A Project Report
On

"Predicting Coronary Heart Disease Using Risk Factors from Framingham Heart Study "

Framingham Heart Study

A Project of the National Heart, Lung, and Blood Institute and Boston University

"The work contained and presented here is my work and my work alone."

Kshitij Sharma, Naresh Vemula, Pallavi Singh, Riddhi Krishnan, Shreya Chandra
MSBAPM Spring 2016

Acknowledgement

As a group we would like to express our heart-felt gratitude to **Dr. James Marsden** for his valuable suggestions and constant support throughout the course. Without his critical evaluation, this report could not have reached its present form.

We would also like to take this opportunity to thank **NIH (National Heart, Lung and Blood Institute)** for their time and consideration for sharing the Framingham Database with us.

Lastly, we are gratefully thankful to each member of our group for sharing fruitful insights from their extensive experience for this project.

Executive Summary

Framingham Heart study data provides substantial insight into the epidemiology of cardiovascular disease and its risk factors. The risk factors such as blood pressure, glucose, total cholesterol and body mass index (BMI) are used to predict Coronary Heart Disease (CHD) in a middle aged population sample. Coronary disease risk prediction model is developed using numerical and categorical data, allowing us to predict CHD risk in patients. The primary purpose of our analysis is to infer risk of Coronary Heart Disease, kind of cardiovascular disease (CVD), from several decision models that are developed by the team using predictive modeling techniques learnt in the class.

A total of 11,627 samples were collected from Framingham Heart Study and refined using such techniques as redefining data types into continuous, nominal or ordinal variables, visualizing data using summary tables, graphs and charts, filling missing data through missing data patterns, identifying redundant variables and outlier detection using Box plot and Mahalanobis distance. The cleaned data set, consisting of 11,546 samples, was segregated into training and validation data set. The models were developed on top of cleaned data set. The team initially prepared Decision Tree model using K fold validation, Logistic Regression model, Discriminant Analysis model, Neural networks model, Bootstrap Forest model. The score sheet of accuracy, true negative and true positive, of these models was captured and compared, and validated with the validation data set. Eventually, it was tested using testing data set. Furthermore, in order to select the best model, we used penalty error matrix which consists different weights for false positive and false negative errors was also calculated. The team finally selected the model based on highest accuracy and lowest penalty error.

Our team believes that this predictive model will increase the efficacy of predicting Coronary Heart Disease (CHD) and population at risk with the help of risk factors or independent variables. We hope this model will be helpful in dealing with the risk factors to prevent deadly heart disease.

Table of Content

(For every topic we will add hyperlinks, so that by just one click it takes him to the desired page)

Introduction

Prior to 1948, cardiovascular disease (CVD) remained grey area even for physicians as they had little knowledge of the causes and risk factors associated with cardiovascular disease. Many physicians did not believe that modifying certain behaviors could enable their patients to avoid or reverse the underlying causes of serious heart and vascular conditions. With cardiovascular disease becoming leading cause of death and serious illness in the United States, National Heart Institute embarked on an ambitious project known as the Framingham Heart Study to identify the factors that contribute to CVD. The study participants who included 5,209 men and women from the town of Framingham, Massachusetts undergone physical examinations such as blood tests, bone scans, eye exams and echocardiograms and lifestyle interviews that helped in analyzing common patterns related to CVD development. Key findings from the study includes:

1. Cholesterol levels, high blood pressure and diabetes are the major risk factors in contributing to cardiovascular disease.
2. Key elements of the American lifestyle that contributed to the high rates of CVD disease. Study researchers found that a lifestyle typified by a faulty diet, sedentary living, and/or unrestrained weight gain acerbated disease risk factors and influenced the occurrence of cardiovascular problems.
3. Smokers were at increased risk of having a myocardial infarction or experiencing a sudden death.
4. Unrestrained weight gain, accompanied by lack of exercise, promoted cardiovascular risk factors such as hypertension and diabetes.

The risk factors associated with CVD as identified from the study are Systolic and Diastolic Blood Pressure, Hypertension, current smoking habits, cigarette per day, total alcohol consumption, Cholesterol level, body mass index and glucose level.

Through this report we have tried to summarize the result of the survey conducted by Framingham Heart Study by applying the learnings from predictive modeling course.

