DEEP LEARNING BASED CARDIOVASCULAR DISEASE RISK PREDICTION USING A MULTI-LAYER PERCEPTRON (MLP)

By

Benard Adala Wanyande

A Project Proposal Submitted as a Partial Requirement for the Award of a Degree in Bachelor of Science in Informatics and Computer Science

School of Computing and Engineering Sciences

Strathmore University

Supervisor: Stephen Obonyo

August 2023

Declaration

I, Benard Adala Wanyande, declare that this project has not been submitted to any other university for the award of a Degree in Informatics and Computer Sciences, is my own original work and has not been submitted to any other institution of higher learning. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references.

SUPERVISOR

Name: Stephen Obonyo

Signature:

Date 18/08/2023

AUTHOR

Name: Benard Adala Wanyande

Signature:

Date: 18/08/2023

Abstract

This project focuses on predicting cardiovascular disease (CVD) using advanced Machine Learning (ML). CVD is a leading global cause of mortality, demanding innovative diagnostic approaches. The project proposes an Artificial Neural Network (ANN) model for enhanced accuracy, overcoming limitations of traditional methods like ECG. The ANN model surpasses conventional approaches, achieving significant accuracy improvements. Through feature selection and ANN-based training, the model's performance is evaluated using various metrics. The objective is to establish a robust ML model for swift and reliable CVD risk predictions, ushering in a new era of predictive healthcare.

Keywords: Cardiovascular Disease, Machine Learning, Artificial Neural Networks, Prediction, IoT, Accuracy, Diagnosis, Performance Metrics.

Table of Contents

List of Figures

List of Tables

Abbreviations

CVD – Cardiovascular Disease

ML – Machine Learning

ANN – Artificial Neural Network

ACC/AHA - American College of Cardiology and the American Heart Association

kNN – k-Nearest Neighbour

AUC-ROC - Area Under the Receiver Operating Characteristic Curve

FNN – Feedforward Neural Network

MLP – Multi-Layer Perceptron

CNN – Convolutional Neural Network

EHR – Electronic Health Record

RETAIN - Reverse Time Attention

BRFSS – Behavioural Risk Factor Surveillance System

1 Chapter 1: Introduction

1.1 Background

The escalating prevalence of cardiovascular diseases (CVDs) presents a formidable global health challenge, accounting for millions of deaths annually. A significant majority of these fatalities are attributed to heart attacks and strokes (World Health Organization, 2021). Low- and middle-income countries are disproportionately burdened by CVD-related deaths, underscoring the urgent need for accurate diagnostic tools. This project serves as a convergence point for medical research, machine learning, and healthcare innovation to address the growing demand for reliable predictive methods (National Health Service, 2022).

Given the high mortality rates associated with CVDs, the importance of early detection and personalised strategies cannot be overstated. To address this, the project capitalises on the potential of Deep Learning, particularly through the utilisation of Artificial Neural Networks (ANNs), to develop a robust predictive model for assessing CVD risk. By bridging the worlds of technology and healthcare, the project endeavours to empower informed decision-making and proactive health management (Schiller, 2015).

The projected outcomes of this undertaking are far-reaching. They encompass elevated accuracy in risk assessment and the potential to drive AI-driven innovation in healthcare (Bajwa et al., 2021). The project's relevance extends to a wide audience including healthcare professionals, researchers, data scientists, and policymakers. For healthcare practitioners, the project offers a tool that complements clinical expertise with data-derived insights, enabling more precise and tailored patient recommendations. Researchers can expand their understanding of CVD risk factors, potentially uncovering novel insights into the disease's mechanisms. Data scientists can explore advanced machine learning techniques in a healthcare context, while policymakers can consider the implications of the project's outcomes for public health initiatives.

Beyond its immediate application to CVDs, the project's collaboration between medical and technological realms demonstrates the potential synergy between these sectors, showcasing innovative strategies to tackle complex health challenges. The transformative nature of deep learning-based predictive models could potentially set a precedent for AI's application in other medical domains, contributing to a broader shift towards data-driven and patient-centric healthcare.

This project has a dual aim: to harness the capabilities of Deep Learning and Artificial Neural Networks to revolutionise the prediction and prevention of cardiovascular diseases. By ingesting patient-specific data and leveraging machine learning, the project aspires to craft a predictive model that empowers healthcare professionals, patients, and policymakers alike. Through this amalgamation of cutting-edge technology and critical healthcare needs, the project sets the stage for redefining the landscape of medical diagnostics, ultimately leading to enhanced patient outcomes and reduced CVD-related morbidity and mortality.

1.2 Problem Statement

Cardiovascular Disease (CVD) poses a global challenge with significant morbidity and mortality, necessitating accurate risk prediction methods. Current diagnostic techniques lack the precision needed for timely interventions, impacting patients, healthcare providers, and systems. Traditional methods like ECGs and conventional machine learning fall short in capturing the complexity of CVDs causative factors. To address this, integrating cutting-edge Deep Learning through Artificial Neural Networks (ANNs) offers a solution. ANNs can uncover intricate risk factor relationships, enabling precise predictions for early interventions and personalised strategies. Given CVD's high mortality rates and the limitations of existing methods, this approach could transform risk prediction and revolutionise CVD management and prevention on a global scale.

1.3 Objectives

This section delineates the overarching goal, specific objectives, and research questions that the proposed study seeks to investigate and address.

1.3.1 General Objective

The general objective of this project is to develop an accurate and reliable machine learning-based model for cardiovascular disease (CVD) risk prediction.

1.3.2 Specific Objectives

i. To review current CVD risk prediction methods.

ii. To analyse existing machine learning techniques used for CVD risk prediction methods.

iii. To develop and implement an Multi-Layer Perceptron (MLP) Model.

iv. To test and validate the developed MLP model.

1.3.3 Research Questions

i. What are the traditional methods currently used for CVD risk prediction, including ECG-based diagnostics and conventional machine learning models?

ii. What are the specific machine learning algorithms commonly employed for CVD risk prediction?

iii. How can an MLP model be specifically tailored to predict CVD risk?

iv. How well does the developed MLP model perform when tested on a Cardiovascular Disease Risk Prediction dataset?

1.4 Justification

Cardiovascular Disease (CVD) presents a pressing global healthcare challenge, demanding improved risk prediction methods due to existing gaps and the potential of Machine Learning (ML). Current diagnostics like ECGs and traditional ML struggle with intricate risk factors, necessitating innovative, precise solutions. Machine Learning, especially Deep Learning, can exploit extensive health data to offer unprecedented risk insights, enhancing preventive care. This research leverages Multi-Layer Perceptron (MLP) to bridge gaps and achieve up to 91-92% accuracy (Pal et al., 2022), aligning with personalised medicine and cost-effective healthcare. This project resonates with healthcare goals, aiding patient outcomes and CVD management.

1.5 Project Scope

This project focuses on developing a Machine Learning (ML) model for precise cardiovascular disease (CVD) risk prediction, emphasising advanced techniques to enhance accuracy. It encompasses data preparation, feature exploration, and creation of an optimised Multi-Layer Perceptron (MLP) to capture intricate risk factors. Rigorous evaluation, including metric comparison, and real-world deployment testing for CVD risk predictions are vital components. The project's goal is to optimise accuracy and contribute to effective CVD risk management.

2 Chapter 2: Literature Review

2.1 Introduction

This chapter serves as a comprehensive exploration of the intricate landscape surrounding the prediction of cardiovascular disease (CVD) risk. By delving into the aetiology, transmission, symptoms, and existing prediction methods of CVD, this chapter lays the foundation for understanding the complex interplay of factors that contribute to cardiovascular health. Additionally, it examines the current employment of Machine Learning (ML) techniques in predicting CVD risk. Through a critical evaluation of these ML techniques for CVD risk prediction and a focus on the capabilities of Multi-Layer Perceptrons (MLPs), this chapter aims to provide a clearer understanding of the nature of the project.

2.2 Aetiology of Respiratory Conditions

According to the World Health Organization (WHO), the most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol (World Health Organization, 2021). Cardiovascular Disease (CVD) is a complex condition with multifaceted origins influenced by a range of interconnected factors. While the exact cause remains elusive, the presence of various risk factors significantly contributes to an individual's likelihood of developing CVD (National Health Service, 2022). These risk factors collectively shape the aetiology of CVD, and their interplay underscores the importance of comprehensive prevention and intervention strategies.

2.2.1 Genetic and Environmental Factors

According to Hajar (2020) genetic predisposition plays a fundamental role in CVD aetiology, with hereditary traits influencing susceptibility to conditions such as hypertension, hyperlipidaemia, and heart defects. However, genetics do not act in isolation. Environmental factors interact with genetic vulnerabilities, accentuating the risk. Exposure to pollutants, toxins, and poor air quality can amplify the impact of genetic factors, leading to the progression of CVD. Genetic variations might also affect how an individual metabolises certain substances, potentially exacerbating CVD risk.

A family history of CVD further contributes to an individual's risk profile. Individuals with parents or siblings diagnosed with CVD at an earlier age are considered at increased risk. Genetic predisposition inherited from family members can exacerbate the impact of other risk factors, compounding the likelihood of CVD development.

2.2.2 Life Choices and Behaviour

Tran et al. (2022) conducted a study exploring the relationship between lifestyle choices and the risk of cardiovascular disease and found that the choices individuals make in their daily lives profoundly influence CVD risk. Smoking and other forms of tobacco use contribute significantly to CVD development. Harmful substances in tobacco damage and narrow blood vessels, increasing the likelihood of atherosclerosis and related complications. Unhealthy diets high in saturated fats, sugars, and processed foods contribute to obesity, diabetes, and high cholesterol levels, all major CVD risk factors. Physical inactivity further compounds the risk, as it often leads to high blood pressure, elevated cholesterol levels, and excess weight.

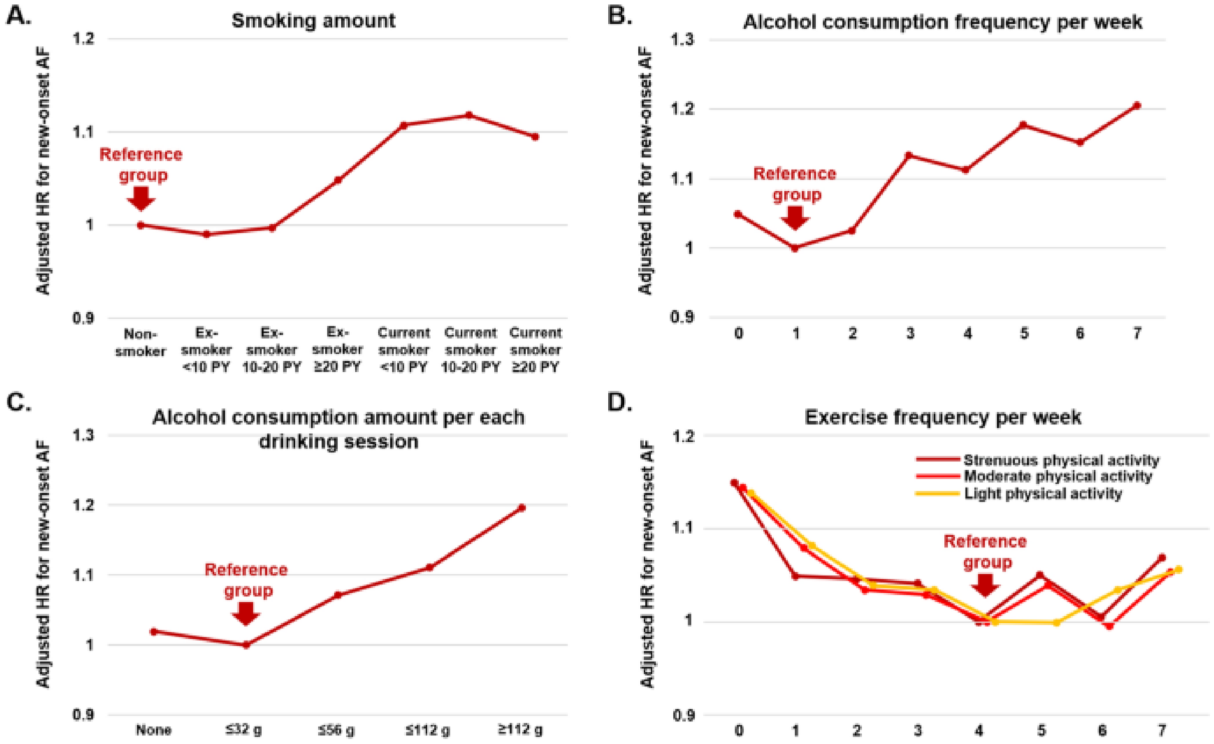


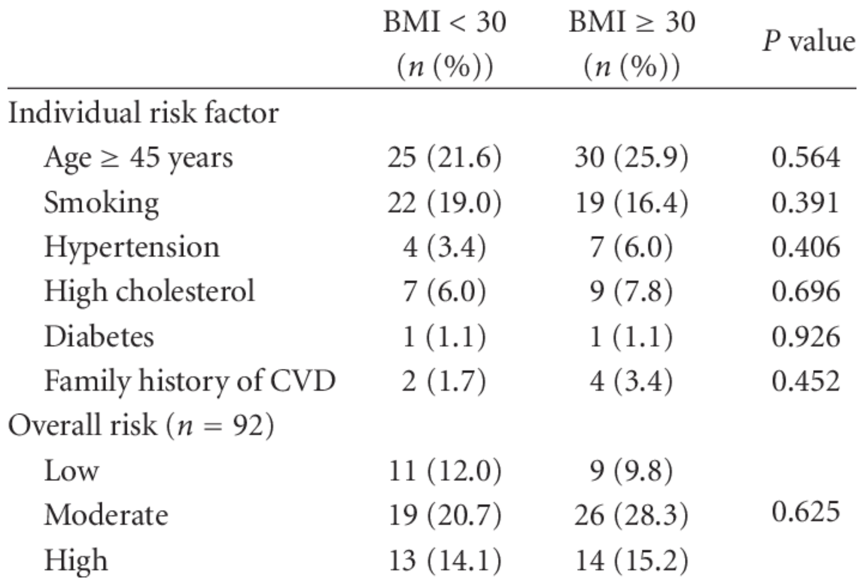
Figure 2. 1 Image illustrating the findings from Lee et al.'s (2020) study on the association between clustering of unhealthy lifestyle factors and the risk of new-onset atrial fibrillation.

2.2.3 Metabolic and Chronic Conditions

Rao et al. (2014) conducted a study exploring the relationship between metabolic syndrome and chronic disease and found that metabolic disorders, particularly diabetes, play a crucial role in CVD aetiology. High blood sugar levels associated with diabetes can damage blood vessels, leading to atherosclerosis and an increased risk of cardiovascular events. Moreover, many individuals with type 2 diabetes are also overweight or obese, further amplifying their susceptibility to CVD. Additionally, stress, often linked to modern lifestyles, contributes to the aetiology by triggering hormonal and inflammatory responses that impact the cardiovascular system.

2.2.4 Weight and Body Composition

Hu et al. (2023) conducted a longitudinal community-based study investigating the relationship between two-year changes in body composition and future cardiovascular events. They found that body weight and composition significantly influence CVD risk. Being overweight or obese increases the likelihood of developing diabetes and high blood pressure, both of which are primary risk factors for CVD. Body Mass Index (BMI) serves as an indicator, with a BMI of 25 or above considered an elevated risk. Central obesity, characterised by excess abdominal fat, is particularly relevant, as it is associated with increased CVD risk, especially among men with a waist measurement of 94 cm or more and women with a waist measurement of 80 cm or more.



	BMI < 30 (n (%))	BMI ≥ 30 (n (%))	P value
Individual risk factor			
Age ≥ 45 years	25 (21.6)	30 (25.9)	0.564
Smoking	22 (19.0)	19 (16.4)	0.391
Hypertension	4 (3.4)	7 (6.0)	0.406
High cholesterol	7 (6.0)	9 (7.8)	0.696
Diabetes	1 (1.1)	1 (1.1)	0.926
Family history of CVD	2 (1.7)	4 (3.4)	0.452
Overall risk (n = 92)			
Low	11 (12.0)	9 (9.8)	
Moderate	19 (20.7)	26 (28.3)	0.625
High	13 (14.1)	14 (15.2)	

Table 1 Tabular representation summarising the findings of Hu et al.'s (2023) longitudinal community-based study.

2.2.5 Ethnic Background and Health Disparities

Ethnic background plays a significant role in CVD aetiology, with people of certain backgrounds being more susceptible due to genetic and cultural factors. Individuals of South Asian and Black African or African Caribbean descent are particularly at risk. These populations often exhibit a higher prevalence of other CVD risk factors such as high blood pressure and type 2 diabetes, further amplifying their overall susceptibility.

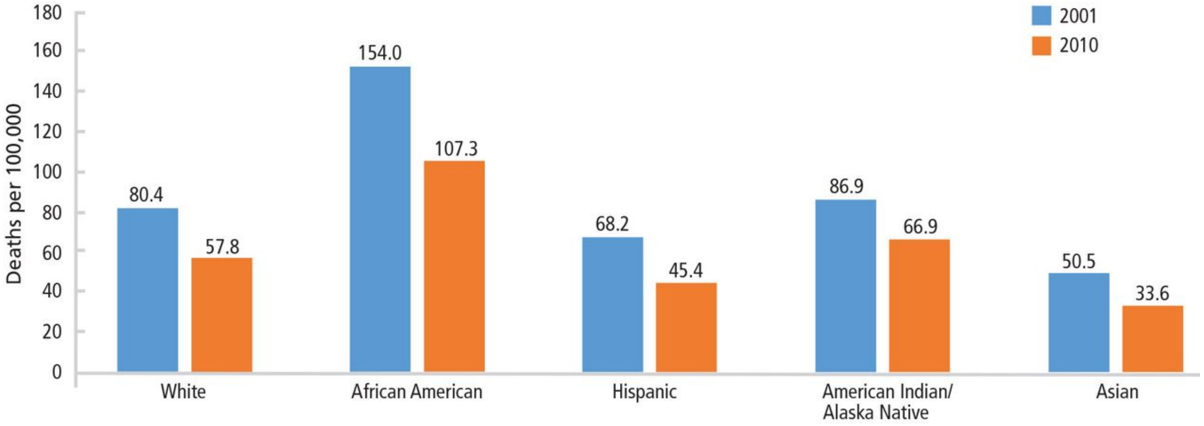


Figure 2. 2 Visual representation illustrating the disparities in cardiovascular disease (CVD) risk among various racial groups, as discussed by Youmans et al. (2019)

2.2.6 Underlying Determinants and Social Factors

Powell-Wiley et al. (2022) explored the impact of social determinants on cardiovascular disease in their study published in Circulation Research. They found that beyond individual risk factors, the underlying determinants of CVD are driven by broader societal forces such as globalisation, urbanisation, population ageing, poverty, and stress. These factors create an environment that either fosters or hinders healthy lifestyle choices. Access to healthcare, economic opportunities, and education also play a pivotal role in shaping an individual's risk profile for CVD.

