HEART DISEASE PREDICTION & DATA EXPLORATION

K.L.U.PERERA

Proposal

Introduction

The heart is one of the vital organs of the human body. It pumps blood through the blood vessels of the circulatory system which transports blood, oxygen, and other materials to different organs of the body. The heart plays the most crucial role in the circulatory system. If the heart does not function properly then it will lead to serious health conditions including death. Change in lifestyle, work-related stress, and bad food habits contribute to the increase in the rate of several heart-related diseases and new heart diseases are rapidly being identified. The health of the heart is to be conserved for healthy living. The health of a human heart is based on the experiences in a person's life and is completely dependent on the professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 17 million deaths are occurring worldwide due to the various types of heart diseases which are also known by the term cardiovascular disease. Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels. According to World Health Organization (WHO), by 2030, almost 23.6 million people will die from CVDs, mainly from heart disease and stroke. It's usually associated with a build-up of fatty deposits inside the arteries (atherosclerosis) and an increased risk of blood clots. It can also be associated with damage to arteries in organs such as the brain, heart, kidneys, and eyes. Cardiovascular disease includes coronary artery diseases (CAD) like angina and myocardial infarction (commonly known as a heart attack). There is another heart disease, called coronary heart disease (CHD), in which a waxy substance called plaque develops inside the coronary arteries. A large blood clot can most of the time completely block blood flow through a coronary artery. If the stopped blood flow isn't restored quickly, some sections of the heart muscle begin to die. Without quick treatment, a heart attack can lead to serious health problems and even death. Cardio-Vascular diseases are the primary cause of death worldwide over the past decade. According to the World Health Organization, it is estimated that out of deaths that occur each year because of cardiovascular diseases, 80% is attributed to coronary artery disease and cerebral stroke. Many habitual factors such as personal and professional habits and genetic predisposition account for heart disease. Various risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, blood pressure, diabetes, high blood cholesterol, and pre-existing heart conditions are often deciding factors for heart disease. Therefore prediction and early diagnosis of CVD are very important.

Objectives

- To identify the various factors that mainly contribute to determine the risk of coronary heart disease.
- To predict the risk of a coronary heart disease occurrence using identified crucial factors which are related to coronary heart disease.

Significance of the study

According to World Health Organization (WHO), heart-related diseases are responsible for taking 17.9 million lives every year, 31% of all global deaths. The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high-risk patients and turn reduce their complications. Most heart diseases are highly preventable and simple lifestyle modifications (such as reducing tobacco use and healthy eating habits) coupled with early treatment greatly improve their prognoses. It is, however, difficult to identify high-risk patients because of the nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, etc. Due to such constraints, in recent years researchers and experts working in the medical field have started realizing the immense knowledge available in medical datasets, thereby inspiring medical analysis of data. They are attempting to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk using some techniques. The prediction of heart disease diagnosis by given some features of users is important to medical fields. The term "cardiovascular disease" includes a wide range of conditions. Diagnosis is a complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on a doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. If such a prediction is accurate enough, we can not only avoid the wrong diagnosis but also save human resources. When a patient is wrongly diagnosed with heart disease, he will fall into unnecessary panic or he will miss the best chance to cure his disease. Such a wrong diagnosis is painful to both patients and hospitals. With accurate predictions, we can solve the unnecessary trouble.

Furthermore, it can also aid in devising a monitory and preventive program for those who might be susceptible to suffering from CVD, based on their medical and family history. Modeling and predicting the CVD with the help of the most influential factors is very vital. The efficient, accurate and early medical diagnosis of cardiovascular disease plays a pivotal role in taking preventive measures to avoid the complications that arise due to such diseases.

Background and Data

The data for this analysis comes from the Framingham Heart Study, a long-term and ongoing research project developed to identify risk factors of cardiovascular disease. This study began in 1948 and it was named for Framingham, a town in eastern Massachusetts that had been selected as the site of the study. The dataset consists of 4238 observations with 16 variables including the 'Ten-year risk of coronary heart disease' as the response variable. The data set is available at https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression

Input Variables			
Variable Category	Variable Name	Description	Data Type
Demographic	male	Male or female	Categorical (Nominal)
	age	Age of the patient	Continuous
	Education	No further information provided	Ordinal/ Continuous
Behavioral	currentSmoker	Current smoker or not?	Categorical (Nominal)
	cigsPerDay	Cigarettes per day?	Continuous
Medical History	BPMeds	Blood pressure medication?	Categorical (Nominal)
	prevalentStroke	Whether previously had a stroke?	Categorical (Nominal)
	prevalentHyp	Whether was hypertensive?	Categorical (Nominal)
	diabetes	Whether had diabetes?	Categorical (Nominal)
Current Medical Status	totChol	Total Cholesterol Level	Continuous
	sysBP	Systolic Blood Pressure	Continuous
	diaBP	Diastolic Blood Pressure	Continuous
	BMI	Body Mass Index	Continuous
	heartRate	Heart Rate	Continuous
	glucose	Glucose Level	Continuous
Target variable to predict	TenYearCHD	The 10-year risk of CHD	Categorical (Binary)

Methodology

An exploratory data analysis containing summary statistics and graphical visualizations such as bar charts, histograms, and boxplots will be done to identify patterns of data. Then an advanced data analysis will be conducted to build up a model to predict the risk of coronary heart disease and for evaluating the fitted model. Since detecting cardiovascular disease involves training a model based on a historical dataset, Machine Learning seems to be an appropriate technology to deal with this problem.

Logistic Regression

Logistic regression is a type of regression analysis in statistics used for the prediction of the outcome of a categorical dependent variable (a dependent variable that can take a limited number of values) from a set of predictor or independent variables. In logistic regression, the dependent variable is always binary (with two categories). In the Logistic Regression model, we apply the sigmoid function. This function successfully maps any number into the value between 0 and 1 and we can regard this value as the probability of predicting classes. Logistic regression is mainly used for prediction and also calculating the probability of success.

References

1. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (2021, May 21)
2. Ho, Kalon & Pinsky, Joan & Kannel, William & Levy, Daniel. (1993). The epidemiology of heart failure: The Framingham Study. Journal of the American College of Cardiology. https://www.sciencedirect.com/science/article/pii/073510979390455A (2021, May 21)
3. Wong, Nathan & Levy, Daniel. (2013). Legacy of the Framingham Heart Study: Rationale, Design, Initial Findings, and Implications. https://globalheartjournal.com/articles/abstract/10.1016/j.gheart.2012.12.001/(2021, May 21)
4. Britannica, T. Editors of Encyclopaedia (2020, October 15). Coronary heart disease. Encyclopedia Britannica. https://www.britannica.com/science/coronary-heart-disease (2021, May 21)

Abstract

Heart disease is a major health burden and the primary cause of death worldwide over the past few decades. Heart diseases also known as cardiovascular diseases encompass a wide range of conditions that affect the heart. These vary from blood vessel diseases, heart rhythm problems to heart defects that one is born with. Coronary Heart Disease (CHD) is the major type of cardiovascular disease which leads to the majority of death all around the world. Also, healthcare expenditures are overwhelming national and corporate spending plans due to asymptomatic diseases including cardiovascular diseases. Therefore, there is an urgent need for early detection and treatment of such diseases to reduce the death rates as well as the health expenditure. The field of medical analysis is often referred to be a valuable source of rich information. The information which is gathered by healthcare studies and data analysis of hospitals are utilizing for analyzing purposes to obtain an accurate and reliable approach to achieving an early diagnosis of the disease. Data Science plays an important in processing large amounts of data in the field of medical sciences. The challenge lies in the complexity of the data and correlations when it comes to prediction using conventional techniques. Therefore the researchers utilize several data mining and machine learning (ML) techniques to analyze large sets of data and aid in the right prediction of heart diseases. This study aims to identify the various factors that mainly contribute to determine the risk of coronary heart disease and to predict the risk of a coronary heart disease occurrence using identified crucial factors which are related to coronary heart disease. The scope of this research is limited to using two supervised learning techniques namely logistic regression and the K-Nearest Neighbour algorithm to discover correlations in CHD data that might help to improve the prediction rate and to carry out a comparative study. Using the Framingham Heart Disease dataset of 4238 initial instances, intelligent models are derived by the considered ML techniques. Further, data analysis is carried out in Python using JupyterLab to validate the accuracy of algorithms. The exploratory data analysis emphasized some observable associations and patterns between the response variable, TenYearCHD, and features such as sysBP, glucose, age, cigsPersDay, totChol, prevalentHyp, male, BPMeds, and diabetes. The features like glucose, age, sysBP, diaBP, cigsPerDay, and totChol have shown positive associations with the response. Also, the data visualizations have identified that conditions like the prevalence of diabetes, prevalence of hypertension, and taking blood pressure medications are associated with the response in a positive manner. Also, total cholesterol and age have implied a positive association between them. There are some major findings on joint effects on the response, 'Ten-year risk of CHD' are identified as well. The joint effect of prevalence of hypertension and higher heart rates, the joint effect of higher age and higher systolic blood pressure and, the joint effect of higher blood pressure levels and smoking habit have shown a higher tendency of getting a heart disease. According to the advanced analysis results, both the logistic regression model and the K-Nearest Neighbour algorithm are somewhat more sensitive than specific. It is found that the K-Nearest Neighbour algorithm provides the best accuracy of 91.89% in comparison to the logistic regression model. Empirical results using different performance evaluation measures report that the K-Nearest Neighbour algorithm is promising in detecting CHD.

Table of Contents

List of Tables

List of Figures

1. Introduction

The heart is one of the main organs of the human body. It pumps blood through the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood, oxygen, and other materials to different organs of the body. The heart plays the most crucial role in the circulatory system. If the heart does not function properly then it will lead to serious health conditions including death. Change in lifestyle, work-related stress, and bad food habits contribute to the increase in the rate of several heart-related diseases and new heart diseases are rapidly being identified. The health of the heart is to be conserved for healthy living. The health of a human heart is based on the experiences in a person's life and is completely dependent on the professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 17 million deaths are occurring worldwide due to the various types of heart diseases which are also known by the term cardiovascular disease.

Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels. It's usually associated with a build-up of fatty deposits inside the arteries (atherosclerosis) and an increased risk of blood clots. It can also be associated with damage to arteries in organs such as the brain, heart, kidneys, and eyes. Cardiovascular disease includes coronary artery diseases (CAD) like angina and myocardial infarction (commonly known as a heart attack). There is another heart disease, called coronary heart disease (CHD), in which a waxy substance called plaque develops inside the coronary arteries. These are the arteries that supply oxygen-rich blood to the heart muscle. When plaque begins to build up in these arteries, the condition is called atherosclerosis. The development of plaque occurs over many years. With time, this plaque can harden or rupture (break open). Hardened plaque eventually narrows the coronary arteries which in turn reduces the flow of oxygen-rich blood to the heart. If this plaque ruptures, a blood clot can form on its surface. A large blood clot can most of the time completely block blood flow through a coronary artery. Over time, the ruptured plaque also hardens and narrows the coronary arteries. If the stopped blood flow isn't restored quickly, some sections of the heart muscle begin to die. Without quick treatment, a heart attack can lead to serious health problems and even death. Heart attack is a common cause of death worldwide. Cardiovascular diseases are the primary cause of death worldwide over the past decade. According to the World Health Organization, it is estimated that out of deaths that occur each year because of cardiovascular diseases, 80% is attributed to coronary artery disease and cerebral stroke. Many habitual factors such as personal and professional habits and genetic predisposition account for heart disease. Various risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, blood pressure, diabetes, high blood cholesterol, and pre-existing heart conditions are often deciding factors for heart disease.

The major challenge faced in the world of medical sciences today is the provision of quality service and efficient and accurate prediction. The whole accuracy in the management of disease lies in the proper time of detection of that disease. Therefore the healthcare industry collects huge amounts of health care data that can be used properly to discover hidden information, to take decisions effectively, to discover the relations that connect patterns. If the data at hand is used to develop screening and diagnostic models, it will not only reduce the strain on medical personnel but also aid early detection and prompt treatment for patients thereby drastically enhancing the health system. The efficient, accurate and early medical diagnosis of cardiovascular disease plays a pivotal role in taking preventive measures to avoid the complications that arise due to such diseases.

Objectives

- To identify the various factors that mainly contribute to determine the risk of coronary heart disease.
- To predict the risk of a coronary heart disease occurrence using identified crucial factors which are related to coronary heart disease.

2. Significance of the study

According to World Health Organization (WHO), heart-related diseases are responsible for taking 17.9 million lives every year, 31% of all global deaths. The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high-risk patients and turn reduce their complications. Most heart diseases are highly preventable and simple lifestyle modifications (such as reducing tobacco use, healthy eating habits, reducing obesity, and exercising) coupled with early treatment greatly improve their prognoses. It is, however, difficult to identify high-risk patients because of the nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, etc. According to World Health Organization (WHO), by 2030, almost 23.6 million people will die from CVDs, mainly from heart disease and stroke. These are projected to remain the single leading causes of death. Due to such constraints, in recent years researchers and experts working in the medical field have started realizing the immense knowledge available in medical datasets, thereby inspiring medical analysis of data. Also, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease. They are attempting to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk using some techniques.

Our problem is that we want to predict whether patients have heart disease by given some features of users. This is important to the medical field. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and how blood is pumped and circulated through the body. Diagnosis is a complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on a doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. If such a prediction is accurate enough, we can not only avoid the wrong diagnosis but also save human resources. When a patient without heart disease is diagnosed with heart disease, he will fall into unnecessary panic and when a patient with heart disease is not diagnosed with heart disease, he will miss the best chance to cure his disease. Such a wrong diagnosis is painful to both patients and hospitals. With accurate predictions, we can solve the unnecessary trouble. Furthermore, it can also aid in devising a monitory and preventive program for those who might be susceptible to suffering from cardiovascular diseases, based on their medical and family history. Modeling and predicting cardiovascular disease with the help of the most influential factors is very vital. Therefore a highly accurate and efficient diagnosis model leads to better patient treatments as well as for early prevention of the disease.

3. Background

Framingham Heart Study, a long-term research project developed to identify risk factors of cardiovascular disease, the findings of which had far-reaching impacts on medicine. Indeed, much common knowledge about heart disease including the effects of smoking, diet, and exercise can be traced to the Framingham study. The study’s findings further emphasized the need for preventing, detecting, and treating risk factors of cardiovascular disease in their earliest stages. The Framingham Heart Study began in 1948. It was named for Framingham, a town in eastern Massachusetts that had been selected as the site of the study. The project was initiated under the direction of the National Heart Institute, which was newly established in 1948 (renamed the National Heart, Lung, and Blood Institute [NHLBI] in 1976). From 1971, through a contract with the institute, the study was carried out in collaboration with the Boston University School of Medicine. At the time, little was known about the underlying causes of heart disease and stroke, but the death rates for cardiovascular disease (CVD) had been increasing steadily since the beginning of the 20th century and had become an American epidemic. Framingham Heart Study began on a cohort of 5,209 men and women between the ages of 30 and 62 recruited from the town of Framingham, Massachusetts. Currently, the study is still ongoing with the participation of the third generation, the grandchildren of the Original Cohort. The study concluded that major CVD risk factors are: high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity. In addition to those major factors, there are several related factors such as blood triglyceride and HDL cholesterol levels, age, gender, and psychosocial issues. The selected dataset from this study consists of 4238 observations with 16 variables including the ten-year risk of coronary heart disease as the variable of interest.

4. Methodology

An exploratory data analysis containing summary statistics and graphical visualization will be done to identify patterns of data. Then an advanced data analysis will be conducted to build up a model to predict the risk of coronary heart disease and for evaluating the fitted model.

4.1 Machine Learning (ML) Techniques

Machine learning techniques allow the use of intelligent methods across different datasets to reveal useful insights. ML is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. This reprogrammable ability of ML in exploring, processing, and interpreting datasets makes it favorable for decision-makers in domains such as medical diagnosis. Since detecting CVD involves training a model based on a historical dataset, ML seems to be an appropriate technology to deal with this problem.

4.2 Logistic Regression

Logistic Regression is supervised learning that computes the probabilities for classification problems with two outcomes. It is a type of regression analysis in statistics used for the prediction of the outcome of a categorical dependent variable (a dependent variable that can take a limited number of values) from a set of predictor or independent variables. In logistic regression, the dependent variable is always binary (with two categories). In the Logistic Regression model, we apply the sigmoid function, which is

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

This function successfully maps any number into the value between 0 and 1 and we can regard this value as the probability of predicting classes. Logistic regression is mainly used for prediction and also calculating the probability of success.

4.3 K-Nearest Neighbour

The K-Nearest Neighbour algorithm is a supervised classification algorithm method. It classifies objects depending on the nearest neighbour. It is a type of instance-based learning. The calculation of the distance of an attribute from its neighbors is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them. K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search.

5. Data

5.1 Data Acquisition

The dataset is collected from the Kaggle website,

available at:
https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression

5.2 Data Description

The dataset consists of 4238 observations with 16 variables including the 'Ten-year risk of coronary heart disease' (TenYearCHD) as the response variable.

Table 5. 1 - Table of data description

Input Variables			
Variable Category	Variable Name	Description	Data Type
Demographic	male	Male or female	Categorical (Nominal)
	age	Age of the patient	Continuous
	Education	No further information provided	Ordinal/ Continuous
Behavioral	currentSmoker	Current smoker or not?	Categorical (Nominal)
	cigsPerDay	Cigarettes per day?	Continuous
Medical History	BPMeds	Blood pressure medication?	Categorical (Nominal)
	prevalentStroke	Whether previously had a stroke?	Categorical (Nominal)
	prevalentHyp	Whether was hypertensive?	Categorical (Nominal)
	diabetes	Whether had diabetes?	Categorical (Nominal)
Current Medical Status	totChol	Total Cholesterol Level	Continuous
	sysBP	Systolic Blood Pressure	Continuous
	diaBP	Diastolic Blood Pressure	Continuous
	BMI	Body Mass Index	Continuous
	heartRate	Heart Rate	Continuous
	glucose	Glucose Level	Continuous
Target variable to predict	TenYearCHD	The 10-year risk of CHD	Categorical (Binary)

5.3 Pre-Processing

In order to build up a more accurate ML model, data pre-processing is required. Data pre-processing is the process of cleaning the data. It will remove all the NAN values from our data. This process is also known as Data Wrangling. This includes the identification of missing data, noisy data, and inconsistent data.

