Title: **Heart Disease Prediction Using Logistic Regression**

Author: Feziwe M Shongwe (Melvin.shongwe@gmail.com)

Abstract

Approximately 12 million deaths worldwide are caused by heart diseases annually and half of the deaths are due to cardiovascular diseases, according to the World Health Organization. Our main goal for this study is to predict which factors are good predictors of patients that have 10-year risk of future coronary heart disease. To reduce the complications from cardiovascular diseases, early diagnosis can help patients make informed lifestyle changes. The data that I will be using for the prediction in my project is from an online source named Kaggle, the data was gathered for a cardiovascular study that is being conducted on residents based in the town of Framingham, Massachusetts. To come up with my conclusion to this study I approached it by using Statistics concepts which are procedures for determining the distribution of my variables and correlation, building a regression model (Logistic Model), and Selection techniques to find an optimal model. I used the backward technique when I was optimising my model because the forward has suppressor effects. Throughout my research, I used the logistic regression model as it is applicable for this study for coming up with a solution to this study of determining factors that are good predictors of whether a patient has a 10-year risk or not. After computing

Research Question: Which factors are good predictors of patients that have 10-year risk of future coronary heart disease?

My research Questions:

1. Does an increase in a patient's total cholesterol increase the risk of coronary heart disease in 10 years?
2. Does one's lifestyle influences the 10-year risk of coronary heart disease?

Literature Review

Introduction

It is generally known that cardiovascular diseases are the leading cause of death globally in addition, the World Health Organisation (WHO) has estimated about 12 million deaths occur worldwide. Most heart diseases can be prevented by detecting them and addressing risk factors before the situation gets critical because it is believed that early prognosis of cardiovascular diseases can assist in making decisions on how the patient's lifestyle changes and in turn reduce the complications so that management with counselling and medicine can begin. The thesis is that diagnosing and reducing the risks does not necessarily give us the factors that are good predictors of patients that have a 10-year risk of future coronary heart disease. This discussion mainly focuses on 15 attributes that are given from a dataset to determine which factors from the dataset attributes that are good predictors of patients that have a 10-year risk of future coronary disease, and I aim to use the attributes from the dataset to determine which of the attributes are good predictors of the 10-year risk of coronary disease. Furthermore, I will also discuss how a patient's exercise activities contribute to the 10-year risk of coronary heart disease, whether a patient's lifestyle influences the 10-year risk of coronary disease, and which age group of patients is more likely to have a risk of 10-year risk of coronary heart disease and which patients' age group has the least chance of a 10-year risk of coronary heart disease. Under this review I will use different sources to observe the trend if there is any in all the studies done by researchers, to have an idea/insight on what results I should expect for this study. Coronary heart disease is hereditary which implies that medical-related information and demographic are more factors than behavioural information that are good predictors of the 10-year risk of coronary disease.

Body

Multiple Logistic regression model will be appropriate to this study as it considers the probability of an event, in this case, we have two binary options. Wilson et al put forward the idea of using logistic regression for predicting models. For the past two decades, it has been possible to estimate CHD risk by use of regression equations derived from observational studies, and the present study demonstrates similar results, predicting later CHD in a middle-aged white population sample (Wilson, D'Agostino, Levy, Albert M. Belanger, & Kannel, 1998). Thus, our model will be able to determine the good predictors by using the logistic regression model as previous studies show the precision of the logistic regression model. This model is not necessarily inverted to replace doctors' decisions and their skills, but its main goal/objective is to provide the factors to assist the professionals to make the right decisions on how to reduce the risk, especially for high-risk patients.

The inverse association between physical activity and incidence of CHD is consistently observed (Powell, Thompson, Caspersen, & Kendrick, 1987). Studies show that people who exercise have a lower risk of coronary heart disease. Thus, this will be useful for one of my research questions as I will be also looking on how patients' exercising contributes to the 10-year risk of coronary disease. Cigarette smoking, low HDL-C levels, and diabetes are less common among those who are physically active (Wilson, D'Agostino, Levy, Albert M. Belanger, & Kannel, 1998). Also, this gives me an insight that there is an association somehow between variables and patient's exercising activities which makes me to conclude that there's also an association between the 10-year risk of coronary and a patient being physical active. Moreover, inclusion of lifestyle characteristics, such as PA habits, in the risk models, increased CVD risk prediction accuracy (Georgousopoulou, et al., 2016).Thus, this will help me on answering my research question whether lifestyle influences the 10-year risk of coronary disease.

According to the World Health Organisation, most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol (World Health Organization, 2021). Drawing from the statement made by WHO, I can conclude that smoking patients have a 10-year risk of coronary disease. Patients with diabetes mellitus and individuals with clinically established cardiovascular diseases are, on average, considered to be at high or very high cardiovascular risk (Rossello, et al., 2019). Furthermore, the Prevalence of smoking is increasing in women in some populations and is a risk factor for coronary heart disease (Huxley & Wiidward, 2011). Drawing from the previous sources, I can conclude that smoking, unhealthy diet and obesity, and diabetes are common factors that contribute to cardiovascular diseases. Thus, this will contribute towards the aim of this study of determining good predictors and finding out whether one's lifestyle influences the 10-year risk of coronary disease.

According to Sallam and Watson, Cardiovascular disease is the leading cause of death in women, contributing to one in three female deaths. Despite improvements in overall cardiovascular outcomes, substantial gender and ethnic disparities remain (Sallam & Watson, 2013). Mass and Appelman put forward that cardiovascular disease develops 7 to 10 years later in women than in men and is still the major cause of death in women (Maas & Appelman, 2010). From these sources I can clearly see that Gender also contributes when predicting for coronary heart disease. Furthermore, a study that was conducted on coronary heart disease shows that the risk of developing coronary heart disease differs by sex, and accumulating evidence suggests that sex differences exist in the effect of coronary risk factors on vascular risk (Mongraw-Chaffin, Peters, Huxley, & Woodward, 2015).

It is generally known that high blood pressure it is a common symptom for heart related diseases and strokes as it forces one's heart to work harder when pumping blood to the rest of the body. From my dataset have only two types of blood pressure which is SBP and DBP. A suggestion was made from a study conducted that reducing Systolic blood pressure below targets may significantly reduce risk of cardiovascular diseases (Bundy, Li, & Stuchlik, 2017). The American heart Association puts forward the idea that more attention is given to systolic blood pressure as a major risk factor for cardiovascular disease (American Heart Assocation, 2021). These sources will help me on comparing the 10-year risk of coronary heart disease with the systolic blood pressure. In addition, the results from the study conducted by Bundy, Li & Stuchlik states that there were linear associations between mean achieved SBP levels and the risk of major CVD, stroke, CHD, all-cause mortality, and CVD mortality (Bundy, Li, & Stuchlik, 2017).Therefore, I can definitely see that there is correlation between some variables from my dataset.

Studies show that Hypertension is considered as another factor for stroke and heart complications. Olafiranye et al. stated that Hypertension is not only a major risk factor for stroke and heart failure (HF), but more importantly for coronary heart disease (Olifiranye, et al., 2011). Thus, statement made by Olafiranye will help me on identifying the good predictors for the 10-year risk of coronary heart disease by comparing the data based on hypertension from my dataset. Approximately a quarter of all adults in the USA suffer from hypertension, a strong predictor of cardiovascular risk. Across all ages, men tend to have higher mean blood pressure than women (Sallam & Watson, 2013). Therefore, I can see that there is somehow a relationship between these two variables (Gender and Hypertension) and the 10-year risk of coronary disease.

For comparison of cholesterol, Mayo Clinic staff went and conducted a study and concluded that one's body needs cholesterol, but high level of cholesterol can increase the risk of heart disease (Mayo Foundation for Medical Education and Research, n.d.). Thus, there is a relationship between cholesterol level and risk of

heart disease. This comparison will contribute towards my research question whether the total cholesterol contribute to the 10- year risk of coronary heart disease.

Conclusion

From the review I have noticed that these findings seem to be consistent from the sources, the information gathered is corresponding from one source to another which makes me to conclude that the logistic model is a good model as it has been used in different areas especially in this context. From my findings gathered from different sources, I am convinced that factors such as Age, Gender, Smoking, Diabetes, Systolic blood pressure, Hypertensive, and Cholesterol level are common factors that contribute when predicting whether a patient has a risk of a 10-year coronary disease. All the studies I went through did not consider the patient's education level as a factor that might be contribute towards prediction of coronary heart disease. These findings will help me to compare my results with the sources' results.

Methodology

introduction

Before I proceed in determining the good factors, I first need to clean my data. I will first visualise my data using SAS as mentioned in my literature review so I can identify missing variables and after that, I will be able to use all the procedures without encountering errors from the statistics package I will be using. These procedures will help me to compare the variables from my dataset in order to make the right decisions when deciding which are good predictors or not. I will use procedures for performing Descriptive Analysis, Correlation Analysis, Simple and Multiple Logistic regression modelling, Analysis of variance, Hypothesis testing, and Techniques to reduce models. These procedures will allow me to gain an insight into my variables on how they are correlated with one another and how they interact with each other.

Body

The first procedure I will perform is descriptive statistics. This procedure will help me to understand the measures of the dataset such as Measures of dispersion and central tendency. Thus, I will gain insight into my variables on how they are related to each other and distributed using kurtosis. I will also use the method of counts which falls under descriptive statistics also to determine the frequency tables of different variables. Lastly, I will also have a graphical comparison for some variables to view them graphically such as a Histogram.

Correlation analysis will help me to measure the association between the variables I would like to. In Addition, it will also measure the closely related regression analysis. This will help me to identify the relationship if it is present (relationship exists) or absent (No relationship) as it tends to be more relevant to everyday life. Variable's correlation can be classified as String, weak or moderate. It is highly possible to visualise these relationships using a scatter matrix plot as it shows the relationship between different variables. Thus, the correlation will help me to see which factors are related to each other in predicting the 10-year risk of coronary diseases.

As studies have shown that logistic regression is fast compared to the other models under supervised learning. Thus, it will be used also for me to create equations that can be used to predict the probability of the outcome of interest. This analysis will also be very useful as it works with variables that are continuous and categorized which makes it more appropriate for our data. Lastly, this analysis will help us to determine the odds ratio for different variables to measure the importance of a predictor variable relative to the response variable. After a few tests for several coefficients, coefficients can be dropped from the model given that they do not improve it for predicting.

I will then compare the means for my analysis of variance (ANOVA) as it will help me to determine whether the differences between the variables are statistically significant with the assumption that my data is normally distributed. To simplify it, This Anova will tell me whether there is a statistical difference between the means of my variables. If the means for specific variables are equal, then it will give us an idea that those variables might be sharing something(related) which means if one of them is one of the factors it is likely those other ones be factors also.

After applying these procedures, I will then follow the testing part, where I will be testing for different hypotheses for each procedure I used. For correlation analysis, I will be testing whether there is multicollinearity between my continuous variables and will be using the odds ratio to measure the strength of

the variables with the response variable. Similarly, for the logistic regression modelling, we will test whether there is statistical evidence that a group of variables are good predictors for the model or not, meaning in this case for my study they will be representing whether they are good predictors for the factors that will predict the 10-year risk of coronary diseases. As mentioned previously for the Anova hypothesis testing I will be testing whether the means for each group in a variable are equal or not so I can determine the relationship between the predictor variables and the response variable.

R-squared goes hand in hand with the correlation between variables. Hence applying it will help me to see the percentage of points that lies within the fitted model. This will help me as said under the correlation part that I will be able to identify how variables are correlated to one another, especially the strength as it is measured differently. The coefficient of determination will help me to measure the proportionate reduction of the total variation in the 10-year risk of coronary diseases associated with the use of all the factors given from the dataset of the residents.

When performing the multiple regression analysis, we can use different types of selection options to specify how the variables will be considered for being in the model. The options include Backward, Forward and Stepwise. The backward will consider all the predictor variables irrespective of whether they improve the model or not, and then it will eliminate the ones that do not meet a criterion as they do not improve the model. Conversely, the FORWARD method brings in the most significant variable that meets the criterion and continues entering variables until none meets the criterion. Lastly, the STEPWISE is a combination of the two (BACKWARD and FORWARD) options as it uses the FORWARD first, but it re-evaluates the variables at each step, and it eliminates a variable that does not meet the criteria. These selection options will help me to determine which variables should be dropped using their value I will then conclude. Also, dropping the variables it will give me estimated regression models that I should use for my prediction as they improve. In each step, the R-squared will change to the one that best fits the model perfectly compared to the previous step.

Lastly, I will apply the concept of model evaluation for classification. This will help me to determine the precision vs Recall of my model. This concept involves 4 things, and this concept is called a confusion matrix. The matrix will store four values (How many are predicted correctly or wrongly that are correct and how many are predicted wrongly or correctly that are wrong). After all these computations have been done the statistics package will give me the results of the curve, and the curve will plot the graph of the precision and Recall of the model. The AUC measures the accuracy(discrimination) of the test. This measure ranges from 0 to 1 where 0 indicates an inaccurate test and on other hand 1 represents a perfectly accurate test. In addition, a moderate scale 0,5 is classified as a no discrimination test. Thus, this will be useful when testing for the accuracy of my model.

Conclusion

After all these tests and procedures are computed I will have all the factors that are good predictors of a 10-year risk. It is highly possible that all the factors I will determine using these procedures will correspond with the ones from the sources as similar procedures were taken just with a different dataset. We are highly confident that we might have all the common factors as they were proven to be factors previously using different concepts such as Machine learning-based studies and other Data science-related studies. The procedures used for the tests are powerful when predicting which makes my argument of being confident in finding the best factors that are good predictors.