Literature Review:

Cardio Vascular Diseases(CVD):

Any disease which affects the cardio vascular process or system are termed as the cardio vascular diseases(CVD) such as Coronary artery disease, High blood pressure, Cardiac arrest, Congestive heart failure, Arrhythmia, Peripheral artery disease, Stroke, Congenital heart disease. Cardiovascular diseases include illnesses that involve the blood vessels (veins, arteries and capillaries) or the heart, or both - diseases that affect the cardiovascular system. Cardio vascular system consists of heart, arteries, veins, and capillaries and is responsible for blood flow throughout human body. This system transports oxygenated blood from the lungs, heart and arteries to the human body, essential for survival. The Table below gives description of important CVD's and their causes.

Table 1- Coronary Vascular Diseases Description

Type of Coronary Vascular Disease	Cause
Coronary artery disease	Damage or disease in the heart's major blood vessels.
High blood pressure	A condition in which the force of the blood against the artery walls is too high.
Cardiac arrest	Sudden, unexpected loss of heart function, breathing, and consciousness.
Congestive heart disease	A chronic condition in which the heart doesn't pump blood as well as it should.
Arrhythmia	Improper beating of the heart, whether irregular, too fast, or too slow.
Peripheral artery disease	A circulatory condition in which narrowed blood vessels reduce blood flow to the limbs.
Stroke	Damage to the brain from interruption of its blood supply.
Congenital heart disease	An abnormality in the heart that develops before birth.

Risk Factors in CVD:

Demographic or non-modifiable risk factors:

CVD becomes increasingly common with advancing age. As a person gets older, the heart undergoes subtle physiologic changes, even in the absence of disease [7]. A man is at greater risk of heart disease than a pre-menopausal woman. Once past the menopause, a woman's risk is similar to a man's. Risk of stroke, however, is similar for men and women [7].

Hypertension (high blood pressure):

Blood pressure is measured as two numbers recorded over mercury, ex 120/80 mm hg. The upper part known as systolic blood pressure (the contraction of arteries pressure) and the lower part is diastolic blood pressure (the relaxation of arteries pressure). High blood pressure is defined as increase in either or both systolic blood pressure>140 and diastolic blood pressure> 90. Globally, more than 1 billion people have high blood pressure(hypertension).

Hypertension is one of the most important causes of premature death worldwide and the problem is growing; in 2025, an estimated 1.56 billion adults will be living with hypertension. Hypertension is the leading cause of CVD worldwide. People with hypertension are more likely to develop complications of diabetes. High blood pressure is called the "silent killer" because it often has no warning signs or symptoms, and many people do not realize they have it; that is why it's important to get blood pressure checked regularly. (Ref: World Health Federation, Cardio Vascular Disease Risk Factors)

Current Smoker:

A research says about 10% of CVD's are caused by smoking tobacco, this risk is higher in female smokers, young men and heavy smokers. Statistics says that more than 1 billion people are smoking tobacco. Within two years of quitting smoking the risk of coronary heart disease is substantially reduced, after 15 years the risk reduced to a non-smoker's risk. (Ref: World Health Federation, Cardio Vascular Disease Risk Factors)

Glucose and Diabetes:

Cardiovascular risk increases with raised glucose values, Lack of early detection and care for diabetes results in severe complications, including heart attacks, strokes, renal failure, amputations and blindness. Diabetes is defined as having a fasting plasma glucose value of 7.0 mmol/l (126 mg/dl) or higher. [8]. In 2008, diabetes was responsible for 1.3 million deaths globally. [8] • In 2008, the global prevalence of diabetes was estimated to be 10 percent [8]. CVD accounts for about 60 per cent of all mortality in people with diabetes. [8] • The risk of cardiovascular events is from two to three times higher in people with type 1 or type 2 diabetes and the risk is disproportionately higher in women. [8] • In some age groups, people with diabetes have a two-fold increase in the risk of stroke. [8]. Patients with diabetes also have a poorer prognosis after cardiovascular events compared to people without diabetes [8].