5.3.1 Data Cleaning

I. **Missing values –** The initial dataset consisted of 582 data records with missing values. The summarized total number of missing values based on the attributes is given below.

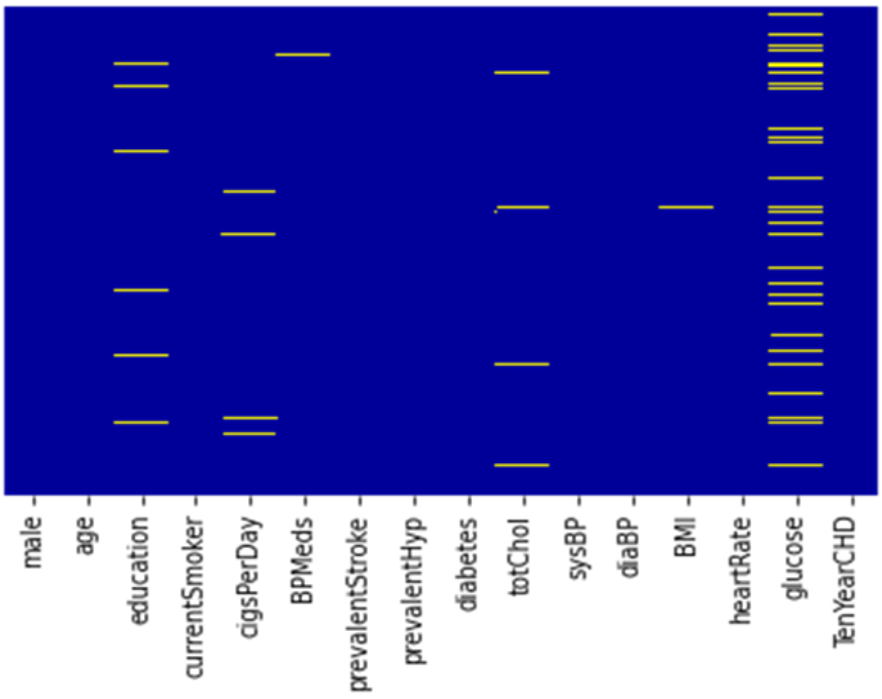

Figure 5. 1- Visualization of missing values

Table 5. 2 - Missing value count

Variable	Missing value count for the variable
education	105
cigsPerDay	29
BPMeds	53
totChol	50
BMI	19
heartRate	1
glucose	388

The percentage of missing values is only 14 percent of the entire dataset. But considering the further reduction of missing values, the variable 'glucose' is imputed since it consists of the majority of missing values. As for the feature 'glucose', it can be noticed that the heat map for correlations indicates a somewhat higher correlation between 'glucose' and 'diabetes' and the dataset is imbalanced. Also, many types of research, medical studies, and web journals (such as Villines ,2019, Medical News Today; Diabetes Care by the American Diabetes Association) have emphasized that 'glucose' level is highly associated with 'diabetes'. Therefore the feature 'diabetes' is used to fill missing values in 'glucose'. After the imputation, the point-biserial correlation is calculated and checked the significance of the association between 'diabetes' and 'glucose' is below.

Table 5.3 – Point-Biserial Result

Point-Biserial Result	
Correlation	0.62519
P-value	0.0

The association is identified as significant and therefore the analysis is continued with this imputation. After the imputation of 'glucose', the total number of rows with missing values is 251 and since it is only 6 percent of the entire dataset the other records with missing values are excluded.

II. **Duplicates** – No duplicate records consisted in the study dataset.

III. **Outliers** – Some removable outliers are identified before the advanced analysis as shown below.

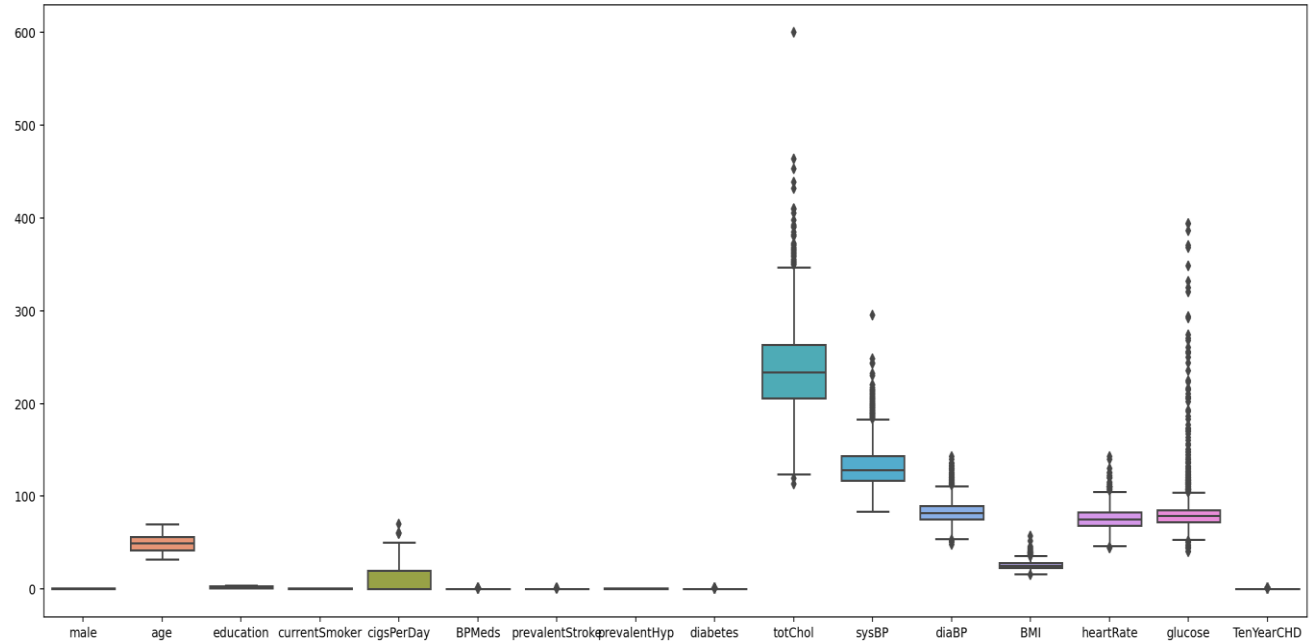

Figure 5. 2 – Visualization of outliers

There were two removable outliers are detected in 'totChol' and 'sysBP' variables of the dataset. Outliers in all other numerical variables are important and thus cannot be removed. The outlier in 'totChol' with the value of 600 is removed since it is highly unlikely compared to the normal human total cholesterol level. Also, the outlier in 'syBP' has shown a highly unlikely value of 295 compared to the normal systolic blood pressure range therefore it is removed from the dataset.

5.3.2 Data Transformation

Resampling

The initial dataset is imbalanced. The visualization of the response variable, the 10-year risk of Coronary Heart Disease (TenYearCHD) has shown below.

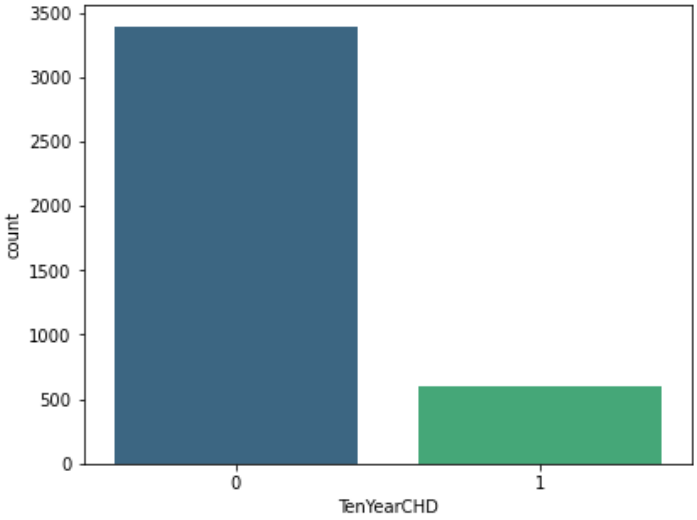

Figure 5. 3 – Visualization of response variable before data balancing

Therefore the imbalanced data is resampled by using the oversampling method in the advance analysis to increase the accuracy and reliability of the model fitting and prediction process. The visualization of the response variable 'TenYearCHD' after balancing the data has shown below.

```
1    3392
0    3392
Name: TenYearCHD, dtype: int64
```

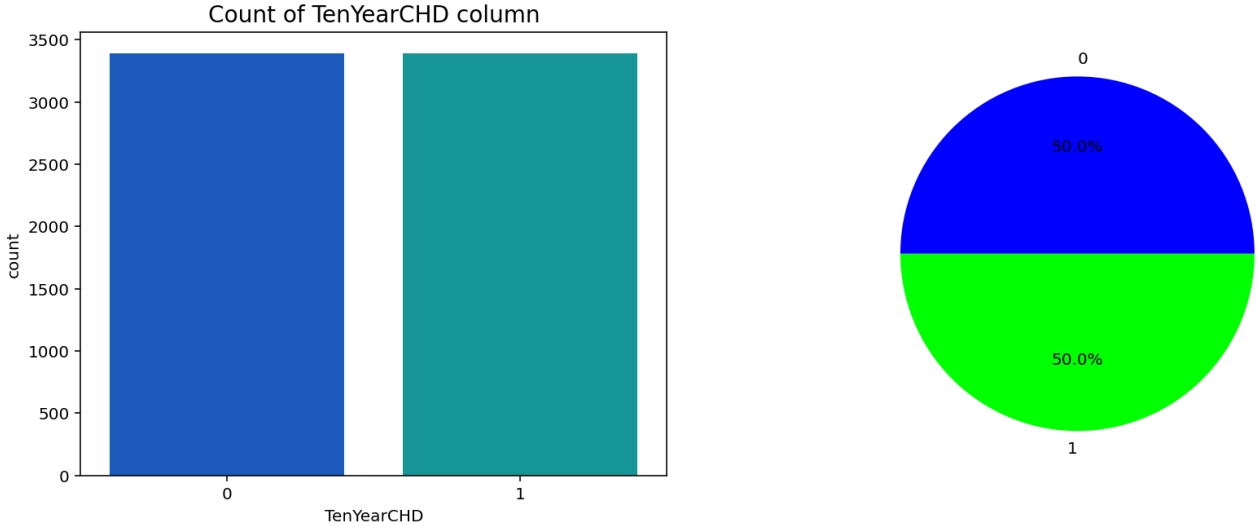

Figure 5. 4 – Visualization of response variable after data balancing

Data Splitting

The data set was separated into training and testing sets for the evaluation process of advanced analysis. 80% of the available data is randomly assigned to the training set and the remaining 20% to the validation set.

Data Scaling

Feature(data) scaling is the method used to standardize the range of features of data. Since the range of values of data may vary widely, it becomes a necessary step in data preprocessing. In this study, the data is scaled using the MinMaxScaler in python for constructing the final model. In this scaler, the minimum of feature is made equal to zero and the maximum of feature equal to one. MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

6. Data Analysis

Data Analysis was carried out on the Jupyter Notebook for further classification, using Python.

6.1 Univariate Analysis

Descriptives

	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
count	3987.000000	3987.000000	3987.000000	3987.000000	3987.000000	3987.000000	3987.000000	3987.00000
mean	49.478806	9.020316	236.620517	132.222724	82.861174	25.774650	75.873840	81.66466
std	8.531588	11.914558	44.019766	21.949243	11.882166	4.079846	12.087463	22.99468
min	32.000000	0.000000	113.000000	83.500000	48.000000	15.540000	44.000000	40.00000
25%	42.000000	0.000000	206.000000	117.000000	75.000000	23.060000	68.000000	72.00000
50%	49.000000	0.000000	234.000000	128.000000	82.000000	25.380000	75.000000	79.00000
75%	56.000000	20.000000	263.000000	143.500000	89.500000	27.990000	83.000000	85.00000
max	70.000000	70.000000	600.000000	295.000000	142.500000	56.800000	143.000000	394.00000

Figure 6. 1 - Descriptives

Table 6. 1 - Table of mode values for continuous variables

Variable	Mode
age	40
cigsPerDay	0
totChol	240
sysBP	120
diaBP	80
BMI	22.54
heartRate	75
glucose	79

6.1.1 male

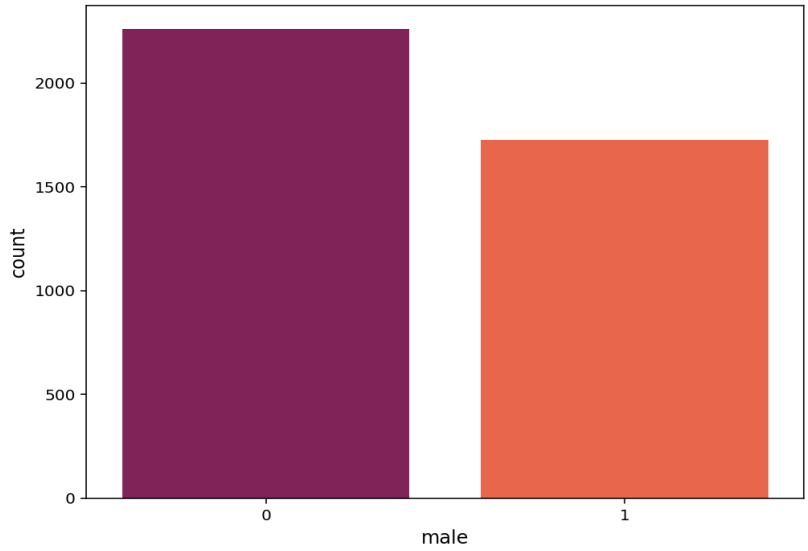

Figure 6. 2 – Bar chart of male

As shown in the above figure, the variable is binary and the study dataset has the majority of observations for females. But the difference in count between male and female is not much higher. Therefore the effect on modeling results and conclusions associated with this feature is applicable in general since it is not biased for one gender category.

6.1.2 education

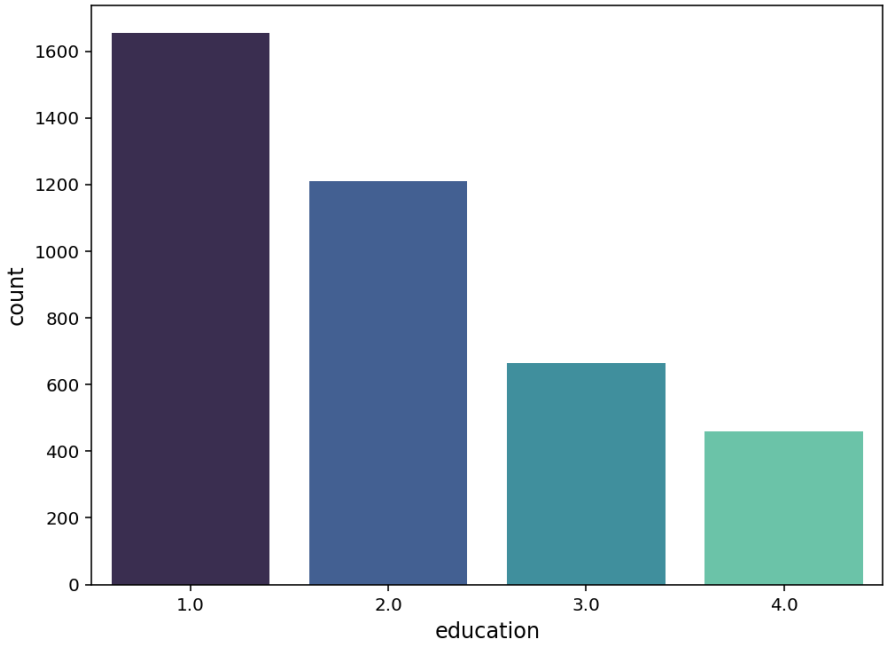

Figure 6. 3 – Bar chart of education

In this study, we do not have much information about the scope of 'education' on heart risk. As shown in the above figure, bars have ordered in descending order concerning the level of education. Sufficient data is not provided about the education categories. But when considering the value counts it seems like an ordinal variable.

6.1.3 currentSmoker (Current smoker or not)

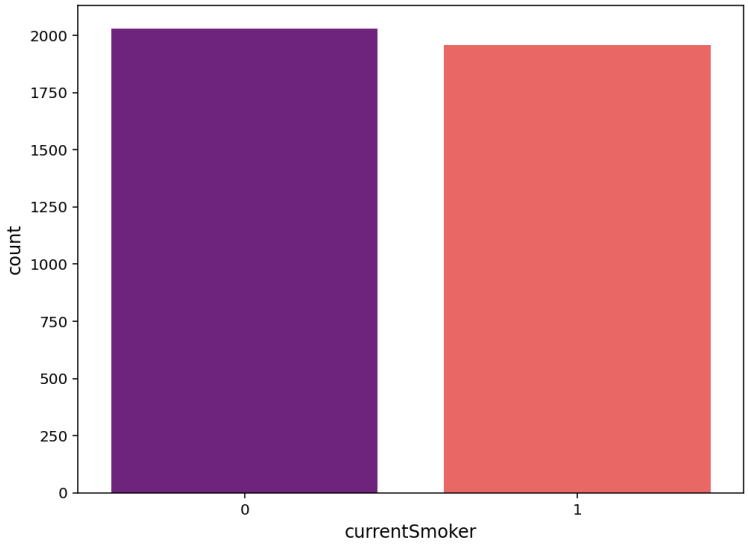

Figure 6. 4 - Bar chart of currentSmoker

The above bar plot depicts that the 'currentSmoker' is binary and roughly balanced. The representation of Smokers and non-Smokers in currentSmoker variable is almost the same. Therefore the study data consists of balanced data for this variable and the effect on modeling results associated with currentSmoker is not biased towards one category. Therefore the conclusions are more general on smoking habit.