Initial Analysis - Results and Discussion

Introduction

Determining these factors will help me to have a better understanding and to know which factors are good predictors of patients who have a 10-year risk of future coronary heart disease. In addition, this study will allow me to identify which type of patient's information contributes to the 10-year risk of coronary diseases and it will also allow me to know which factor contributes the most compared to the other ones. The original residents' dataset I had from the Kaggle website had 3390 entries and the data cleaning process resulted the dataset that I will be working with to have 2919 entries as my analysis will encounter problem with the null values. Approximately 14% of the data had nulls, therefore it can be removed from the dataset to use for modelling without losing the meaning of the data from it. Under this section I will go on with my analysis for these variables to determine Descriptive Analysis, Correlation Analysis, Simple and Multiple Logistic regression modelling, Analysis of variance, Hypothesis testing, and Techniques to reduce models. These procedures will allow me to gain an insight into my variables on how they are correlated with one another and how they interact with each other. After using the logistic regression to determine the factors that are good predictors, then the ones that are not good factors will then be removed (not necessary)/noted from the list of variables as they are not good factors. This study will also benefit me to enhance my industrial knowledge as it uses a concept of Machine learning which is logistic regression.

Body

Descriptive Analysis

For Demographic, I computed a procedure for obtaining the distribution of the gender variable and patient's age. This analysis will help me by giving me an insight in finding the measures and how the patient's data is distributed per demographic variable. The analysis shows that the age variable is positively skewed, as the measures of central tendency indicate $\boldsymbol{mean > median > mode = 0.44261733 > 0 > 0}$ with a $\boldsymbol{Skewness = 0.23}$, and the $\boldsymbol{kurtosis = 1.95}$ which tells me that the normal curve is platykurtic and it has lighter tails as the $\boldsymbol{kurtosis < 0.}$ Moving to the variance between the patients' gender, we obtained a $\boldsymbol{variance = 0.247}$ in this case the variance won't be large as we only had two values but statistically speaking a platykurtic curve enhances more variation among data. Lastly, from the histogram (see figure 1.1.1) it is clear that the bar that indicated the females is higher than the males' one which can be verified with the mode as we obtained that the most occurring gender is $\boldsymbol{mode = 0}$ and from the dataset we know that 0 represent females. Concluding from this analysis of gender, the obtained results will help me on finding results whether the claim made by my sources was true that women have higher chances of coronary disease compared to males.

The descriptive statistics for the patients' Age was obtained as following, for measures of central tendency $\boldsymbol{mean = 49.435}$, $\boldsymbol{median = 49}$, $\boldsymbol{Mode = 40.}$ These results show that the average patients who participated in this study were 49 years old and $\boldsymbol{median = 49}$ which tells me that half of the patients are below the age of 49, and half of the patients are above the age of 49. Furthermore, most occurring patients' age who participated in this study is 40. The skewness obtained for patient's age was $\boldsymbol{skewness = 0.236\ and\ Kurtosis = -1}$, from these results I can tell that patients' age is positively skewed (Skewed to the right and the normal curve is platykurtic. The $\boldsymbol{varience = 73.216,}$ as stated previously that we could not interpret it properly as we only had two values for the gender of patients. Thus, in this case we can see that

the kurtosis resulted in a flat normal curve (Platykurtic) which means there's variation among the data. Hence our variance is large in this case as it also supports that there is variation among patients' age who participated in this study. This analysis will assist me to on finding which age group has more risk of coronary heart disease as it well distributed.

For behavioural information, I computed descriptive statistics for finding out the summary statistics of the patients on their behavioural information. I obtained that out of the 2919 participants 1429 are smoking which makes it **49%** and **51%** of patients are not smoking. The distribution for the variable that displays whether a patient smokes or not (CurrentSmoker) is said to be positively skewed as its skewness measure equals to 0.042($\boldsymbol{skewness = 0.042}$) and the kurtosis equals to -0.2 ($\boldsymbol{kurtosis = -2 < 3}$) which tells me that the curve is platykurtic for this variable also. Furthermore, the $\boldsymbol{variance = 0.25}$ which won't be effective in terms of the variation in the curve in this case as we only have two values 0 and 1. To avoid inaccuracy for the number of cigarettes a smoker smokes a day I only considered the $\boldsymbol{49}\%$ that smokes because the ones that indicated that they are not currently smoking hence it is unnecessary to include them in this descriptive statistics.

From this sample of 1429, the average of cigarettes smoked by the participants who smoke is approximately 19 cigarettes ($\boldsymbol{mean = 18.59}$), the common number of cigarettes smoked by each participant is 20 (and half of the smoker's smoke less than 20 cigarettes a day and halt of the patients who smoke more than 20 cigarettes a day. By comparing the measures of central tendency $\boldsymbol{mean = 18.59 < 20 < 20}$ which makes it hard to tell whether the distribution is left or symmetrical or right. The skewness ($\boldsymbol{skewness = 0.733}$) suggests that the distribution is positively distributed. Going further to the variance and kurtosis,$\boldsymbol{varience = 120.11\ and\ kurtosis = 0.924}$ drawing from these values, I can clearly conclude that the curve is platykurtic and there is lesser variance among the number of cigarettes smoked by patients in a day. This analysis will help me on determining whether the number of cigarettes smoked by a patient is a good predictor of the 10-year risk of coronary disease or not. Lastly, this will also allow me to prove the claim made by my sources that smoking is a good predictor for the prediction of a 10-year risk of coronary heart disease.

For Information on medical history, the descriptive analysis of the blood pressures (Systolic and Diastolic blood pressure) will help me towards my research question as I will obtain how these variables are distributed among patients which will help me on concluding which patients have the 10-year risk of coronary heart disease. For systolic I obtained the following descriptive statistics, for measures of central tendency $\boldsymbol{mean = 132.413, median = 128\ and\ mode = 130}$ which makes me to conclude that the average of a patients' systolic blood pressure is approximately 132, half of the patients have systolic blood pressure that is less than 128 and the other half have a systolic blood pressure more than 128. Furthermore the $\boldsymbol{skewness = 1.173\ and\ kurtosis = 2.4 < 3}$ of the patients' systolic blood pressure which implies that the systolic blood pressure of patients is positively distributed, and the curve is platykurtic. From the analysis of the systolic blood pressure, the $\boldsymbol{varience = 485.5}$ which tells me that there's a greater variation among the patients' systolic blood pressure. Similarly for the Diastolic blood pressure, $\boldsymbol{mean = 82.93, median = 82\ and\ 80}$ which shows that the average diagnostic blood pressure for among the patients is 82.93, half of the patients have a diastolic blood pressure that is less than 82 and a half have diastolic pressure that Is greater than 82, and the common reading of the diastolic pressure among the patients is 80. The measure of central tendency shows that the diastolic bloop pressure is positively skewed ($\boldsymbol{mean = 82.93 > meadian = 82 > mode = 80}$) and ($\boldsymbol{skewness = 0.714}$). The measure of variability, $\boldsymbol{variance = 140.58\ and\ kurtosis = 1.33 < 3}$ which tells me that there is variation among the diastolic blood pressure of the patients and the curve is said to platykurtic. (*See figures for the results used above)*.

Similarly, to the heart rate and diabetes, these variables are also skewed to the right as its measures of central tendency using the following results, for heart rate ($\boldsymbol{mean = 75.82 > median = 75 > mode = 75\ and\ Skewness = 0.731}$) and for a patient whether had diabetes or not ($\boldsymbol{skewness = 5.955}$). Drawing from the measures of central tendency, the average heart rate for patients is approximately 76, half of the patients had a heart rate less than 75 and half of the patients had a heart rate greater than 75, and the most common heart rate among the patients is 75. For the measure of variability whether a patient had diabetes ($\boldsymbol{variance = 0.025\ and\ kurtosis = 33.49 > 3}$) which supports that there is lesser variation among the data as I only have 0 and 1(leptokurtic) and for patients' heart rate ($\boldsymbol{variance = 144.023\ and\ kurtosis = 1.186 < 3}$) which implies that there is greater variation among the patients' heart rate and the curve is platykurtic which supports that there is variation. These results can be observed also from the histograms provided in the appendix.

Looking at patients' total cholestenone level, the measures of central tendency indicate that the $\boldsymbol{mean = 236.91, median = 234,\ mode = 240\ and\ skewness = 0.552}$ which can be interpreted as, the average of the total cholesterol among the patients was 236.91, half of the patients' total cholesterol level are less than 234 and half of the patients' total cholestenone are more than 234, the most occurring total cholesterol is 240 and the distribution of the patients' total cholesterol level is positive Furthermore, the measure of variability indicates that the $\boldsymbol{variance = 1984\ and\ kurtosis = 0.80 < 3,}$ the measure of variability corresponds as the variance indicates that there is greater variance among my data and the kurtosis shows that the curve is flatter(platykurtic) which indicates that there is variation among the patients' total cholesterol level. This analysis will help me on understanding the distribution of the total cholesterol level among patients as it is one of the questions for this study I am conducting. From this analysis conducted for medical information, I can make an assumption for now that the variables of the same information type are distributed the same, which will help me to determine which type of information contributes to the prediction of the 10-year risk of Coronary diseases.

Correlation Analysis

Under this section, I have computed the correlation procedure to observe how my predictor variables are correlated to one another and how are they correlate to the response variable (TenYearCHD). Firstly, I computed a procedure to obtain whether my predictor variables are correlated to one another (Multicollinearity). Using the Pearson correlation coefficient, I can clearly see from Table 1 that the predictor variables are somehow associated with the response variable ($P - value < 0.05$) which favours the alternative hypothesis. More tests will be done under the hypothesis testing section. This analysis will help me to determine how the predictor are corrected to one another (Multicollinearity). It is not recommended to use both highly correlated predictor variables in a model as they contain similar information. In this case my results show me that (glucose and diabetes are highly correlated with a $\boldsymbol{correlation\ Coef = 0.6}$), (systolic and diastolic blood pressures are highly correlated with a $\boldsymbol{correlation\ coef = 0.783}$), (Patients who was hypertensive and systolic blood pressure are highly correlated with a $\boldsymbol{correlation\ coef = 0.692}$), (Patients who was hypertensive and diastolic blood pressure are highly correlated with a $\boldsymbol{correlation\ coef = 0.612}$) , and (cigarette smoking and the number of cigarettes smoked per day are highly correlated with a $\boldsymbol{correlation\ coef = 0.772}$). From the results I obtained from the correlation between the predictor variables, I can see that the variables that are highly correlated contain similar information drawing from glucose and diabetes, followed by the blood pressure and the patient's smoking information.

I also compared each predictor variable with the response variable to get an idea of how the relationship and strength is using the odds ratio and simple logistic regression (see figures in the appendix). The odds ratio will help me to measure the strength of association between each predictor variable and the response variable. The following comparison will be applied Odd Ratio>1 implies that the greater odds of association with the predictor variable and the response, Odds Ratio=1 means there is no association between the predictor variable and the response variable, and Odd Ratio<1 Implies that there is a lower odds of association between the predictor variable and the response variable. The odds ratios are as follows: $\boldsymbol{Gender = 1.662, Age = 1.083, education = 0.83, currentSmoker = 1.132,\ cigPerDay = 1.013, BPMeds = 2.274,}$ $\boldsymbol{prevalentStroke = 3.608, prevalentHyp = 2.397, diabetes = 3.423, totChol = 1.005, sysBP = 1.024,}$ $\boldsymbol{diaBP = 1.032, BMI = 1.055,\ heartRate = 1.003\ and\ Glucose = 1.011}$ .Hence, Using the tables and the results in the appendix I can conclude that there are greater odds that an association exist between the response variable and all predictor variables except the patient's education. What I have noticed from this analysis, high odds were observed from the predictor variables that contain medical information.

Multiple Logistic Model (Full Model)

My dataset can be classified as a binary logistic model as it consists of only two options on my response variable. This analysis will help me to create an equation that I can use to predict the probability occurrence of the patient's results for the 10-year risk of coronary heart disease. The multiple logistic models for all predictor variables and the response were successfully created. This model will give me the equation for obtaining the probability of the possible outcome for each patient. The preceding selection will go briefly on how I came up with the model and how it will contribute to my research questions more specifically the determining of the factors. A lot of the coefficients will not be included in the reduced model because their null hypothesis is not rejected as their p-values are greater than the level of significance. Hence, more in-depth tests will be computed in the next section. Using SAS procedure (Logistic), I obtained the following estimated logistic model:

$$\boldsymbol{\hat{Y} = -8.445 + 0.49X_1 + 0.0689X_2 - 0.051X_3 + 0.040X_4 + 0.022X_5 - 0.066X_6 + 0.841X_7 + 0.064X_8 + 0.049X_9 + 0.002X_{10} + 0.015X_{11} - 0.002X_{12} + 0.013X_{13} - 0.007X_{14} + 0.007X_{15}}$$

Where:

$\boldsymbol{X_1 = Gender, X_2 = Age,\quad X_3 = Education,\quad X_4 = CurrentSmoker,\ X_5 = CigsPerDay, X_6 = BPMeds,}$
$\boldsymbol{X_7 = prevalentStroke,\quad X_8 = prevalentHyp,\quad X_9 = Diabetes,\quad X_{10} = TotChol,}$
$\boldsymbol{X_{11} = sysBP,\quad X_{12} = DiaBP,\quad X_{13} = BMI\ ,\quad X_{14} = HeartRate,}$
$\boldsymbol{X_{15} = Glucose\ and\ \overline{Y} = 10 - year\ risk\ of\ coronary\ heart\ disease}$

See the estimators under appendix

Hypothesis testing

It is essential for one when interpreting finding to assess whether these finding are relevant or not towards your research. Under this section I will use the systematic procedure for deciding whether the results I came with supports the theory which applies to a population. I will start by testing the findings from my correlation analysis. Different correlations were computed under the correlation analysis for finding the relationship between the specified variables per test. Using the Pearson correlation coefficients, the correlation between the predictor variables computed in Table 1 shows that most of the null hypothesis for most predictors is rejected($\boldsymbol{P - value < 0.05}$) which implies that there is relationship between the predictor variables and other predictor variable . The test for the correlation between the gender variable and the other variables shows that gender is correlated with (currentSmoker, CigsPerDay, BPMeds, totChol, diaBP, BMI and heartrate) as their

p-values are less that the level of significance $\boldsymbol{p-values<0.05}$ for the Pearson correlation coefficient , which makes it uncorrelated with the other predictor variables (Age, education, prevalentStrike, prevalentHyp, diabetes, sysBP and Glucose) as their p-values are more than the level of significance $\boldsymbol{p-values>0.05}$.

