluation-metrics-classification-regression.html>

* What models can be used?
  + Pros and Cons
* Methods of combining models
* Include looking at methods of measurement.

# 3 Methodology

The methodology section of this study aims to provide a comprehensive understanding of the selected dataset through Exploratory Data Analysis (EDA) and adhering to the CRISP-DM (Cross-Industry Standard Process for Data Mining) cycle. In addition to understanding the dataset, our goals also include applying modelling techniques and evaluating the performance of the developed models.

Following the CRISP-DM cycle, we will incorporate various phases to ensure a systematic and rigorous approach to data analysis. While 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.

## 3.1 Data Exploration and Extraction

## 3.2 Data Preparation

|  |
| --- |
| ***df = df.fillna(df.mean())*** |

Listing 1: Filling Missing Values with Column Names

## 3.3 Data Comprehension

A screenshot of a graph

Description automatically generated with low confidence

**Figure 4**: Correlation Heatmap: Framingham Dataset (Pre-Filtration)

A screenshot of a graph

Description automatically generated with low confidence

**Figure 5.1**: Heatmap Correlation for Columns 31-38

A screenshot of a graph

Description automatically generated with low confidence

Figure 6: Correlation Heatmap: Post Filtration

## 3.4 Data Modelling

## 3.5 Data Evaluation

* EDA and understanding the dataset.
* Correlation of data, and understanding what is what?
* Feature Extraction
* Data Collection and Pre-Processing
* Data Mining Platform

# 4 Results + Testing

# 5 Discussion

# 6 Conclusion & Future Work

# Reflections

# References

Abraham, A., 2013. Comparison of Supervised and Unsupervised Learning. *Algorithms for Pattern Classification,* Volume 2.

American College of Cardiology, 2016. *Project Risk Reduction by Therapy.* [Online]   
Available at: https://tools.acc.org/ascvd-risk-estimator-plus/#!/calculate/estimate/  
[Accessed 6 June 2023].

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

British Heart Foundation, 2023. *UK Factsheet.* [Online]   
Available at: https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-uk-factsheet.pdf?rev=e771367bf0654a4dae85cbc9dbefae17&hash=76C0182379BB6EE118EC6F76FA35A158#:~:text=It%20was%20also%20the%20leading%20cause%20of%20death%20  
[Accessed 23 May 2023].

De Filippis, A. P. et al., 2011. Journal of the American College of Cardiology. *Journal of the American College of Cardiology,* 58(20), pp. 2076-2083.

Kawada, T., 2023. Reynolds Risk Score as a Risk Assessment Tool for Cardiovascular Disease After 10 Years: Its Strong Relationship with Blood Presssure. *The Journal of Clinical Hypertension,* 14(8), pp. 571-572.

Klisić, A., 2018. 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), pp. 20-30.

Lehr, J. et al., 2021. Supervised learning vs. unsupervised learning: A comparison for optical inspection applications in quality control. *IOP Conference Series: Materials Science and Engineering,* 1140(1).

Lloyd-Jones, D., 2011. Concepts of screening for cardiovascular risk factors and disease. *Preventive Cardiology: Companion to Braunwald's Heart Disease,* pp. 433-442.

Nottingham, U. o., 2018. *Welcome to the QRISK®3-2018 risk calculator https://qrisk.org.* [Online]   
Available at: https://qrisk.org  
[Accessed 6 June 2023].

Ridker, D. P. M., 2018. *Reynolds Risk Score.* [Online]   
Available at: http://www.reynoldsriskscore.org  
[Accessed 6 June 2023].

World Health Organisation , 2021. *Cardiovascular diseases.* [Online]   
Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab\_1  
[Accessed 23 May 2023].

**Appendices**

**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 from the article developed and written by Tomoyuki Kawada MD, PhD. (Kawada, 2023)

A picture containing text, screenshot, font, number

Description automatically generated