Cholesterol:

Raised blood cholesterol increases the risk of heart disease and stroke [8]. Globally, one third of ischaemic heart disease is attributable to high cholesterol. [8] • Overall, raised cholesterol is estimated to cause 2.6 million deaths (4.5 per cent of total) and 29.7 million DALYS, or 2 per cent of total DALYS globally. [8]

Overweight and obesity (high BMI):

Obesity (high Body Mass Index-BMI) is one of the major reason for CVD risk factors like high BP, glucose, type 2 diabetes. Worldwide, at least 2.8 million people die each year as a result of being overweight or obese, and an estimated 35.8 million (2.3 per cent) of global DALYs are caused by overweight or obesity [8]. The prevalence of raised BMI increases with income level of countries, up to upper-middle-income levels. The prevalence of overweight in high-income and upper-middle-income countries was more than double that of low- and lower-middle-income countries. [8]

* Dalys - The disability-adjusted life year (DALY) is a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death.

Existing Risk Analysis Tools:

There are more than hundred risk assessment tools for predicting CVD, couple of assessments are the most successful among them, Framingham scoring system(FSS) and Joint British Societies (JBS).

JBS has its own risk assessment factors but their main focus was to predict the coronary disease stroke. However, it inherited some scoring inputs from FSS which has a score assigned to each risk factor and the summation of these scores yields the probability of getting CVD. Scottish Intercollegiate helped JBS in adding few additional risk factors to their scoring sheet like family history, social deprivation in predicting Coronary disease stroke.

With availability of different statistical tools there has been an increase in risk assessment systems. However, not always risk assessments yielded positive results. There are many factors to consider like location, demographics. However, there are some limitations to this model as studies reveal that it does not perform as per the standards with regards to Japanese population as predicted for American population.

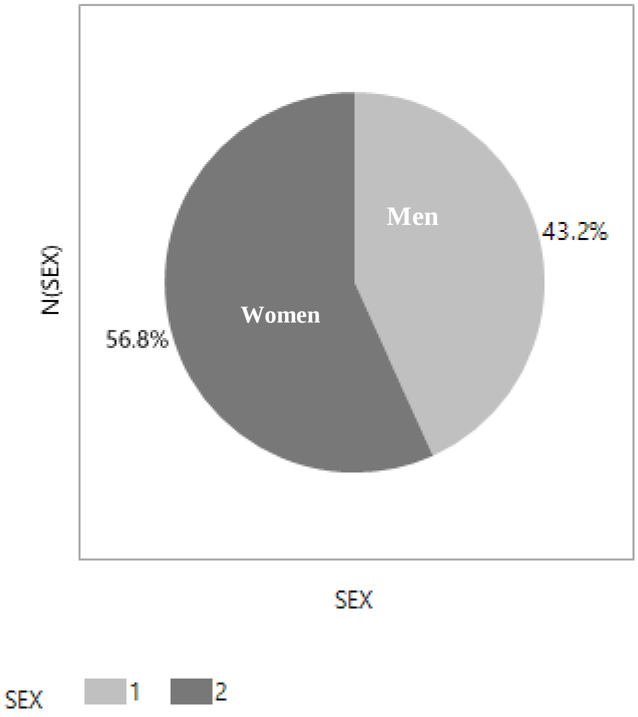
Hence, through this project we have tried to build few predictive models based on the learnings to achieve the accuracy in determining most important risk factors causing Coronary Heart Disease.

SEMMA Methodology

SEMMA- Sampling

The Framingham Heart study is long term prospective study of the etiology of cardio vascular disease among Framingham population. The study began in 1948 and 5,209 subjects were initially enrolled in the study. However, the dataset we used is a subset of data collected as part of Framingham heart study consist of 4,434 participants. Participant clinic data was collected during three examination periods, approximately 6 years apart, from roughly 1956 to 1968. An extensive follow up was done for each participant for 24 years to keep a check on CHD

Figure 1: Percentage of Men and Women participated in the survey



sEMMA- Data Exploration

Below mentioned table consists of demographics risk factors from the dataset such as Sex (male or female), Age (current age of the participant at the time of the test), educ (educational history of participant).

Table 1.1 Demographic Risk Factors

Field Name	Data Type	Risk factor Category	Description	Value/ Format	Units of Measure
SEX	Nominal/ Dichotomous variable	Demographic Factor	This field collected to determine how many of men and women are at the risk of having coronary heart disease	1=Men 2=Women	NA
AGE	Continuous	Demographic Factor	To determine which age bracket has probability of having coronary heart disease more.	32-81	Years
EDUC	Ordinal	Demographic Factor	To measure the level of education the participants has	1=0-11 years 2=High School Diploma, GED 3=Some College, Vocational School 4=College (BS, BA) degree or more	NA

In this dataset, behavioral risk factors are any particular behavior or behavior pattern which adversely affects the health of the individual and can vary.