The test for the correlation between the age variable and the other variables shows that age is correlated with all the other predictor variables except for the heartrate as their p-values are less that the level of significance $\boldsymbol{p-values<0.05}$ for the Pearson correlation coefficient , which makes it uncorrelated with the other predictor variable which is Heartrate as its p-value is more than the level of significance $\boldsymbol{p-value>0.05.}$

Testing for correlation for the education and the other predictor variables, the results show that the education predictor variable is correlated with (Age, prevalentHyp, sysBP, diaBP, BMI, HeartRate) as their p-values are less than the level of significance $\boldsymbol{p-values<0.05.}$ which makes the education predictor variable uncorrelated with (Gender, currentSmoker, CigsPerDay, PrevalentStroke, diabetes, totChol, and glucose) as their p-values are greater than the one of significance $\boldsymbol{p-values>0.05.}$

Using the Pearson correlation coefficients, the currentSmoker predictor variable is correlated with all the other predictor variables except for the variable (Education) because its $\boldsymbol{P-value=0.2062>0.05}$ which makes is uncorrelated as we will accept the null hypothesis that there is no relationship between these variables. Se the table for more results for correlation.

R-square and Coefficient of determination

The R-square is a measure in a regression model that will help me to determine the proportion of variance in the response variable that can be explained by the predictor variables. It can be simplified as statistic that gives an insight on how well the data fit the regression model. This measure ranges from 0-1, where 0 represent poorly fit and 1 represents perfect fit. Note that that R-square being small does not necessary the problem, and high R-square are not necessarily good. I obtained a value of $\boldsymbol{R-square=0.1011}$ (see Table 4), this implies that the variation in patient's 10-year risk of coronary heart disease is reduced by 10.11% when all the predictor variables are considered. As said previously that less value does not mean the model is not precise with the prediction, in some cases a lower R-square value is recommended.

Multiple Logistic Model (Reduced Model)

These options were explained under the methodology section under multiple logistics model. I will apply these selection methods as explained to drop factors that are not improving my model.

- Forward (see Table 5 in Appendix)
  The model formed: $\boldsymbol{\widehat{Y}=-8.6560-0.246X_1+0.0716X_2+0.022X_3+0.016X_4+0.008X_5}$

  Where: $X_1=Gender(Female), X_2=Age, X_3=CigsPerDay, X_4=sysBP, X_5=glucose$

  Odds Ratio: $\boldsymbol{male\ 0\ vs\ 1\ =\ 0.612}$ For females equals 0.612, which means that females are 0.612 times more likely to have the coronary heart disease than males.

  Odds ratio: $\boldsymbol{Age=1.074}$, the odds of a patient's having coronary disease after 10 years. This means that the odds ration Age indicates that patients who are older are more likely to have 10-year risk of CHD.

Odds Ration: $\boldsymbol{CigsPerDay = 1.022}$: the odds of a patient's having coronary disease after 10 years. This means that the odds ratio CigsPerDay indicates that patients who smokes 1,022 times more are more likely to have 10-year risk of CHD.

Odds Ration: $SysBP = 0.0158$: This means that patience who have less sysBP are more likely not to have the coronary heart disease in 10 years.

Odds ratio:$Glucose = 0.0077$ this means patience who have lower glucose level are more likely not to have the risk of coronary heart disease in 10 years

- Backward (see Table 6)

The model formed: $\boldsymbol{\hat{Y} = -7.994 - 0.245X_1 + 0.071X_2 + 0.221X_3 - 0.439X_4 + 0.016X_5 + 0.008X_6}$

Where: $X_1 = Gender(Female)$ $X_2 = Age$ $X_3 = CigsPerday$, $X_4 = prevalentStrike(no)$, $X_5 = sysBP$ $X_6 = Glucose$

Odds Ratio: $\boldsymbol{male\ 0\ vs\ 1 = 0.612}$ For females equals 0.612, which means that females are 0.612 times more likely to have the coronary heart disease than males.

Odds ratio: $\boldsymbol{Age = 1.074}$, the odds of a patient's having coronary disease after 10 years. This means that the odds ration Age indicates that patients who are older are more likely to have 10-year risk of CHD.

Odds Ration: $\boldsymbol{CigsPerDay = 1.022}$: the odds of a patient's having coronary disease after 10 years. This means that the odds ratio CigsPerDay indicates that patients who smokes 1,022 times more are more likely to have 10-year risk of CHD.

Odds Ration: $\boldsymbol{SysBP = 0.0158}$: This means that patients who have less sysBP are more likely not to have the coronary heart disease in 10 years.

Odds ratio:$Glucose = 0.008$ this means patience who have lower glucose level are more likely not to have the risk of coronary heart disease in 10 years

- Stepwise (see Table 7)
  - The model formed: $\boldsymbol{\hat{Y} = -8.411 - 0.246X_1 + 0.072X_2 + 0.022X_3 + 0.016X_4 + 0.008X_5}$

Where: $X_1 = Gender(Female)., X_2 = Age, X_3 = cigsPerday, X_4 = sysBP, X_5 = Glucose$

ROC Chart

As stated under my methodology that ROC curve is a plot of the sensitivity vs 1-specity of a test. I used the ROC procedure from SAS to determine a curve for each model.

- For the Full Model (Not Reduced): Area under curve :0.7393
- For the Reduced Model using (Forward technique): Area under curve=0.7363
- For the Reduced Model using (Backward technique): Area under curve=0.7364
- For the Reduced Model Using (Stepwise technique): Area under curve =0.7363

The indexes for all my model represent a strong ROC index as they are all greater than 0.7. Noticing that the index of the Forward and the Stepwise technique are the same, it is not surprising because stepwise combines both the forward and backward technique. As stated under the methodology that this curve will help me in

determining how accurate are my model. The AUC measures the accuracy(discrimination) of the test. This measure ranges from 0 to 1 where 0 indicates an inaccurate test and on other hand 1 represents a perfectly accurate test. In addition, a moderate scale of 0,5 is classified as a no discrimination test. Thus, for this case, I can conclude that the logistic model was accurate for this study as its index represent a strong accurate test.

Summary and Conclusion

From all the computations I have performed for this study and all my findings from different sources. The gender claims made by my sources that women are more likely to have the 10-year risk of coronary heart disease from my findings were true. Therefore, it is critical that women become more aware of their own risk factors as that would prevent them from having the 10-year risk of coronary heart disease. Furthermore, Age was also found to be a good predictor of the 10-year risk of coronary disease which also supports the claim made by one of my sources that this model works well with middle-aged patients which means age is another good factor for the prediction. Smoking also contributes to the prediction of the risk of coronary heart disease after 10 years. Also, from my analysis, I see that smoking does not contribute but the number of cigarettes a patient smokes a day determines whether the patient has a risk of coronary disease after 10 years. Thus, I can conclude that a patient's lifestyle has an influence on the risk of coronary disease in 10 years. For the study I conducted total cholesterol level, my model did not identify it as a factor that is a good factor in a 10-year risk of coronary disease. I can conclude that an increase in total cholesterol does not contribute to the risk of coronary disease in 10 years. Glucose was also found to be a factor that contributes to the prediction of a 10-year risk of coronary, I can confidently say that women that have a high level of glucose and smokes more than the average of 20 cigarettes as this study indicated are more likely to have the risk of coronary disease after 10 years.

Limitations: The study was done in a short period of time; I did not get enough time for gathering data and analysing more data.

Recommendation: I would recommend that this study should have more factors to analyse from so a researcher can be able to determine also factors that are good predictors, and the sample size should be larger than these one so we can get results that will be based on a large sample which means it will be applicable to a large d=group of people.

Appendix

1.1 Descriptive Analysis Graphs and Statistics for Demographic information(click here to see SAS code for descriptive stats)

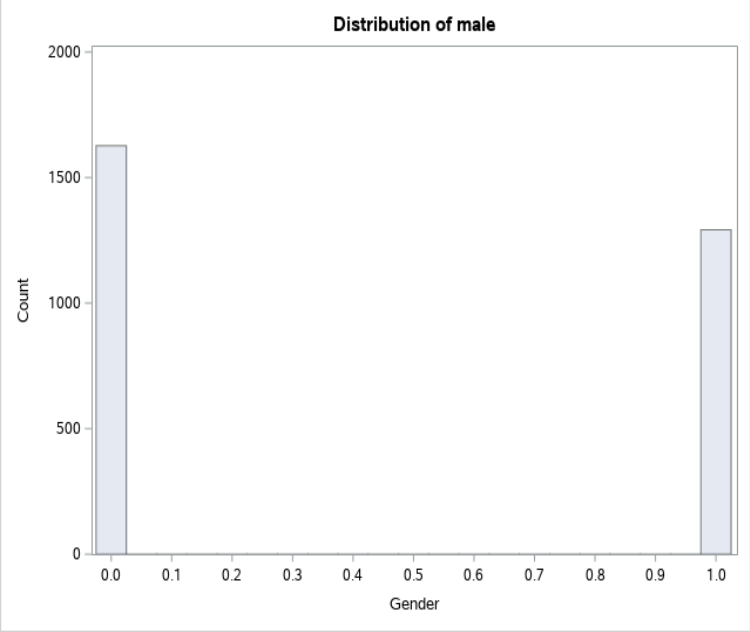


Figure 1.1.1: Distribution of Patients' Gender

THE Discriptive Statistics for the Patients Gender

The UNIVARIATE Procedure
Variable: male (Gender)

Moments			
N	2919	Sum Weights	2919
Mean	0.44261733	Sum Observations	1292
Std Deviation	0.49678142	Variance	0.24679178
Skewness	0.23117615	Kurtosis	-1.9478927
Uncorrected SS	1292	Corrected SS	720.138404
Coeff Variation	112.237226	Std Error Mean	0.00919493

Figure 1.1.2: Descriptive Statistics for Patients' Gender

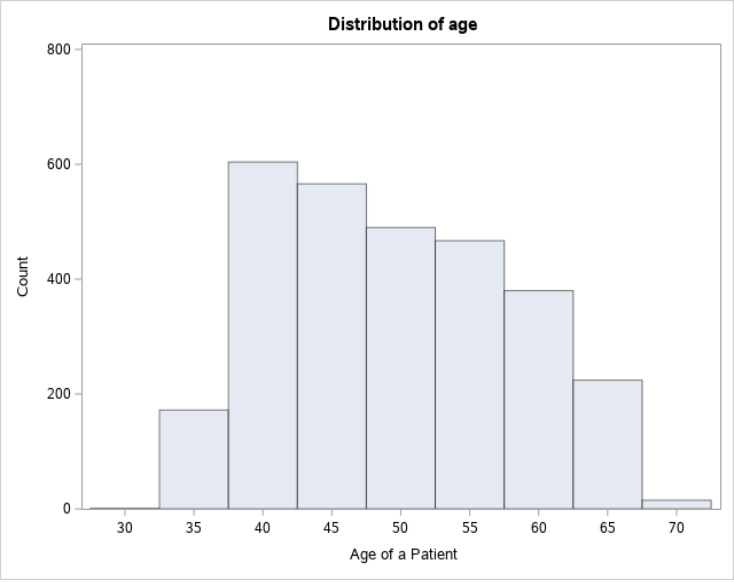


Figure 1.1.3: Distribution of Patients' Age

Descriptive Statistics for the Patients Age

The UNIVARIATE Procedure
Variable: age (Age of a Patient)

Moments			
N	2919	Sum Weights	2919
Mean	49.4354231	Sum Observations	144302
Std Deviation	8.55662062	Variance	73.2157564
Skewness	0.23615862	Kurtosis	-0.9984649
Uncorrected SS	7347274	Corrected SS	213643.577
Coeff Variation	17.3086829	Std Error Mean	0.15837449

Figure 1.1.4: Descriptive Statistics for Patients' Age

1.2 Descriptive Analysis Graphs and Statistics for behavioural information

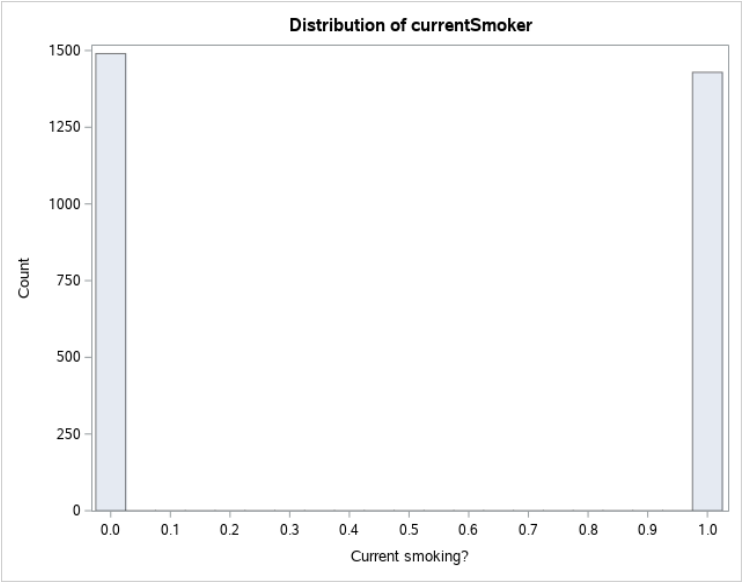


Figure 1.2.1: Distribution for patients whether are they smoking currently or not

Descriptive Statistics Whether a Patient is Smoking or not

The UNIVARIATE Procedure
Variable: currentSmoker (Current smoking?)