2.3 Transmission of Respiratory Conditions

According to Burlaton, Gourbat, and Seigneuric (1991), the transmission of cardiovascular disease (CVD) risk factors can occur through direct genetic and hereditary factors such as sex, race, antecedents, dyslipidemia, high blood pressure, and diabetes. However, the article also highlights that the transmission of CVD risk factors can extend beyond genetics to encompass the transmission of a particular lifestyle. The manifestation and progression of cardiovascular disease (CVD) are intricate processes driven by a cascade of biological events that stem from the interplay of various risk factors. These factors interact synergistically, creating a conducive environment for the initiation and progression of cardiovascular conditions.

2.4 Symptoms of Respiratory Conditions

According to the World Health Organization (2021), cardiovascular disease (CVD) spans a spectrum of symptoms and conditions, from acute chest pain and angina to chronic fatigue and heart failure. Acute symptoms include chest pain, shortness of breath, and cold sweat during a heart attack, while strokes can cause sudden weakness, speech difficulty, and confusion. Chronic CVD conditions like heart failure lead to fatigue, breathlessness, and fluid retention, while Peripheral Arterial Disease (PAD) results in leg pain and ulcers.

Importantly, gender-specific differences in symptom presentation exist. Women may experience atypical CVD symptoms such as nausea and back pain (World Health Organization, 2021). Awareness of this diversity is crucial for timely diagnosis and intervention. Overall, understanding the range of CVD symptoms is essential for recognizing warning signs and enabling effective medical management (National Health Service, 2022).

2.5 Current Techniques Used for CVD Risk Prediction

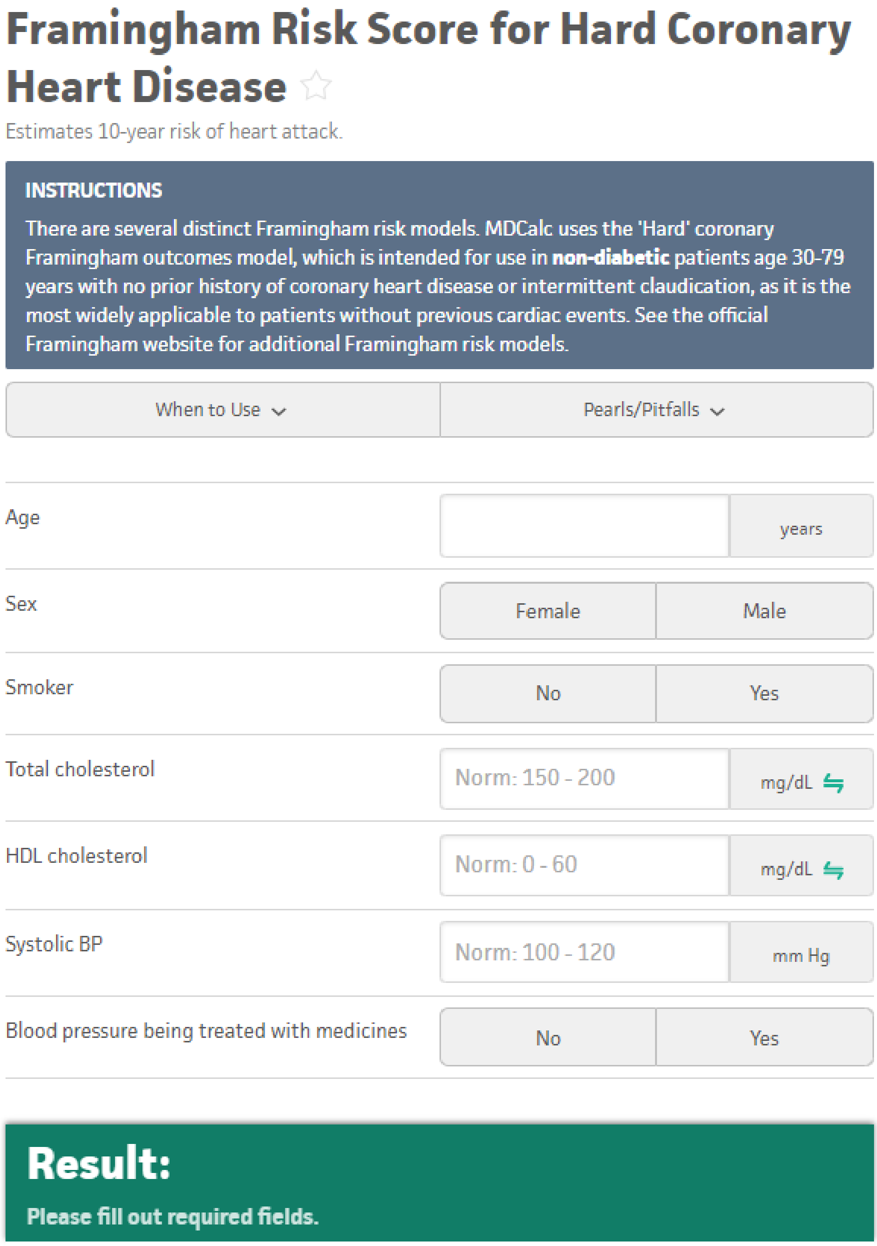
2.5.1 Framingham Risk Score

The Framingham Risk Score, also known as the Framingham Cardiovascular Disease (CVD) Risk Score, is a widely used tool for assessing an individual's 10-year risk of developing cardiovascular diseases, particularly coronary heart disease (CHD) and stroke. It was developed based on a long-term study called the Framingham Heart Study, which began in 1948 and involved the residents of Framingham, Massachusetts. The study aimed to identify risk factors associated with cardiovascular diseases and develop predictive models to estimate an individual's likelihood of experiencing such events.

The Framingham Risk Score considers several risk factors that have been identified through the Framingham Heart Study and subsequent research as being strongly associated with the development of cardiovascular diseases. These risk factors include age, gender, total cholesterol, HDL cholesterol, blood pressure, smoking, diabetes, and body mass index (BMI) (Jahangiry, Farhangi, & Rezaei, 2017).

Using these risk factors, the Framingham Risk Score calculates an individual's 10-year risk of developing a cardiovascular event. The calculated risk percentage provides an estimate of the likelihood of experiencing a heart attack or stroke within the next 10 years. Healthcare professionals use this score as a guideline to inform treatment and preventive strategies. Generally, the higher the risk score, the more aggressive the interventions might be, such as lifestyle changes (diet, exercise, smoking cessation), medication, or other medical interventions.

It's important to note that the Framingham Risk Score has limitations. It was developed based on data from a specific population (Framingham residents), and while it has been validated in various populations, there may be differences in risk factors and disease prevalence that could affect its accuracy in different demographic groups. Additionally, there are other risk assessment tools available that consider a broader range of risk factors, including family history, ethnicity, and more advanced blood tests.



Framingham Risk Score for Hard Coronary Heart Disease ☆

Estimates 10-year risk of heart attack.

INSTRUCTIONS

There are several distinct Framingham risk models. MDCalc uses the 'Hard' coronary Framingham outcomes model, which is intended for use in **non-diabetic** patients age 30-79 years with no prior history of coronary heart disease or intermittent claudication, as it is the most widely applicable to patients without previous cardiac events. See the official Framingham website for additional Framingham risk models.

Age		years
Sex	Female	Male
Smoker	No	Yes
Total cholesterol	Norm: 150 - 200	mg/dL ⇆
HDL cholesterol	Norm: 0 - 60	mg/dL ⇆
Systolic BP	Norm: 100 - 120	mm Hg
Blood pressure being treated with medicines	No	Yes

Result:

Please fill out required fields.

Figure 2. 3 Screenshot illustrating the user interface of a web app demonstrating the application of the Framingham Risk Score

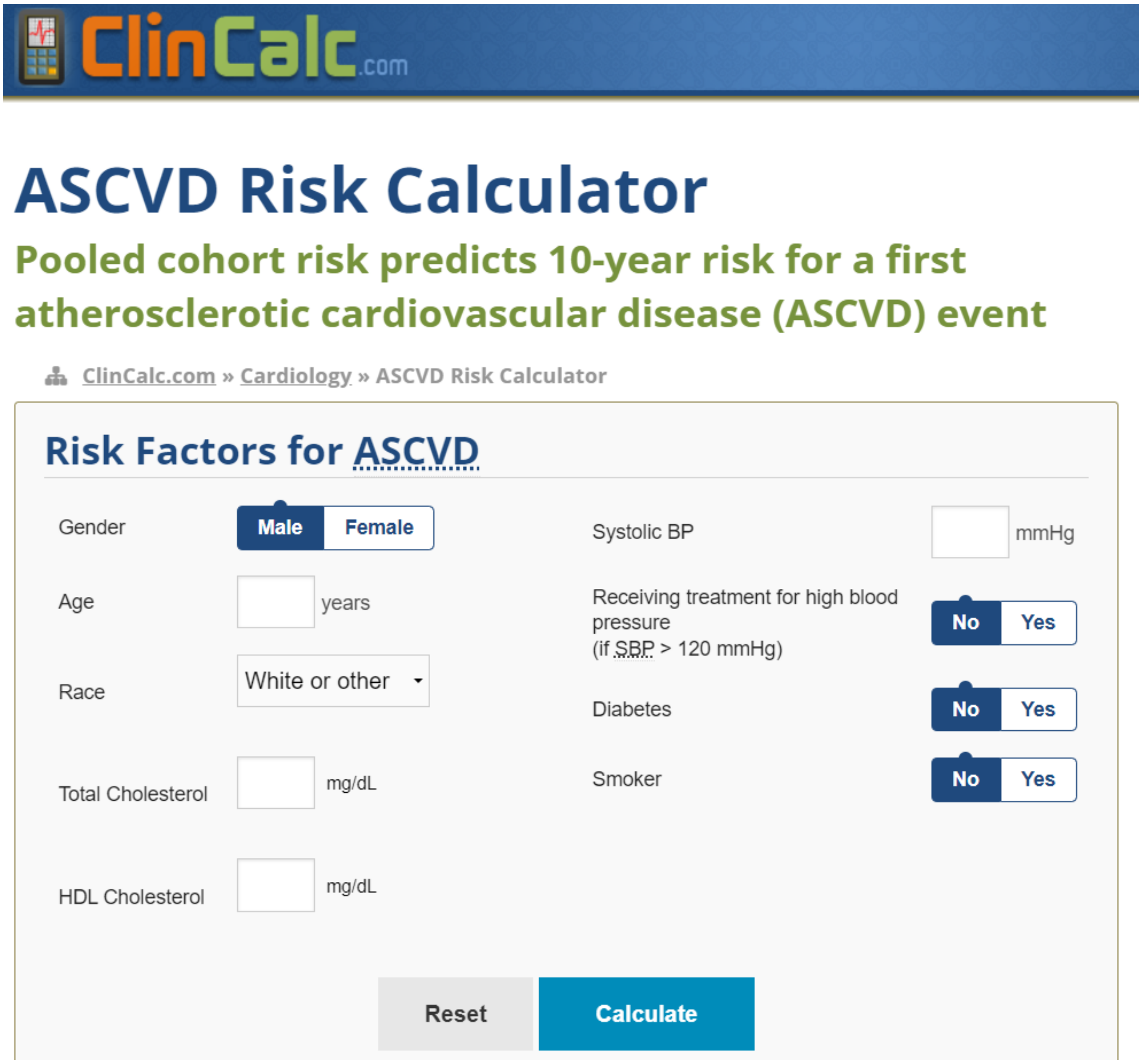
Users can input relevant data to calculate their 10-year cardiovascular risk based on the Framingham Risk Score algorithm.

2.5.2 ACC/AHA Pooled Cohort Equations

The ACC/AHA Pooled Cohort Equations, or Pooled Cohort Equations, are widely used cardiovascular disease (CVD) risk assessment tools developed by the American College of Cardiology (ACC) and the American Heart Association (AHA). They estimate an individual's 10-year risk of atherosclerotic cardiovascular events, such as heart attacks and strokes (Campos-Staffico et al., 2021). Introduced in 2013, these equations build upon the Framingham Risk Score, considering a broader array of risk factors including age, gender, race, cholesterol levels, blood pressure, and diabetes status.

These equations serve as a guide for healthcare professionals to make informed decisions about interventions and preventive strategies. They estimate an individual's likelihood of experiencing a first-time ASCVD event within a decade. Like the Framingham Risk Score, the Pooled Cohort Equations provide a framework for discussions on lifestyle changes, medication, and interventions. However, the Pooled Cohort Equations are not without limitations. They are most applicable to those aged 40 to 79 without previous ASCVD events, potentially less accurate for certain demographic groups, and do not cover all possible risk factors.

As part of a comprehensive assessment, healthcare professionals use the Pooled Cohort Equations alongside an individual's medical history and overall health to tailor their approach to cardiovascular risk management.


ClinCalc.com

ASCVD Risk Calculator

Pooled cohort risk predicts 10-year risk for a first atherosclerotic cardiovascular disease (ASCVD) event

ClinCalc.com » Cardiology » ASCVD Risk Calculator

Risk Factors for ASCVD

Gender	[✓ Male] Female	Systolic BP	mmHg
Age	years	Receiving treatment for high blood pressure (if SBP > 120 mmHg)	[✓ No] Yes
Race	White or other	Diabetes	[✓ No] Yes
Total Cholesterol	mg/dL	Smoker	[✓ No] Yes
HDL Cholesterol	mg/dL		

Figure 2. 4 Screenshot illustrating the interface of a web app demonstrating the practical application of the ACC/AHA Pooled Cohort Equations.

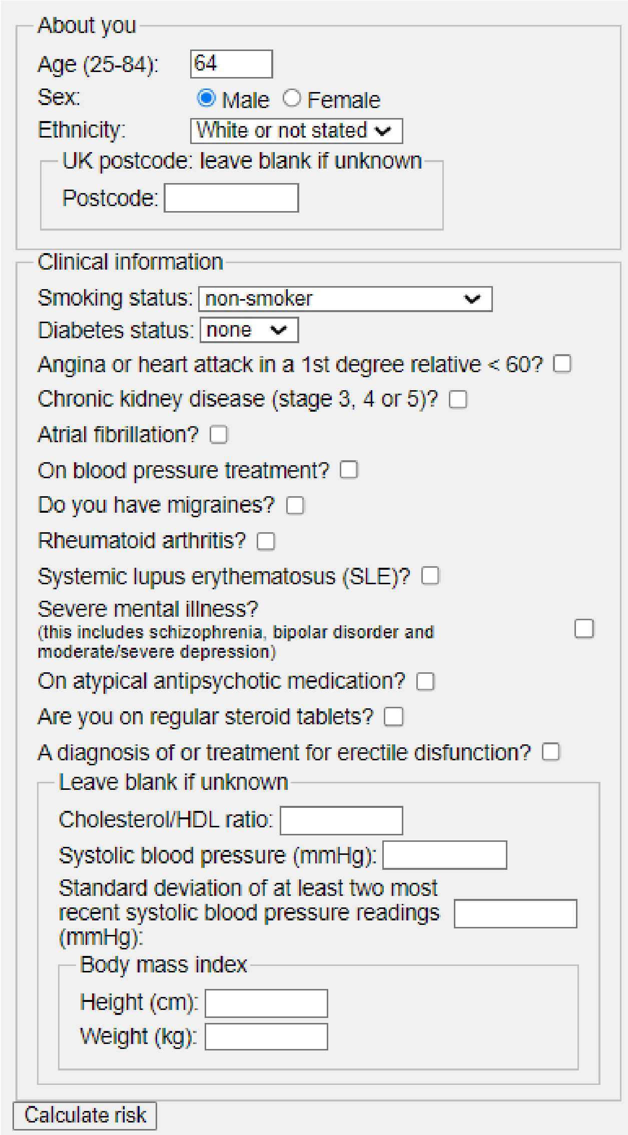
Users input relevant data to estimate their 10-year risk of atherosclerotic cardiovascular events, providing insights into heart disease and stroke risk.

2.5.3 QRISK

QRISK is a cardiovascular disease (CVD) risk assessment tool that is widely used in the United Kingdom. It was developed to estimate an individual's 10-year risk of experiencing a cardiovascular event, including heart attacks and strokes (Hippisley-Cox et al., 2007). QRISK is unique in that it considers a broader range of risk factors, including clinical, lifestyle, and socioeconomic variables, making it a comprehensive tool for risk assessment.

QRISK has been praised for its inclusive approach to risk assessment by considering a wide range of risk factors, including socioeconomic factors and ethnicity. This can lead to more tailored and accurate risk assessments for a diverse population. However, it's important to note that QRISK, like any risk assessment tool, has limitations and should be used in conjunction with clinical judgement. It's advisable for healthcare professionals to evaluate an individual's overall health, medical history, and specific circumstances before making treatment recommendations based on QRISK scores.

Users input a variety of factors, including clinical and lifestyle variables, to estimate their 10-year risk of cardiovascular events.