6.1.4 BPMeds (Blood pressure medication)

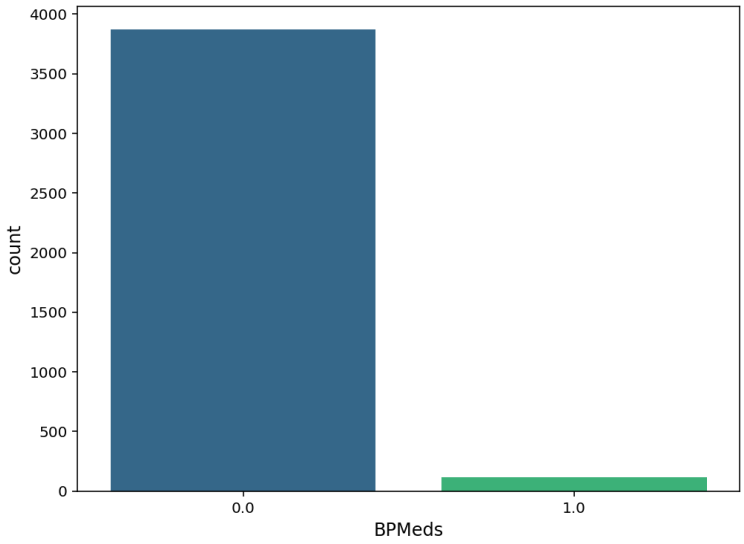

Figure 6. 5 - Bar chart of BPMeds

The above bar plot depicts that the 'BPMeds' variable is binary and highly imbalanced. The number of people who take blood pressure medication is very less compared to the people who do not take blood pressure medications. Therefore the modeling and prediction results associated with this feature will be biased due to the imbalanced nature of data. The conclusions are more applicable for people who do not take blood pressure medication more than the people who take blood pressure medication.

6.1.5 prevalentStroke (Whether previously had a stroke)

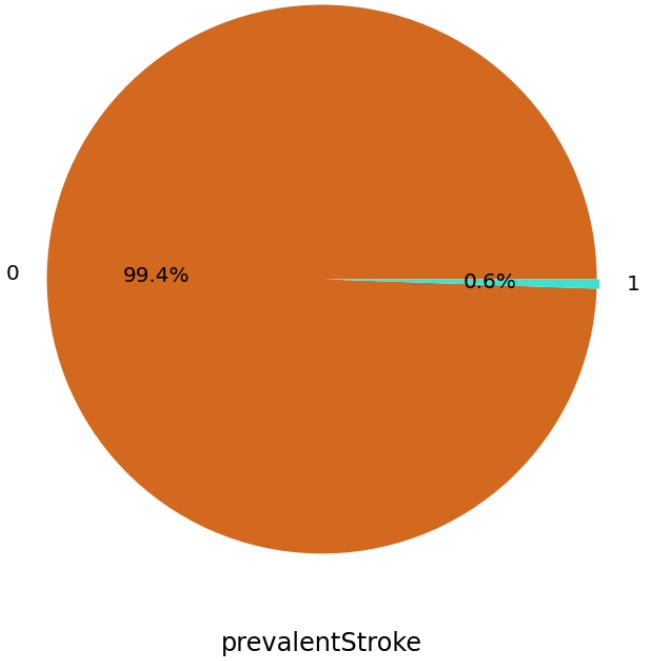

Figure 6. 6 - Pie chart of prevalentStroke

As shown in the above pie chart, 'prevalentStroke' is a binary variable. The percentage of people who previously had a stroke is very less. It is almost negligible since it is less than 1%. On the other side, 99.4% of the majority have not previously had a stroke. Therefore the data in this variable is unbalanced.

6.1.6 prevalentHyp (Whether was hypertensive)

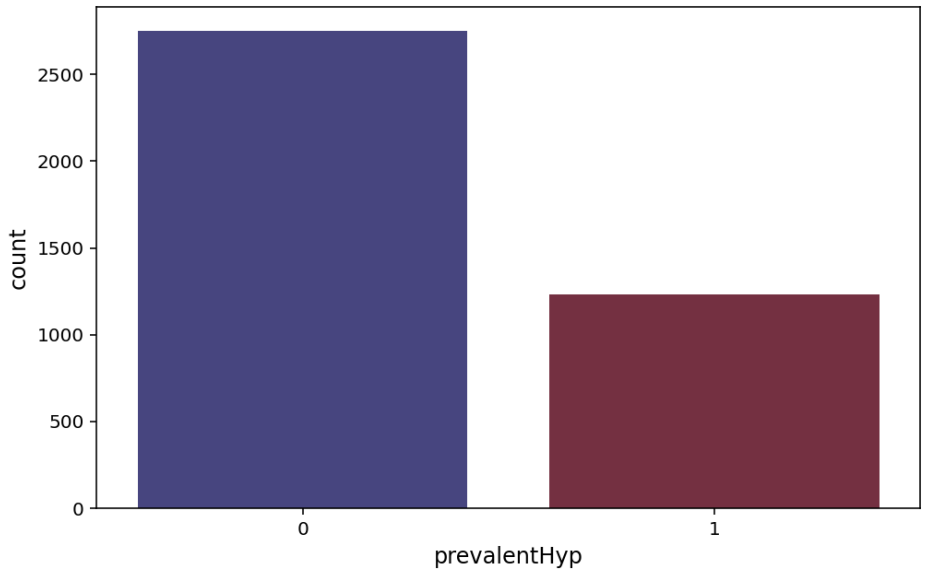

Figure 6. 7 – Bar chart of prevalentHyp

It can be observed in the above bar plot, the majority of people were not hypertensive since the 'prevalentHyp = 0' category represents the majority of observations. The number of people who were hypertensive seems to be roughly half of the number of people who were not hypertensive. The variable is binary.

6.1.7 diabetes (Whether had diabetes)

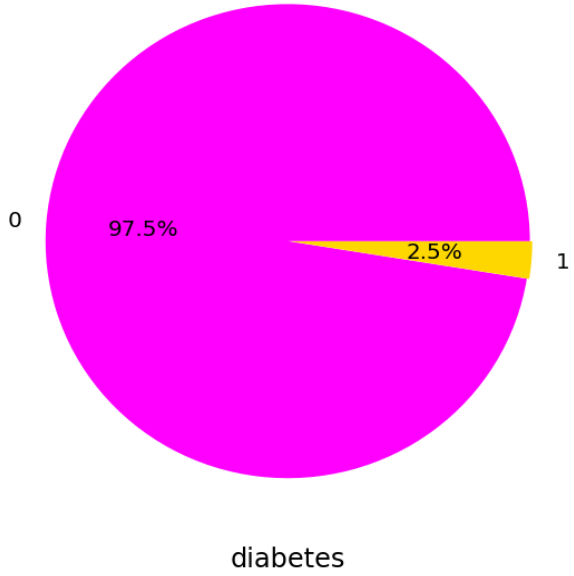

Figure 6. 8- Pie chart of diabetes

The above figure depicts that 'diabetes' is a binary variable with the majority from the category of people who had no diabetes. It is roughly 2.5% of the minority had diabetes. Therefore this variable consists of unbalanced data.

6.1.8 cigsPerDay (Cigarettes per day)

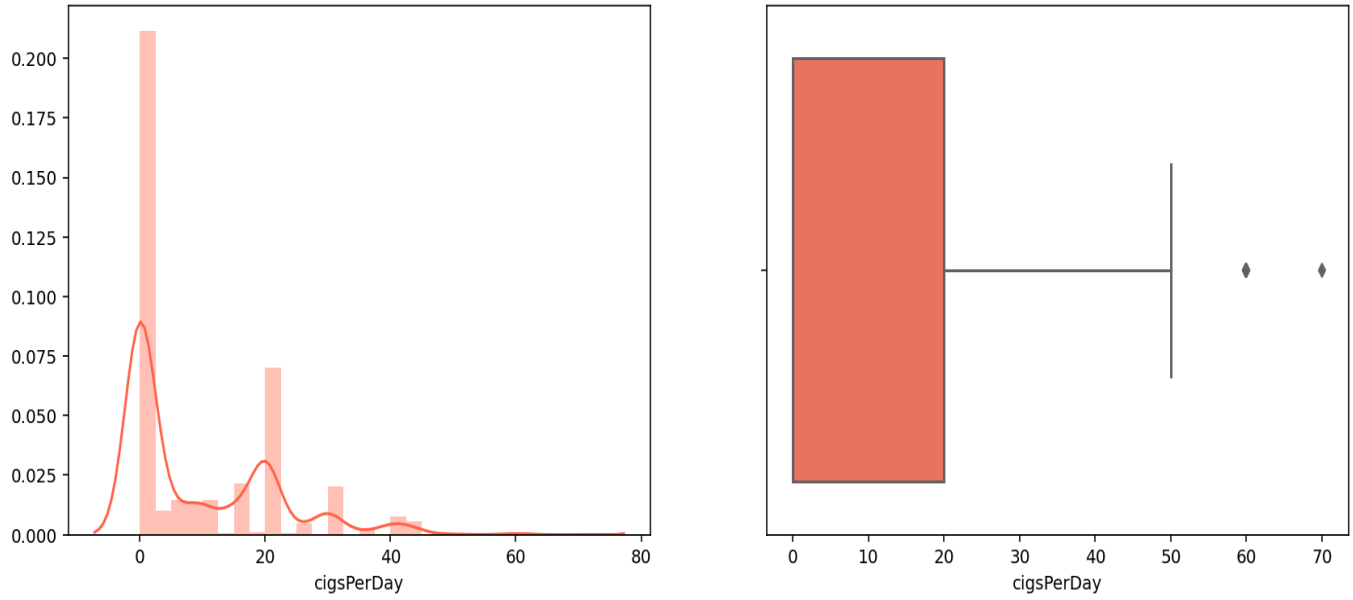

Figure 6. 9 – Distribution plot and boxplot of cigsPerDay

As shown in the distribution plot in the above figure, the distribution of 'cigsPerDay' seems to be right-skewed and multimodal with most data present in zero. The above boxplot depicts that the data distribution has two outliers. The median of the distribution seems to be zero. Two outliers around 60 and 70 indicate that those persons are having 60 and 70 cigarettes per day which is unusual but there is some possibility to happen. Therefore these outliers should be investigated.

6.1.9 age

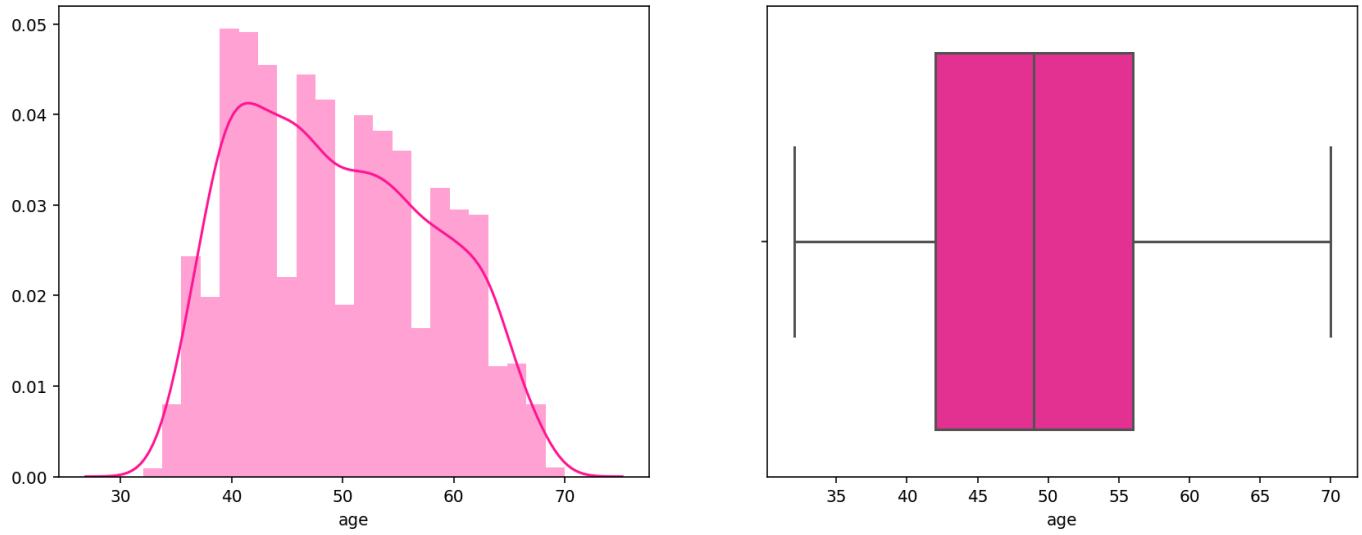

Figure 6. 10 - Distribution plot and boxplot of age

The two plots in the above figure indicate that the age distribution is somewhat normal with a slight skewness. The mean and the median values around 49. The data distribution of age is concentrated between 35 and 65. It seems to be the observed data range of age is much suitable for the study since that age range consists of the majority of people with a higher tendency of getting heart disease in general.

6.1.10 BMI (Body Mass Index)

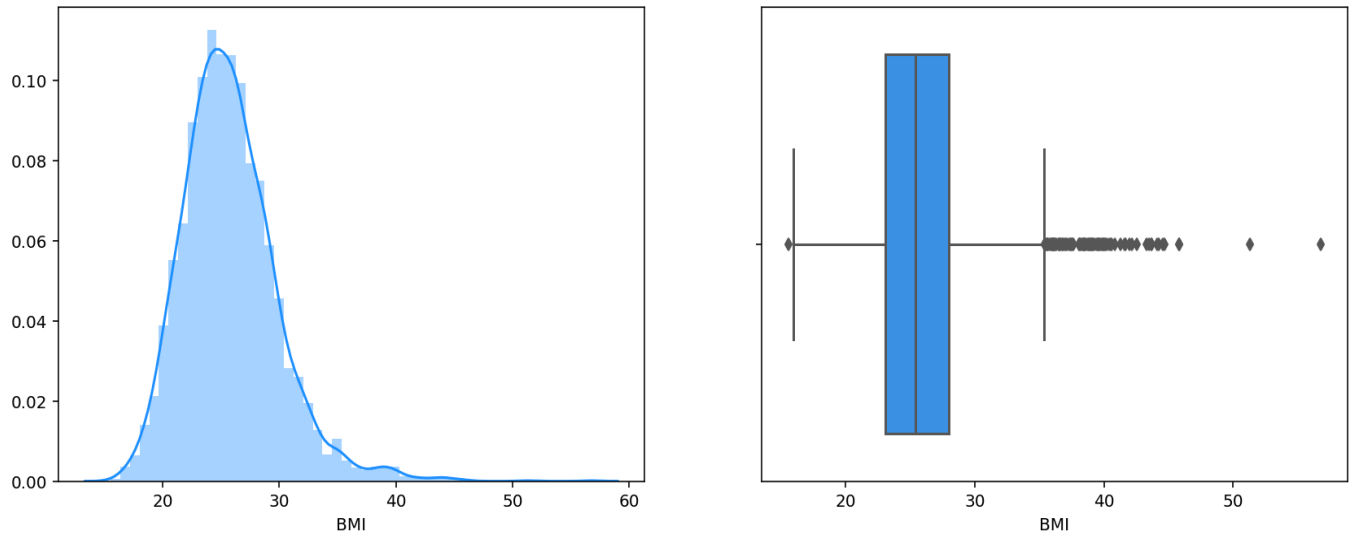

Figure 6. 11 - Distribution plot and boxplot of BMI

As shown in the above figure, the distribution of 'BMI' appears to be unimodal and roughly normal. But there is a slight skewness of data at the right tail. The above boxplot indicates the majority of outliers exist at the right tail of the distribution while the left tail consists of a few. This may be the reason for the skewness of the distribution. But these outliers are possible since 'BMI' values can be affected by some physical conditions. The mean and the median appears to be roughly equal to 25 and the majority of the data is concentrated in the 20-32 range.