Moments			
N	2919	Sum Weights	2919
Mean	0.48955122	Sum Observations	1429
Std Deviation	0.49997646	Variance	0.24997646
Skewness	0.04182576	Kurtosis	-1.9996211
Uncorrected SS	1429	Corrected SS	729.431312
Coeff Variation	102.129551	Std Error Mean	0.00925406

Figure1.2.2: Descriptive statistics for patients whether they are currently smoking or not

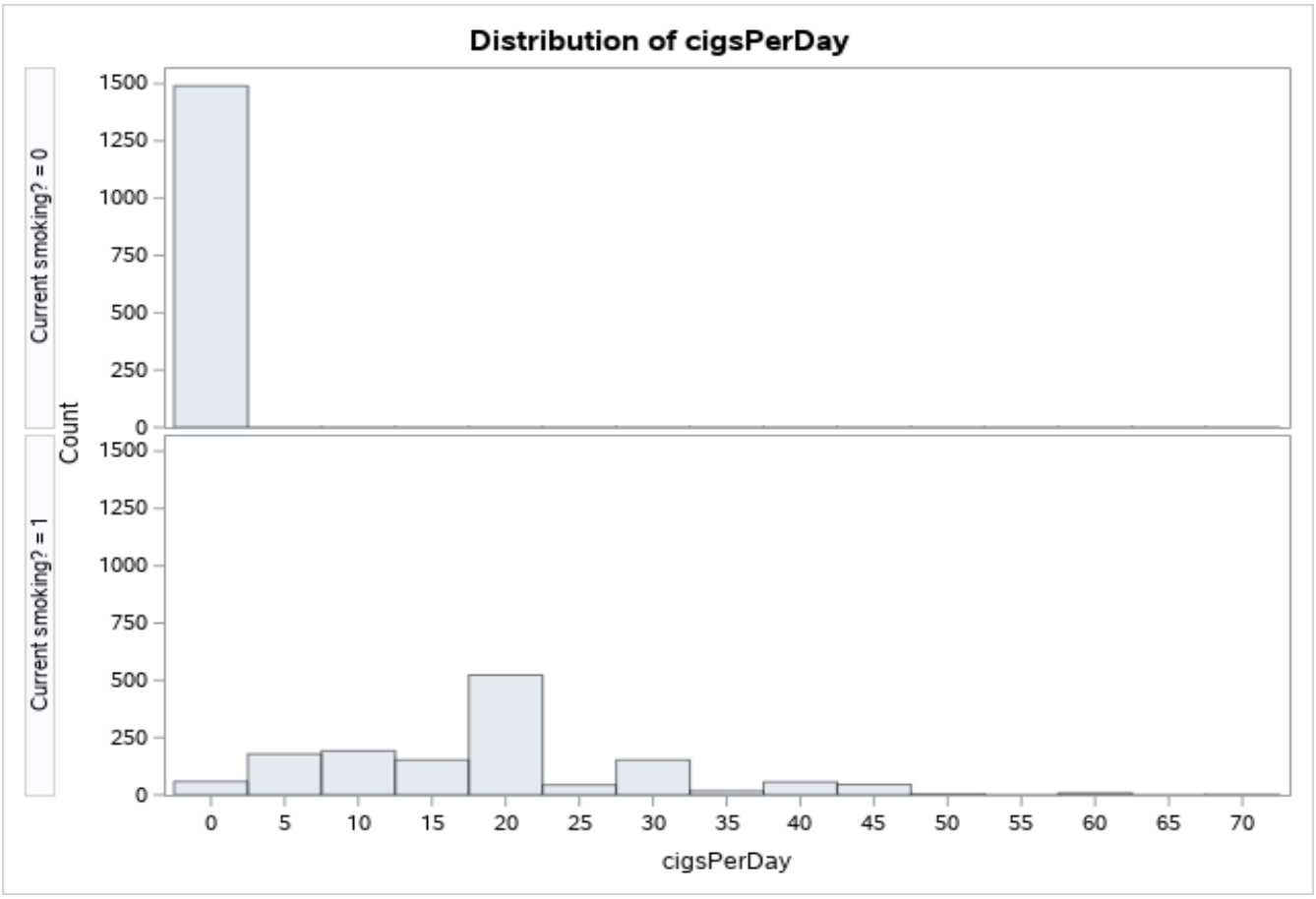


Figure1.2.3: Histogram for Number of Cigarettes a Patient smokes a Day

Descriptive Statistics of Number of Cigarattes that a patient Smokes a day

The UNIVARIATE Procedure
Variable: cigsPerDay
currentSmoker = 1

Moments			
N	1429	Sum Weights	1429
Mean	18.5948216	Sum Observations	26572
Std Deviation	10.9595769	Variance	120.112326
Skewness	0.73298444	Kurtosis	0.9235695
Uncorrected SS	665622	Corrected SS	171520.402
Coeff Variation	58.9388658	Std Error Mean	0.28991967

Figure 1.2.4: Descriptive Statistics pf number of Cigarettes that a patient smokes a Day

1.3 Descriptive Analysis Graphs and Statistics for Medical Condition related information (Nominal variable)

Descriptive Statistics whetehr a patient wa on Blood Pressure Medication

The UNIVARIATE Procedure
Variable: BPMeds

Moments			
N	2919	Sum Weights	2919
Mean	0.02911956	Sum Observations	85
Std Deviation	0.16817045	Variance	0.0282813
Skewness	5.60387728	Kurtosis	29.4236002
Uncorrected SS	85	Corrected SS	82.5248373
Coeff Variation	577.517121	Std Error Mean	0.00311267

Figure 1.3.1: Descriptive statistics whether a patient was on Blood pressure medication or not

Descriptive Statistics whetether a patient had stroke

The UNIVARIATE Procedure
Variable: prevalentStroke (previously had as Stoke)

Moments			
N	2919	Sum Weights	2919
Mean	0.0061665	Sum Observations	18
Std Deviation	0.07829796	Variance	0.00613057
Skewness	12.6228606	Kurtosis	157.444486
Uncorrected SS	18	Corrected SS	17.8890031
Coeff Variation	1269.73186	Std Error Mean	0.00144922

Figure 1.3.3: Descriptive statistics whether a patient had a stroke or not

Descriptive Statistics whetether a patient was Hypertensive

The UNIVARIATE Procedure
Variable: prevalentHyp (was hypertensive)

Moments			
N	2919	Sum Weights	2919
Mean	0.31209318	Sum Observations	911
Std Deviation	0.46342702	Variance	0.2147646
Skewness	0.8115006	Kurtosis	-1.342387
Uncorrected SS	911	Corrected SS	626.683111
Coeff Variation	148.489953	Std Error Mean	0.00857757

Figure 1.3.3: Whether a patient was hypertensive or not

Descriptive Statistics whetether a patient had Diabetes

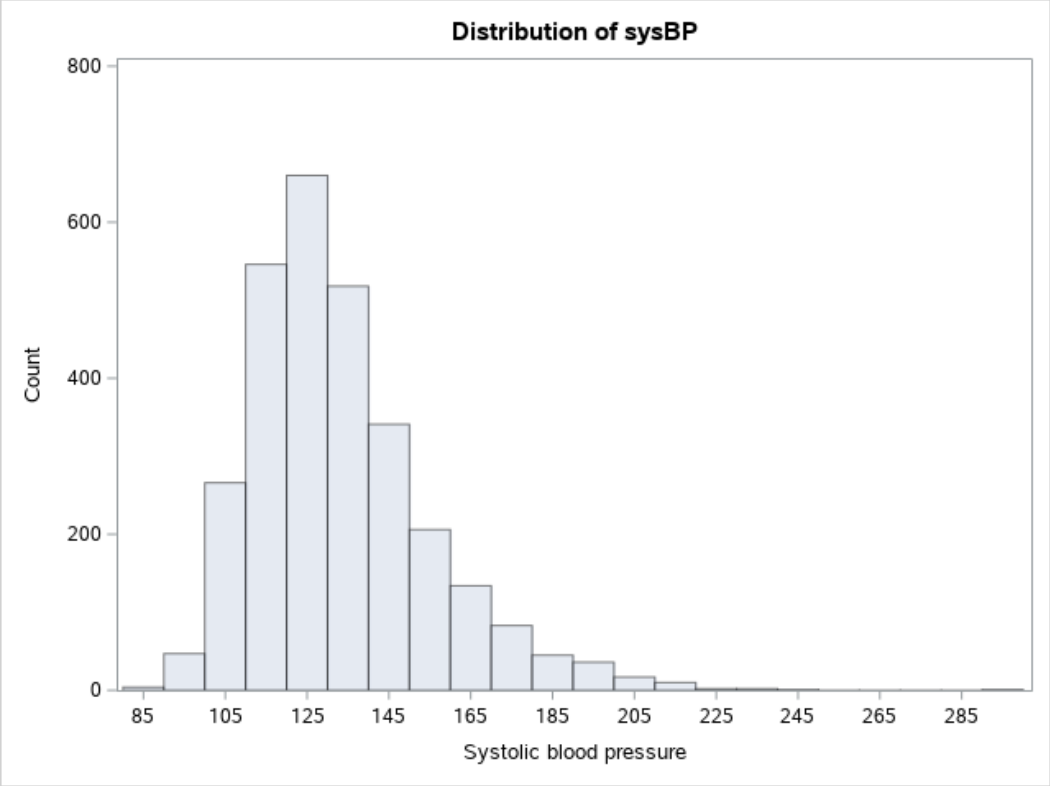
The UNIVARIATE Procedure
Variable: diabetes (had diabetes)

Moments			
N	2919	Sum Weights	2919
Mean	0.02603631	Sum Observations	76
Std Deviation	0.15927057	Variance	0.02536711
Skewness	5.95576009	Kurtosis	33.4940268
Uncorrected SS	76	Corrected SS	74.0212402
Coeff Variation	611.724729	Std Error Mean	0.00294794

Figure 1.3.4: Distribution for patients who had diabetes or not

1.4 Descriptive Analysis Graphs and Statistics for Medical Condition related information (Continuous variable)

1.4.1 This page will only display the descriptive statistics for Systolic Blood Pressure



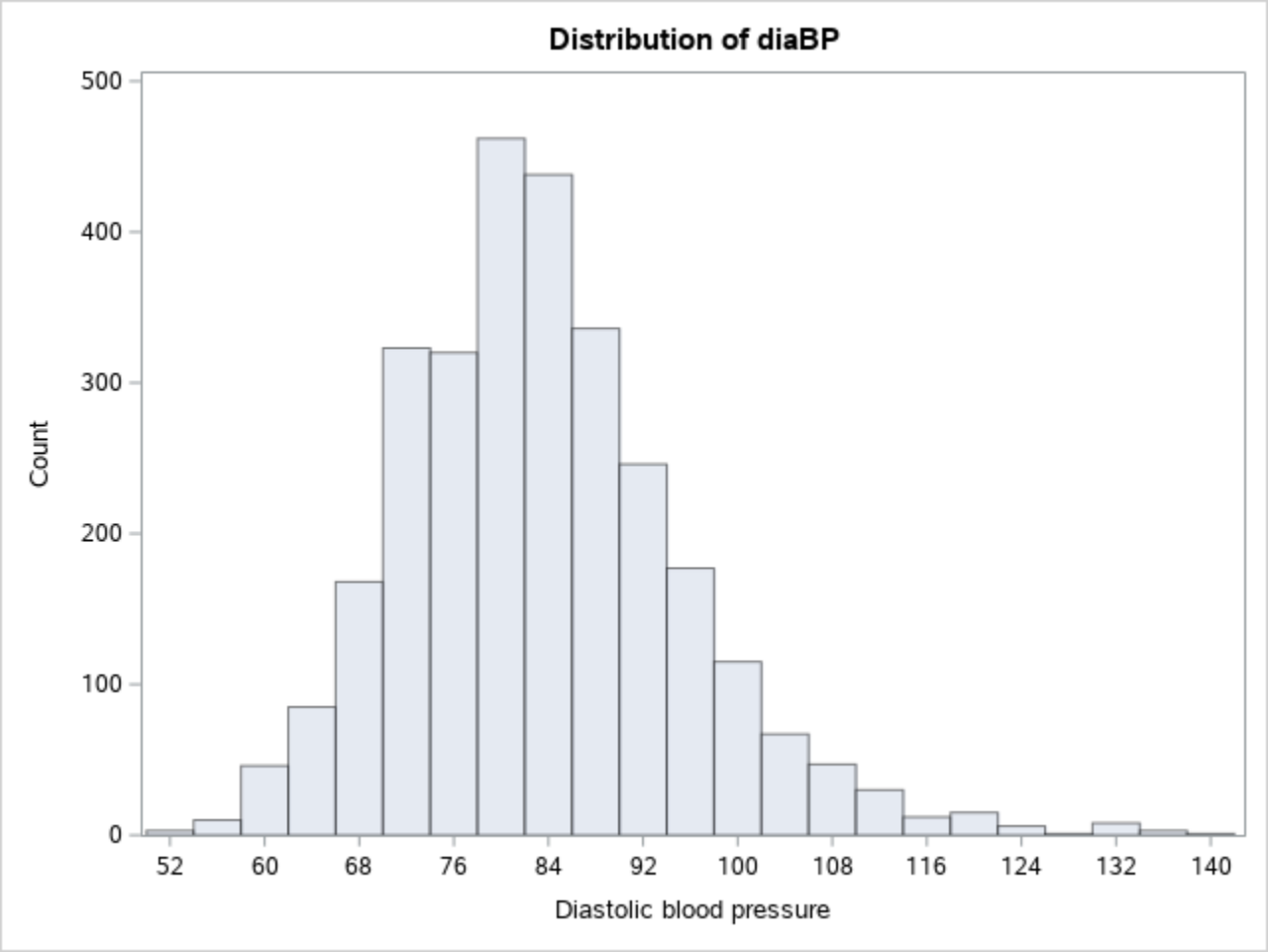
Descriptive Statistics for Systolic Blood pressure