About you

Age (25-84): 64

Sex: ● Male ○ Female

Ethnicity: White or not stated

UK postcode: leave blank if unknown

Postcode:

Clinical information

Smoking status: non-smoker

Diabetes status: none

Angina or heart attack in a 1st degree relative < 60? □

Chronic kidney disease (stage 3, 4 or 5)? □

Atrial fibrillation? □

On blood pressure treatment? □

Do you have migraines? □

Rheumatoid arthritis? □

Systemic lupus erythematosus (SLE)? □

Severe mental illness? (this includes schizophrenia, bipolar disorder and moderate/severe depression) □

On atypical antipsychotic medication? □

Are you on regular steroid tablets? □

A diagnosis of or treatment for erectile disfunction? □

Leave blank if unknown

Cholesterol/HDL ratio:

Systolic blood pressure (mmHg):

Standard deviation of at least two most recent systolic blood pressure readings (mmHg):

Body mass index

Height (cm):

Weight (kg):

Figure 2. 5 Screenshot showcasing the user interface of a web app demonstrating the practical application of the QRISK cardiovascular risk assessment tool.

2.6 Current Machine Learning Techniques Used for CVD Risk Prediction

2.6.1 kNN

k-Nearest Neighbours (kNN) is a simple, yet effective machine learning algorithm used for classification and regression tasks. In the context of classification, the algorithm assigns a new data point to a class based on the majority class among its k nearest neighbours in the training dataset. In the context of regression, kNN predicts a continuous value by averaging or taking the weighted average of the target values of its k nearest neighbours. The choice of k, the number of neighbours to consider, is a crucial parameter that influences the algorithm's performance.

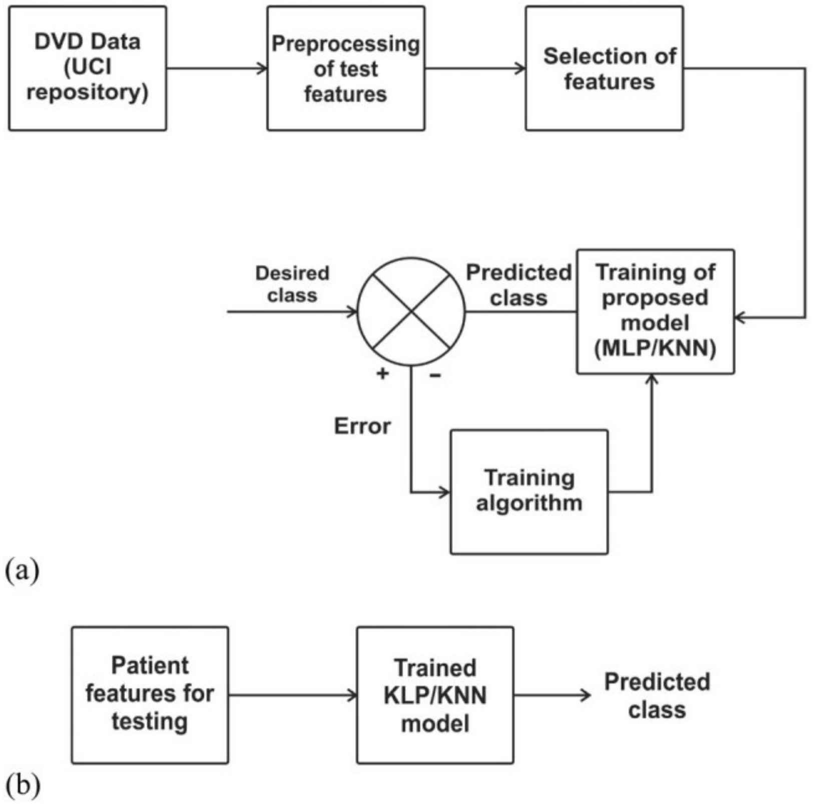

DVD Data (UCI repository)
Preprocessing of test features
Selection of features
Desired class
Predicted class
Training of proposed model (MLP/KNN)
+
-
Error
Training algorithm
(a)
Patient features for testing
Trained KLP/KNN model
Predicted class
(b)

Figure 2. 6: Methodology employed in developing a kNN CVD detection model.

(a) Depicts the training phase of machine learning models, where algorithms learn patterns from a dataset. (b) Represents the testing phase, where the trained models are evaluated for their performance in predicting CVD risk.

K-nearest neighbours (kNN) stands as an influential machine learning algorithm in the realm of data analysis and classification. In the context of predicting cardiovascular disease (CVD) risk, kNN operates on the principle of similarity. Just as people with similar lifestyles and health attributes tend to share common risk profiles, kNN leverages this concept to make predictions based on the proximity of an individual's data to known instances within a training dataset (Pal et al., 2022). For instance, consider an individual who presents certain health markers and attributes resembling those of known CVD cases in the dataset. The algorithm identifies the k-nearest data points – those with the closest resemblance – and extrapolates the individual's risk of experiencing a cardiovascular event based on the outcomes of these neighbours.

The operational mechanics of kNN involve selecting a value for k (the number of nearest neighbours to consider) and a distance metric, often Euclidean distance, to measure the proximity between data points. For instance, if k is set to 5, the algorithm would identify the five most similar cases to the target individual. If three of these neighbours have experienced CVD events, while two have not, the algorithm would predict a higher risk for the target individual, reflecting the majority class among the neighbours. However, while kNN holds promise in capturing nuanced patterns that might evade conventional analysis, it's important to note that its performance hinges on the quality of data, the choice of k, and the distance metric employed (Pal et al., 2022).

In a study conducted by Pal et al. (2022), kNN is investigated alongside other machine learning techniques for CVD detection. The study employs publicly available University of California Irvine repository data, where kNN's ability to assess an individual's risk of cardiovascular events is evaluated. Through identifying the closest neighbours within the dataset, kNN can predict an individual's likelihood of developing CVD based on similarities with known cases. However, the study ultimately finds that the multi-layer perceptron (MLP) model outperforms kNN in accuracy and predictive performance for CVD risk assessment (Pal et al., 2022).

2.6.2 Random Forests

Random Forests is a powerful ensemble learning algorithm widely used in machine learning for classification and regression tasks. It's based on the concept of creating multiple decision trees and combining their predictions to improve accuracy and reduce overfitting. Each decision tree in a random forest is constructed using a subset of the training data and a subset of features chosen randomly. The algorithm aggregates the predictions of these individual trees to make a final prediction, resulting in a more robust and accurate model.

Random Forests can be employed for CVD risk prediction by utilising patient data and risk factors to classify individuals into different risk categories or predict their risk scores. Patient data, including demographic information, medical history, blood pressure, cholesterol levels, and lifestyle factors, can be used as input features. The algorithm learns the relationships between these features and the likelihood of CVD events from historical patient data with known outcomes.

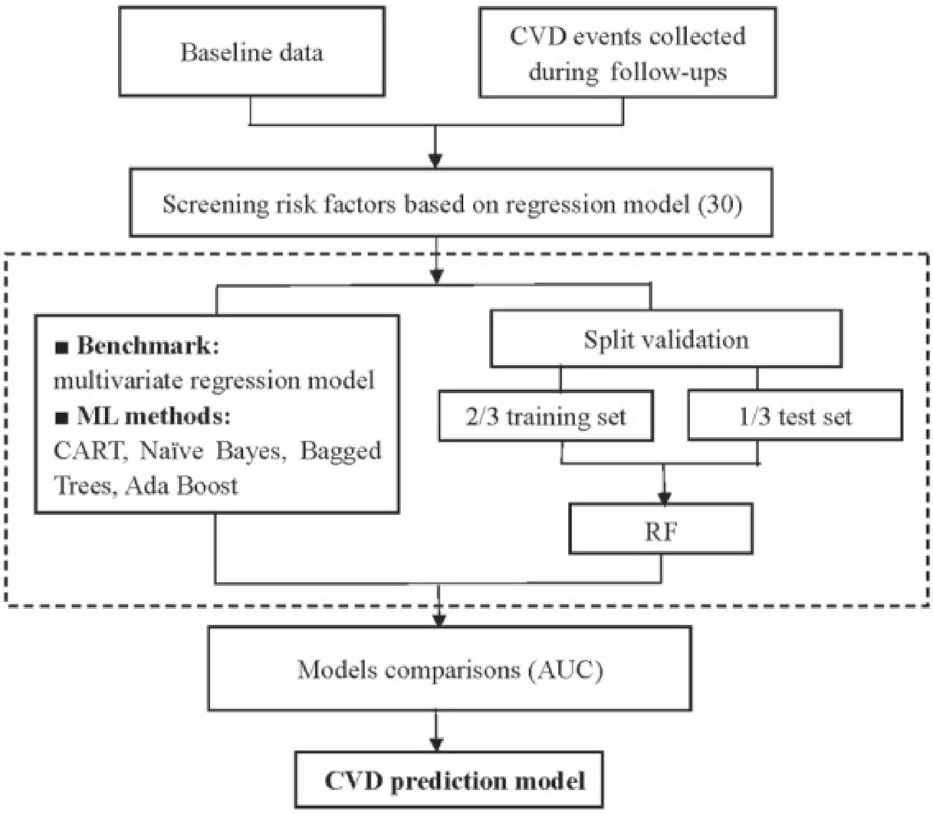

Baseline data
CVD events collected during follow-ups
Screening risk factors based on regression model (30)
■ Benchmark:
multivariate regression model
■ ML methods:
CART, Naïve Bayes, Bagged Trees, Ada Boost
Split validation
2/3 training set
1/3 test set
RF
Models comparisons (AUC)
CVD prediction model

Figure 2. 7 Schematic illustration representing the cardiovascular disease (CVD) Random Forests prediction model.

The diagram outlines the conceptual framework and components of the model utilised for predicting CVD risk.

Random Forests are particularly suited for CVD risk prediction due to their ability to handle complex interactions between risk factors and non-linear relationships. They can capture feature interactions that might be missed by traditional linear models. The algorithm also performs well with noisy and incomplete data, which is common in healthcare datasets. To use Random Forests effectively, data preprocessing and feature selection are important. Additionally, hyperparameter tuning can optimise the algorithm's performance by adjusting parameters such as the number of trees in the forest and the maximum depth of each tree.

One significant benefit of Random Forests is their ability to provide insights into feature importance. This can help clinicians and researchers identify which risk factors play the most significant role in CVD risk prediction. Random Forests are also robust against overfitting, as the ensemble nature of the algorithm reduces variance. However, they can be computationally intensive for large datasets and might not be as interpretable as simpler models. They also require proper tuning to optimise their performance, and the results might vary based on the selected hyperparameters. Overall, Random Forests offer a versatile and accurate approach for CVD risk prediction, capturing complex relationships and improving accuracy compared to individual decision trees or linear models.

2.6.3 Naïve Bayes

Naive Bayes, a well-established machine learning approach, has found applications beyond traditional classification tasks, including its utilisation in risk prediction involving censored time-to-event data, as exemplified in the study by Wolfson et al. (2015). This approach serves as a valuable tool for assessing the likelihood of future clinical outcomes, with implications for clinical decision-making and public health strategies. While conventional longitudinal cohort studies have been employed for risk model construction, modern electronic health records (EHRs) present an alternative data source that offers extensive patient information and sample sizes. Nonetheless, the complexity of EHR data, characterised by missing values and incomplete follow-up, necessitates innovative statistical techniques to unveil intricate associations and interactions within the data.

The Naive Bayes technique, as adapted by Wolfson et al. (2015), showcases its versatility in handling time-to-event outcomes subject to censoring. It leverages a probabilistic framework to estimate the likelihood of specific outcomes, considering the interaction of covariates. In comparison, the commonly used Cox proportional hazards model is a well-established method for risk prediction in healthcare populations. By utilising the Naive Bayes approach, the authors seek to enhance risk prediction, particularly in cardiovascular disease (CVD), using an EHR dataset from a comprehensive healthcare system in the Midwest.

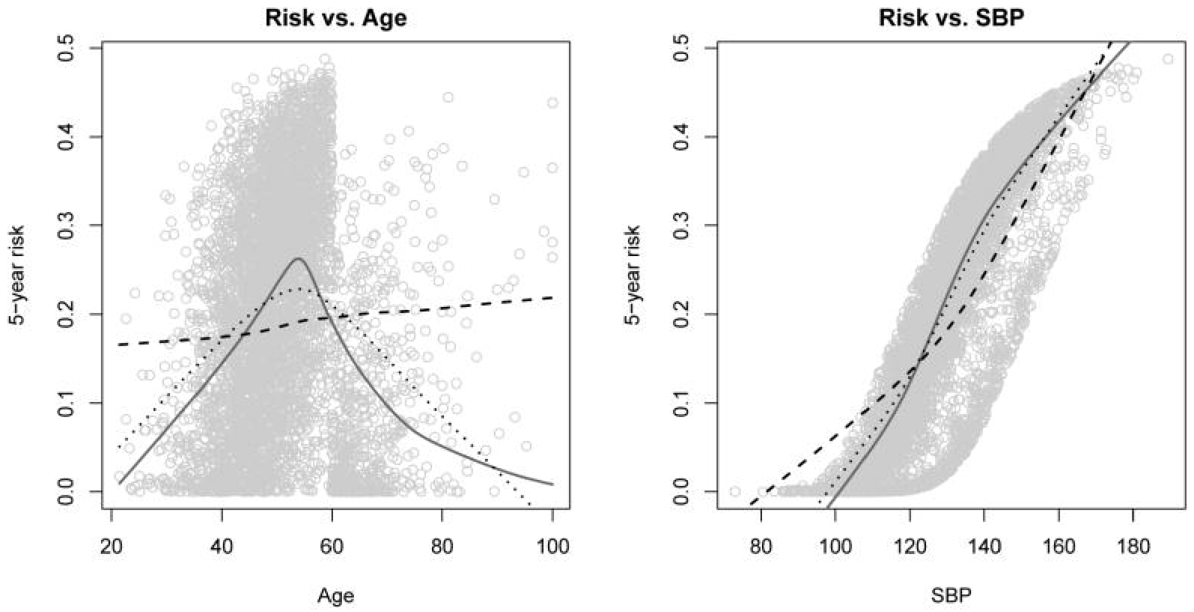


Figure 2. 8: Graph illustrated risk prediction vs age of a Naive Bayes based prediction model.

The image visually presents risk prediction based on a Log-logistic model, exploring the interaction between age and systolic blood pressure in cardiovascular risk assessment. The left panel illustrates age's influence on the 5-year risk prediction, while the right panel focuses on systolic blood pressure's impact. True risks are shown as plotted points, and a solid line provides a smoothed representation. The image compares Cox Proportional Hazards and Naive Bayes models, revealing distinct predictive trends for varied methodologies and enhancing insight into cardiovascular risk assessment dynamics.

The study highlights the potential of Naive Bayes in handling censored time-to-event data for risk prediction. In practice, this technique enables healthcare professionals and researchers to evaluate an individual's probability of experiencing specific clinical outcomes, such as cardiovascular events, based on their medical history and covariate patterns. The research underscores the significance of employing advanced machine learning approaches to navigate the intricacies of modern healthcare data, ultimately contributing to improved risk assessment and informed decision-making.

2.6 ANNs for CVD Prediction

Artificial Neural Networks (ANNs) are a class of machine learning algorithms inspired by the structure and functioning of the human brain's neural networks (Grossi & Buscema, 2008). ANNs consist of interconnected processing units, or "neurons," organised into layers: an input layer, one or more hidden layers, and an output layer. The different classes of neural networks will be discussed here.

ANNs are capable of handling diverse data types, such as categorical, numerical, and even image data if relevant. In the context of CVD risk prediction, ANNs can consider various factors like age, gender, cholesterol levels, blood pressure, smoking status, and more. Additionally, ANNs can incorporate novel biomarkers, genetic information, and even wearable device data for more accurate predictions (Grossi & Buscema, 2008; Mandeep et al., 2022).

The complexity of cardiovascular risk necessitates models that can handle intricate interactions, and ANNs excel at this. Their multi-layer architecture enables them to learn hierarchical representations of data, automatically extracting relevant features. ANNs are also highly adaptable; they can be tailored to different data types and complexities by adjusting the number of hidden layers and neurons (Grossi & Buscema, 2008).

Despite their power, ANNs do require large datasets for optimal performance, as their complexity can lead to overfitting with small datasets. Proper preprocessing and regularisation techniques are crucial to address these issues. However, when provided with sufficient data and expertly designed architectures, ANNs have demonstrated state-of-the-art performance in CVD risk prediction tasks (Mandeep et al., 2022). While ANNs might require more computational resources and expertise to build and fine-tune, their ability to uncover intricate risk factor relationships makes them one of the best tools for accurate and personalised CVD risk prediction.

2.6.1 Multi-Layer Perceptron (MLP)/ Feedforward Neural Network (FNN)

In an MLP, each connection between neurons has a weight associated with it, which determines the strength of the connection. ANNs learn from data by adjusting these weights through a process called training, aiming to minimise the difference between predicted and actual outcomes.

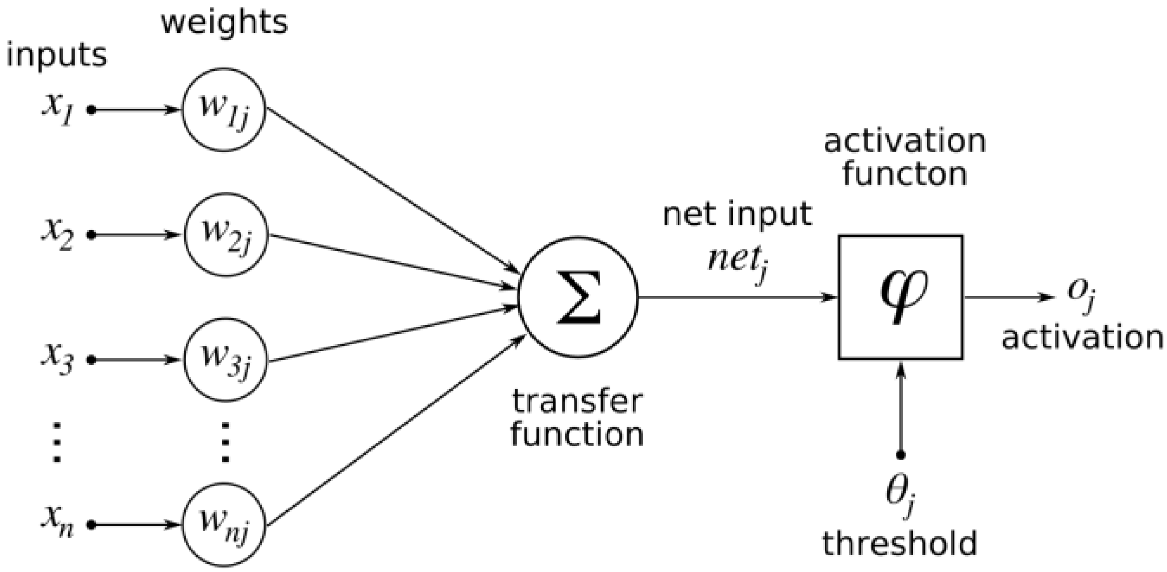


Figure 2. 9 Visual representation of the fundamental structure of a feedforward neural network (FNN)

The image illustrates the interconnected layers of neurons with input, hidden, and output layers.