6.1.11 totChol (Total Cholesterol Level)

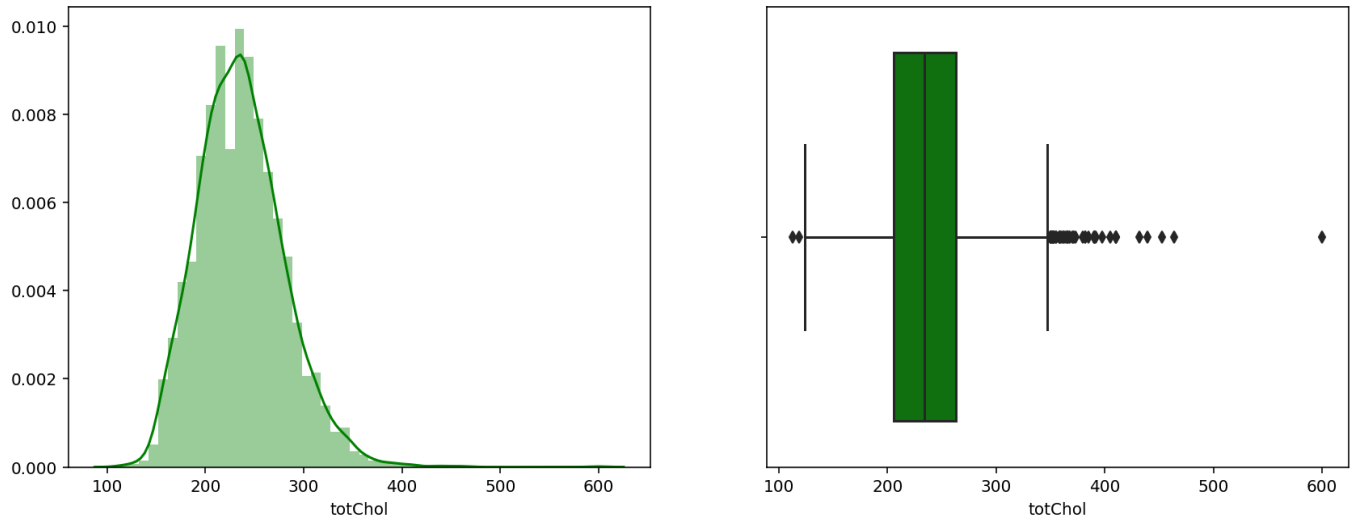

Figure 6. 12 - Distribution plot and boxplot of totChol

The distribution of 'totChol' appears to be roughly normal and unimodal. But the distribution indicates a slight skewness at the right tail. The boxplot indicates some outliers at both tails of the distribution but the right tail consists of some significant outliers. This may be a possible reason for the slight skewness of 'totChol' data. There may be people who have some higher or lesser unusual total cholesterol levels. But there is a very significant outlier at the right tail which is at 600 and it is somewhat far away from other outliers. This indicates the necessity of an investigation since that observation is very unusual in general. The data distribution is highly concentrated between 150 and 350.

6.1.12 sysBP (Systolic Blood Pressure)

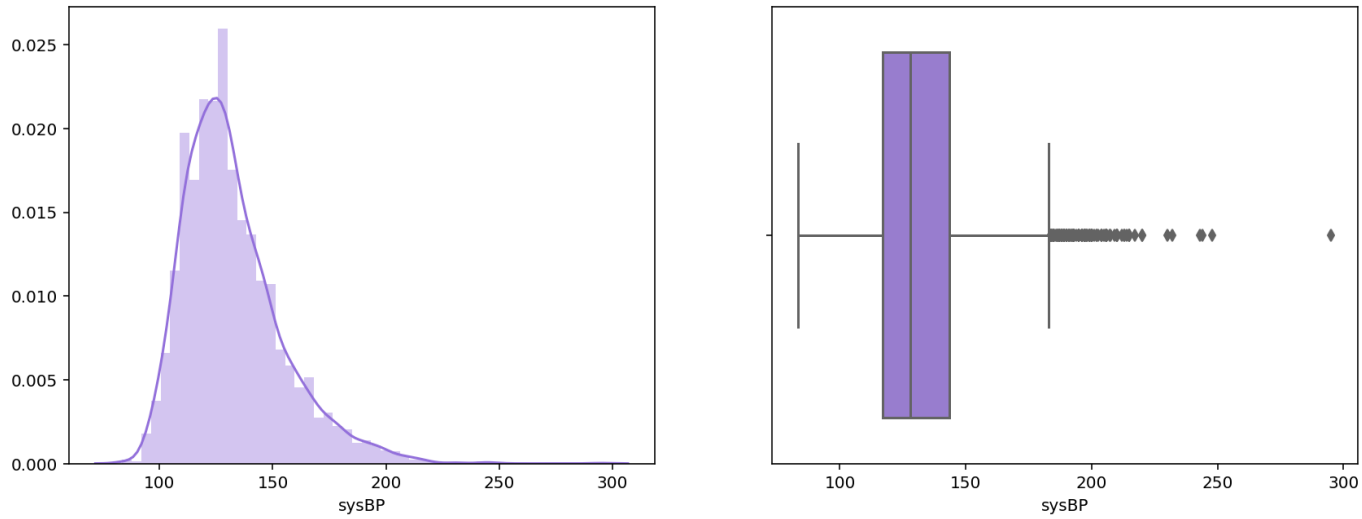

Figure 6. 13 - Distribution plot and boxplot of sysBP

As shown above, the distribution of 'sysBP' is a bit rightly skewed and unimodal. The majority of values lie between 100 and 175. The above boxplot depicts that the right tail consisted of some outliers. It may be possible since some people may have high systolic blood pressure values. There is a very significant outlier at the right end around 300 and it should be investigated since it is somewhat far away from other outliers. It appears to be the median of the distribution is in the 125-130 range.

6.1.13 diaBP (Diastolic Blood Pressure)

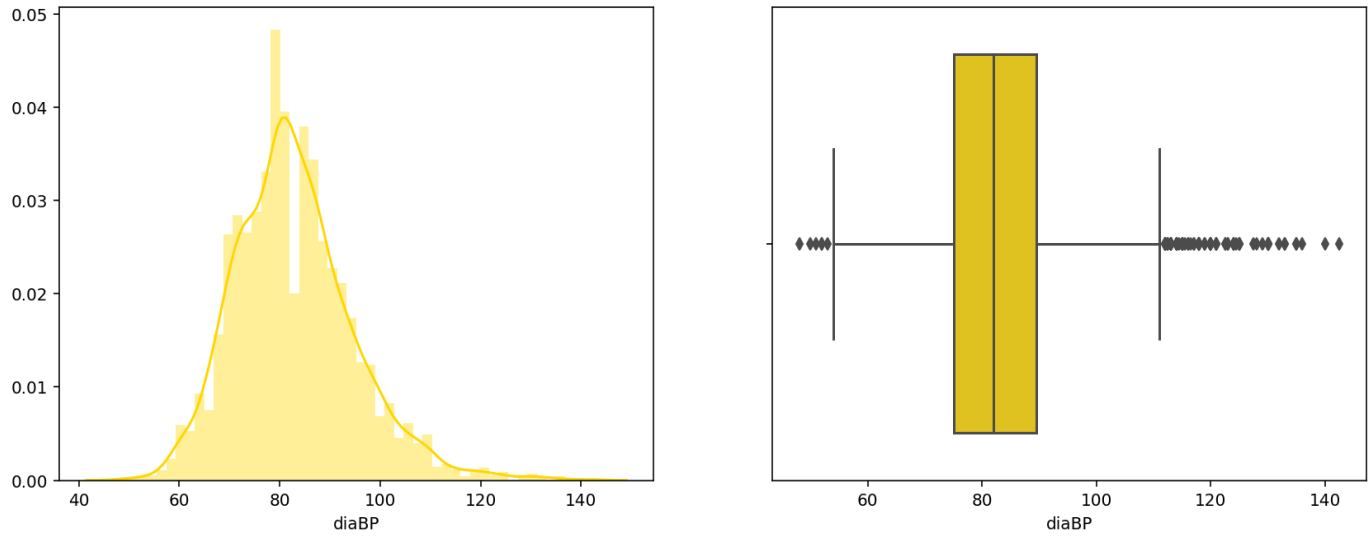

Figure 6. 14 - Distribution plot and boxplot diaBP

The distribution of 'diaBP' appears to be roughly normal with some values concentrated in both tails. As shown in the above boxplot, some outliers can be observed at both ends. It may be possible since some people can have higher or lesser diastolic blood pressure values with their health conditions. The majority of 'diaBP' values lie in between the 65-105 range. The mean and the median appears to be lying around 82.

6.1.14 heartRate

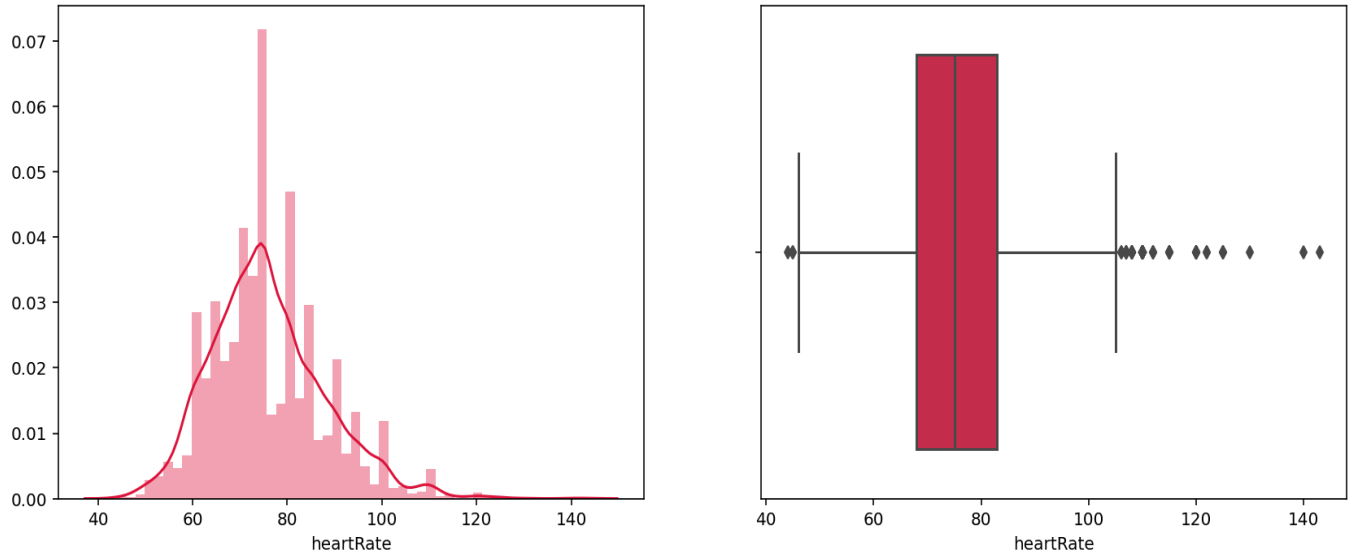

Figure 6. 15 - Distribution plot and boxplot of heartRate

The 'heartRate' distribution appears to be roughly normal and unimodal. But it consists of a somewhat skewness at the right tail of the distribution. The boxplot depicts that some outliers exist at both tails which justifies the skewness of the distribution. The majority of outliers exist at the right tail indicating very high heart rates. These situations may occur due to various health conditions. It seems like both the mean and the median are around 75 and the distribution is concentrated between 60 and 90.

6.1.15 glucose (Glucose Level)

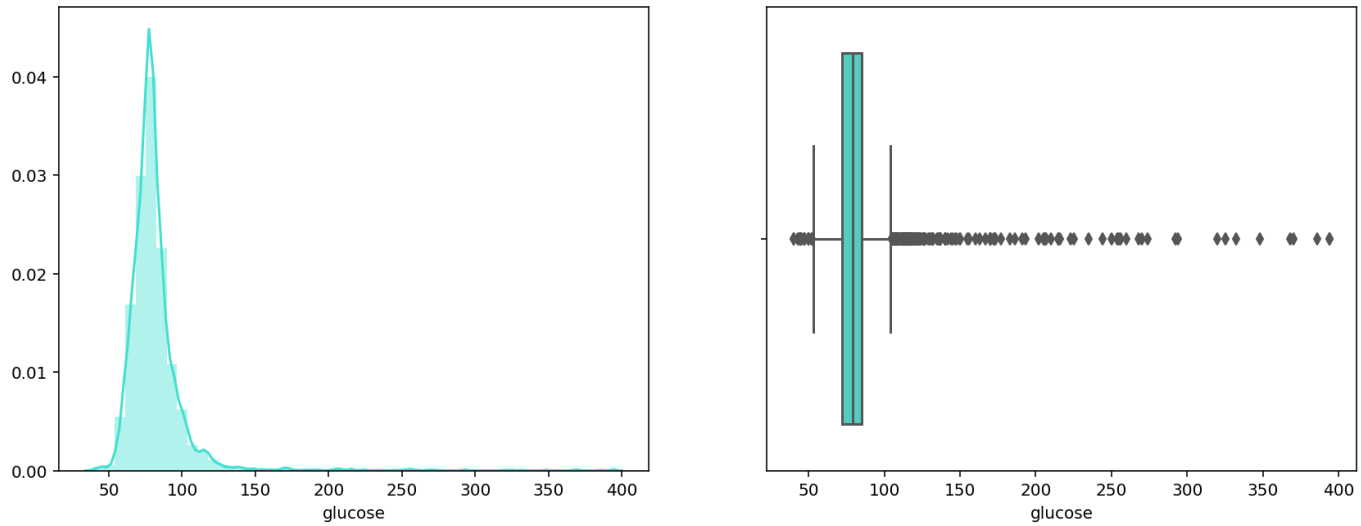

Figure 6. 16 - Distribution plot and boxplot of glucose

As shown above, the distribution of 'glucose' is somewhat normal but slightly right-skewed. The boxplot indicates a series of outliers at the right tail and some at the left tail. This may be a possible reason for the observed skewness of the distribution. But these values may be possible since glucose values can be affected by the time which the measurement was taken (fasting or non-fasting glucose levels) and by the various health conditions. The distribution is concentrated within the 50-105 range. It appears to be both the mean and the median are between 77 and 82.

6.1.16 TenYearCHD (The 10-year risk of CHD)

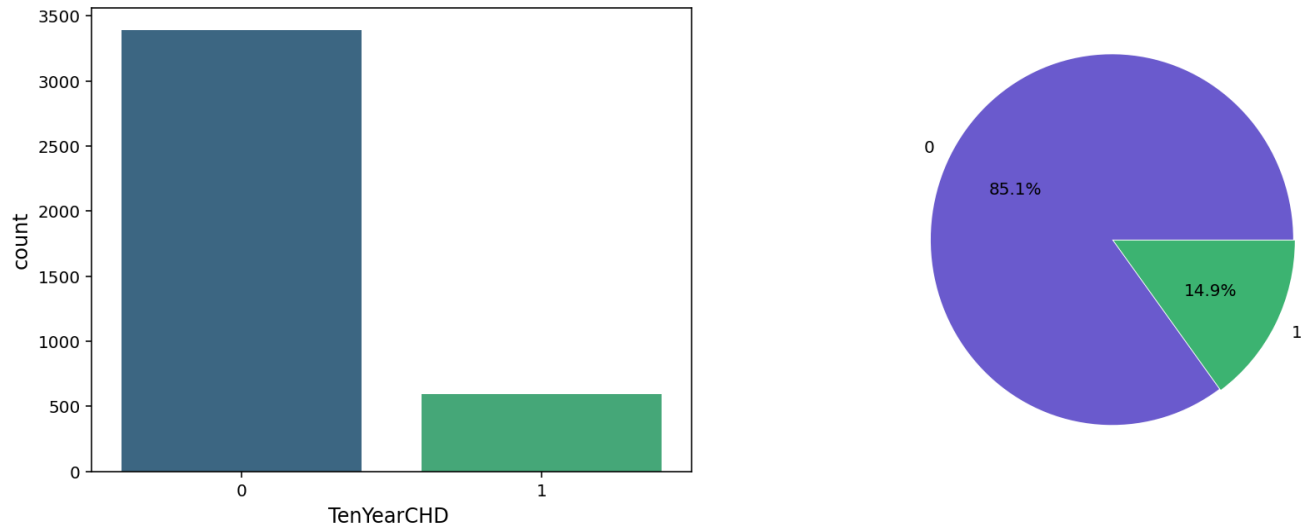

Figure 6. 17 – Bar chart and pie chart of heartRate

The above plots depict that the response variable of the study is binary and highly imbalanced. The majority belongs to the category of 'not having a 10-year risk of CHD' which is around 85%. The category of 'having a 10-year risk of CHD' consists only around 15% of study data. Therefore the modeling and prediction results are biased for the group of not having a 10-year risk of CHD. Since the bias nature, the conclusions are more applicable to that group. Therefore, this problem needs to be addressed and taken care of to obtain more generally applicable results and conclusions.

6.2 Bivariate Analysis

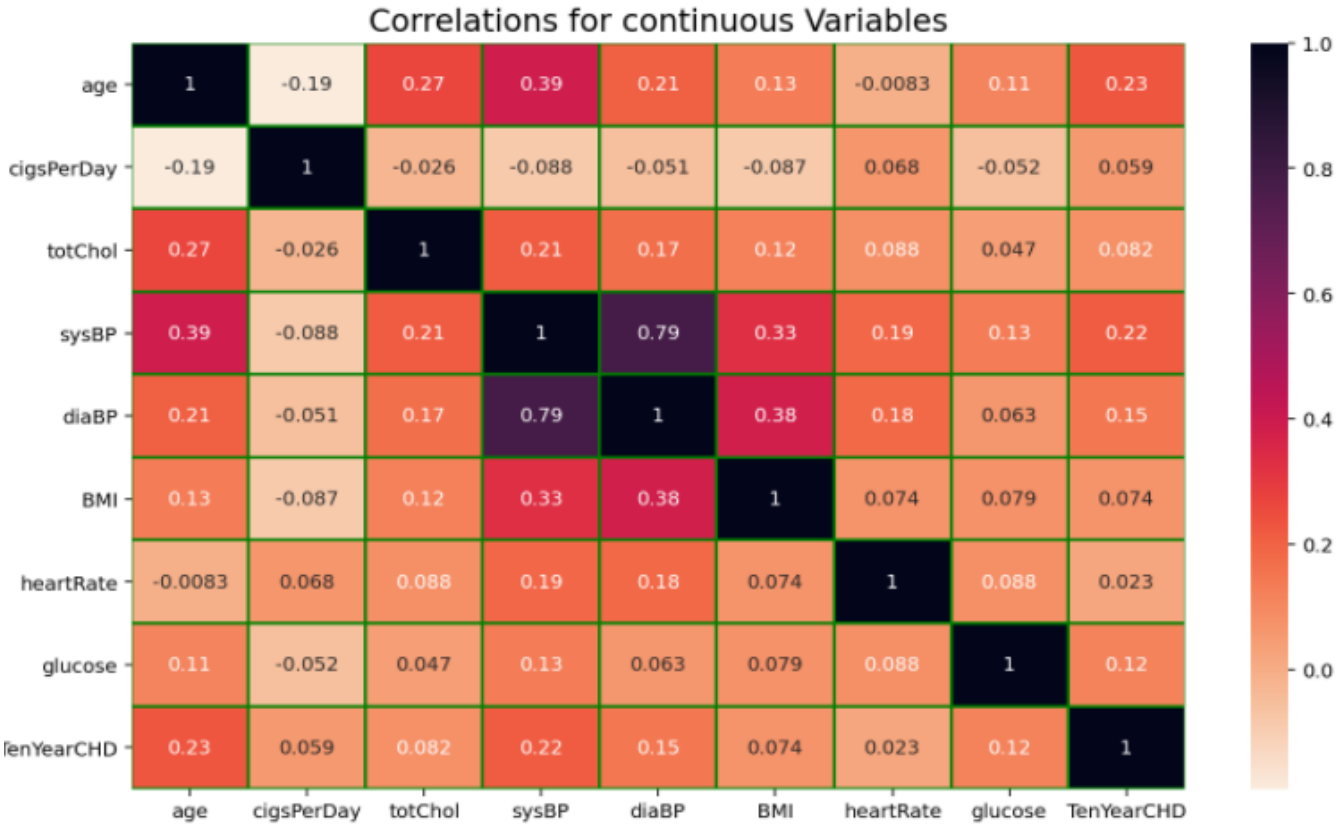

Figure 6. 18 – Visualization of correlations for continuous variables including the response variable

Only the systolic blood pressure and the diastolic blood pressure have shown a considerable association.