The UNIVARIATE Procedure
Variable: sysBP (Systolic blood pressure)

Moments			
N	2919	Sum Weights	2919
Mean	132.412641	Sum Observations	386512.5
Std Deviation	22.0339535	Variance	485.495107
Skewness	1.1731263	Kurtosis	2.39926385
Uncorrected SS	52595815.8	Corrected SS	1416674.72
Coeff Variation	16.6403701	Std Error Mean	0.40782644

Basic Statistical Measures			
Location		Variability	
Mean	132.4126	Std Deviation	22.03395
Median	128.0000	Variance	485.49511
Mode	130.0000	Range	211.50000
		Interquartile Range	26.50000

1.4.2 This page will only display the descriptive statistics for Diastolic Blood Pressure



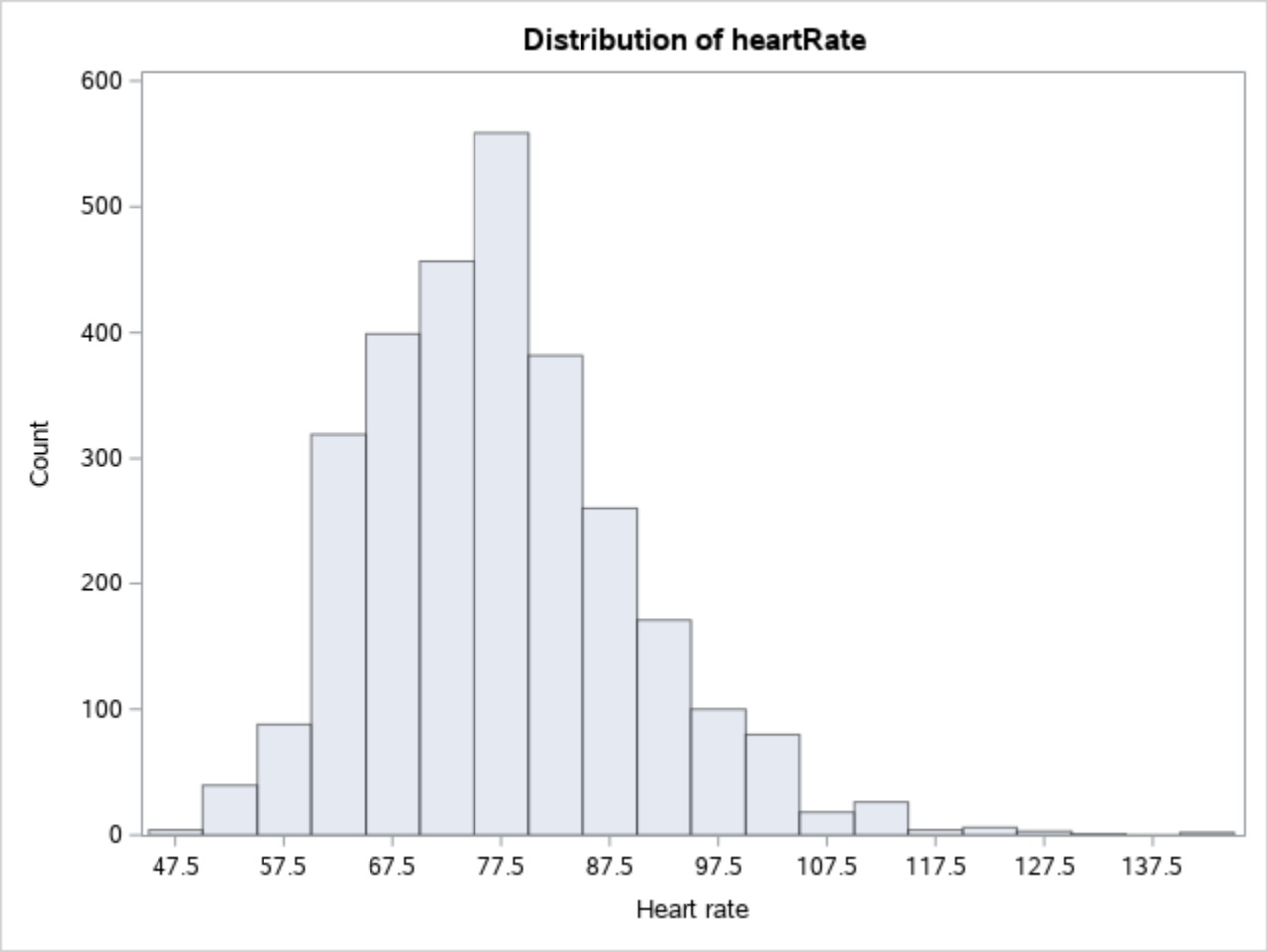
Descriptive Statistics for Diastolic Blood pressure

The UNIVARIATE Procedure
Variable: diaBP (Diastolic blood pressure)

Moments			
N	2919	Sum Weights	2919
Mean	82.9284001	Sum Observations	242068
Std Deviation	11.8565439	Variance	140.577634
Skewness	0.71432955	Kurtosis	1.33065817
Uncorrected SS	20484517.5	Corrected SS	410205.536
Coeff Variation	14.2973263	Std Error Mean	0.21945277

Basic Statistical Measures			
Location		Variability	
Mean	82.92840	Std Deviation	11.85654
Median	82.00000	Variance	140.57763
Mode	80.00000	Range	89.00000
		Interquartile Range	14.50000

1.4.3 This page will only display the descriptive statistics for patients' heart rate



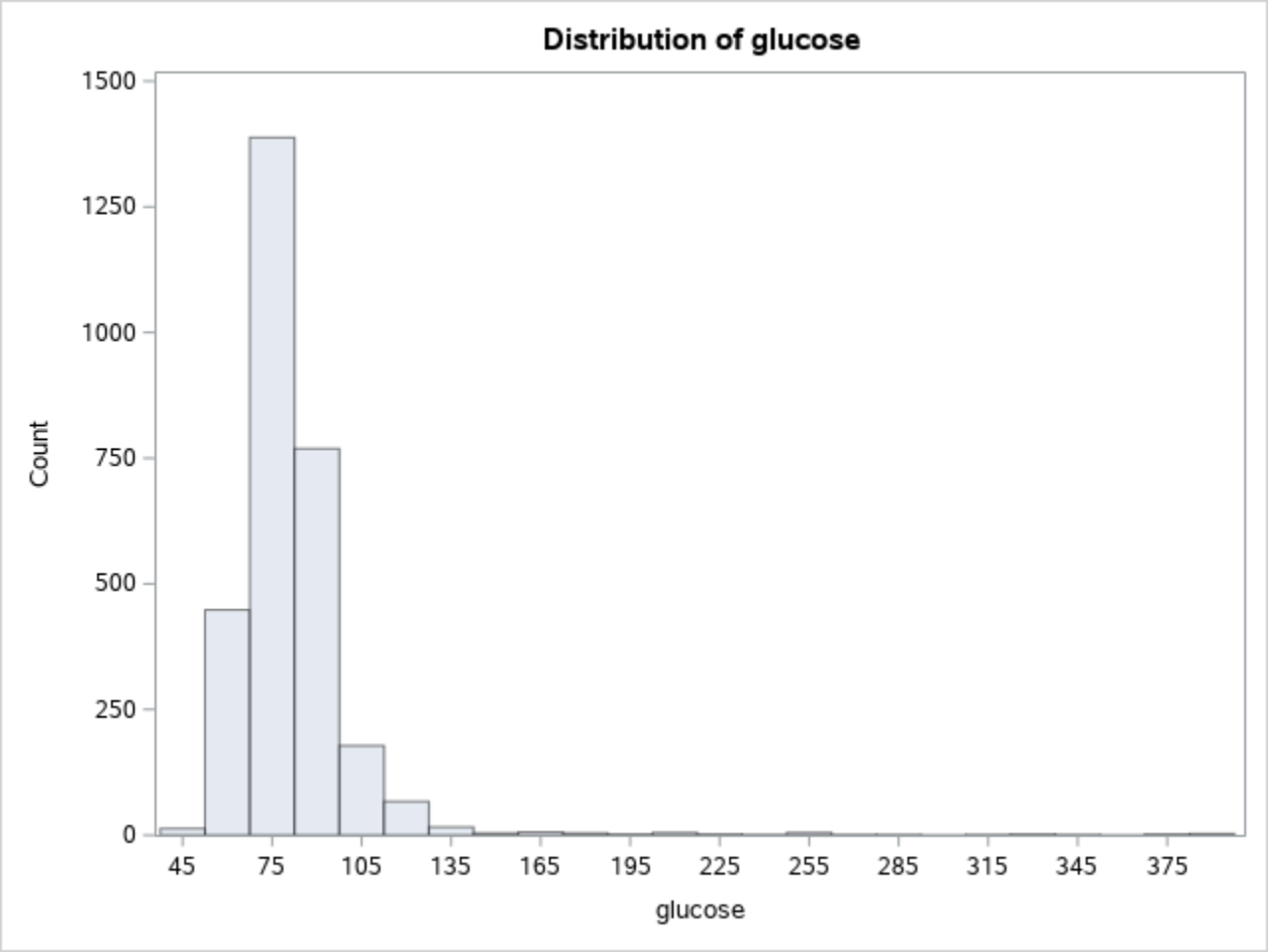
Descriptive Statistics for Diastolic Blood pressure

The UNIVARIATE Procedure
Variable: heartRate (Heart rate)

Moments			
N	2919	Sum Weights	2919
Mean	75.8208291	Sum Observations	221321
Std Deviation	12.0009894	Variance	144.023747
Skewness	0.73128009	Kurtosis	1.18559082
Uncorrected SS	17201003	Corrected SS	420261.294
Coeff Variation	15.8280905	Std Error Mean	0.22212631

Basic Statistical Measures			
Location		Variability	
Mean	75.82083	Std Deviation	12.00099
Median	75.00000	Variance	144.02375
Mode	75.00000	Range	98.00000
		Interquartile Range	14.00000