FNNs are particularly suited for CVD risk prediction due to their ability to capture complex and non-linear relationships among risk factors (Mandeep et al., 2022). Cardiovascular risk is influenced by a combination of traditional and non-traditional risk factors, some of which may interact in intricate ways. ANNs can learn these intricate relationships from data, allowing them to model the nuances of risk factor interactions that might be missed by simpler models (Mandeep et al., 2022).

2.6.2 Convolutional Neural Network

According to Vaz and Balaji (2021), Convolutional Neural Networks (CNNs) are a class of neural networks that are specifically designed for processing multi-dimensional data, such as images and arrays. CNNs are known for their ability to adaptively learn patterns of varying complexity as they progress through layers, making them particularly effective in tasks like image recognition and analysis.

CNNs possess two key characteristics that distinguish them from other types of neural networks: weight sharing and local connectivity. Weight sharing means that within a given layer, the same set of weights is applied to all neurons. This shared weight structure allows CNNs to efficiently capture common patterns across the input data, enabling them to recognize features like edges, textures, and shapes at lower layers before identifying more complex patterns at higher layers.

Local connectivity, as defined by Vaz and Balaji, refers to the concept that each neuron in a CNN receives input from only a localised region of the input data or the previous layer. In other words, each neuron is connected to a limited subset of neurons in the preceding layer or region of the input. This arrangement is inspired by the way the visual cortex in the brain processes visual information by focusing on small receptive fields.

In summary, Convolutional Neural Networks (CNNs) are a class of neural networks designed for processing multi-dimensional data, with a primary focus on image-related tasks. They are characterised by weight sharing, which promotes the efficient learning of common patterns, and local connectivity, which mimics the receptive field concept in the brain's visual processing. These unique properties make CNNs highly effective in various applications, including image analysis, computer vision, and, as the paper discusses, their applications in pharmacogenomics.

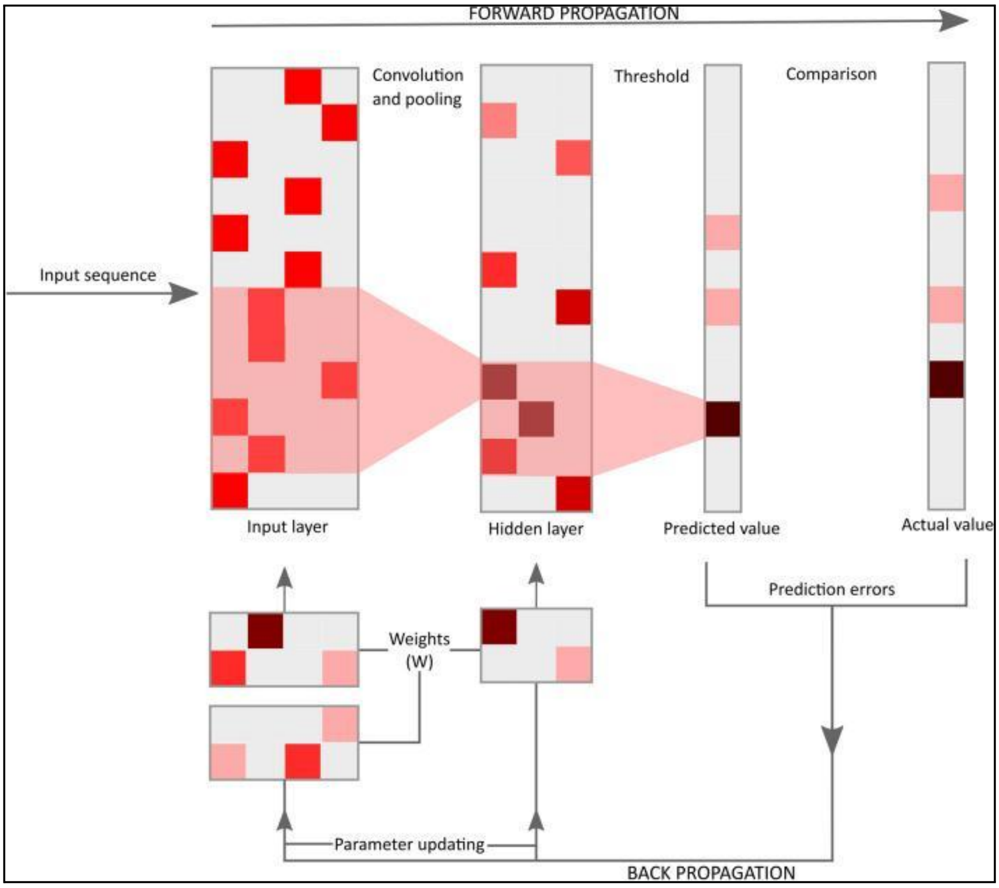


Figure 2. 10 Visual representation of a CNN according to Vaz and Balaji.

2.6.3 Recurrent Neural Network

According to Chen et al. (2019), Recurrent Neural Networks (RNNs) are a class of neural networks that excel in handling sequential data, such as time-stamped events in Electronic

Health Record (EHR) data. RNNs are particularly well-suited for tasks that involve modelling and capturing temporal dependencies within the data.

One of the key features of RNNs is their ability to encode and process sequences of data while maintaining an internal hidden state. This hidden state acts as a memory that can capture information from previous time steps and influence the model's predictions at the current time step. In the context of EHR data, RNNs can capture how past medical events or measurements relate to the current state of a patient's health, making them valuable for tasks like early detection of heart failure.

Chen et al. mention that RNNs, including gated recurrent unit (GRU) variants, have been applied in the field of healthcare, such as for early detection of heart failure. RNNs are capable of learning latent representations from longitudinal EHR data, which can be used for classification tasks. Additionally, RNN-based models with attention mechanisms, like RETAIN, offer interpretability by highlighting relevant information within the patient's history.

Furthermore, the authors emphasise that RNNs are known to excel in capturing temporal information, which is crucial for predicting events that unfold over time. This temporal modelling capability allows RNNs to outperform traditional machine learning models in scenarios where the sequence of events matters, as is often the case in healthcare data analysis.

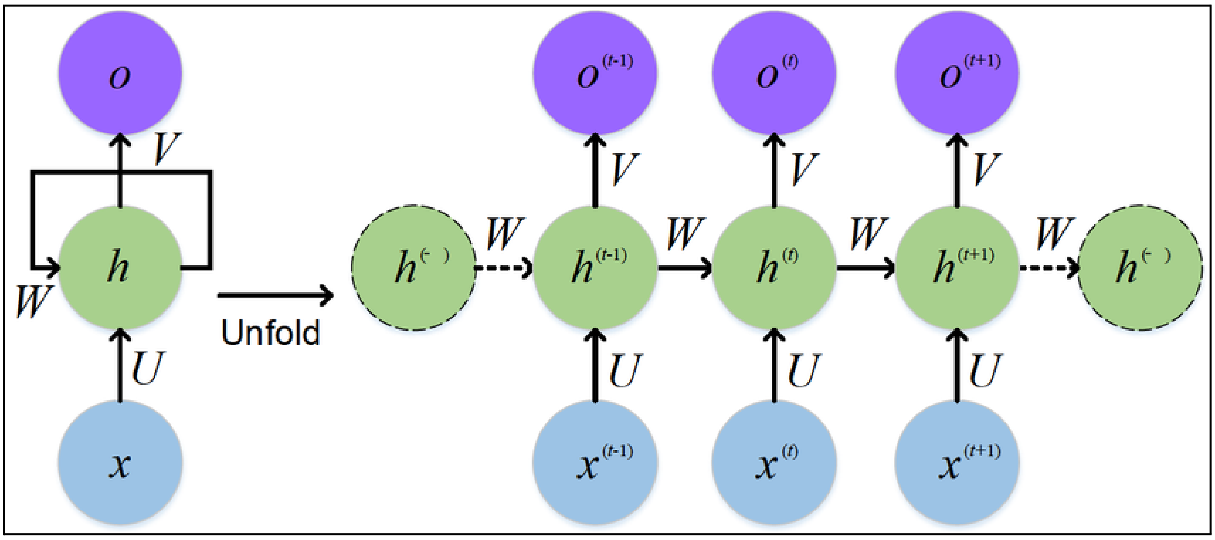


Figure 2. 11 General structure of an RNN according to Chen et al.

2.7 Evaluating Performance of MLP for CVD Prediction

When evaluating the performance of an Artificial Neural Network (ANN) for CVD risk prediction, several common performance measures can be used to assess how well the model is performing (Ahangari et al., 2023).

True Positive (TP) is the number of instances that are actually positive and are correctly predicted as positive by the model. True Negative (TN) is the number of instances that are actually negative and are correctly predicted as negative by the model. False Positive (FP), also known as a Type I error, is the number of instances that are actually negative but are incorrectly predicted as positive by the model. False Negative (FN), also known as a Type II error, is the number of instances that are actually positive but are incorrectly predicted as negative by the model.

Accuracy calculates the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset.

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{Total\ Instances}$$

Precision calculates the ratio of true positives to the total number of predicted positives (true positives + false positives). It indicates how many of the predicted positive cases were actually positive.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall, also known as sensitivity or true positive rate, calculates the ratio of true positives to the total number of actual positives (true positives + false negatives). It indicates how many of the actual positive cases were correctly predicted.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, giving an overall measure of a model's performance. The F1 score is especially useful when the classes are imbalanced, as it considers both false positives and false negatives.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Area Under the ROC Curve (AUC-ROC) is a graphical representation of a model's performance across different thresholds. The AUC-ROC score quantifies the area under the ROC curve, which represents the model's ability to distinguish between positive and negative cases.

It is calculated by plotting the True Positive Rate (Recall) against the False Positive Rate at different classification thresholds and calculating the area under the curve.

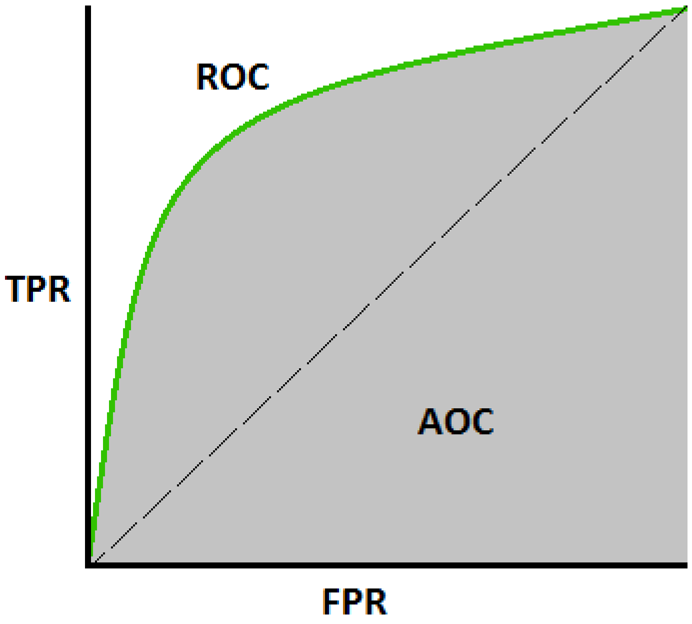
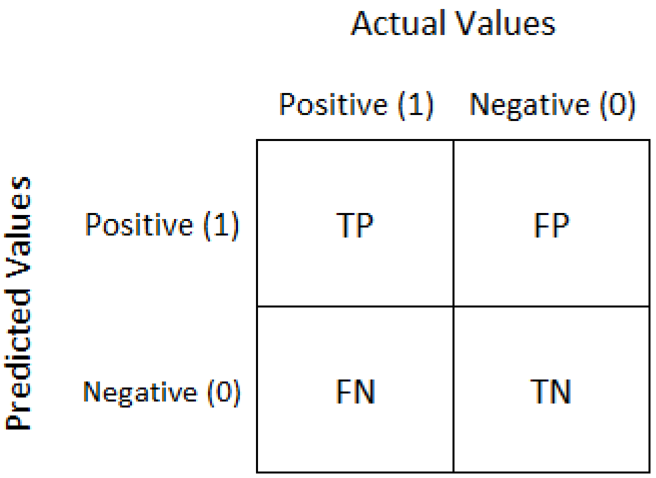


Figure 2. 12 The AUC-ROC curve graphically depicts a model's performance.

The AUC-ROC curve graphically depicts a model's performance in distinguishing between positive and negative cases.

The confusion matrix is a tabular representation that provides a detailed breakdown of the model's predictions. It includes metrics like true positives, true negatives, false positives, and false negatives. From the confusion matrix, various performance metrics can be calculated, including accuracy, precision, recall, and F1 score.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 2. 13 The confusion matrix visualises model predictions.

The confusion matrix visualises model predictions, displaying true positives, true negatives, false positives, and false negatives.

2.8 Conceptual Framework

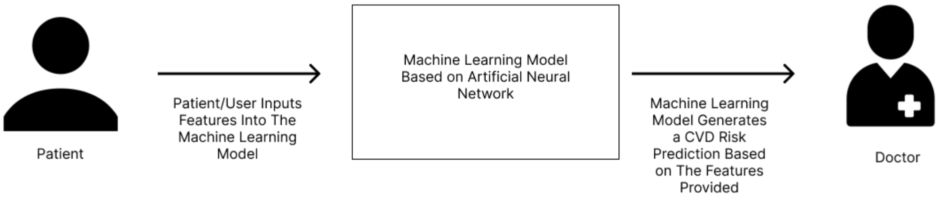

Patient
Patient/User Inputs Features Into The Machine Learning Model
Machine Learning Model Based on Artificial Neural Network
Machine Learning Model Generates a CVD Risk Prediction Based on The Features Provided
Doctor

Figure 2. 14 Conceptual framework for the proposed system

The proposed conceptual framework introduces a novel approach to cardiovascular disease (CVD) risk prediction by leveraging artificial neural networks (ANNs) and collaborative healthcare. At the forefront of this framework is the active involvement of users or patients who seek to understand their risk of developing CVD. Through an interactive interface, users input various relevant features, including personal information, clinical measurements, and lifestyle factors. These inputs serve as the foundation for the subsequent risk prediction process.

The core of the framework resides in the application of artificial neural networks (ANNs). These advanced machine learning models possess the capability to comprehend complex relationships among diverse features, making them ideal for predicting intricate outcomes like CVD risk. Trained on historical health records and corresponding CVD outcomes, the ANN learns to recognize patterns and correlations within the input data. This training enhances the model's predictive accuracy, allowing it to generate personalised CVD risk predictions based on the user's input.

The output of the framework is a personalised CVD risk prediction that empowers users with valuable insights into their health status. This prediction serves as a proactive tool, enabling individuals to make informed decisions about their lifestyle choices and potential medical interventions. Moreover, the framework promotes collaboration between users and medical professionals by integrating risk predictions into healthcare consultations. Armed with accurate risk assessments, medical practitioners can offer tailored advice and interventions, fostering a collaborative and patient-centred approach to CVD prevention and management. Ultimately, the conceptual framework bridges the gap between AI-driven risk prediction and expert medical guidance, empowering individuals to take proactive control of their cardiovascular health.

3 Chapter 3: Methodology

3.1 Introduction

This section explains the research methodology towards the initial intentions of this study. Research methodology is a systematic way to solve a problem. The procedures by which researchers go about their work of describing, explaining, and predicting phenomena (Kumari et al., 2023).

3.2 Research Design

Research design is the overall plan for connecting the conceptual research problems to the pertinent and achievable empirical research. It is an inquiry which provides specific direction for procedures in research (Asenahabi, 2019).

The research design is intended to provide an appropriate framework for a study. A very significant decision in the research design process is the choice to be made regarding the research approach since it determines how relevant information for a study will be obtained (Jilcha, 2020).

In a descriptive design in relation to the study, the paper seeks to gather information on who the study involves, when the study applies to them, and where the study is applicable.

3.3 Model Development

The development of the CVD risk prediction model utilised state-of-the-art machine learning techniques, particularly Feedforward Neural Networks (FNNs). FNNs are well-suited for handling complex patterns and relationships within the data, making them ideal for predicting intricate health outcomes like cardiovascular risks. The model architecture was designed to accommodate the input features from the dataset, with multiple hidden layers to capture the underlying complexities of CVD risk factors.

The model is a feedforward neural network that will be developed as follows:

I. Obtaining data
II. Processing of data
III. Extraction of features
IV. Development of the model
V. Validation of the model

3.3.1 Data Collection

The data was obtained from the publicly available Behavioural Risk Factor Surveillance System (BRFSS). The BRFSS is a premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviours, chronic health conditions, and use of preventive services.

The primary collectors of the data that made the data available publicly adhered to the Declaration of Helsinki and the patients under study gave written and informed consent to participate in the study.

The data is in the form of a comma separated values (.csv) file. The dataset contains 304 unique variables (columns) collected from 308,855 individuals (rows) that relate to the various CVD risk factors that contribute to the likelihood of developing CVD.

Some of the factors that were included in the dataset include but are not limited to the following: general health, last check up, exercise, heart disease, skin cancer, any other form of cancer and even depression.