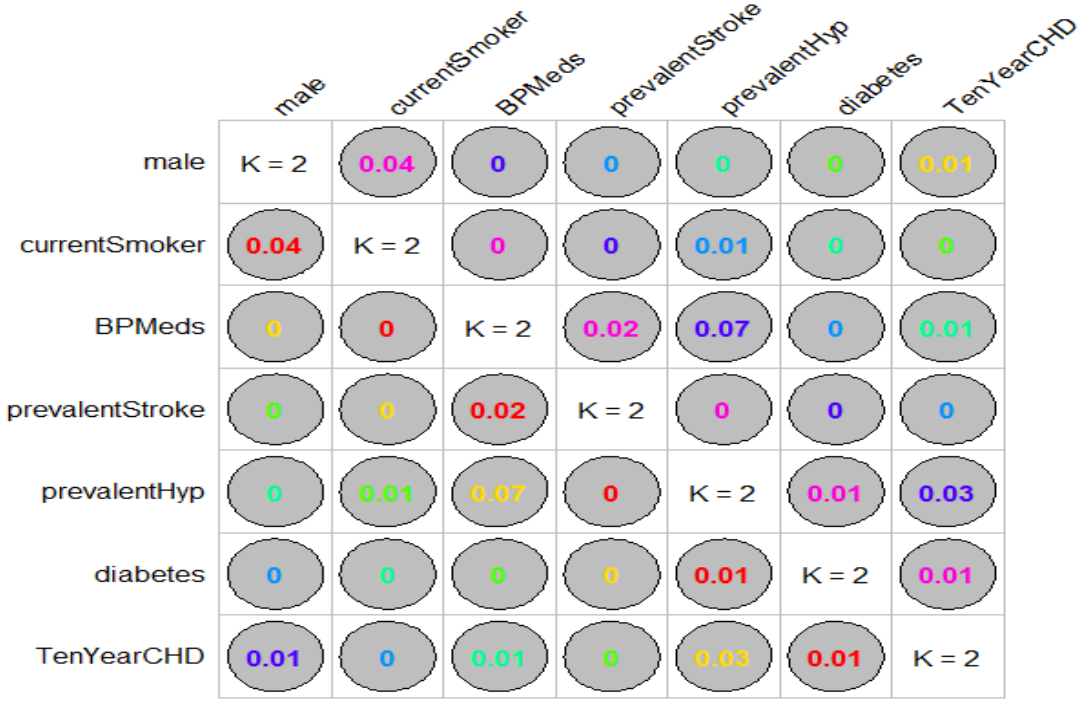

Figure 6. 19 - Visualization of Goodman-Kruskal correlations for categorical variables

No considerably significant associations can be observed between categorical variables.

6.2.1 male vs TenYearCHD

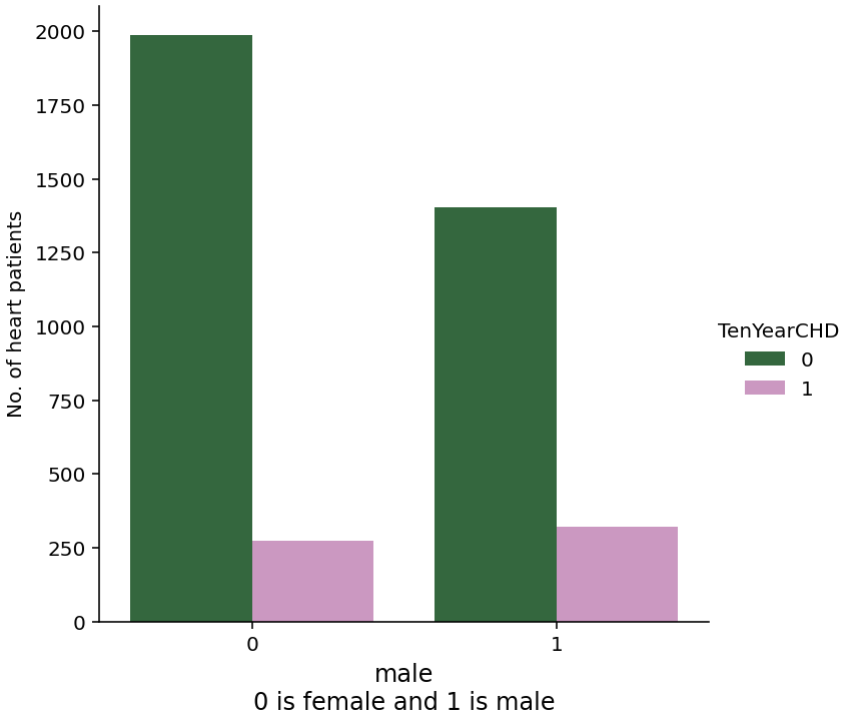

Figure 6. 20 – Multiple bar chart of male vs TenYearCHD

It can be observed that the majority of males are having a risk of CHD compared to females. The cause for this observation may be the bad health habits in the day-to-day life of males like daily smoking and frequent liquor consumption. Also, men's coping with stressful events may be less adaptive physiologically, behaviorally, and emotionally, contributing to their increased risk for CHD in general.

6.2.2 age vs TenYearCHD

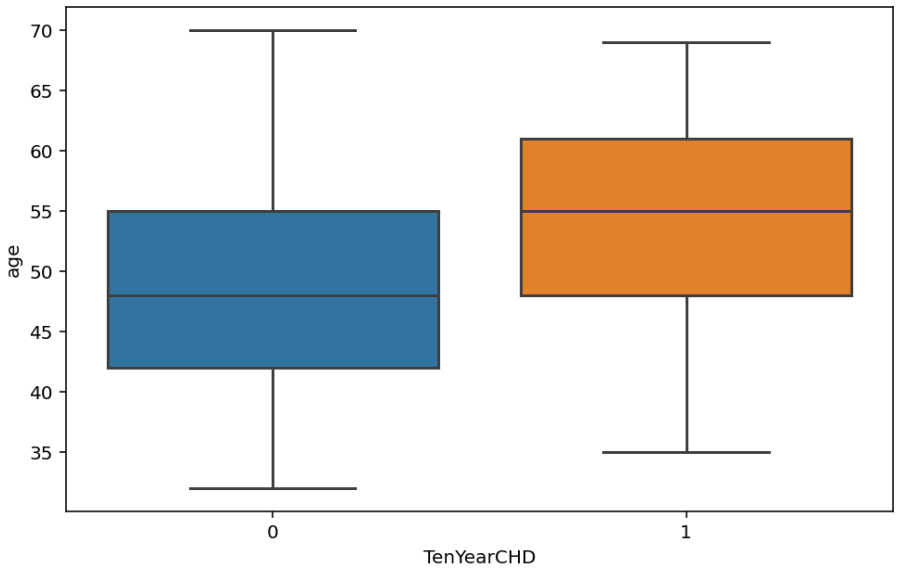

Figure 6. 21 – Boxplots of age vs TenYearCHD

It can be observed that the people with a 10-year risk of CHD are having a higher age distribution compared to the people with no 10-year risk of CHD. This observation is more general since older people are having a higher tendency of getting heart disease due to their long-term bad habits and weak condition of the cardiovascular system.

6.2.3 currentSmoker vs TenYearCHD

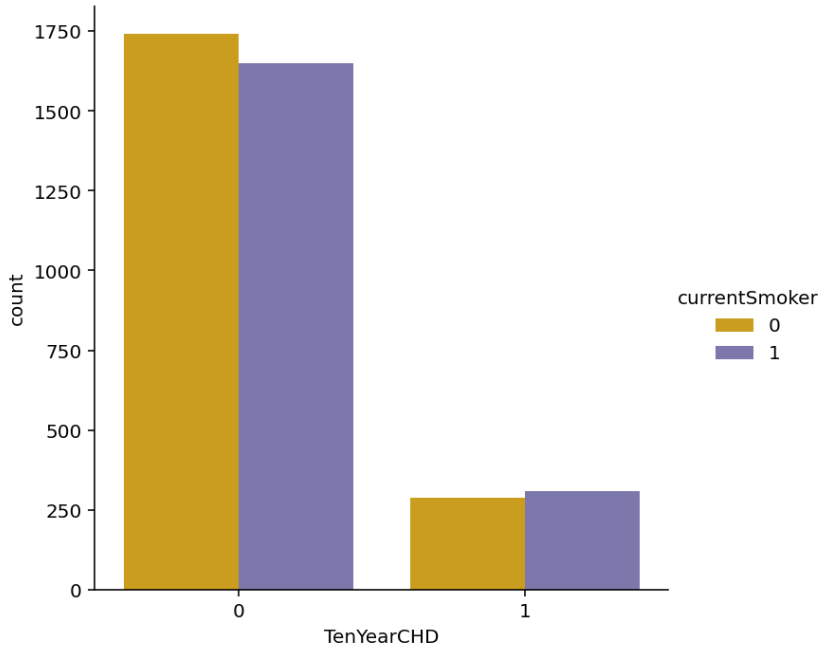

Figure 6. 22 – Multiple bar chart of currentSmoker vs TenYearCHD

As shown in the above figure, smoking habits are not much difference between the two categories of 'TenYearCHD'. But in the '10-year risk of CHD' category, current smokers are slightly higher than the non-smokers, and in the 'no 10-year risk of CHD' category, non-smokers are slightly higher than the smokers. Therefore it justifies the current health study findings since smoking habits affect heart diseases in general.

6.2.4 cigsPerDay vs TenYearCHD

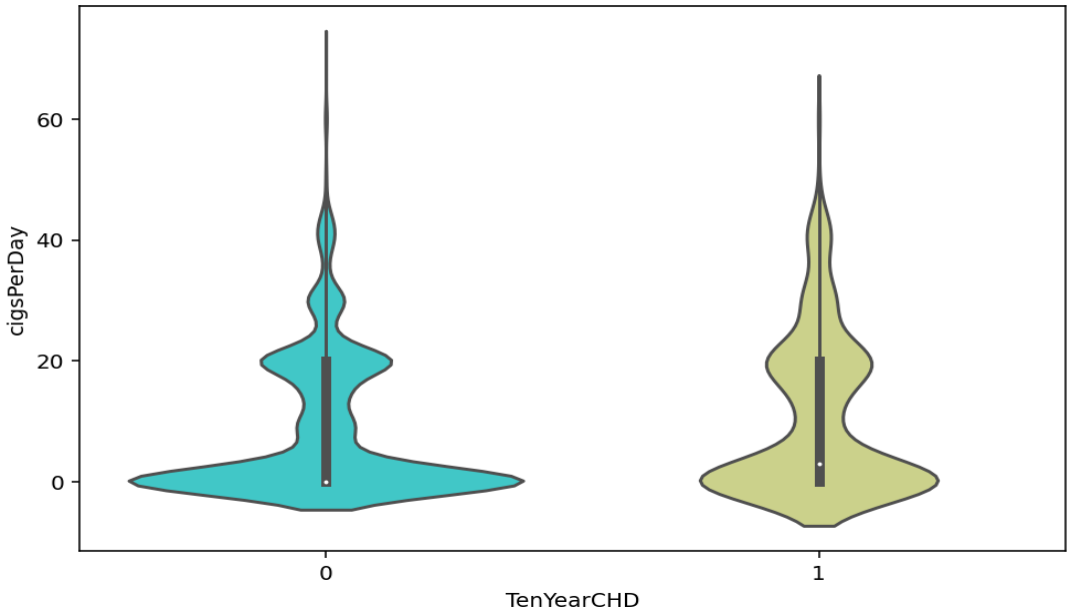

Figure 6. 23 – Violin plots of cigsPerDay vs TenYearCHD

The above violin plots depict that the 'cigsPerDay' has uneven distributions for both categories. The median number of cigarettes per day is higher for the people with a risk of CHD compared to the people with no risk of CHD. For both groups, most of the observations lie between 0-5 and for the risk group, the distribution has a slightly higher concentrated distribution for higher 'cigsPerDay' values compared to the non-risk group. This is the general case since a smoking higher number of cigarettes per day may lead to getting early heart diseases.

6.2.5 BPMeds vs TenYearCHD

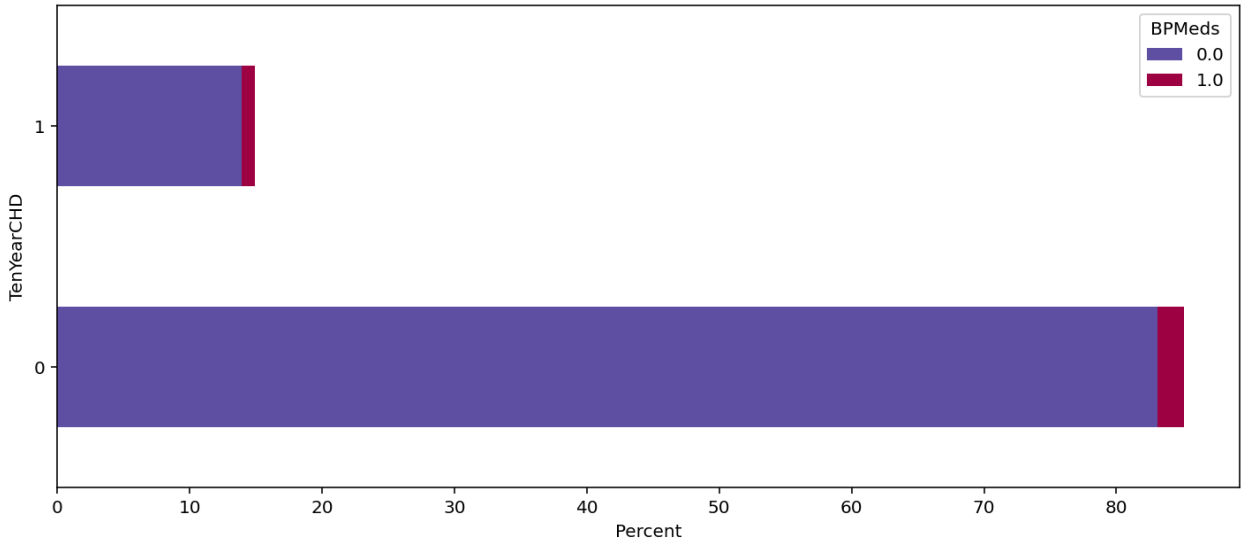

Figure 6. 24 – Stacked bar chart of BPMeds vs TenYearCHD

As shown in the above stacked bar plot, both groups consist of the majority of people who do not take blood pressure medications. Since the dataset is unbalanced, the difference between the two categories is not much observable. But it appears to be the proportion of people who take blood pressure medications within the risk category is somewhat higher than the proportion of people in the non-risk group. Therefore it leads to the general conclusion since blood pressure abnormalities may affect heart diseases.

6.2.6 prevalentHyp vs TenYearCHD

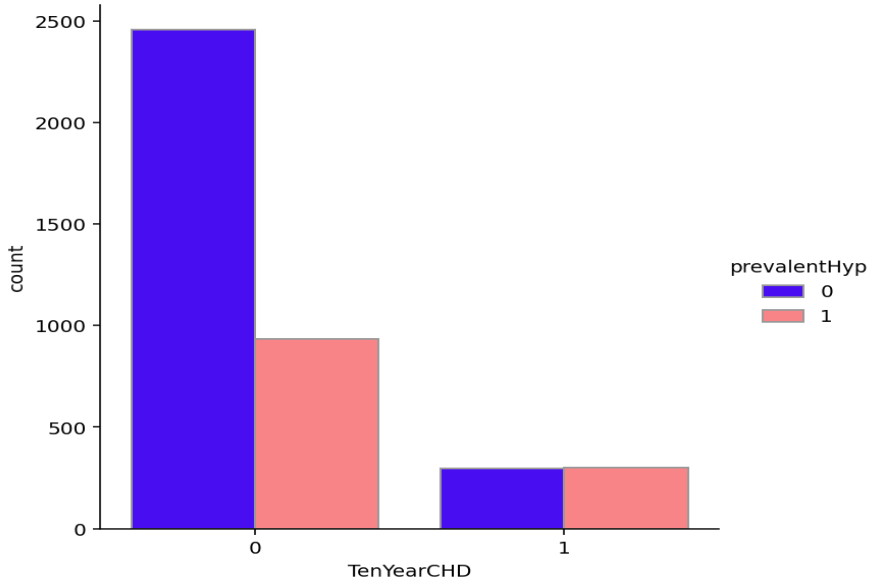

Figure 6. 25 – Multiple bar chart of prevalentHyp vs TenYearCHD

The above multiple bar plot depicts that the number of people who were hypertensive and who weren't hypertensive are roughly equal for the group of 'having a risk of CHD'. But in the 'non-risk of CHD' group, the number of people who were not hypertensive is more than twice of the hypertensive people. These observations suggest that the risk of CHD is higher for hypertensive people. According to the day-to-day cases of heart disease and findings in the medical field, this observation is more general since hypertension may lead to heart diseases.