1.4.4 This page will only display the descriptive statistics for patients' glucose level



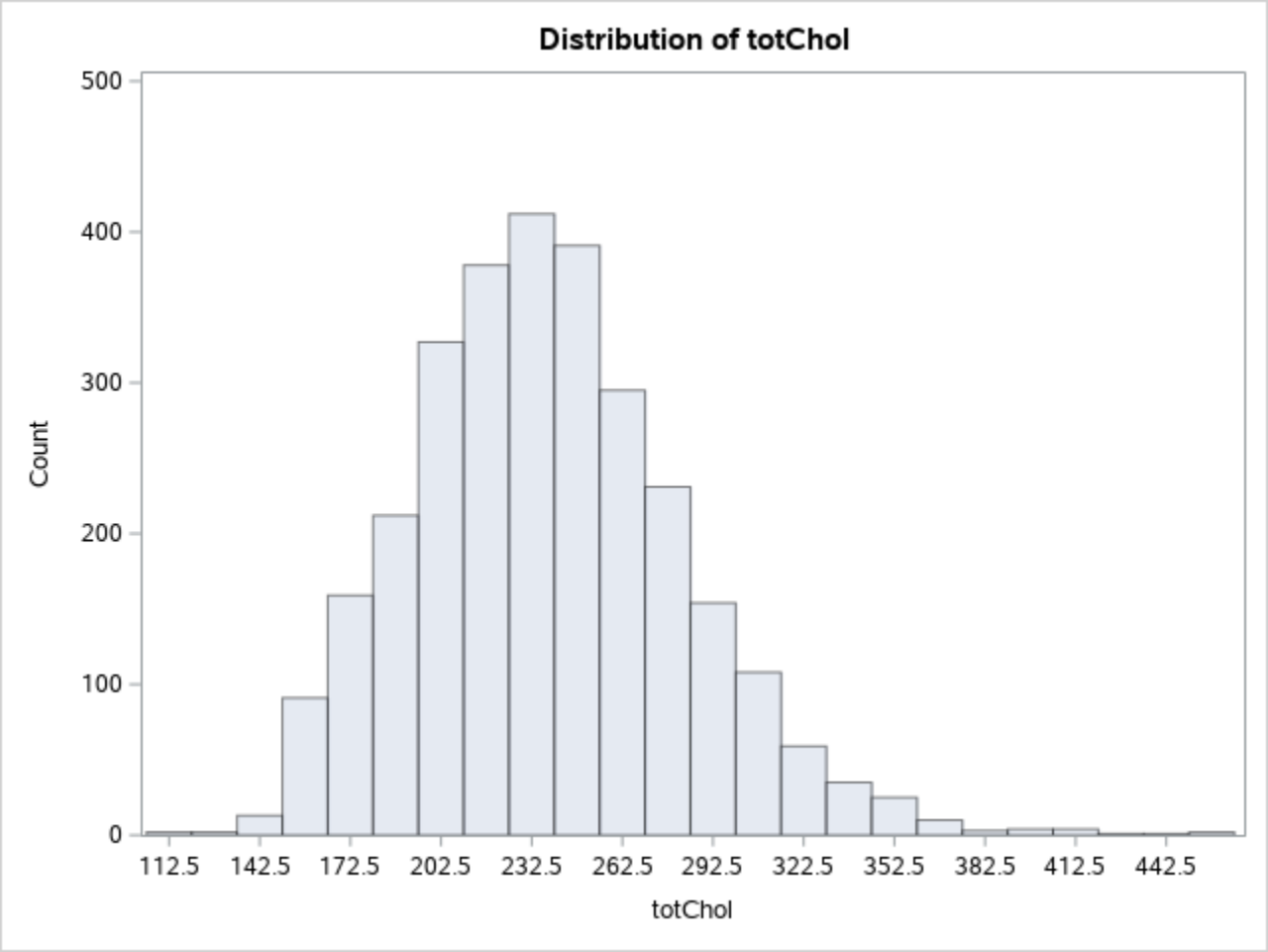
Descriptive Statistics for Patients Glucose level

The UNIVARIATE Procedure
Variable: glucose

Moments			
N	2919	Sum Weights	2919
Mean	81.785543	Sum Observations	238732
Std Deviation	24.3540522	Variance	593.119859
Skewness	6.5538226	Kurtosis	64.1951722
Uncorrected SS	21255550	Corrected SS	1730723.75
Coeff Variation	29.7779428	Std Error Mean	0.45076915

Basic Statistical Measures			
Location		Variability	
Mean	81.78554	Std Deviation	24.35405
Median	78.00000	Variance	593.11986
Mode	75.00000	Range	354.00000
		Interquartile Range	16.00000

1.4.4 This page will only display the descriptive statistics for patients' Total cholesterol Level



Descriptive Statistics for Patients Toatal cholesterol level

The UNIVARIATE Procedure
Variable: totChol

Moments			
N	2919	Sum Weights	2919
Mean	236.906817	Sum Observations	691531
Std Deviation	44.5425714	Variance	1984.04066
Skewness	0.55170972	Kurtosis	0.79535954
Uncorrected SS	169617839	Corrected SS	5789430.65
Coeff Variation	18.8017263	Std Error Mean	0.82443845

Basic Statistical Measures			
Location		Variability	
Mean	236.9068	Std Deviation	44.54257
Median	234.0000	Variance	1984
Mode	240.0000	Range	351.00000
		Interquartile Range	58.00000

2 Correlation Analysis

Table 1: Correlation Analysis for Multicollinearity

Pearson Correlation Coefficients, N = 2919 Prob > |r| under H0: Rho=0															
	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
male Gender	1.00000	-0.02520 0.1735	0.02844 0.1245	0.20903 <.0001	0.33213 <.0001	-0.05588 0.0025	-0.00852 0.6454	0.00860 0.6424	0.01889 0.3076	-0.05727 0.0020	-0.03624 0.0502	0.06470 0.0005	0.06493 0.0004	-0.11080 <.0001	0.00660 0.7214
age Age of a Patient	-0.02520 0.1735	1.00000	-0.16670 <.0001	-0.20997 <.0001	-0.18730 <.0001	0.12455 <.0001	0.05482 0.0031	0.30588 <.0001	0.11540 <.0001	0.27145 <.0001	0.39805 <.0001	0.21777 <.0001	0.12998 <.0001	0.02043 0.2699	0.12376 <.0001
education Education level	0.02844 0.1245	-0.16670 <.0001	1.00000	0.02340 0.2062	0.00897 0.6283	-0.01452 0.4328	-0.03271 0.0772	-0.06774 0.0002	-0.03467 0.0611	-0.00897 0.6280	-0.12355 <.0001	-0.05417 0.0034	-0.12829 <.0001	-0.07411 <.0001	-0.02588 0.1622
currentSmoker Current smoking?	0.20903 <.0001	-0.20997 <.0001	0.02340 0.2062	1.00000	0.77150 <.0001	-0.04325 0.0194	-0.04212 0.0229	-0.09019 <.0001	-0.04823 0.0092	-0.04459 0.0160	-0.12665 <.0001	-0.10381 <.0001	-0.15880 <.0001	0.05843 0.0016	-0.05484 0.0030
cigsPerDay	0.33213 <.0001	-0.18730 <.0001	0.00897 0.6283	0.77150 <.0001	1.00000	-0.04224 0.0225	-0.04135 0.0255	-0.05694 0.0021	-0.04425 0.0168	-0.02284 0.2172	-0.08635 <.0001	-0.04470 0.0157	-0.08100 <.0001	0.07317 <.0001	-0.05895 0.0014
BPMeds was on Blood Pressure Med	-0.05588 0.0025	0.12455 <.0001	-0.01452 0.4328	-0.04325 0.0194	-0.04224 0.0225	1.00000	0.11649 <.0001	0.25712 <.0001	0.07404 <.0001	0.09895 <.0001	0.25835 <.0001	0.18521 <.0001	0.11429 <.0001	0.02653 0.1519	0.06947 0.0002
prevalentStroke previously had as Stoke	-0.00852 0.6454	0.05482 0.0031	-0.03271 0.0772	-0.04212 0.0229	-0.04135 0.0255	0.11649 <.0001	1.00000	0.06972 0.0002	0.01460 0.4303	0.01481 0.4239	0.05683 0.0021	0.04920 0.0078	0.04027 0.0296	-0.01524 0.4106	0.02406 0.1938
prevalentHyp was hypertensive	0.00860 0.6424	0.30588 <.0001	-0.06774 0.0002	-0.09019 <.0001	-0.05694 0.0021	0.25712 <.0001	0.06972 0.0002	1.00000	0.08952 <.0001	0.15654 <.0001	0.69211 <.0001	0.61071 <.0001	0.31387 <.0001	0.16201 <.0001	0.09283 <.0001
diabetes had diabetes	0.01889 0.3076	0.11540 <.0001	-0.03467 0.0611	-0.04823 0.0092	-0.04425 0.0168	0.07404 <.0001	0.01460 0.4303	0.08952 <.0001	1.00000	0.05048 0.0064	0.10939 <.0001	0.05325 0.0040	0.07240 <.0001	0.04709 0.0110	0.60328 <.0001
totChol Toatl Cholesterol Level	-0.05727 0.0020	0.27145 <.0001	-0.00897 0.6280	-0.04459 0.0160	-0.02284 0.2172	0.09895 <.0001	0.01481 0.4239	0.15654 <.0001	0.05048 0.0064	1.00000	0.21881 <.0001	0.17749 <.0001	0.10784 <.0001	0.10197 <.0001	0.05036 0.0065
sysBP Systolic blood pressure	-0.03624 0.0502	0.39805 <.0001	-0.12355 <.0001	-0.12665 <.0001	-0.08635 <.0001	0.25835 <.0001	0.05683 0.0021	0.69211 <.0001	0.10939 <.0001	0.21881 <.0001	1.00000	0.78301 <.0001	0.34435 <.0001	0.19852 <.0001	0.14498 <.0001
diaBP Diastolic blood pressure	0.06470 0.0005	0.21777 <.0001	-0.05417 0.0034	-0.10381 <.0001	-0.04470 0.0157	0.18521 <.0001	0.04920 0.0078	0.61071 <.0001	0.05325 0.0040	0.17749 <.0001	0.78301 <.0001	1.00000	0.39829 <.0001	0.18708 <.0001	0.07533 <.0001
BMI Body Mass Index	0.06493 0.0004	0.12998 <.0001	-0.12829 <.0001	-0.15880 <.0001	-0.08100 <.0001	0.11429 <.0001	0.04027 0.0296	0.31387 <.0001	0.07240 <.0001	0.10784 <.0001	0.34435 <.0001	0.39829 <.0001	1.00000	0.07101 0.0001	0.07969 <.0001
heartRate Heart rate	-0.11080 <.0001	0.02043 0.2699	-0.07411 <.0001	0.05843 0.0016	0.07317 <.0001	0.02653 0.1519	-0.01524 0.4106	0.16201 <.0001	0.04709 0.0110	0.10197 <.0001	0.19852 <.0001	0.18708 <.0001	0.07101 0.0001	1.00000	0.09502 <.0001
glucose Glucose Level	0.00660 0.7214	0.12376 <.0001	-0.02588 0.1622	-0.05484 0.0030	-0.05895 0.0014	0.06947 0.0002	0.02406 0.1938	0.09283 <.0001	0.60328 <.0001	0.05036 0.0065	0.14498 <.0001	0.07533 <.0001	0.07969 <.0001	0.09502 <.0001	1.00000

Table 2: Correlation between each predictor variables and the response variable

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
male	1.0000	1.662	1.356	2.038

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
age	1.0000	1.083	1.070	1.097

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
education	1.0000	0.830	0.748	0.922

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
currentSmoker	1.0000	1.132	0.924	1.386

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
cigsPerDay	1.0000	1.013	1.005	1.021

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
BPMeds	1.0000	2.274	1.402	3.688

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
prevalentStroke	1.0000	3.608	1.391	9.357

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
prevalentHyp	1.0000	2.397	1.951	2.946

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
diabetes	1.0000	3.423	2.123	5.518

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
totChol	1.0000	1.005	1.003	1.007

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
sysBP	1.0000	1.024	1.019	1.028

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
diaBP	1.0000	1.032	1.024	1.040

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
BMI	1.0000	1.055	1.030	1.080

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
heartRate	1.0000	1.003	0.995	1.012

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
glucose	1.0000	1.011	1.007	1.014

SAS Code Used to generate the Multicollinearity logistic

```
*MULTICOLLINEARITY;
ODS PDF;
ODS GRAPHICS ON;
PROC CORR DATA=WORK.CLEANDATA ;
VAR MALE       AGE     EDUCATION      CURRENTSMOKER
CIGSPERDAY     BPMEDS          PREVALENTSTROKE       PREVALENTHYP   DIABETES       TOTCHOL

SYSBP   DIABP   BMI     HEARTRATE      GLUCOSE;
RUN;
ODS GRAPHICS OFF;
ODS PDF CLOSE;


*SIMPLE LOGISTIC PER PREDICTOR VARIBALE;
*PREDICTOR VARIABLE>> MALE  AGE     EDUCATION      CURRENTSMOKER
CIGSPERDAY     BPMEDS          PREVALENTSTROKE       PREVALENTHYP   DIABETES       TOTCHOL
SYSBP   DIABP   BMI     HEARTRATE      GLUCOSE;
ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING;
MODEL TENYEARCHD = MALE/ RISKLIMITS;
RUN;
ODS HTML CLOSE;
```

This code only has one simple logistic model (Gender) to save space to view full code click : here!