General Health	Last Checkup	Exercise	… (298 other factors)	Skin Cancer	Depression	Heart Disease (Label)
Poor	Within past year	No	…	No	No	No
Very Good	Within last 2 years	No	…	No	No	Yes
Very Good	Within past year	Yes	…	No	No	No
Poor	Within past year	Yes	…	No	No	Yes
Good	Within last 2 years	No	…	No	Yes	No

Table 2 Data Characteristics

3.3.2 Data Processing

The data processing phase involved several key steps to prepare the dataset for CVD risk prediction. Initial exploratory data analysis (EDA) provided insights into the dataset's characteristics. Highly correlated columns were identified and addressed to reduce redundancy. The data was split into training (75%) and testing (25%) sets, and the target variable, "Heart disease," was established. Categorical data was encoded into numerical format, a critical step for neural network processing. A comprehensive preprocessing pipeline was created to ensure data compatibility with the model. These steps collectively optimised the dataset, facilitating accurate and efficient CVD risk prediction.

3.3.3 Model Development

In the model development phase, a Multilayer Perceptron (MLP) was meticulously crafted. Early stopping callbacks were defined to prevent overfitting. The MLP was initialised, and its layers were carefully structured, considering the intricacies of CVD risk prediction. Following this, the MLP was compiled, configuring the model for training. Subsequently, rigorous training sessions were conducted to ensure the model's proficiency in learning complex patterns from the data, a critical step in achieving accurate CVD risk predictions.

3.3.4 Model Validation

To ensure the reliability and accuracy of the developed model, a rigorous validation process was carried out. The dataset was divided into training, validation, and testing sets using appropriate techniques like stratified sampling. The training set was used to train the model, while the validation set facilitated hyperparameter tuning and optimization. The model's performance was then evaluated on the testing set to assess its ability to generalise to new, unseen data. Common performance metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) were calculated to quantify the model's predictive capabilities.

3.4 System Development Methodology

3.4.1 Application of the Data-Driven Modelling Methodology

The study will utilise the Data-driven Modelling (DDM) in the development of the system. This technique is useful in computational intelligence, machine learning and data mining. Computational intelligence includes neural networks (which this study will rely on), fuzzy systems and evolutionary computing (Solomatine and Ostfeld, 2008). The methodology encompasses several stages, including data collection, preprocessing, model development, validation, and evaluation. This systematic process ensures that the model is built on a robust foundation of relevant and accurate data. The use of a data-driven methodology allows for flexibility and adaptability, enabling the model to capture the dynamic nature of cardiovascular risk factors and their interactions.

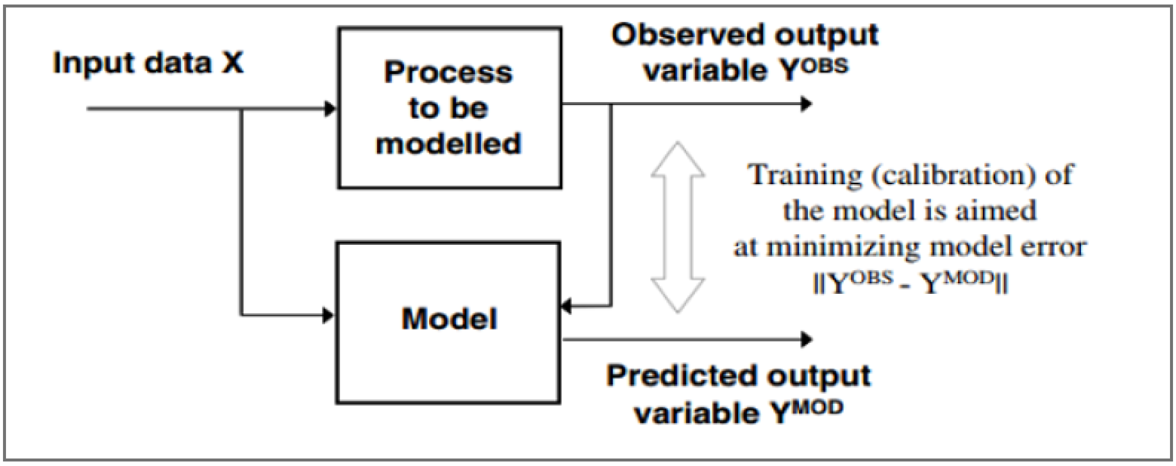

Input data X
Process to be modelled
Observed output variable Y^{OBS}
Training (calibration) of the model is aimed at minimizing model error $||Y^{OBS} - Y^{MOD}||$
Model
Predicted output variable Y^{MOD}

Figure 3.0.1 Data-driven modelling (Adapted from (Solomatin and Ostfeld, 2008))

3.5 System Development Tools and Techniques

3.5.1 Python

The development of the CVD risk prediction model was facilitated by Python, a widely used programming language known for its versatility and extensive libraries. Python provided a robust environment for data manipulation, preprocessing, and model construction. Leveraging its vast ecosystem, the project team efficiently managed various aspects of the model's development.

3.5.2 Google Colaboratory

Google Colaboratory (Colab), a cloud based Jupyter notebook environment, played a pivotal role in the development process. It enabled seamless collaboration among team members and provided a platform for coding, data exploration, and model training. Colab's integration with Google Drive and GPU acceleration expedited model development and experimentation, ensuring an efficient workflow.

3.5.3 Scikit-Learn

Scikit-learn, a comprehensive machine learning library in Python, was a cornerstone tool for this project. It facilitated data preprocessing, feature engineering, and model evaluation. Scikit-learn's user-friendly APIs streamlined the development of the prediction model by offering a range of algorithms, enabling the team to explore and select the most suitable approach.

3.5.4 TensorFlow

The project incorporated TensorFlow, a popular open-source deep learning framework. TensorFlow enabled the construction and training of complex neural network architectures for the CVD risk prediction model. Its scalability and compatibility with GPUs expedited the model training process, allowing for the exploration of intricate network designs.

3.5.5 Keras

Keras, a high-level neural network API built on top of TensorFlow, simplified the implementation of deep learning models. Its user-friendly interface allowed the project team to design, configure, and train the Multi-Layer Perceptron (MLP) model efficiently. Keras abstracted intricate details, enabling a focus on model architecture and performance.

3.5.6 GitHub

GitHub served as the version control system and collaboration platform for the project. It enabled efficient code management, version tracking, and collaborative code reviews among team members. The platform's features facilitated seamless integration of updates, ensured code transparency, and provided a unified repository for the entire development lifecycle.

3.6 Ethical Considerations

Incorporating these tools and techniques, ethical considerations were at the forefront of the project's development. From data collection to model deployment, privacy and data security were meticulously upheld. Measures were taken to anonymize and protect sensitive patient data, ensuring compliance with ethical standards.

Bias mitigation was a crucial aspect, particularly when utilising machine learning tools. The project team addressed algorithmic biases and worked to ensure equitable performance across different demographic groups. Transparency in model training and evaluation was maintained, fostering accountability and responsible AI practices.

The interpretability of the CVD risk prediction model was prioritised, aligning with ethical guidelines. Model outputs were designed to be easily comprehensible for medical professionals and patients, enhancing the overall trustworthiness and usefulness of the predictions. Throughout the project's lifecycle, the ethical implications of the model's predictions were carefully considered.

4 Chapter 4: System Design and Architecture

4.1 Introduction

This section outlines the architecture of the developed prediction model for CVD risk prediction. The architecture builds up on the conceptual model developed in Figure 2.5. This section covers the requirements that need to be satisfied, the components of the developed system, and the interaction between the end user and the developed system. In addition, it covers the interactions between the components of the developed model. Use case diagrams, sequence diagrams, system sequence diagrams as well as flow chart diagrams were used to model the system.

4.2 Requirement Analysis

The requirements will be broken down into four namely: functional requirements, non-functional requirements, usability requirements, and reliability requirements.

4.2.1 Functional Requirements

i. The system should allow the user to input data as a CSV file.

ii. The system should be able to predict the risk of CVD by use of deep learning.

iii. The predicted CVD risk should be valid based on the input supplied from the user.

4.2.2 Non-functional Requirements

i. The system should be easy to re-train and adjust the weights.

ii. The system should be able to robustly generalise new instances and avoid overfitting.

iii. The system should be secure to avoid unauthorised changes to the model parameters.

iv. The system should have persistent weight storage to avoid re-training every time a prediction is done.

4.2.3 Usability Requirements

The system is intended to be used in a clinical setting; hospitals and clinics. The system should therefore be simple to use. It should also be straightforward to ensure that it can be easily accepted by the users. The system should also provide accurate predictions and explanations to the predictions as it may directly impact the lives of the patients.

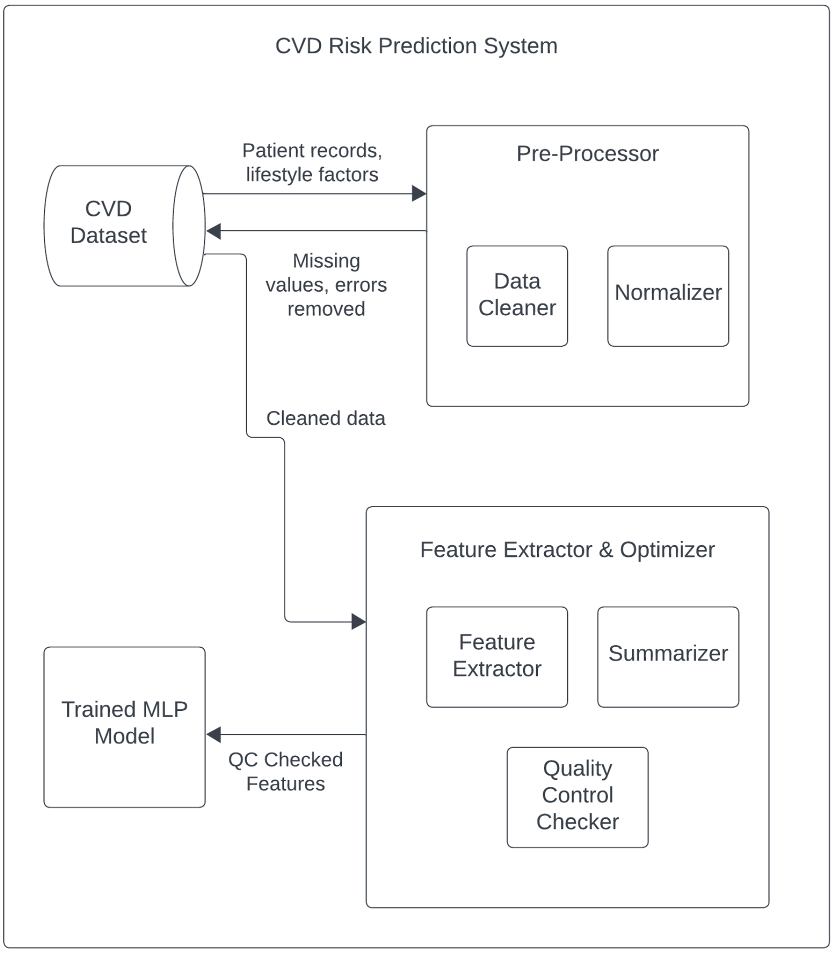
4.2.4 Reliability Requirements

i. The system should always interface with the existing database containing clinical information.

ii. The administrator should be able to correctly restore the system in the occurrence of a failure.

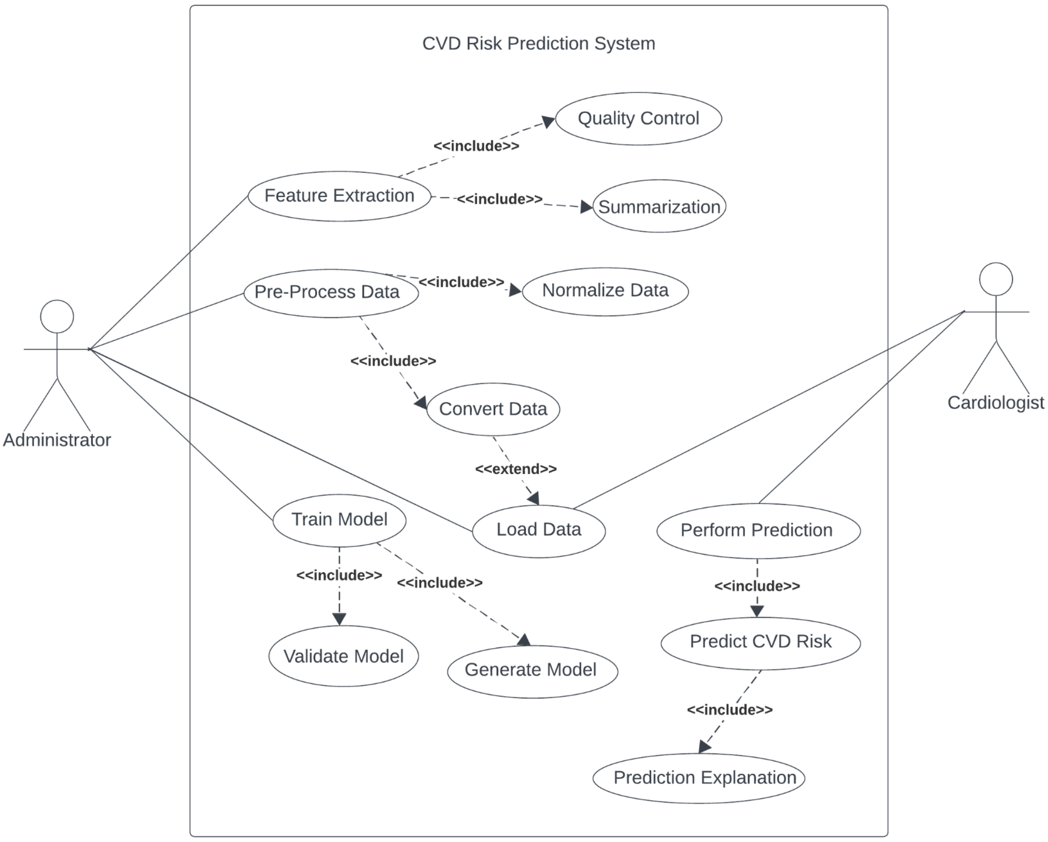
4.3 Architecture

The system architecture gives a generalised layout of the CVD risk prediction system and its individual components as shown in Figure 4.1. The process begins by extracting data from the dataset that is supplied by the user. The data is then pre-processed to clean the data, detect outliers as well as handle missing values. The data is then fed to the feature extracter where the suitable features are determined for further processing. The data with the extracted features is then fed into the trained model and prediction obtained.



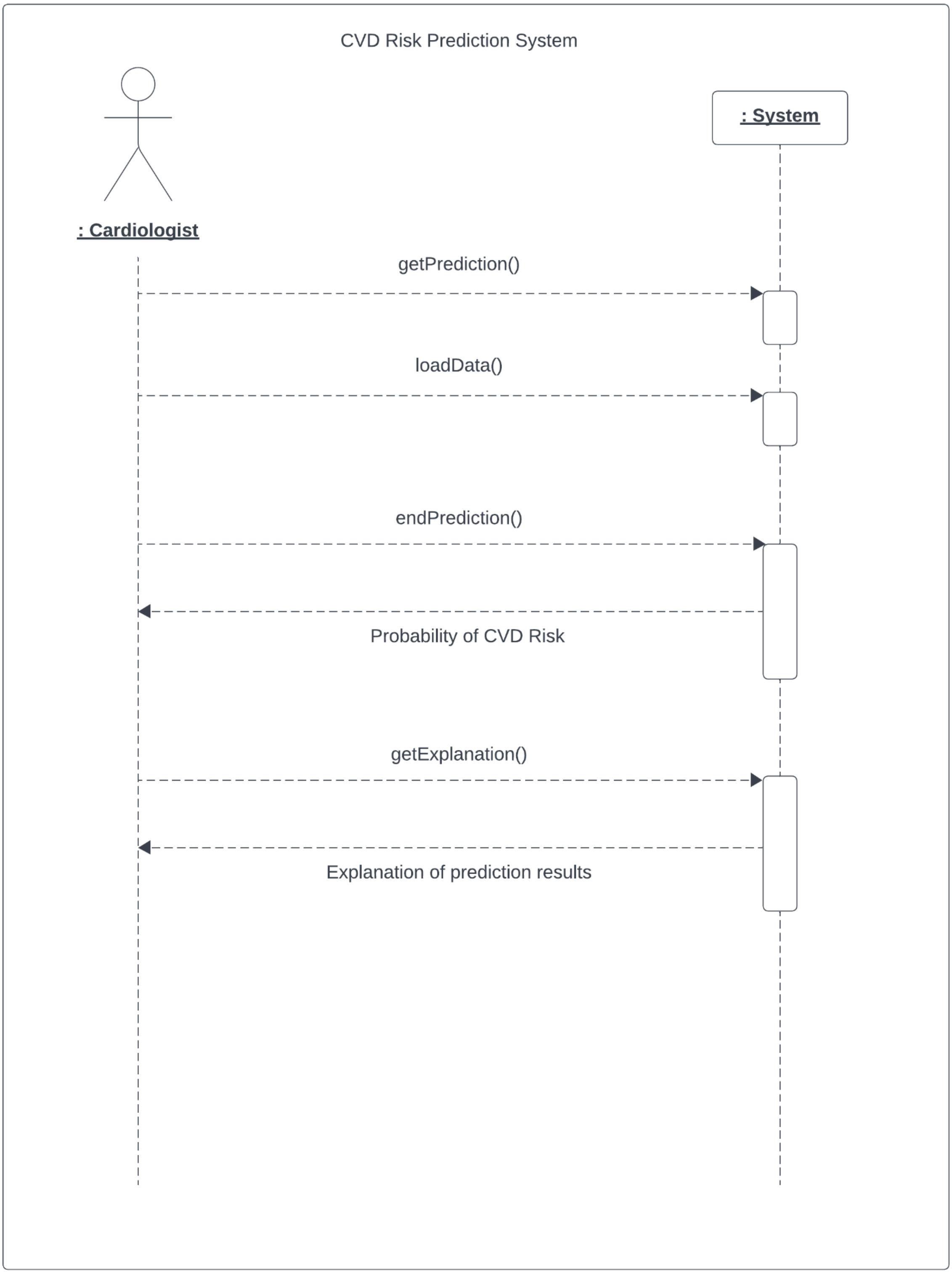
4.4 Use Case Diagram

The use case diagram is used to depict the interaction between actors and the system. Figure 4.2 illustrates this interaction as well the proposed functionality that the system should have.