6.2.7 totChol vs TenYearCHD

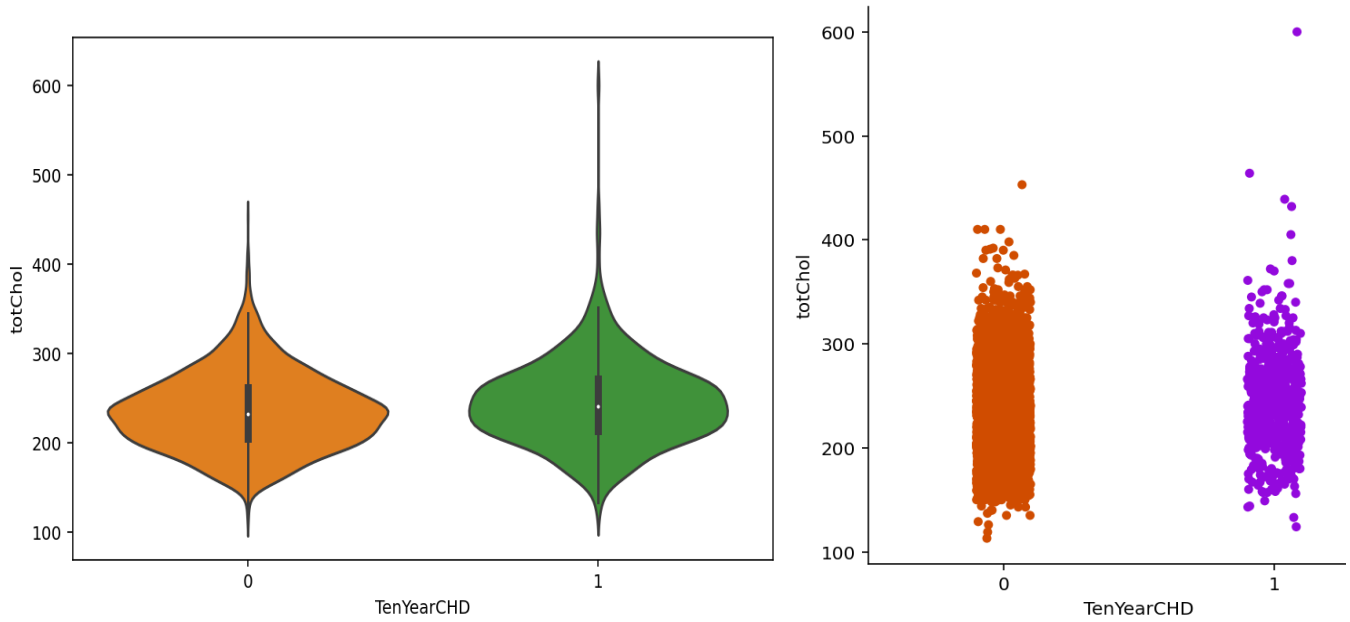

Figure 6. 26 – Violin charts and strip plot of totChol vs TenYearCHD

The above plots depict that the median for the CHD risk group is slightly higher than the median for the CHD non-risk group. Both distributions indicate high density within the 200-300 cholesterol level while the risk group density is somewhat higher than the non-risk group. The CHD risk group has shown a highly skewed distribution with some higher total cholesterol level observations compared to the non-risk group when considering the data imbalance. These observations can be generalized since the cholesterol level is a widely identified factor for heart diseases in many medical studies (Peters et al., 2016). But the effect of total cholesterol is not much significant here since total cholesterol contains some good cholesterol.

6.2.8 sysBP vs TenYearCHD

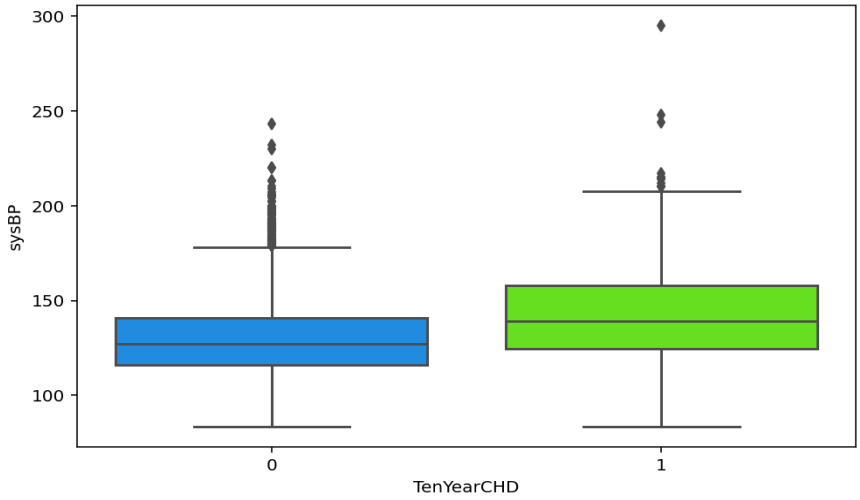

Figure 6. 27 – Boxplots of sysBP vs TenYearCHD

It can be observed that when comparing the two distributions of 'sysBP' for the 'risk of CHD' group and the 'no-risk of CHD' group, the distribution for the 'risk of CHD' group is slightly higher than the 'distribution for non-risk of CHD' group. Also, the median value is higher for the risk group compared to the non-risk group. Both groups contain some higher systolic blood pressure values but there are some significantly higher values in the 'CHD risk' group. It seems like the 'sysBP' has shown some tendency towards the CHD. This observation is justifiable since systolic blood pressure damages arteries, making them more vulnerable to narrowing and may lead to heart diseases in general.

6.2.9 diaBP vs TenYearCHD

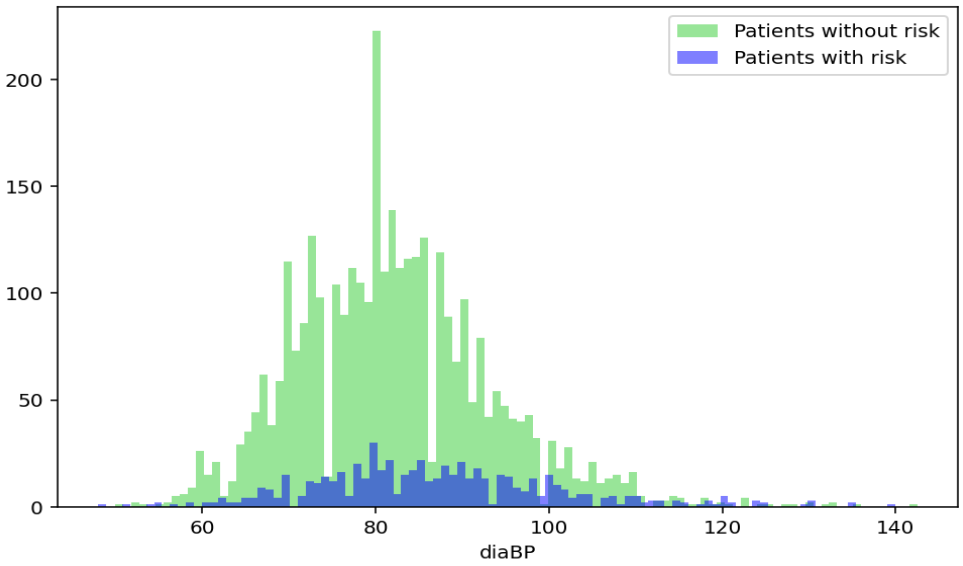

Figure 6. 28 – Histogram of diaBP vs TenYearCHD

As shown in the above figure, both risk and non-risk group distributions for 'diaBP' have shown a high density in 60-105. The patients in the non-risk group have shown a roughly normal distribution for 'diaBP'. But it can be observed a slight skewness in the risk group distribution for 'diaBP'. The 'diaBP' distribution is rightly skewed for the risk group. This implies that patients with risk have indicated a tendency of having high diastolic blood pressure values. This can be generalized since high diastolic blood pressure puts added force against the artery walls and can lead to a high risk of CHD in general.

6.2.10 BMI vs TenYearCHD

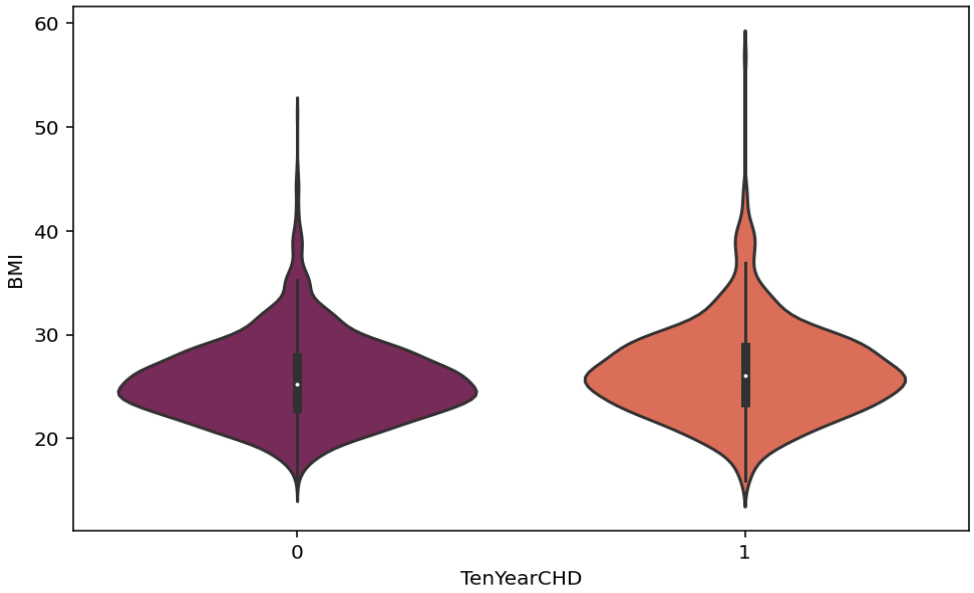

Figure 6. 29 - Violin plots of BMI vs TenYearCHD

It can be observed that BMI values for both risk and non-risk groups have shown a high density of BMI values between 20 and 30. The median BMI value for the risk group is slightly higher than for the non-risk group. Also, it is observable that the risk group has shown a high density for higher BMI values compared to the non-risk group. This suggests that higher BMI values may lead to a higher risk of having CHD. Normally, the majority of the people who have high BMI values are obese. High obesity leads to more diseases in general especially in the case of heart disease. Therefore these observations justify the general case.

6.2.11 glucose vs TenYearCHD

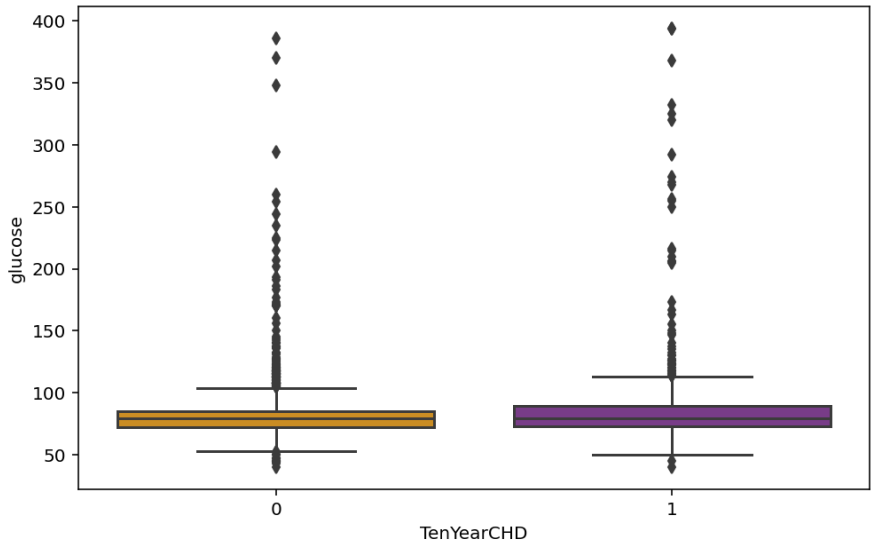

Figure 6. 30 – Boxplots of glucose vs TenYearCHD

The above figure depicts that both risk and non-risk groups have shown some high glucose values. But when the figure is investigated carefully, it can be observed that the glucose distribution and the median values are slightly higher for the risk group compared to the distribution and the median of the non-risk group. Therefore it implies that the higher glucose levels may have a somewhat tendency for higher risk of CHD. In general, higher glucose levels lead to a higher risk of heart and other types of diseases.

6.2.12 age vs totChol

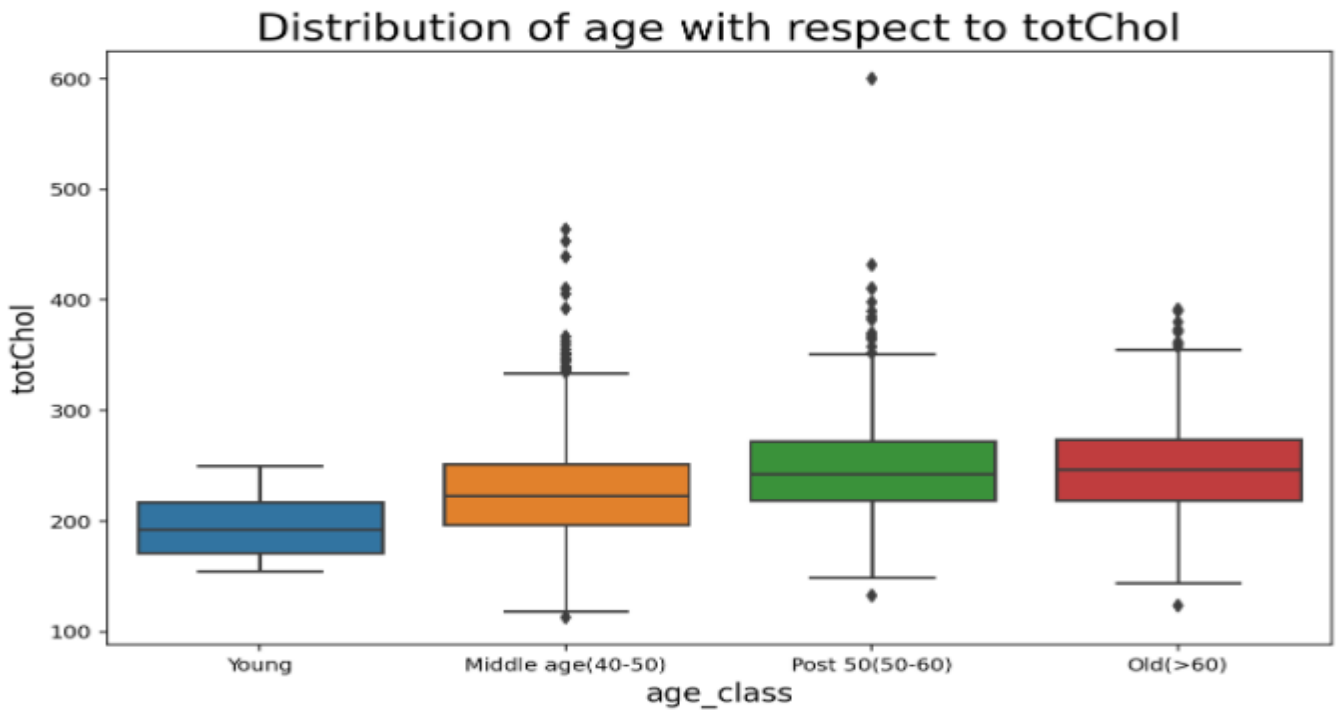

Figure 6. 31 – Boxplots of age vs totChol

As shown in the above series of boxplots, it can be observed that the median values of boxplots are shifted in an upward manner suggesting that aged people have more cholesterol (bad cholesterol in general). Some higher total cholesterol values are indicated in age groups for people older than 40 years. This justifies the general situation since when the age is increasing majority of people are getting diseases like high cholesterol.

6.3 Multivariate Analysis

6.3.1 heartRate with respect to TenYearCHD and prevalentHyp

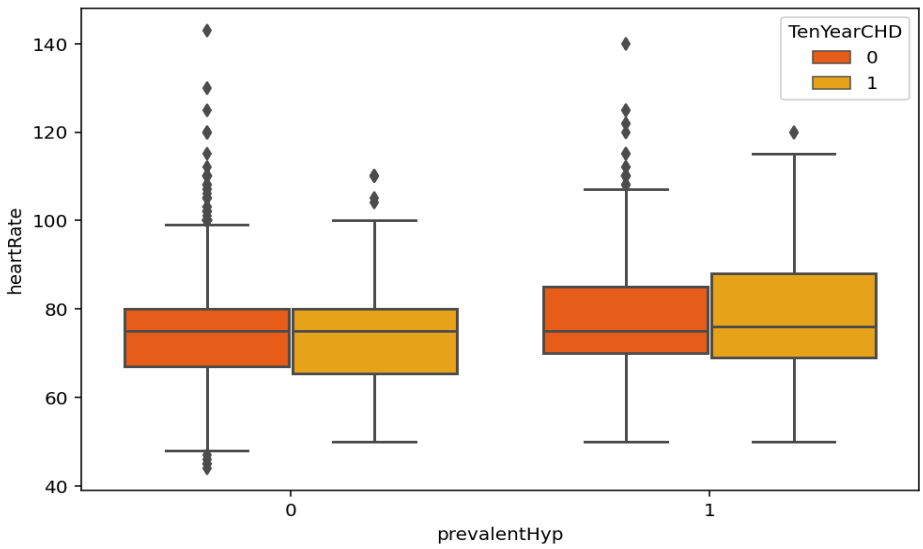

Figure 6. 32 - Boxplots of heartRate by TenYearCHD and prevalentHyp

As shown above, for 'prevalentHyp'=0 the distributions of 'heartRate' are roughly the same for both 'risk of CHD' and 'non-risk of CHD' groups. But for 'prevalentHyp'=1 the distribution of 'heartRate' is somewhat higher for the CHD risk group. Therefore hypertension may cause higher heart rates and that condition may increase the risk of CHD.