Table 3: Multiple Logistic Model Full model

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	291.8037	15	<.0001
Score	295.0442	15	<.0001
Wald	244.9444	15	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.4446	0.7966	112.3703	<.0001
male	1	0.4900	0.1226	15.9848	<.0001
age	1	0.0688	0.00756	82.8979	<.0001
education	1	-0.0509	0.0555	0.8399	0.3594
currentSmoker	1	0.0401	0.1755	0.0521	0.8194
cigsPerDay	1	0.0219	0.00685	10.1937	0.0014
BPMeds	1	-0.0658	0.2829	0.0541	0.8160
prevalentStroke	1	0.8406	0.5307	2.5087	0.1132
prevalentHyp	1	0.0636	0.1553	0.1679	0.6820
diabetes	1	0.0494	0.3558	0.0193	0.8897
totChol	1	0.00167	0.00127	1.7359	0.1877
sysBP	1	0.0153	0.00431	12.6769	0.0004
diaBP	1	-0.00198	0.00733	0.0732	0.7867
BMI	1	0.0130	0.0143	0.8191	0.3654
heartRate	1	-0.00696	0.00478	2.1242	0.1450
glucose	1	0.00749	0.00244	9.4491	0.0021

SAS Code Used to generate the Multiple logistic model

```
*Multiple logistic model (Full model);

ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING ;
MODEL TENYEARCHD = MALE     AGE     EDUCATION     CURRENTSMOKER        CIGSPERDAY
BPMEDS         PREVALENTSTROKE      PREVALENTHYP  DIABETES      TOTCHOL  SYSBP
       DIABP  BMI     HEARTRATE     GLUCOSE
;
RUN;
ODS HTML CLOSE;
```

Table 4: Anova and R-square

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	37.91117	2.52741	21.76	<.0001
Error	2903	337.16043	0.11614		
Corrected Total	2918	375.07160			

Root MSE	0.34080	R-Square	0.1011
Dependent Mean	0.15142	Adj R-Sq	0.0964
Coeff Var	225.06430		

SAS CODE USED TO GENERATE THE ANOVA TABLE AND R-SQUARED

```
ODS HTML;
PROC REG DATA=WORK.CLEANDATA;
MODEL TENYEARCHD = MALE     AGE     EDUCATION     CURRENTSMOKER        CIGSPERDAY
BPMEDS        PREVALENTSTROKE      PREVALENTHYP  DIABETES      TOTCHOL  SYSBP
       DIABP  BMI    HEARTRATE     GLUCOSE;
RUN;
ODS HTML CLOSE;
```

Table 5: Selection (Forward)

Note: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Forward Selection						
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	age	1	1	167.6890	<.0001	Age of a Patient
2	sysBP	1	2	41.4641	<.0001	Systolic blood pressure
3	cigsPerDay	1	3	43.1438	<.0001	
4	glucose	1	4	20.2926	<.0001	Glucose Level
5	male	1	5	17.3095	<.0001	Gender

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
male	1	17.1384	<.0001
age	1	98.4104	<.0001
cigsPerDay	1	22.4004	<.0001
sysBP	1	42.5767	<.0001
glucose	1	16.9467	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-8.4105	0.4620	331.4635	<.0001	0.000
male	0	1	-0.2455	0.0593	17.1384	<.0001	0.782
age		1	0.0716	0.00721	98.4104	<.0001	1.074
cigsPerDay		1	0.0218	0.00461	22.4004	<.0001	1.022
sysBP		1	0.0158	0.00242	42.5767	<.0001	1.016
glucose		1	0.00766	0.00186	16.9467	<.0001	1.008

SAS Code Used to generate the model using the forward selection

```
ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING;
CLASS MALE EDUCATION CURRENTSMOKER PREVALENTSTROKE PREVALENTHYP DIABETES BPMEDS;
MODEL TENYEARCHD = MALE     AGE     EDUCATION     CURRENTSMOKER        CIGSPERDAY
BPMEDS        PREVALENTSTROKE      PREVALENTHYP  DIABETES      TOTCHOL  SYSBP
       DIABP  BMI    HEARTRATE     GLUCOSE
/EXPB SELECTION=FORWARD SLENTRY= 0.05 SLSTAY=0.1 ;
RUN;
ODS HTML CLOSE;
```

Table 6: Selection Method (Backward)

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	diabetes	1	14	0.0111	0.9159	had diabetes
2	BPMeds	1	13	0.0453	0.8314	was on Blood Pressure Med
3	currentSmoker	1	12	0.0544	0.8155	Current smoking?
4	diaBP	1	11	0.0621	0.8032	Diastolic blood pressure
5	prevalentHyp	1	10	0.1341	0.7142	was hypertensive
6	education	3	9	2.5171	0.4722	Education level
7	BMI	1	8	0.9788	0.3225	Body Mass Index
8	totChol	1	7	1.5792	0.2089	Toatl Cholesterol Level
9	heartRate	1	6	1.7085	0.1912	Heart rate

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
male	1	17.0528	<.0001
age	1	97.3151	<.0001
cigsPerDay	1	22.9190	<.0001
prevalentStroke	1	2.8175	0.0932
sysBP	1	41.6642	<.0001
glucose	1	16.8269	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-7.9440	0.5377	218.2798	<.0001	0.000
male	0	1	-0.2451	0.0593	17.0528	<.0001	0.783
age		1	0.0713	0.00722	97.3151	<.0001	1.074
cigsPerDay		1	0.0221	0.00462	22.9190	<.0001	1.022
prevalentStroke	0	1	-0.4386	0.2613	2.8175	0.0932	0.645
sysBP		1	0.0156	0.00242	41.6642	<.0001	1.016
glucose		1	0.00762	0.00186	16.8269	<.0001	1.008

SAS Code Used to generate the model using the backward selection

```
ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING;
CLASS MALE EDUCATION CURRENTSMOKER PREVALENTSTROKE PREVALENTHYP DIABETES BPMEDS;
MODEL TENYEARCHD = MALE     AGE     EDUCATION     CURRENTSMOKER        CIGSPERDAY
BPMEDS        PREVALENTSTROKE      PREVALENTHYP  DIABETES      TOTCHOL  SYSBP
       DIABP  BMI    HEARTRATE     GLUCOSE
/EXPB SELECTION=BACKWARD SLENTRY= 0.05 SLSTAY=0.1 ;
RUN;
ODS HTML CLOSE;
```

Table 7: Selection (Stepwise)

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	age		1	1	167.6890		<.0001	Age of a Patient
2	sysBP		1	2	41.4641		<.0001	Systolic blood pressure
3	cigsPerDay		1	3	43.1438		<.0001	
4	glucose		1	4	20.2926		<.0001	Glucose Level
5	male		1	5	17.3095		<.0001	Gender

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
male	1	17.1384	<.0001
age	1	98.4104	<.0001
cigsPerDay	1	22.4004	<.0001
sysBP	1	42.5767	<.0001
glucose	1	16.9467	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-8.4105	0.4620	331.4635	<.0001	0.000
male	0	1	-0.2455	0.0593	17.1384	<.0001	0.782
age		1	0.0716	0.00721	98.4104	<.0001	1.074
cigsPerDay		1	0.0218	0.00461	22.4004	<.0001	1.022
sysBP		1	0.0158	0.00242	42.5767	<.0001	1.016
glucose		1	0.00766	0.00186	16.9467	<.0001	1.008

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
male 0 vs 1	0.612	0.485	0.772
age	1.074	1.059	1.089
cigsPerDay	1.022	1.013	1.031
sysBP	1.016	1.011	1.021
glucose	1.008	1.004	1.011

SAS Code Used to generate the model using the stepwise selection

```
ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING;
CLASS MALE EDUCATION CURRENTSMOKER PREVALENTSTROKE PREVALENTHYP DIABETES BPMEDS;
MODEL TENYEARCHD = MALE     AGE     EDUCATION     CURRENTSMOKER        CIGSPERDAY
BPMEDS         PREVALENTSTROKE       PREVALENTHYP  DIABETES      TOTCHOL  SYSBP
       DIABP  BMI     HEARTRATE     GLUCOSE
/EXPB SELECTION=STEPWISE SLENTRY= 0.05 SLSTAY=0.1 ;
RUN;
ODS HTML CLOSE;
```

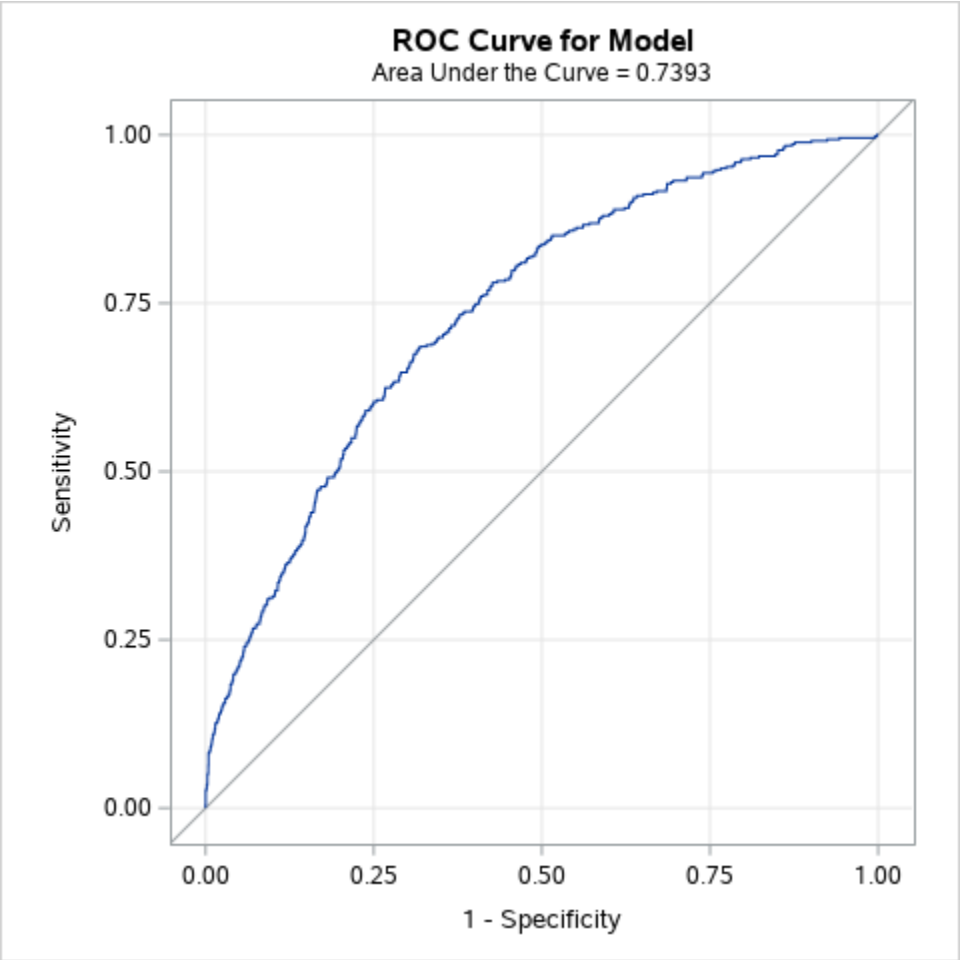


Figure 2: ROC Chart for the Full model

SAS Code Used to Generate the Curve:

```
*ROC CHART FOR FULL MODEL;

ODS HTML;

PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING PLOTS=ROC;

MODEL TENYEARCHD = MALE    AGE   EDUCATION  CURRENTSMOKER    CIGSPERDAY

BPMEDS     PREVALENTSTROKE  PREVALENTHYP    DIABETES   TOTCHOL  SYSBP   DIABP      BMI
     HEARTRATE  GLUCOSE ;

RUN;

ODS HTML CLOSE;
```

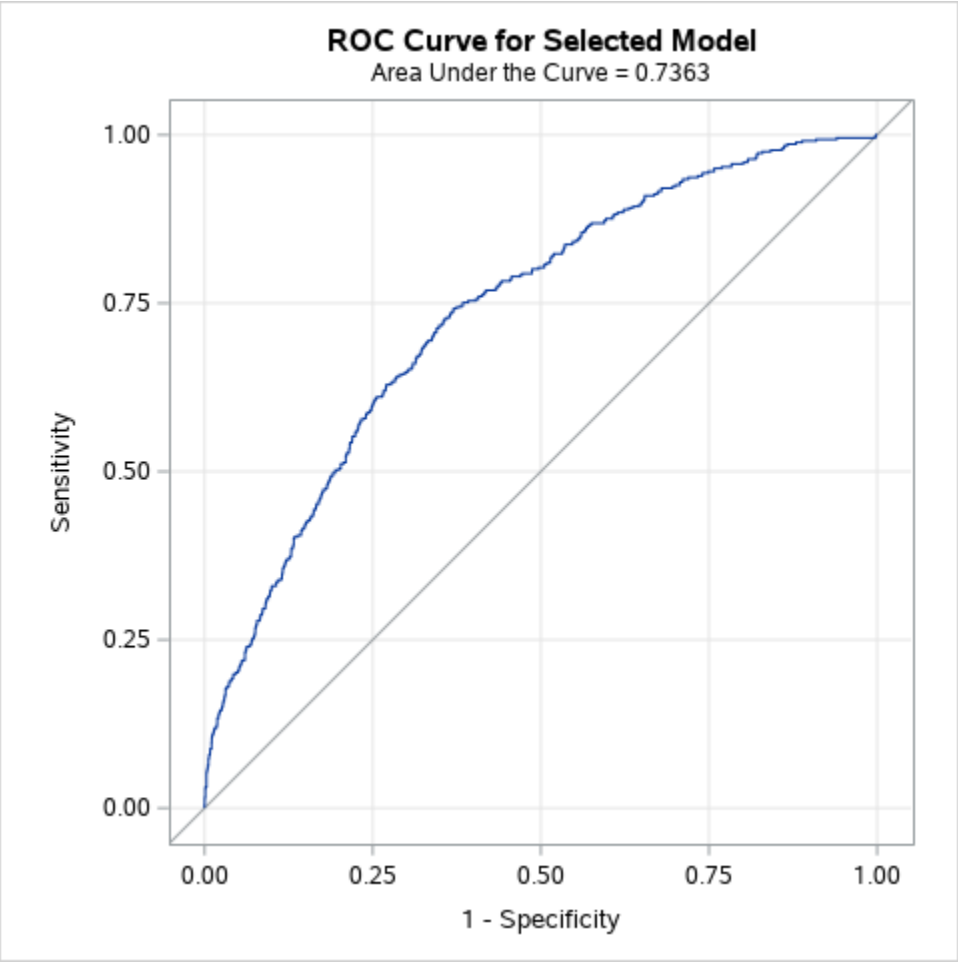


Figure 3: Roc Curve for the forward technique model

SAS Code Used to Generate the Curve:

```
*ROC CHART FOR REDUCED MODEL: FORWARD SELECTION;
ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING PLOTS=ROC;
CLASS MALE EDUCATION CURRENTSMOKER PREVALENTSTROKE PREVALENTHYP DIABETES BPMEDS;
MODEL TENYEARCHD = MALE     AGE   EDUCATION  CURRENTSMOKER    CIGSPERDAY
BPMEDS     PREVALENTSTROKE  PREVALENTHYP     DIABETES   TOTCHOL SYSBP    DIABP      BMI
      HEARTRATE  GLUCOSE
/EXPB SELECTION=FORWARD SLENTRY= 0.05 SLSTAY=0.1;
RUN;
ODS HTML CLOSE;
```