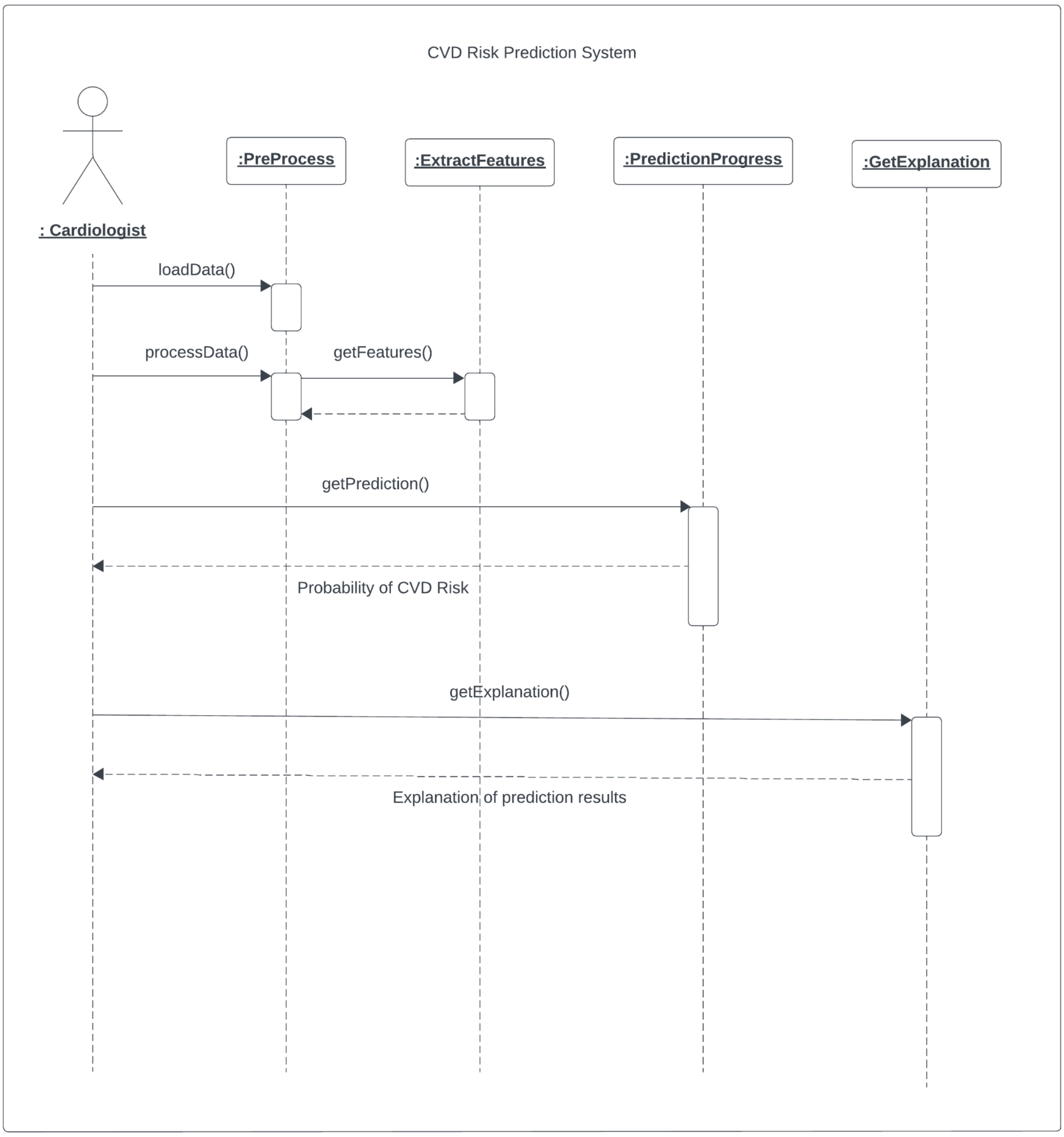
4.5 System Sequence Diagram

The system sequence diagram illustrates how the primary user and the system interact. Figure 4.3 illustrates this interaction. The cardiologist will first initiate the process of prediction. He/She will load the data that they intend to get prediction results for and then end the prediction. The system will then output the predicted values to show the patient's probability of CVD. Lastly, the cardiologist will ask for an explanation of how the prediction was arrived at and the system will output the explanation.



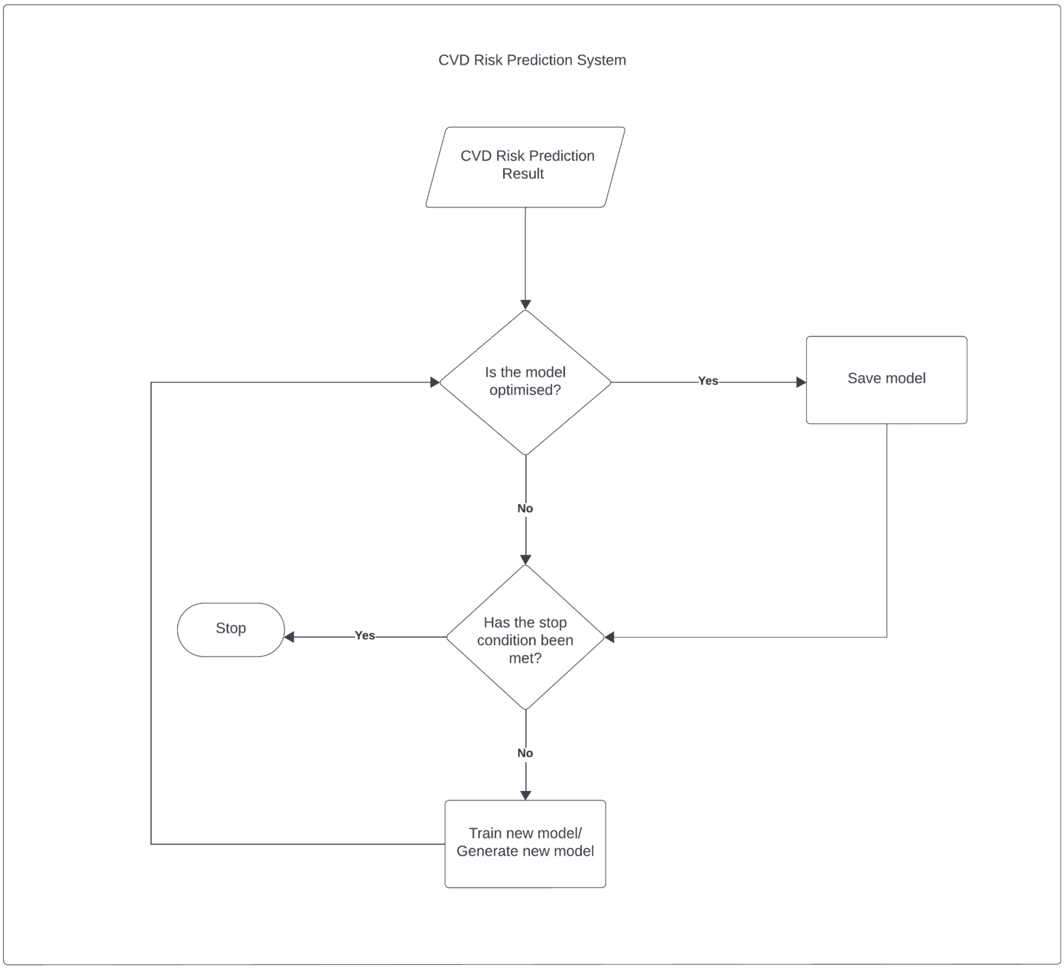
4.6 Sequence Diagram

The sequence diagram shown in Figure 4.4 shows the sequence of interactions between the user and the internal components of the system. The data uploaded is first processed and the processed data undergoes feature extraction. The data is then split into training and test cases and fed into the multi-layer perceptron to create the model. The system then performs the prediction and sends it to the cardiologist with the percentage accuracy. The cardiologist then asks for an explanation to the prediction and the system sends the explanation as to how the prediction was obtained.



4.7 Flow Chart Diagram

The process flow of how the training model is obtained is shown in Figure 4.5. Patient data is entered and the data is then pre-processed. The data is then tested based on the existing generated model based on the predefined number of epochs (stopping condition). If the model is not optimal, then it checks if the number of epochs that have to be run while training the model has been met. If the number of epochs has been met, the system stops. If the number has not been met, the weights are readjusted and a new model is trained. The system repeats this process until the stopping condition has been met. The optimal model is then saved and will be used at runtime.


CVD Risk Prediction System
CVD Risk Prediction Result
Is the model optimised?
Yes
Save model
No
Stop
Yes
Has the stop condition been met?
No
Train new model/ Generate new model

5 Chapter 5: System Implementation

5.1 Introduction

This chapter describes the process of how the prototype was implemented, trained, tested, validated and optimised. The process starts by extracting data and normalising that data. Preprocessing is then examined and afterwards the model is trained. The model is then validated by using the validation data set and it is optimised by adjusting its variables. Once the model is optimised, it is tested against the test data and the results evaluated. To validate the approach,the model's performance is evaluated using various metrics such as accuracy. The model with the highest accuracy was chosen.

5.2 Data Acquisition

The dataset used in this project was obtained from the Behavioural Risk Factor Surveillance System (BRFSS), a well-established system of health-related telephone surveys in the United States. The BRFSS collects comprehensive state-level data on health-related risk behaviours, chronic health conditions, and the utilisation of preventive services among U.S. residents.

5.2.1 Data Source and Description

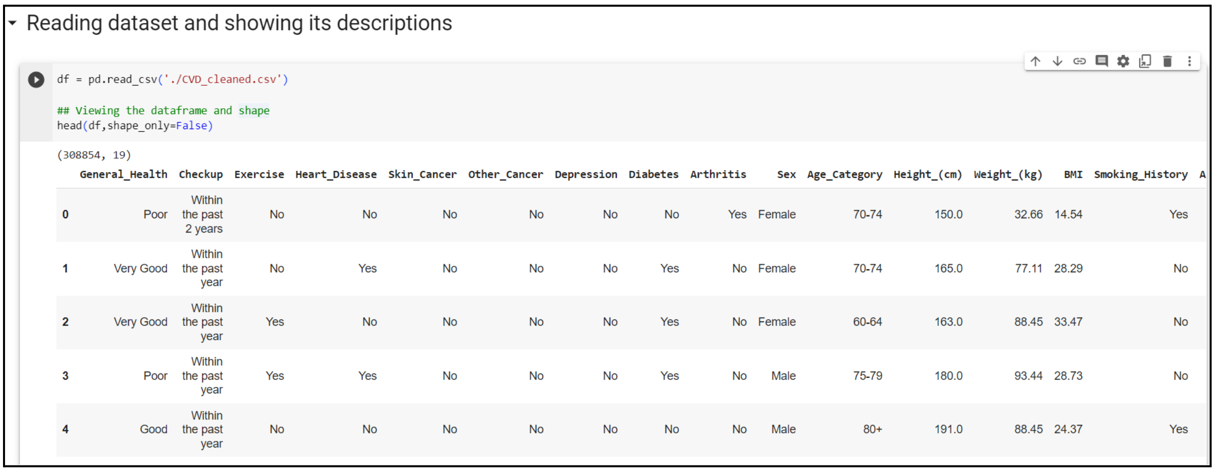
The data used was publicly available from BRFSS, ensuring its accessibility and utility for research purposes. The dataset is provided in comma-separated values (CSV) format, a common and convenient format for data analysis. The dataset consists of 304 unique variables (columns) collected from 308,855 individuals (rows). These variables encompass various factors related to cardiovascular disease (CVD) risk.

5.2.2 Ethics and Permissions

The primary data collectors responsible for making this dataset available adhered to the Declaration of Helsinki (2007), reflecting ethical principles in medical research. All individuals under study provided written and informed consent to participate in the survey, ensuring ethical data collection.

5.2.3 Code Snippet and Data Retrieval

This code snippet demonstrates how the dataset was read into a Python environment and displays a sample of the data. It marks the initial step in the data acquisition process.



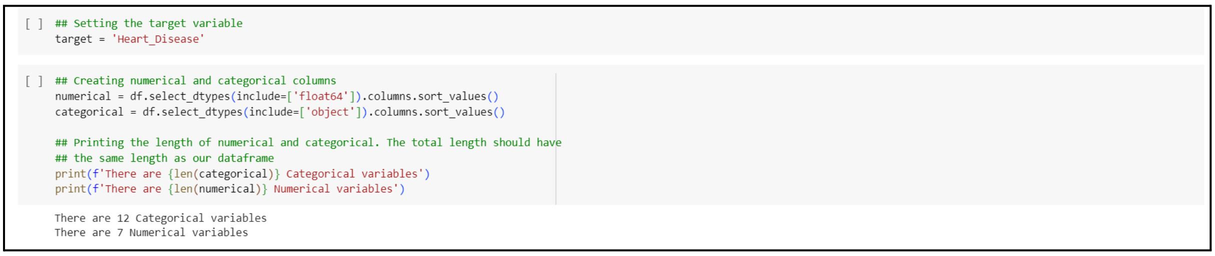
Reading dataset and showing its descriptions

```
df = pd.read_csv('./CVD_cleaned.csv')

## Viewing the dataframe and shape
head(df,shape_only=False)
```

```
(308854, 19)
```

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History	A
0	Poor	Within the past 2 years	No	No	No	No	No	No	Yes	Female	70-74	150.0	32.66	14.54	Yes	
1	Very Good	Within the past year	No	Yes	No	No	No	Yes	No	Female	70-74	165.0	77.11	28.29	No	
2	Very Good	Within the past year	Yes	No	No	No	No	Yes	No	Female	60-64	163.0	88.45	33.47	No	
3	Poor	Within the past year	Yes	Yes	No	No	No	Yes	No	Male	75-79	180.0	93.44	28.73	No	
4	Good	Within the past year	No	No	No	No	No	No	No	Male	80+	191.0	88.45	24.37	Yes	

This code snippet shows how the target variable is set, numerical and categorical columns are identified, saved as variables and their lengths are outputted. There are 12 categorical variables and 7 numerical variables.



```
[ ] ## Setting the target variable
    target = 'Heart_Disease'
```

```
[ ] ## Creating numerical and categorical columns
    numerical = df.select_dtypes(include=['float64']).columns.sort_values()
    categorical = df.select_dtypes(include=['object']).columns.sort_values()

    ## Printing the length of numerical and categorical. The total length should have
    ## the same length as our dataframe
    print(f'There are {len(categorical)} Categorical variables')
    print(f'There are {len(numerical)} Numerical variables')
```

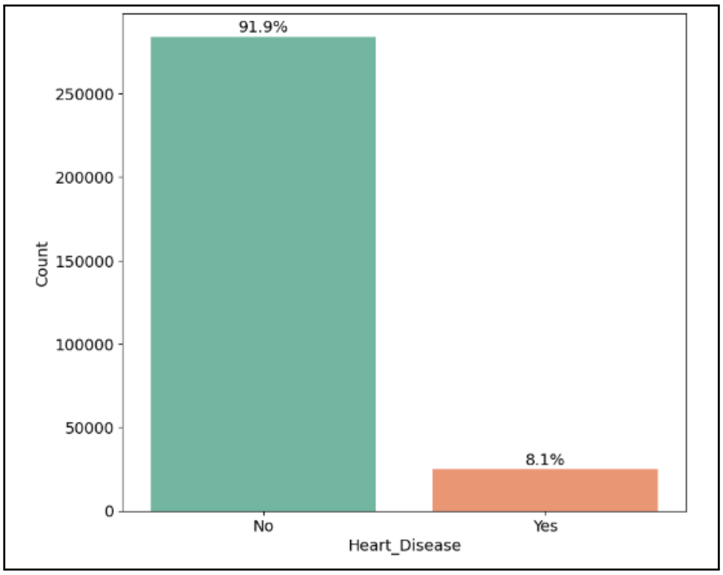
```
    There are 12 Categorical variables
    There are 7 Numerical variables
```

5.3 Exploratory Data Analysis (EDA)

In this section, we perform comprehensive Exploratory Data Analysis (EDA) to gain a deep understanding of the CVD dataset. EDA involves the analysis of the target variable, univariate, bivariate, and multivariate data analysis.

5.3.1 Target Variable Analysis

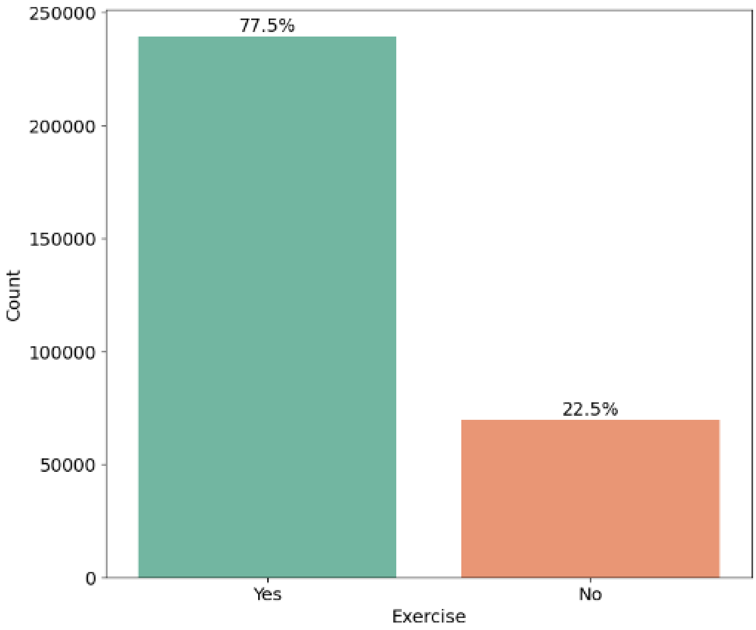
Using a bar chart, we analyse the distribution of the 'heart_disease' target variable.