6.3.2 sysBP with respect to age and TenYearCHD

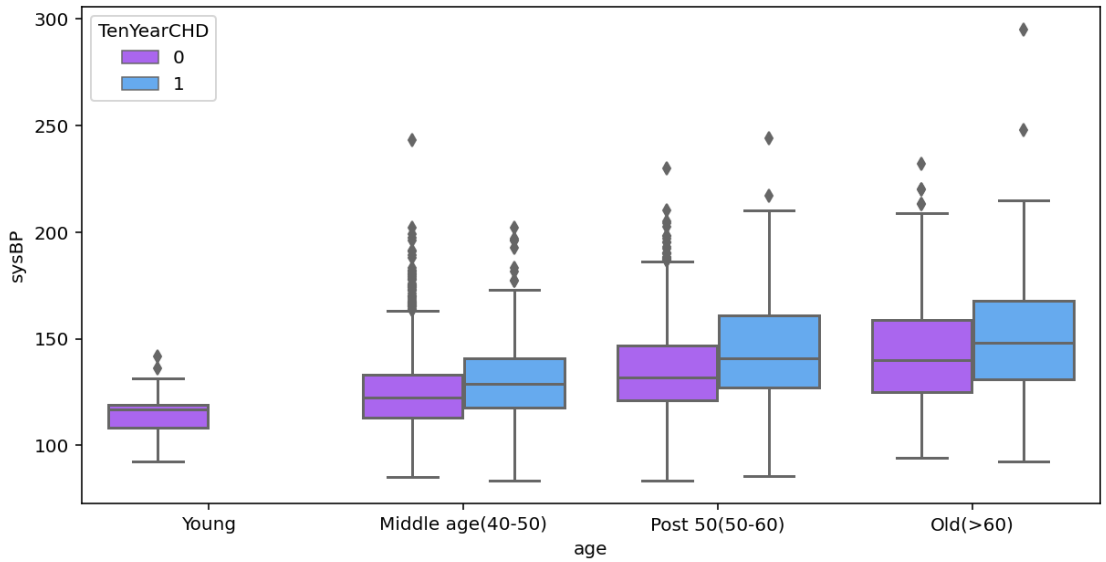

Figure 6. 33 – Boxplots of sysBP by age and TenYearCHD

As shown in the above figure, for the higher age categories systolic blood pressure level distributions are higher. Also, it can be observed that within each age category, systolic blood pressure distributions are higher for the risk group compared to the distributions for the non-risk group. Therefore these observations imply that when the higher age and higher systolic pressure levels combined, it leads to an increased risk of CHD. This is a reliable observation since for an older person with high systolic blood pressure, the risk of CHD is higher in general.

6.3.3 cigsPerDay, totChol and glucose with respect to age

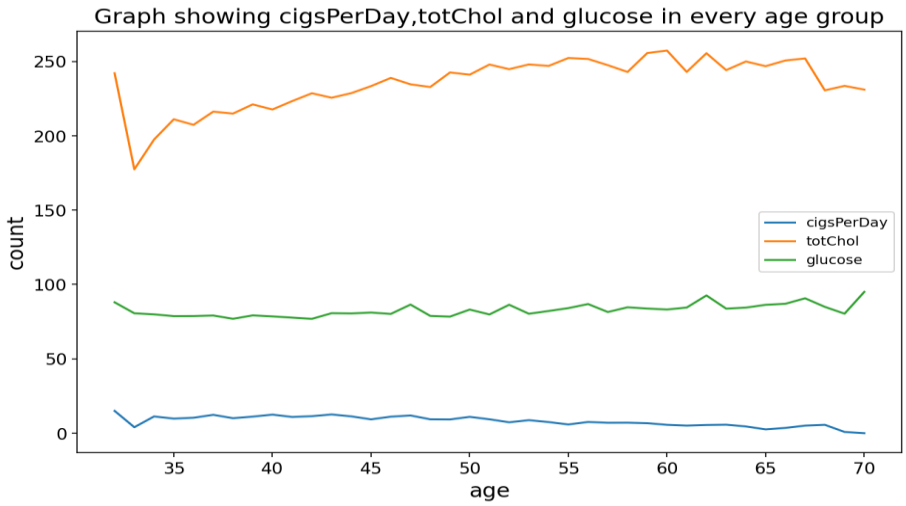

Figure 6. 34 – Line plot of cigsPerDay, totChol, and glucose by age

The above line plot depicts that there is a minor relation between 'totChol' and 'glucose' when considering the fluctuations. The 'totChol' has a steep, linear, and inverse graph for lower ranges of age. It appears to be 'cigsPerDay' has a fairly parallel relationship with age. 'cigsPerDay' and 'glucose' have shown somewhat similar distributions in early ages and for higher ages, glucose levels have shown some fluctuations. This may happen since 'cigsPerDay' continues as a habit and glucose levels may have affected by age in general.

6.3.4 sysBP vs diaBP with respect to currentSmoker and male attributes

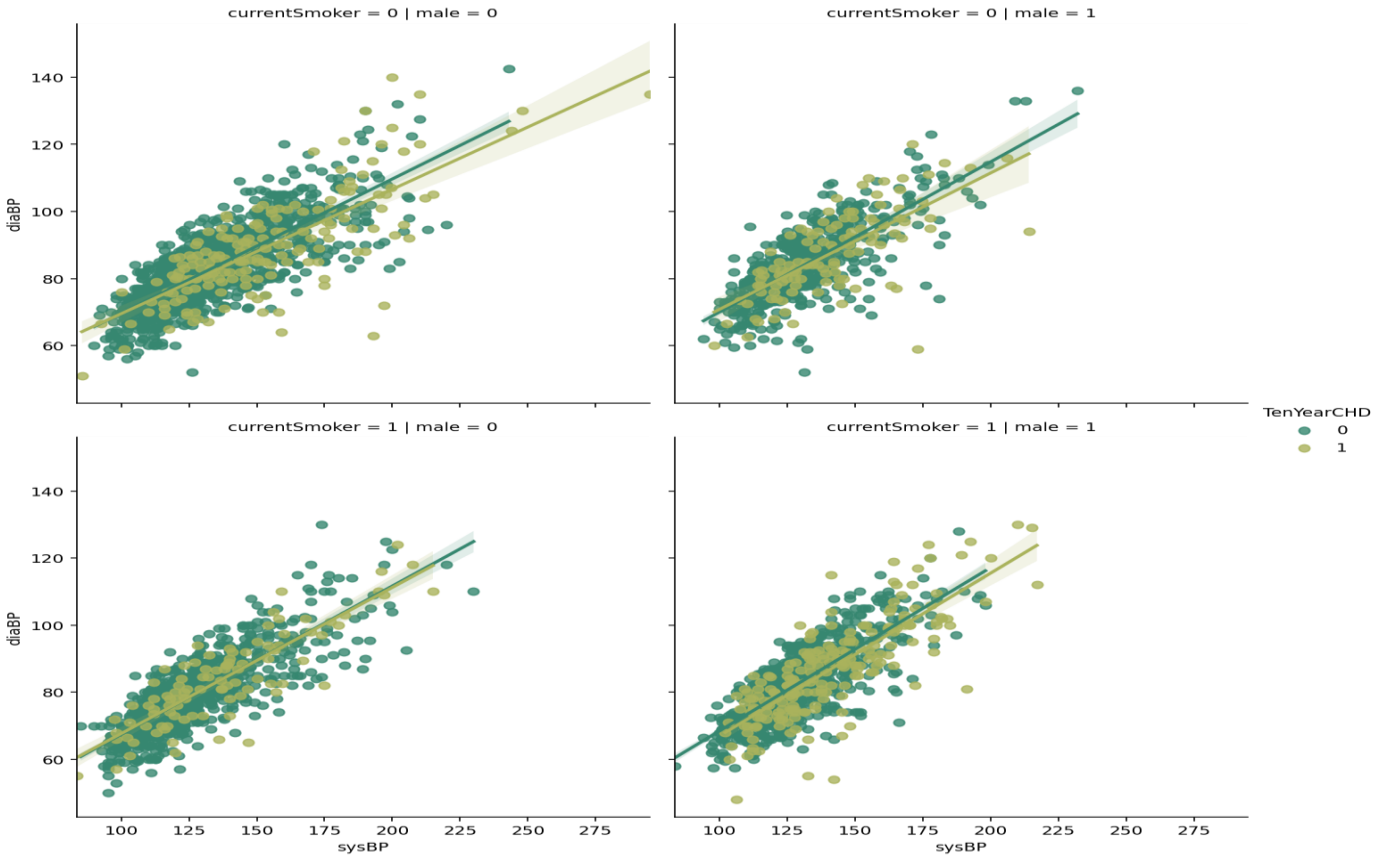

Figure 6. 35 – Scatterplots of sysBP vs diaBP by currentSmoker and male attributes

The above graph plots the relationship between systolic blood pressure and diastolic blood pressure for patients based on their gender and whether they are current smokers or not and plots the best fit line. It can be observed that for males who smoke, most points at the top of the graph are the patients at risk of CHD. Therefore it may be the case that for the males with smoking habit, blood pressure values are higher and this joint effect can cause a higher risk of CHD than others.

6.4 Advanced Analysis

The imbalanced data is balanced before carryout the advanced analysis as mentioned in section 5.3.2 under resampling.

6.4.1 Feature Selection

A feature selection can be done before the model fitting to identify the features that have a larger contribution towards the outcome variable, 'TenYearCHD'. Feature selection methods can be used to identify and remove unneeded, irrelevant, and redundant attributes from data that do not contribute to the accuracy of a predictive model or may decrease the accuracy of the model. Fewer attributes are desirable because it reduces the complexity of the model, and a simpler model is simpler to understand and explain.

In this study, the 'SelectKBest' method in python is used for extracting the best features of the dataset. The scores of features are shown below in an orderly manner.

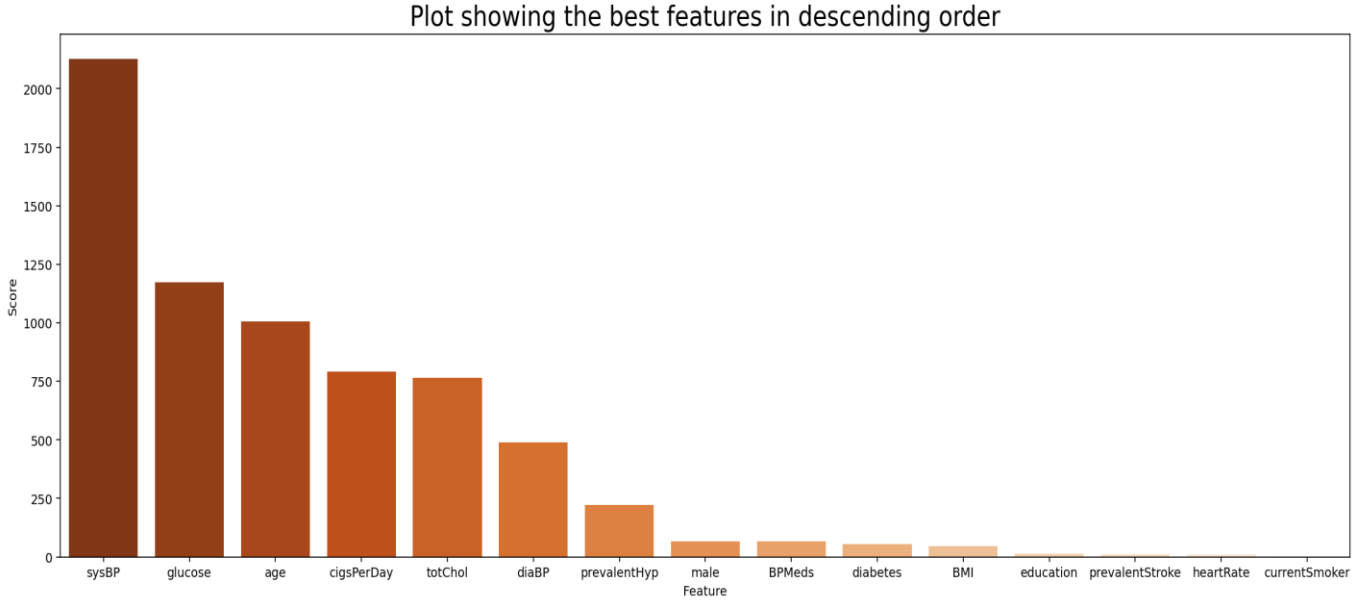

Figure 6. 36 – Visualization of feature selection

The features are selected by using these score values and by investigating associations determined in correlations.

6.4.2 Multicollinearity

When observing the correlations among the selected features in correlation visualization, it can be identified that 'sysBP' and 'diaBP' have shown a higher association between them. Also note that the association between 'glucose' and 'diabetes' is moderately high and significant as mentioned under the missing values in section 5.3.1, data cleaning. These associations will be used in model evaluation. But the initial models are fitted with all selected features to check if one feature from both associated pairs will be dropped in the following iterations in the model fitting process.

6.4.3 Logistic Regression Model Fitting.

The logistic regression model is fitted and evaluated since the response variable, 'TenYearCHD' is a binary variable. Here logistic regression algorithm is mainly used for prediction and also calculating the probability of success through the mathematical equation. Initially, the logistic regression model is fitted for selected variables and then the model

is refitted multiple times by inspecting p values for the variables until all the variables become significant. The significance of variables is checked at a 5% significance level.

The forward selection, backward elimination, and stepwise selection methods are tested and the backward elimination process is included here since it has shown a better model fitting with the selected features in the presence of collinearity. Because the backward elimination method may be forced to keep all those features in the initial model, unlike the forward and stepwise selection where none of them might be entered.

Model 1

Table 6. 2 - Table of model 1 statistics

Parameter	P>|z|
const	0.000
sysBP	0.000
glucose	0.002
age	0.000
cigsPerDay	0.000
totChol	0.000
diaBP	0.628
prevalentHyp	0.006
male	0.000
BPMeds	0.121
diabetes	0.478

When observing the above table, it can be observed that 'diaBP' has shown the highest p-value among the variables which have a p-value higher than 0.05. Note that as mentioned in above 6.4.2, multicollinearity section, 'sysBP' and 'diaBP' have indicated a higher correlation and between these two 'sysBP' has shown the highest association with the response. Therefore this may be the reason for the insignificance of 'diaBP'. The model will be refitted by removing the 'diaBP' in the next step.

Model 2

Table 6. 3 - Table of model 2 statistics

Parameter	P>|z|
const	0.000
sysBP	0.000
glucose	0.002
age	0.000
cigsPerDay	0.000
totChol	0.000
prevalentHyp	0.004
male	0.000
BPMeds	0.123
diabetes	0.471

It can be observed that 'diabetes' is the most insignificant variable in the above statistics. As mentioned in the multicollinearity section, 'diabetes' and 'glucose' have indicated a higher correlation. From these two variables 'glucose' has shown the highest association with the response variable. Therefore this may lead to the insignificance of 'diabetes'. The model will be refitted in the next step by removing the 'diabetes' from the existing model. Also note that with dropping 'diabetes' from the model, one feature from each identified pair has dropped for both associated pairs which are observed in the multicollinearity section.

Model 3

Table 6. 4 - Table of model 3 statistics

Parameter	P>|z|
const	0.000
sysBP	0.000
glucose	0.000
age	0.000
cigsPerDay	0.000
totChol	0.000
prevalentHyp	0.004
male	0.000
BPMeds	0.119

It can be observed that this model is the same as the model which can gain by dropping features from selected best features using multicollinearity and associations with the response variable. Therefore the effects of observed associated pairs are eliminated. Only the 'BPMeds' has shown insignificance in the above statistics. Notice that 'BPMeds' has the lowest association with the response variable from existing variables in the model. This may be the reason for the insignificance of the variable. Therefore the model will be refitted by removing the 'BPMeds' from the existing model.

Model 4

Table 6. 5 - Table of model 4 statistics

Parameter	P>|z|
const	0.000
sysBP	0.000
glucose	0.000
age	0.000
cigsPerDay	0.000
totChol	0.000
prevalentHyp	0.002
male	0.000

It can be observed that all of the variables in the above table of statistics are significant. Therefore the variables in the above model are selected for the final model fitting.

The final model is fitted after scaling the variables as mentioned in section 5.3.2 under data scaling by using the significant variables as identified above. The model evaluation is conducted as shown below.

Final Model with Standardized Variables

Classification Report

Table 6. 6 – Logistic regression classification report

Logistic Regression	
Parameter	**Value**
Accuracy	66.986
f1-score for category '0'	0.67
f1-score for category '1'	0.67

The fitted logistic regression model has shown around a 67% of accuracy. The f1-score is a single metric that combines recall and precision using the harmonic mean. It appears to be that f1-scores are the same and moderately high for both groups which indicate both precision and recall of the classifier indicate somewhat good results.

Confusion Matrix

This has been used to indicate the summary of prediction results including correct and incorrect on a classification problem. Further, this was used to not only errors but also types of errors. The segments of the confusion matrix indicate the following parameters.

- True Positives (TP): cases which are predicted yes (they have the disease), and they do have the disease.
- True Negatives (TN): cases which are predicted no, and they do not have the disease.
- False Positives (FP): cases which are predicted yes, but they do not have the disease (Type I error).
- False Negatives (FN): cases which are predicted no, but they do have the disease (Type II error).