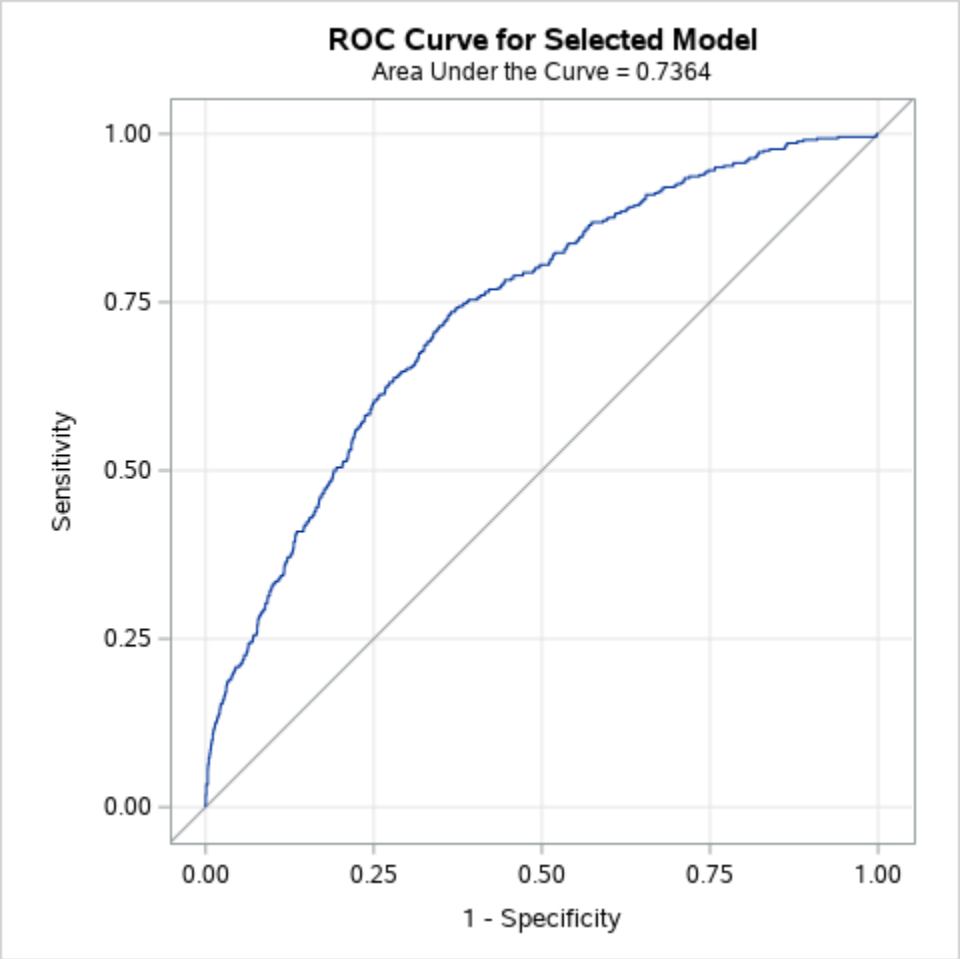


Figure 4: ROC curve for the backward technique

SAS Code Used to Generate the Curve:

```
*ROC CHART FOR REDUCED MODEL: BACKWARD SELECTION;
ODS HTML;
PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING PLOTS=ROC;
CLASS MALE EDUCATION CURRENTSMOKER PREVALENTSTROKE PREVALENTHYP DIABETES BPMEDS;
MODEL TENYEARCHD = MALE     AGE     EDUCATION     CURRENTSMOKER        CIGSPERDAY
BPMEDS        PREVALENTSTROKE      PREVALENTHYP  DIABETES      TOTCHOL SYSBP
       DIABP  BMI    HEARTRATE     GLUCOSE
/EXPB SELECTION=BACKWARD SLENTRY= 0.05 SLSTAY=0.1 ;
RUN;
ODS HTML CLOSE;
```

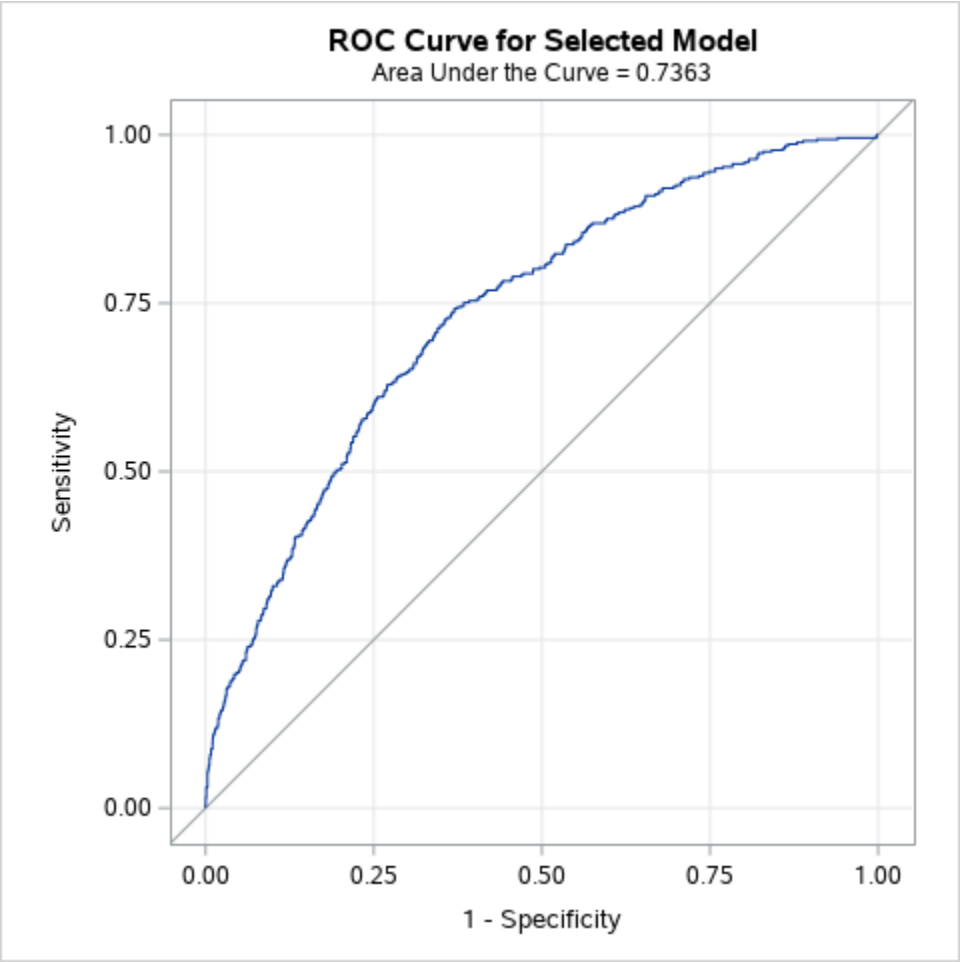


Figure 5: ROC Curve for the stepwise selection technique

SAS Code Used to Generate the Curve:

```
*ROC CHART FOR REDUCED MODEL: STEPWISE SELECTION;

 ODS HTML;

PROC LOGISTIC DATA=WORK.CLEANDATA DESCENDING PLOTS=ROC;

CLASS MALE EDUCATION CURRENTSMOKER PREVALENTSTROKE PREVALENTHYP DIABETES BPMEDS;

MODEL TENYEARCHD = MALE    AGE   EDUCATION  CURRENTSMOKER    CIGSPERDAY

BPMEDS     PREVALENTSTROKE  PREVALENTHYP    DIABETES   TOTCHOL  SYSBP   DIABP      BMI
     HEARTRATE  GLUCOSE

/EXPB SELECTION=STEPWISE SLENTRY= 0.05 SLSTAY=0.1 ;

RUN;

ODS HTML CLOSE;
```

Table 3: Project Timeline

Objective/Task	Schedule
Proposal and Timeline	24 August 2021
Construction of Methodology Literature Review	7 September 2021
Revised Research proposal and initial Analysis	27 September 2021
Draft report	11 October 2021
Final Report	18 October 2021

Datasets Links:

- Cleaned Dataset
- Complete Dataset
- Nulls Dataset

GitHub Code link: https://github.com/FeziweMelvin/SAS-project.git

Google Drive Code Link: Click! To open the code folder

PLAGIARISM DECLARATION

I hereby declare that:		**YES**	**NO**
a.	I have perused and understood the relevant sections relating to plagiarism, citation and referencing;	√	
b.	I know that plagiarism is wrong;	√	
c.	I did not attempt to present the ideas of another as if they were my own;	√	
d.	I did not attempt to represent the words or work of another as if they were my own;	√	
e.	I did not utilize the ideas, words or work of another without acknowledgement;	√	
f.	I did not use the printed text, electronic text, images, computer programme, sound, performance or creative works of another without proper acknowledgement;	√	
g.	Where I engaged with group of student to create a particular piece of work, the work correctly reflected the contribution made (where a single piece of work is collected generated, all of the group carries the responsibility for that piece of work);	√	
h.	I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.	√	
i.	I have not copied another person's assignment, essay or take-home test or any part thereof.	√	
j.	I have not plagiarized.	√	

I acknowledge that in that if I commit the offence of plagiarism, disciplinary proceedings will be instituted against me.

In the event of the court finding me guilty of the said offence, the sentence that will be imposed on me will be as follows:

- **i) exclusion from the University for a specific period;**
- **ii) cancellation of examination marks, semester marks, year marks and other form of credit earned in examinations, tests or otherwise;**
- **iii) endorsement of my academic record; and**
- **iv) publication of my conviction and sentence on the Official Notice Board.**

References

American Heart Assocation. (2021, August 10). *Understanding Blood Pressure Readings*. Retrieved from Heart Organisation: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Bundy, J. D., Li, C., & Stuchlik, P. (2017). Systolic Blood Pressure Reduction and Risk of Cardiovascular Disease and Mortality. 775-781.

Di Giosia, P., Passacquale, G., Petrarca, M., Giorgini, P., Marra, A. M., & Ferro, A. (2017, May). *Gender differences in cardiovascular prophylaxis: Focus on antiplatelet treatment*. Retrieved from ScienceDirect: https://www.scopus.com/record/display.uri?eid=2-s2.0-85010993749&doi=10.1016%2fj.phrs.2017.01.025&origin=inward&txGid=0d3b4a0d290e5e63a2f22a2f3e359591

Georgousopoulou, E. N., Panagiotakos, D. B., Bougatsas, D., Chatzigeorgiou, M., Kavouras, S. A., Chrysohoou, C., & Pitsavos, C. (2016, March 9). *Physical Activity Level Improves the Predictive Accuracy of Cardiovascular Disease Risk Score: The ATTICA Study (2002–2012)*. Retrieved from US National Lbrary of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4809127/

Huxley, R., & Wiidward, M. (2011, August 10). *Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies* . Retrieved from PubMed: https://pubmed.ncbi.nlm.nih.gov/21839503/

Maas, A., & Appelman, Y. (2010, December 18). *Gender differences in coronary heart disease*. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605/

Mayo Foundation for Medical Education and Research. (n.d.). *High Cholesterol*. Retrieved from mayo clinic : https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800?p=1

Mongraw-Chaffin, M. L., Peters, S. A., Huxley, R. R., & Woodward, M. (2015, May 7). *The sex-specific association between BMI and coronary heart disease: a systematic review and meta-analysis of 95 cohorts with 1·2 million participants* . Retrieved from NCBI: https://pubmed.ncbi.nlm.nih.gov/25960160/

Olifiranye, O., Zizi, F., Brimah, P., Jean-louis, G., Makaryus, A. N., McFarlane, S., & Ogedegbe, G. (2011, July 13). *Management of Hypertension among Patients with Coronary Heart Disease*. Retrieved from Hindawi: https://www.hindawi.com/journals/ijhy/2011/653903/

Powell, K., Thompson, P., Caspersen, C., & Kendrick, J. (1987). *Pysical Activity and incidence of coronary heart disease.* Retrieved from Annual reviews Web site: https://www.annualreviews.org/doi/pdf/10.1146/annurev.pu.08.050187.001345

Rossello, X., Dorresteijn, J. A., Janssen, A., Lambrinou, E., Scherrenberg, M., Bonnefoy-Cudraz, E., & Cobain, M. (2019, June 25). *Prediction tools in cardiovascular disease prediction*. Retrieved from Sage jornals: https://journals.sagepub.com/doi/10.1177/2048872619858285

Sallam, T., & Watson, K. E. (2013, September 17). *Predictors of cardiovascular risk in women*. Retrieved from US National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6097244/

Wilson, P. W., D'Agostino, R. B., Levy, D., Albert M. Belanger, H. S., & Kannel, W. B. (1998, May 12). *Prediction of Coronary Heart Disease Using Risk Factor Categories*. Retrieved from ahajornals: https://www.ahajournals.org/doi/10.1161/01.CIR.97.18.1837

World Health Organization. (2021, June 11). *Cardiovascular diseases(CVDs)*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)