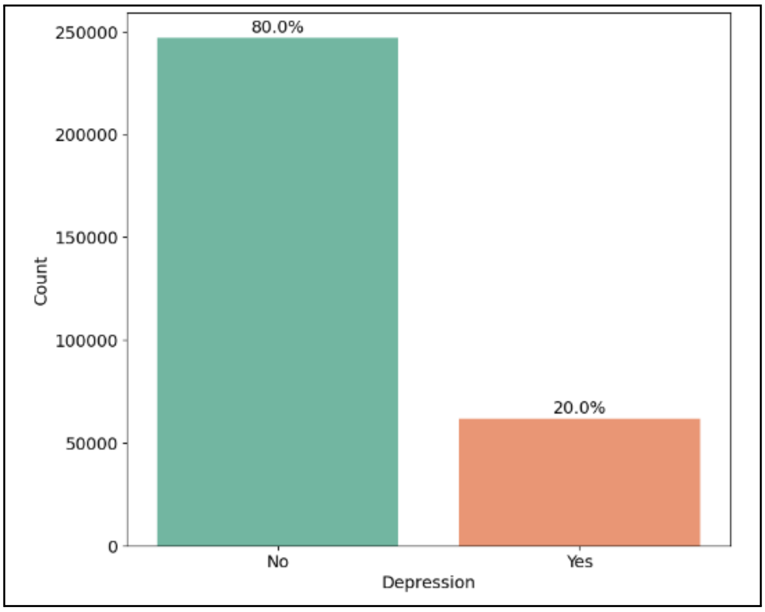
We discover that 8.1% of patients have heart disease, while 91.9% do not, indicating an imbalanced dataset.

5.3.2 Univariate Data Analysis

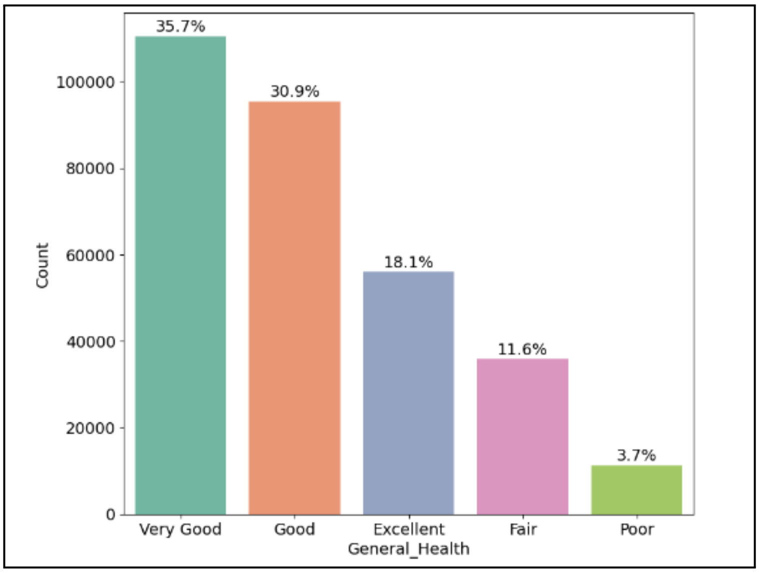
In this section we conduct univariate analysis by exploring different columns/features to understand their distributions and percentages. Included here are bar charts for columns like exercise, general health, and depression to show percentages of individuals with and without specific characteristics.



Here we can see that 77.5% of individuals in the dataset exercise while only 22.5% of individuals do not actively engage in exercise.



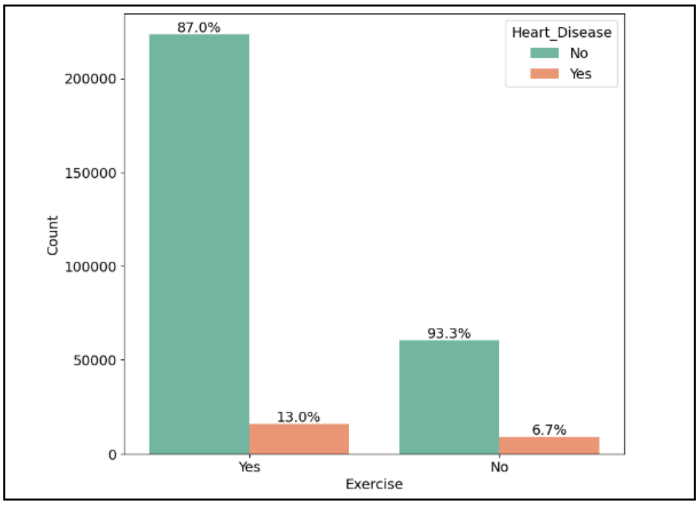
Here we can see that 80% of individuals in the dataset are not depressed while 20% of the individuals are clinically depressed.



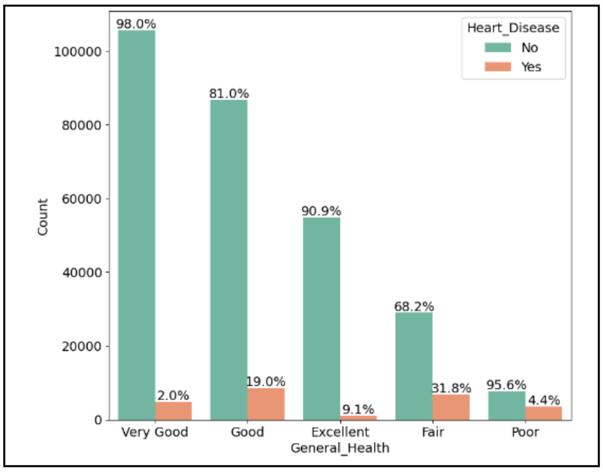
Here we can see the different percentages of individuals across different general health brackets including; very good, good, excellent, fair and poor.

5.3.3 Bivariate Data Analysis

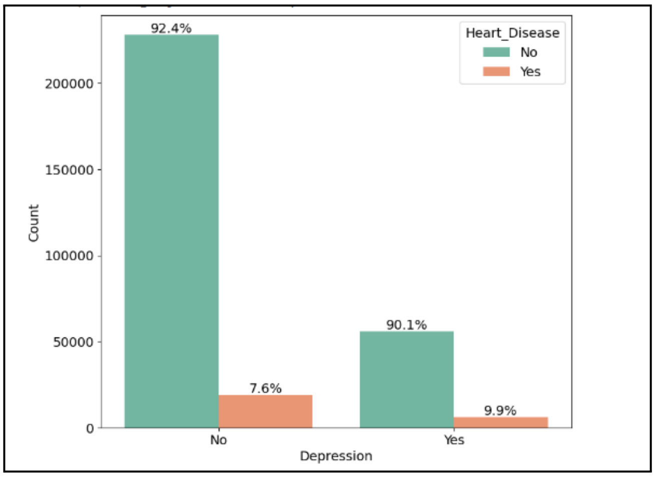
In this section we perform bivariate analysis by examining the relationship between various features and the target variable ('heart_disease'). Bar charts are utilised to show how features like exercise, general health, and depression relate to the presence or absence of heart disease.



Here the relationship between exercise and the target variable (‘heart_disease’) is made clear using a bar chart illustrating the different percentages of people who do or do not exercise and whether they have heart disease or not.



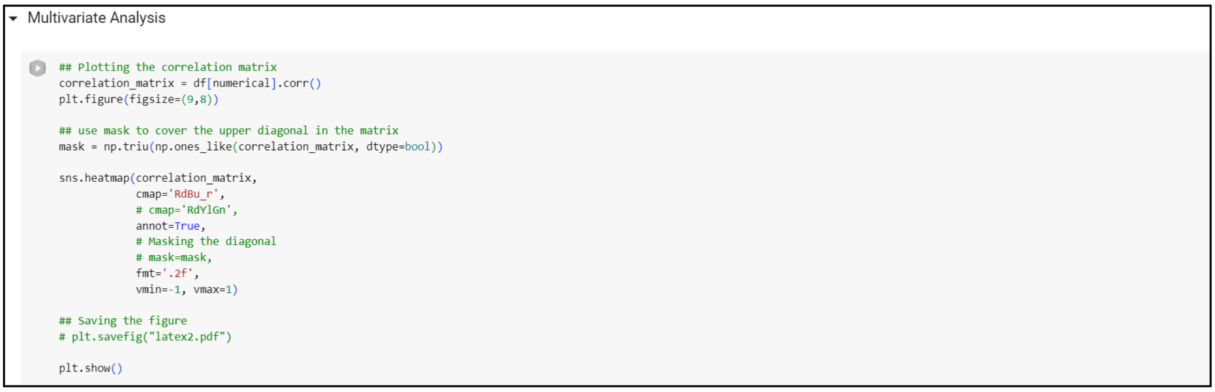
Here the relationship between general health and the target variable (‘heart_disease’) is made clear using a bar chart illustrating the different percentages of general health brackets and whether they have heart disease or not.



Here the relationship between depression and the target variable (‘heart_disease’) is made clear using a bar chart illustrating the different percentages of people who do or do not suffer from depression and whether they have heart disease or not.

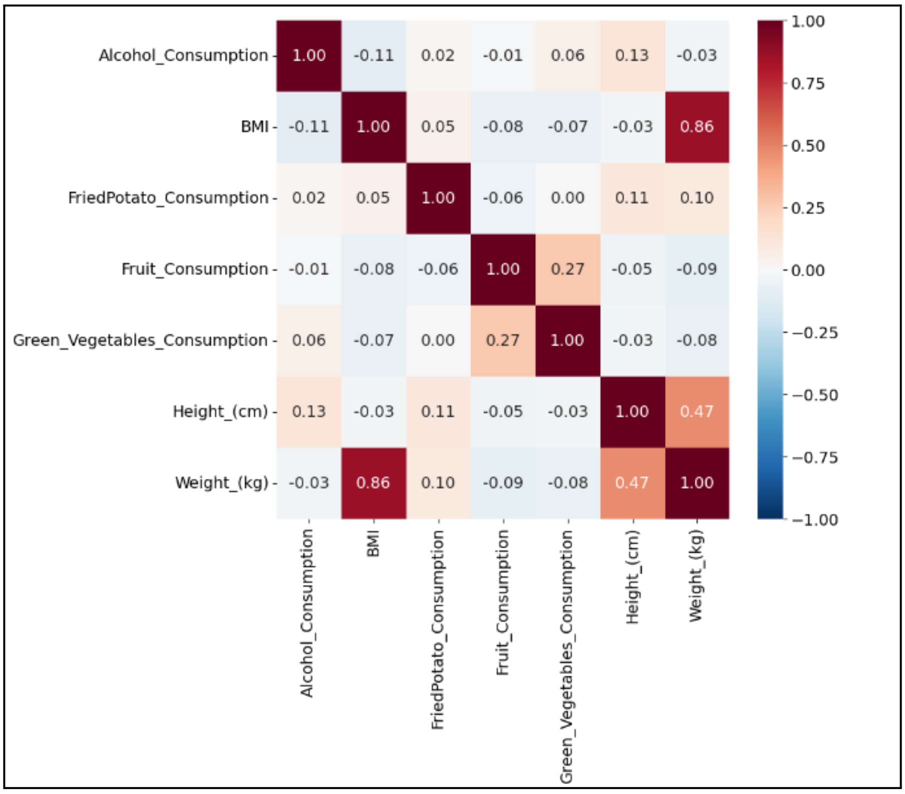
5.3.4 Multivariate Data Analysis

Here a correlation matrix is created to analyse the relationships between selected variables and other selected variables.



```
## Plotting the correlation matrix
correlation_matrix = df[numerical].corr()
plt.figure(figsize=(9,8))

## use mask to cover the upper diagonal in the matrix
mask = np.triu(np.ones_like(correlation_matrix, dtype=bool))

sns.heatmap(correlation_matrix,
            cmap='RdBu_r',
            # cmap='RdYlGn',
            annot=True,
            # Masking the diagonal
            # mask=mask,
            fmt='.2f',
            vmin=-1, vmax=1)

## Saving the figure
# plt.savefig("latex2.pdf")

plt.show()
```

The correlation matrix is used to explore how variables' correlations may provide insights into CVD risk factors.

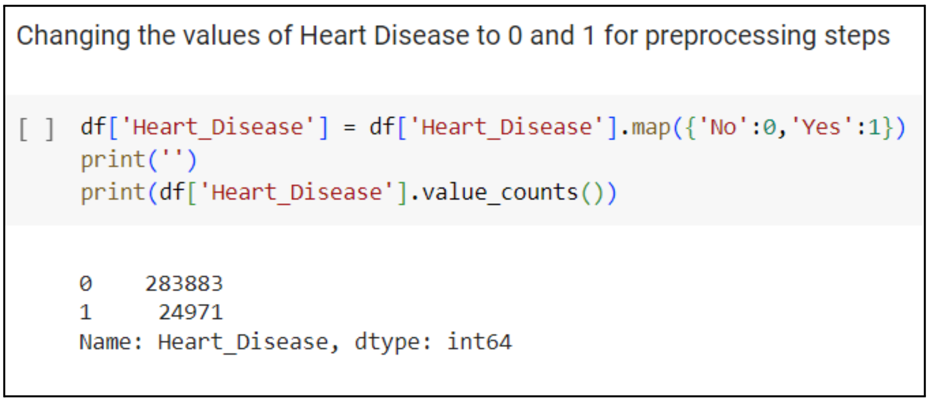


5.4 Data Processing (Pre-processing)

In this section, we detail the pre-processing steps applied to prepare the CVD dataset for modelling. These steps include encoding the target variable, data splitting, creating pipelines for categorical, numerical, and ordinal features, and demonstrating the effect of pre-processing on the dataset's shape.

5.4.1 Encoding the Target Variable

Here we transform the 'heart_disease' target variable into binary format, mapping 'yes' or 'no' values to 1 and 0, respectively. This binary encoding makes it suitable for the MLP classification model.



Changing the values of Heart Disease to 0 and 1 for preprocessing steps

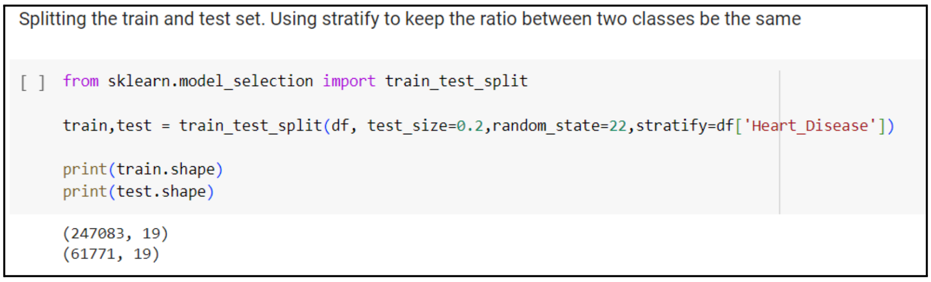
```
[ ]  df['Heart_Disease'] = df['Heart_Disease'].map({'No':0,'Yes':1})
     print('')
     print(df['Heart_Disease'].value_counts())
```

```
0    283883
1     24971
Name: Heart_Disease, dtype: int64
```

5.4.2 Data Splitting

Here we split the dataset into training and testing sets using the stratify parameter to maintain the class distribution balance in both sets.

Stratified splitting is essential when dealing with imbalanced data because it helps maintain the relative proportions of different classes in both the training and testing datasets. This helps to preserve the class distribution, avoid overfitting, and for the model to generalise better.

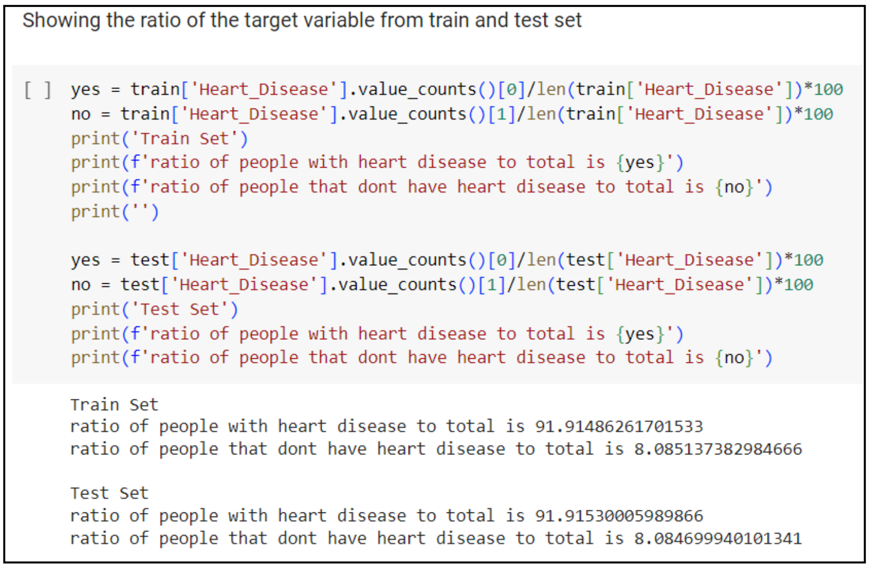


Splitting the train and test set. Using stratify to keep the ratio between two classes be the same

```
[ ] from sklearn.model_selection import train_test_split

    train,test = train_test_split(df, test_size=0.2,random_state=22,stratify=df['Heart_Disease'])

    print(train.shape)
    print(test.shape)

    (247083, 19)
    (61771, 19)
```

5.4.3 Class Distribution in Train and Test Sets

Here we present the class distribution of the target variable ('heart_disease') in both the training and testing sets. It can also be seen that the stratified split successfully maintained the ratio between the two classes.