The following outcome indicates the confusion matrix of the dataset.

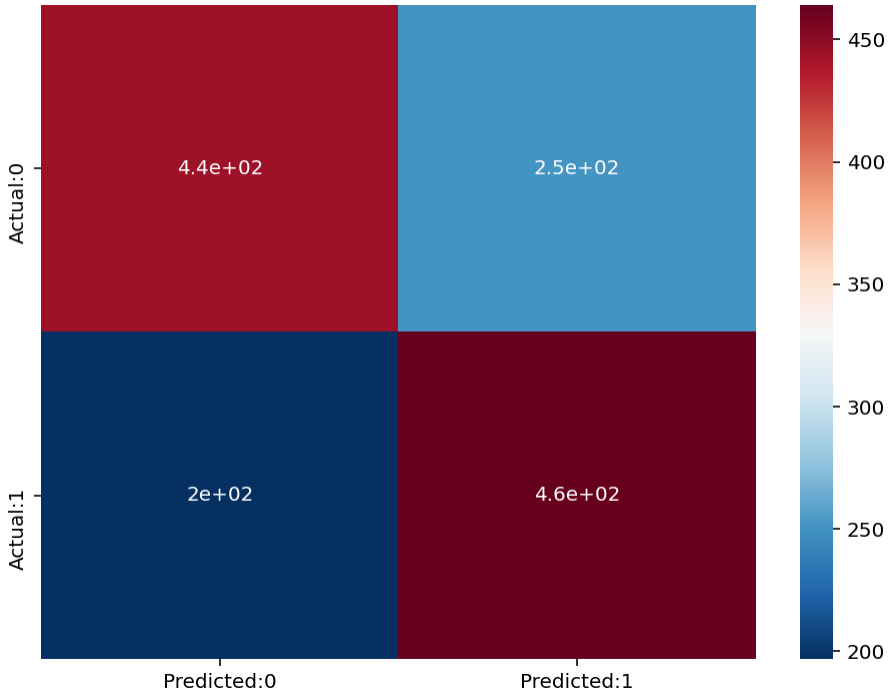

Figure 6. 37 - Visualization of Logistic regression model confusion matrix

It can be observed from the above figure, true positives and true negatives are roughly twice as false positives and false negatives respectively. Therefore the model is moderately accurate but it should be improved to enhance the accuracy.

Table 6. 7 – Logistic regression model sensitivity and specificity

Model sensitivity	0.702
Model specificity	0.6394

The model sensitivity summarizes and provides a metric to interpret that when the actual value is positive, how often is the prediction correct and the model specificity summarizes that when the actual value is negative, how often is the prediction correct. With analyzing confusion matrix data, it is evident that the model is a little bit highly sensitive than specific. Further, the positive values in the model are predicted more accurately than the negatives.

ROC Curve

The ROC Curve is a simple plot used to visualize the performance of a binary classifier. Further, this shows the tradeoff between the true positive rate and the false positive rate of a classifier for various choices of the probability threshold. The area under the ROC curve quantifies model classification accuracy. The ROC curve for the final model is shown below.

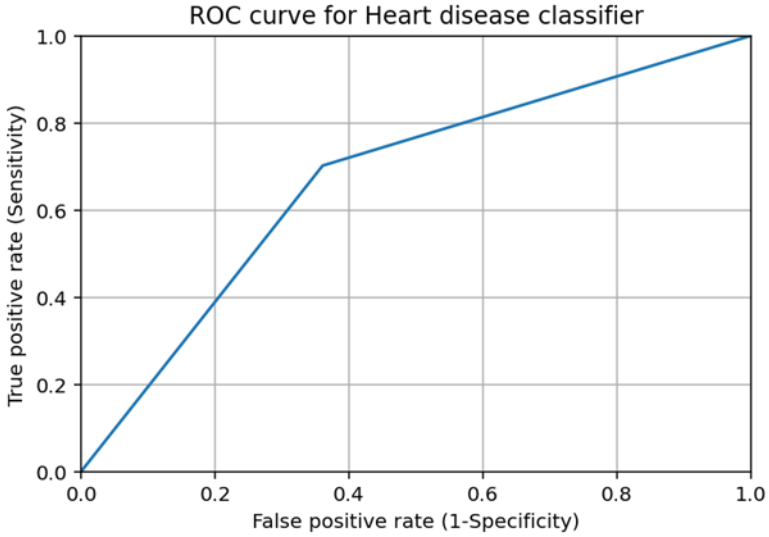

Figure 6. 38 – ROC curve

The AUC score is identified as 0.671 for the above ROC curve. This implies that the disparity between true and false positives is fairly greater, and the model is somewhat stronger in classifying members of the training dataset. But the model should be improved to have a higher AUC score since the closer the AUC to 1 is the better.
The evaluated model is shown below.

Model Coefficients

Table 6. 8 – Logistic regression final model coefficients

Parameter	Coeffecient
Intercept	2.98444783
age	2.518200
glucose	2.110861
sysBP	1.815927
cigsPerDay	1.628150
totChol	0.678745
male	0.531865
prevalentHyp	0.348755

The accuracy of the logistic regression model is not sufficient. Therefore it is needed to determine a model with higher accuracy. In that case, The K-Nearest Neighbors algorithm is carried out and evaluated as mentioned in the next section.

6.4.3 KNeighbors Classifier

The K-Nearest Neighbors algorithm is carried out and the accuracy of outcomes is analyzed as below.

Classification Report

Table 6. 9 - K-Nearest Neighbors model classification report

Logistic Regression	
Parameter	**Value**
Accuracy	91.8939
f1-score for category '0'	0.91
f1-score for category '1'	0.92

It appears to be the accuracy of this classifier is very high since it is around 92%. The f1-scores are roughly the same for both outcome groups as well as the scores are very high compared to the logistic regression model. Therefore it can be observed that the model accuracy is developed by a considerable margin.

Confusion Matrix

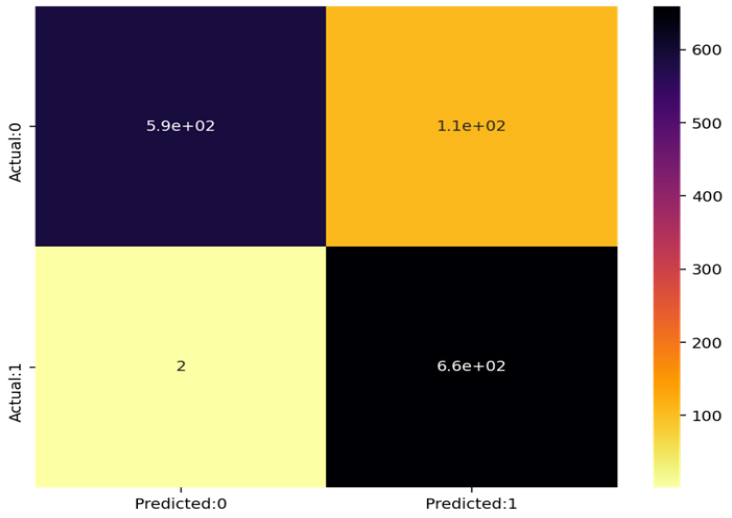

Figure 6. 39 - Visualization of K-Nearest Neighbors algorithm confusion matrix

The above outcome depicts that the true negatives and true positives are higher compared to the false negatives and false positives respectively. This indicates the presence of good classification.

Table 6. 10 - K-Nearest Neighbors model sensitivity and specificity

Model sensitivity	0.997
Model specificity	0.8448

As shown in the above table, it can be observed that the model sensitivity and the model specificity have increased to a very good level compared to the evaluated logistic regression model. Here, the model sensitivity is a bit higher than the model specificity which indicates that the positive values in the model are predicted more accurately than the negatives.

When comparing both models, it can be observed that the K-Nearest Neighbors algorithm (KNeighbors Classifier) is better with higher accuracy. Also, It can be concluded that according to the evaluated results, the KNeighbors Classifier has indicated a good model fitting with sufficiently higher accuracy.

7. General Discussion and Conclusion

The features, systolic blood pressure, glucose level, age, number of cigarettes per day, total cholesterol level, hypertension, and being male are identified as factors that mainly contribute to determining the risk of coronary heart disease.

According to this dataset, the empirical results show that males have shown a slightly higher risk of coronary heart disease (TenYearCHD). This may happen due to the comparatively higher bad personal habits of males than females and negligence of good health practice with the more work-stressed environment. Therefore they may tend to pay less attention to follow a healthy diet and could be less aware of their symptoms and wellbeing.

Older people have shown a higher tendency of getting coronary heart disease. This is the general case since older people are more vulnerable to heart diseases due to their weak physical conditions and long-term practice of bad habits.

When considering the number of cigarettes smoking per day (cigsPerDay) and the heart risk, low cigsPerDay comes with a lower risk of CHD. Although that is the case, low cigsPerDay doesn't guarantee a much lower risk of CHD. It suggests that cigsPerDay influences the risk of heart disease but the risk heavily depends on the joint effect of several other factors. But, when considering the smoking status (currentSmoker) people who smoke have shown a higher risk of CHD and those who don't smoke have a low risk of contracting the disease. It is an obvious and an identified fact that the chemicals in tobacco smoke harm blood cells, damage the function of the heart and the structure and function of blood vessels. Eventually, smoking contributes to atherosclerosis and increases the risk of having and dying from heart disease. These facts justify the study observations and therefore they are applicable in general.

The people who take blood pressure medications (BPMeds) show a somewhat higher risk of CHD compared to the people who do not take blood pressure medications. Generally, it is a commonly accepted fact that there is an increased risk of heart disease for a person who has blood pressure issues compared to a person who has not those issues. This observation is reliable since high blood pressure causes the blood vessels to become narrow and blood flow to the heart can slow. Therefore these can have an impact on the risk of CHD.

The study observations suggest that the risk of CHD is higher for hypertensive people. This is the general case according to many medical studies (Hu, 2017), since hypertension damages arteries that can become blocked and prevent blood flow to the heart muscle.

The cholesterol shows a sort of positive impact on the risk of CHD. The people with higher total cholesterol levels show a slightly higher risk of CHD compared to the patients with lower total cholesterol levels. But total cholesterol does not show a much observable difference between risk and non-risk patients. This could be due to the presence of good cholesterol (HDL) in the total cholesterol reading.

The systolic blood pressure (sysBP) shows a higher risk of CHD for those who have higher sysBP levels compared to the ones who have lower sysBP levels. This is more general since most medical studies (American Heart Association, 2016) have found that having high systolic blood pressure for a long period can increase the risk of strokes and heart disease. Also, diastolic pressure (diaBP) has shown a sort of positive impact on the risk of CHD. The results show that the risk of CHD is marginally higher for people with higher diaBP levels. When considering the joint effect of both blood pressure categories, high blood pressure puts added force against the artery walls. Over time, this extra pressure can damage the arteries. Therefore both blood pressure levels are increasing the risk of CHD in general.

The study results emphasize that higher glucose levels have a considerable tendency for the increased risk of heart diseases. Many medical studies (Zawn, 2019) have found that over time, high blood glucose can damage the blood vessels and the nerves that control the heart and blood vessels. The diabetes feature in the study is correlated with glucose and also it has shown a sort of impact on heart disease. It is justifiable since the higher blood glucose levels lead to diabetes and it is a sign of diabetes in general. In the medical field, it is a common observation that the longer someone has diabetes, the higher the chances of developing heart disease. Some studies (National Institute of Diabetes and Digestive and Kidney Diseases, 2017) have identified that in adults with diabetes, the most common causes of death are heart disease and stroke. Therefore the results are more general for glucose and diabetes.

When considering cholesterol and age, older people tend to have higher cholesterol levels compared to younger people. This observation is also more general since the risk of having high cholesterol may increase as someone gets older due to some causes such as decreased physical activity and long-term bad eating habits.

Also, the study findings emphasize that the presence of hypertension with the association of a higher heart rate shows an increased risk of heart disease. This is the general case in real life since elevated heart rate is usually associated with increased risk for hypertension and this combination can cause heart diseases.

The results have implied that the risk of CHD is higher for an older person who has a higher systolic blood pressure value compared to a younger person who has a lower systolic blood pressure value. This simply emphasizes the fact that a higher systolic blood pressure value at a higher age tends to a higher risk of CHD. When this is analyzed from the medical perspective it may be the case that with the age, the vascular system changes, and arteries get stiffer, so blood pressure goes up. This is true even for people who have heart-healthy habits. Therefore these facts justify the study observation.

The total cholesterol (totChol) has indicated an inverse association with age for lower ranges of age and then an overall gradual increase for higher ranges of age. This may happen since the majority of people focus on their health and diet at an early age and then they neglect it with the time due to the increased workload and busy life.

Also, the study results reveal that for the males with smoking habit, the blood pressure levels and the risk of CHD are relatively higher than others. This observation is expected since smoking and high blood pressure levels are major causes of CHD. Also, the males show a higher risk compared to the females as mentioned above in conclusions. Therefore the combined effect may also direct to the same conclusion as we observed.

The logistic regression model which is derived through P values of the variables has shown around 67% of accuracy. The model sensitivity and the model specificity are determined as 0.702 and 0.6394 respectively. The positive values in the model are predicted more accurately than the negatives. Further, the model accuracy is investigated by using f1-scores and ROC curve and it is observed that the model is somewhat satisfactory but not sufficient enough.

The results obtained by K-Nearest Neighbors algorithm through the KNeighbors Classifier have indicated around 92% of accuracy. The model sensitivity and specificity are determined as 0.997 and 0.8448 respectively. In this model also the positive values in the model are predicted more accurately than the negatives. This model has shown high f1-scores for both response categories. Therefore the model appears to be reasonably good when considering the accuracy metrics.

When comparing the logistic regression and K-Nearest Neighbors algorithm (KNeighbors Classifier), the K-Nearest Neighbors algorithm shows a significantly higher accuracy compared to the logistic regression model. Also, there is a considerable improvement of f1-scores in the K-Nearest Neighbors algorithm. Both models are somewhat more sensitive than specific. But the model sensitivity and specificity are also higher for the K-Nearest Neighbors algorithm compared to the logistic regression model. Therefore when considering these results, it can be concluded that the K-Nearest Neighbors algorithm is better and it indicates a good model fitting.

8. Recommendations

A Possible shortcoming of this study is that there were not enough instances for one CHD class hence the dataset was imbalanced. Even though the issues are treated it may result in some reduction of accuracy when generalizing to the public. Therefore, the study can be developed by better data and enhanced performance of ML algorithms. Based on the obtained machine learning results from the study dataset, future research needs to be carried out to improve the performance of the model, especially to increase the sensitivity and specificity rates. One such attempt could be to check if using unsupervised learning techniques before undertaking prediction, will enhance the model furthermore in terms of its prediction performance. Thereafter, the evaluated features and prediction models obtained through the study conducted can be used to develop some applications which will help people to track their health and thereby lead to early detection for CHD.

9. References

1. American Heart Association. (2016). How High Blood Pressure Can Lead to a Heart Attack. https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attack (2021, June 01)

2. Britannica, T. Editors of Encyclopaedia (2020, October 15). Coronary heart disease. Encyclopedia Britannica. https://www.britannica.com/science/coronary-heart-disease (2021, May 31)

3. Conget, I., & Giménez, M. (2009). Glucose control and cardiovascular disease: is it important? No. *Diabetes care*, *32 Suppl 2*(Suppl 2), S334–S336. https://doi.org/10.2337/dc09-S334 (2021, May 30)

4. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care **33**(Suppl 1): S62–S69(2010). https://doi.org/10.2337/dc10-S062 (2021, May 30)

5. Ho, Kalon & Pinsky, Joan & Kannel, William & Levy, Daniel. (1993). The epidemiology of heart failure: The Framingham Study. Journal of the American College of Cardiology. https://www.sciencedirect.com/science/article/pii/073510979390455A (2021, June 01)

6. Hu, Lihua & Huang, Xiao & You, Chunjiao & Li, Juxiang & Hong, Kui & li, Ping & Wu, Yanqing & Qinhua, wu & Bao, Huihui & Cheng, Xiaoshu. (2017). Prevalence and Risk Factors of Prehypertension and Hypertension in Southern China. PLOS ONE. 12. e0170238. 10.1371/journal.pone.0170238. (2021, June 01) https://www.researchgate.net/publication/312506470_Prevalence_and_Risk_Factors_of_Prehypertension_and_Hypertension_in_Southern_China (2021, June 01)

7. https://www.cdc.gov/diabetes/index.html (2021, June 01)

8. https://www.webmd.com/cholesterol-management/default.htm (2021, June 01)

9. https://www.healthline.com/health/high-cholesterol/levels-by-age#adults (2021, June 01)

10. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (2021, June 01)

11. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).(2017) https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke (2021, June 01)

12. Peters, S. A., Singhateh, Y., Mackay, D., Huxley, R. R., & Woodward, M. (2016). Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis. Atherosclerosis, 248, 123–131. https://doi.org/10.1016/j.atherosclerosis.2016.03.016 (2021, June 01)

13. Southern Cross. (2018). Coronary heart disease - causes, symptoms, prevention. Retrieved from Southern Cross: https://www.southerncross.co.nz/group/medicallibrary/coronary-heart-disease-causes-symptoms-prevention (2021, June 01)

14. Villines Zawn (April 25, 2019) - Medically reviewed by Maria Prelipcean, M.D. Medical News Today. https://www.medicalnewstoday.com/articles/249413#Takeaway (2021, May 30)

15. Wong, Nathan & Levy, Daniel. (2013). Legacy of the Framingham Heart Study: Rationale, Design, Initial Findings and Implications. https://globalheartjournal.com/articles/abstract/10.1016/j.gheart.2012.12.001/(2021, June 01)