Showing the ratio of the target variable from train and test set

```
[ ] yes = train['Heart_Disease'].value_counts()[0]/len(train['Heart_Disease'])*100
    no = train['Heart_Disease'].value_counts()[1]/len(train['Heart_Disease'])*100
    print('Train Set')
    print(f'ratio of people with heart disease to total is {yes}')
    print(f'ratio of people that dont have heart disease to total is {no}')
    print('')

    yes = test['Heart_Disease'].value_counts()[0]/len(test['Heart_Disease'])*100
    no = test['Heart_Disease'].value_counts()[1]/len(test['Heart_Disease'])*100
    print('Test Set')
    print(f'ratio of people with heart disease to total is {yes}')
    print(f'ratio of people that dont have heart disease to total is {no}')

    Train Set
    ratio of people with heart disease to total is 91.91486261701533
    ratio of people that dont have heart disease to total is 8.085137382984666

    Test Set
    ratio of people with heart disease to total is 91.91530005989866
    ratio of people that dont have heart disease to total is 8.084699940101341
```

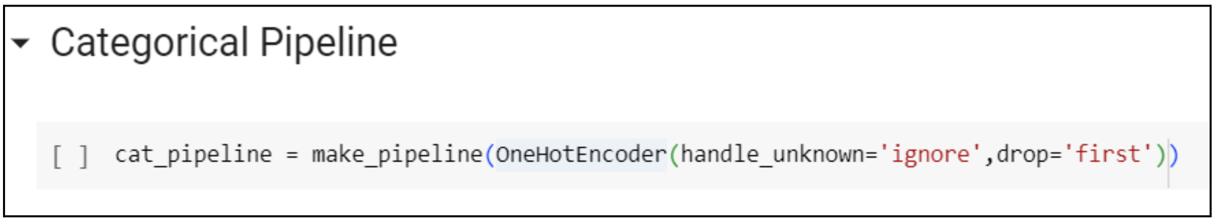
5.4.4 Creating Pipelines

This section outlines the creation of separate pipelines for different feature types: categorical, numerical, and ordinal.

The motivation for using pipelines in this case is to create a structured and efficient workflow for data pre-processing. Pipelines allow for a seamless integration of data transformation steps, making the code more organised and easier to maintain. Additionally, they enable consistent and reproducible pre-processing, reducing the risk of errors in the data preparation phase.

5.4.4.1 Categorical Pipeline

For the categorical pipeline, only OneHotEncoder will be implemented. Since this dataset has been cleaned and there are no missing values.



Categorical Pipeline

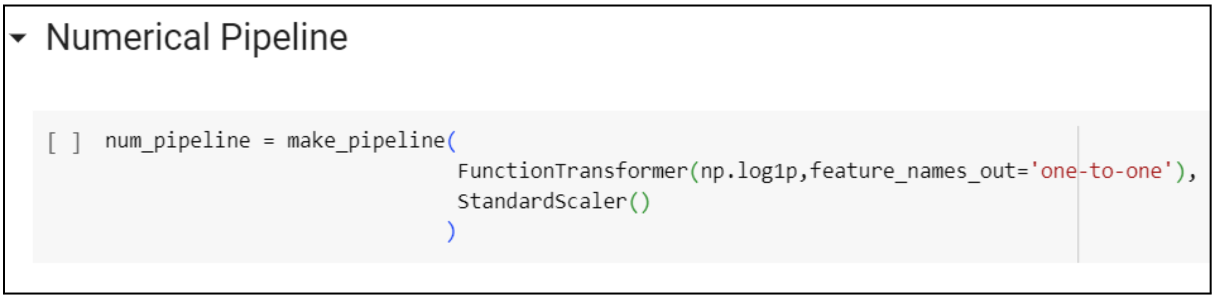
```
cat_pipeline = make_pipeline(OneHotEncoder(handle_unknown='ignore',drop='first'))
```

5.4.4.2 Numerical Pipeline

In the numerical pipeline, two techniques are applied to enhance the quality of the dataset.

First, the log transformation is used to address right-skewed numerical features, a common issue in real-world data. By taking the logarithm of these variables, the transformation mitigates the impact of extreme values, effectively reducing the skewness and making the data conform more closely to a normal distribution. This helps prevent the model from being overly influenced by a few outliers.

Second, standard scaling is employed to ensure that all numerical variables are on the same scale. This process involves centering the data (mean subtraction) and scaling it to have unit variance, which is essential for many machine learning algorithms, particularly those sensitive to the scale of input features. Standard scaling ensures that each variable contributes equally to the model's learning process, making the data more suitable for modelling and improving the model's performance.

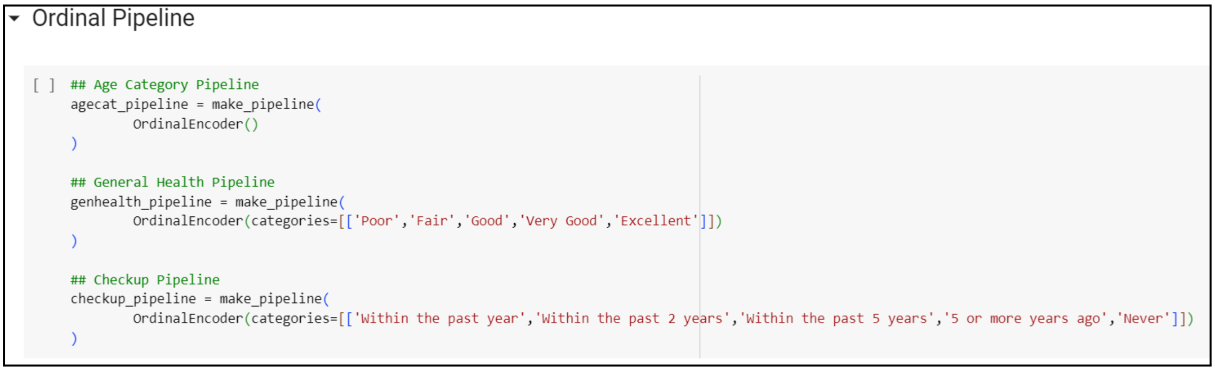


Numerical Pipeline

```
[ ]  num_pipeline = make_pipeline(
                                  FunctionTransformer(np.log1p,feature_names_out='one-to-one'),
                                  StandardScaler()
                                 )
```

5.4.4.3 Ordinal Pipeline

For the ordinal pipeline, the OrdinalEncoder() is used to encode each of the three ordinal variables present in the dataset; age, general check up and patient's last checkup. Here, the transformation process involves mapping the lowest values starting with zero and increasing them by one.

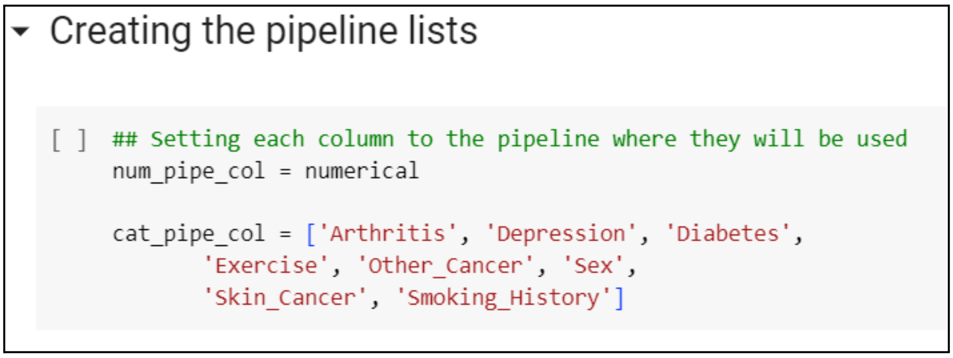


Ordinal Pipeline

```
[ ]  ## Age Category Pipeline
     agecat_pipeline = make_pipeline(
             OrdinalEncoder()
     )

     ## General Health Pipeline
     genhealth_pipeline = make_pipeline(
             OrdinalEncoder(categories=[['Poor','Fair','Good','Very Good','Excellent']])
     )

     ## Checkup Pipeline
     checkup_pipeline = make_pipeline(
             OrdinalEncoder(categories=[['Within the past year','Within the past 2 years','Within the past 5 years','5 or more years ago','Never']])
     )
```

5.4.5 Creating Pipeline Lists

Here we create lists that assign each column to the appropriate pre-processing pipeline, based on its type (categorical, numerical, or ordinal).

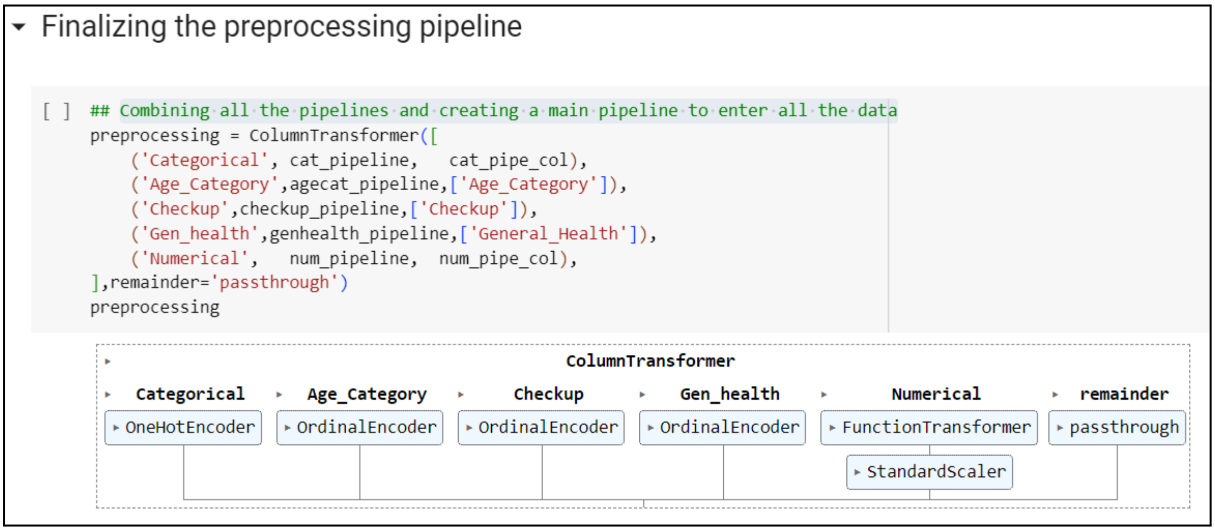


Creating the pipeline lists

```
[ ]  ## Setting each column to the pipeline where they will be used
     num_pipe_col = numerical

     cat_pipe_col = ['Arthritis', 'Depression', 'Diabetes',
             'Exercise', 'Other_Cancer', 'Sex',
             'Skin_Cancer', 'Smoking_History']
```

5.4.6 Finalising the Preprocessing Pipeline

Here we combine all the individual pipelines into a main preprocessing pipeline that can process the entire dataset.

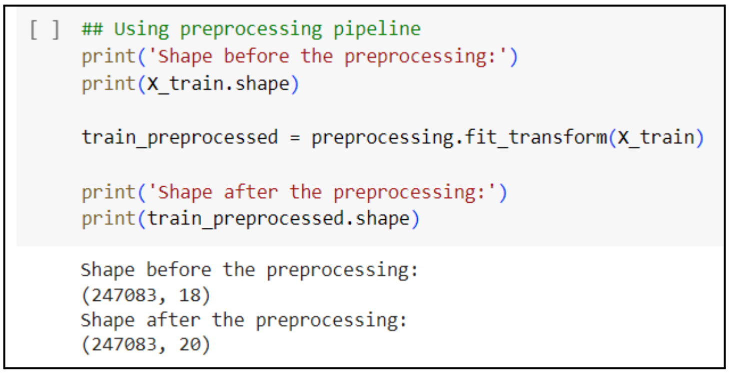


Finalizing the preprocessing pipeline

```
[ ] ## Combining all the pipelines and creating a main pipeline to enter all the data
    preprocessing = ColumnTransformer([
        ('Categorical', cat_pipeline,   cat_pipe_col),
        ('Age_Category',agecat_pipeline,['Age_Category']),
        ('Checkup',checkup_pipeline,['Checkup']),
        ('Gen_health',genhealth_pipeline,['General_Health']),
        ('Numerical',   num_pipeline,  num_pipe_col),
    ],remainder='passthrough')
    preprocessing
```

5.4.7 Previewing the Impact on Dataset Shape

Here we show the shape of the dataset before and after preprocessing to illustrate the impact of these steps.



```
[ ] ## Using preprocessing pipeline
    print('Shape before the preprocessing:')
    print(X_train.shape)

    train_preprocessed = preprocessing.fit_transform(X_train)

    print('Shape after the preprocessing:')
    print(train_preprocessed.shape)
```

```
Shape before the preprocessing:
(247083, 18)
Shape after the preprocessing:
(247083, 20)
```

5.5 Model Development

This section outlines the steps involved in building, configuring, and training the Multi-Layer Perceptron (MLP) model for cardiovascular disease risk prediction.

5.5.1 Defining the MLP Model

5.5.2 Defining the Number of Iterations

5.5.3 Initialise Lists to Store Evaluation Metrics

5.5.4 Output Training Progress

5.6 Model Validation

5.7 Model Optimization

5.8 Model Testing

References

Ahangari N., Yaser, Ansari, Gufran Ahmad, Bhat, Salliah Shafi, Ansari, Mohd Dilshad, Ahmad, Sultan, Nazeer, Jabeen, & Eljialy, A. E. M. (2023). Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction. Computational and Mathematical Methods in Medicine, 2023, 8191261. https://doi.org/10.1155/2023/8191261

Albuquerque, V., Oliveira, A., Barbosa, J. L., Rodrigues, R. S., Andrade, F., Dias, M. S., & Ferreira, J. C. (2021). Smart Cities: Data-Driven Solutions to Understand Disruptive Problems in Transportation—The Lisbon Case Study. Energies, 14(11), Article 11. https://doi.org/10.3390/en14113044

Asenahabi, B. M. (2019). Basics of research design: A guide to selecting appropriate research design. International Journal of Contemporary Applied Research, 6(5), 76–89.

Campos-Staffico, A. M., Cordwin, D., Murthy, V. L., Dorsch, M. P., & Luzum, J. A. (2021). Comparative performance of the two pooled cohort equations for predicting atherosclerotic cardiovascular disease. Atherosclerosis, 334, 23–29. https://doi.org/10.1016/j.atherosclerosis.2021.08.034

Chen, R., Stewart, W. F., Sun, J., Ng, K., & Yan, X. (2019). Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data: Implications for Temporal Modelling With Respect to Time Before Diagnosis, Data Density, Data Quantity, and Data Type. Circulation. Cardiovascular quality and outcomes, 12(10), e005114. https://doi.org/10.1161/CIRCOUTCOMES.118.005114

Dickson, B. (2019, August 5). What are artificial neural networks (ANN)? bdtechtalks.com. https://bdtechtalks.com/2019/08/05/what-is-artificial-neural-network-ann/

Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). The Declaration of Helsinki. BMJ (Clinical research ed.), 335(7621), 624–625. https://doi.org/10.1136/bmj.39339.610000.BE

Grossi, E., & Buscema, M. (2008). Introduction to Artificial Neural Networks. European Journal of Gastroenterology & Hepatology, 19(2), 1046-1054. https://doi.org/10.1097/MEG.0b013e3282f198a0

Han, S. H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dementia and Neurocognitive Disorders, 17(3), 83–89. https://doi.org/10.12779/dnd.2018.17.3.83

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ (Clinical research ed.), 335(7611), 136. https://doi.org/10.1136/bmj.39261.471806.55

Hu, T., Shen, Y., Cao, W., et al. (2023). Two-year changes in body composition and future cardiovascular events: A longitudinal community-based study. Nutrition & Metabolism, 20(1), 4. https://doi.org/10.1186/s12986-023-00727-2

Jahangiry, L., Farhangi, M. A., & Rezaei, F. (2017). Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome. Journal of Health, Population, and Nutrition, 36(1), 36. https://doi.org/10.1186/s41043-017-0114-0

Kumari, S., Kala, V., Lavanya, K., Vidhya, V., Premila, S., & Lawrence, B. (2023). Research Methodology (Vol. 1). Darshan Publishers. URL: https://books.google.co.ke/books?hl=en&lr&id=obG1EAAAQBAJ&oi=fnd&pg=PA9&dq=research+methodology&ots=-32g56XptFGE&sig=fTXOmI4zAMAtq8tDTVP8_Y_M7Zs&redir_esc=y&pli=1#v=onepage&q=research%20methodology&f=false

Lee, S.R., Choi, E.K., Ahn, H.J., Han, K.D., Oh, S., & Lip, G. Y. H. (2020). Association between clustering of unhealthy lifestyle factors and risk of new-onset atrial fibrillation: a nationwide population-based study. Scientific Reports, 10(1), 19224. https://doi.org/10.1038/s41598-020-75822-y

Mandeep, J. S., Madhiarasan, M., & Louzazni, M. (2022). Analysis of Artificial Neural Network: Architecture, Types, and Forecasting Applications. Journal of Electrical and Computer Engineering, 2022, 5416722. https://doi.org/10.1155/2022/5416722

Narkhede, S. (2018, June 26). Understanding AUC - ROC Curve. Towards Data Science. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

National Health Service. (2022, April 22). Cardiovascular disease. https://www.nhs.uk/conditions/cardiovascular-disease/

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. Open Medicine (Warsaw, Poland), 17(1), 1100–1113. https://doi.org/10.1515/med-2022-0508

Powell-Wiley, T. M., Baumer, Y., Baah, F. O., Baez, A. S., Farmer, N., Mahlobo, C. T., Pita, M. A., Potharaju, K. A., Tamura, K., & Wallen, G. R. (2022). Social determinants of cardiovascular disease. Circulation Research, 130(5), 782–799. https://doi.org/10.1161/CIRCRESAHA.121.319811

Rao, D. P., Dai, S., Lagacé, C., & Krewski, D. (2014). Metabolic syndrome and chronic disease. Chronic diseases and injuries in Canada, 34(1), 36–45.

Schiller, D. (2015, April 13). Bridging the Gap Between Healthcare and Technology. Healthcare Innovation Group. https://www.hcinnovationgroup.com/home/article/13024929/bridging-the-gap-between-healthcare-and-technology

Solomatine, D.P., Ostfeld, 2008. Data-driven modelling: some past experiences and new approaches 10, 3–22.

Sunasra, M. (2017, November 11). Performance Metrics for Classification Problems in Machine Learning. Medium. https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Tran, D. T., Silvestri-Elmore, A., & Sojobi, A. (2022). Lifestyle Choices and Risk of Developing Cardiovascular Disease in College Students. International journal of exercise science, 15(2), 808–819.

Vaz, J. M., & Balaji, S. (2021). Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics. Molecular diversity, 25(3), 1569–1584. https://doi.org/10.1007/s11030-021-10225-3

Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Vock, D. M., Musgrove, D., Adomavicius, G., Johnson, P. E., & O'Connor, P. J. (2015). A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. Statistics in Medicine, 34(21), 2941–2957