Table 1.1 Behavioral Risk Factors

Field Name	Data Type	Risk factor Category	Description	Value/ Format	Units of Measure
CURSMOKE	Nominal	Behavioral Risk Factor	To examine the current cigarettes smoking habit of individuals participated in exam	0 = Not a Current Smoker. 1 = Current Smoker	NA
CIGPDAY	Continuous	Behavioral Risk Factor	Number of cigarettes a participant smoked each day	0 -90 Cigarettes /day	Number

Participant's Medical history was also considered as it's an important indicator of individual with a heart disease. These are as indicated in the table below:

Table 1.2 Participant's Medical History

Field Name	Data Type	Risk factor Category	Description	Value/ Format
DIABETES	Nominal	Medical History Risk Factor	Diabetes is a disease in which body is unable to produce enough insulin hormone causing elevated levels of glucose in the blood. This variable measures the diabetic individuals from the exam.	0 = Non - Diabetic 1 = Diabetic
BPMEDS	Nominal	Medical History Risk Factor	Participants who take anti-hypertensive medication at exam	0 = No dosage 1 = On dosage

Field Name	Data Type	Risk factor Category	Description	Value/ Format
PREVCHD	Nominal	Medical History Risk Factor	Prevalent Coronary Heart Disease caused by buildup of plaque inside the coronary arteries which supply oxygen rich blood to the heart. It is examined as preexisting Angina Pectoris and Myocardial Infarction	0 = Free of disease 1= Prevalent Disease
PREVAP	Nominal	Medical History Risk Factor	Prevalent Angina Pectoris is a medical term for chest pain or discomfort due to coronary heart disease when heart muscles do not get as much blood as it needs	0 = Free of disease 1= Prevalent Disease
PREVMI	Nominal	Medical History Risk Factor	Prevalent Myocardial Infraction means heart attack. If Angina Pectoris lasts too long, the starved heart tissue dies.	0 = Free of disease 1= Prevalent Disease
PREVSTRK	Nominal	Medical History Risk Factor	Prevalent Stroke is a brain attack, which stops the vital blood flow and oxygen to the brain	0 = Free of disease 1= Prevalent Disease
PREVHYP	Nominal	Medical History Risk Factor	Prevalent Hypertensive is when an individual is having systolic BP greater than 140 or diastolic BP less than 90 or the person is taking antihypertensive medications	0 = Free of disease 1= Prevalent Disease

High risk factors such as total cholesterol level, systolic BP, diastolic BP, BMI (body mass index to determine obesity), Heart Rate, Glucose were also taken into consideration. Below table shows the same.

Field Name	Data Type	Risk factor Category	Description	Value/ Format	Units of Measure
TOTCHOL	Continuous	High Risk Factor for CHD	Total Cholesterol assess several types of fat in the blood (in mg/DL). It is the sum of measures of Low- Density lipoprotein(LDL) cholesterol and High Density lipoprotein (HDL) cholesterol. It measures the cholesterol level of individuals participated in exam.	107-696	mg/dl
SYSBP	Continuous	High Risk Factor for CHD	When heart beats, it contracts & pushes blood through arteries to rest of body. It creates pressure called Systolic blood pressure. A normal systolic BP is below 120.A systolic BP of 140 or higher is considered to be hypertension. It measures the systolic BP of the participants	83.5-295	mm Hg
DIABP	Continuous	High Risk Factor for CHD	When heart rests between beats, the pressure created in arteries is called Diastolic blood pressure. A normal diastolic BP is less than 80.A diastolic BP between 80-89 is considered to be prehypertension.	37-150	mm Hg
BMI	Continuous	High Risk Factor for CHD	Determines whether one is in a healthy weight range as per the height. Body-Mass Index is a value derived from body mass divided by square root of the body height of the participants. Higher the BMI, the greater the risk of coronary artery disease	14.43 - 56.8	kg/m^2

Field Name	Data Type	Risk factor Category	Description	Value/ Format	Units of Measure
HEARTRTE	Continuous	High Risk Factor for CHD	Measure of heart rate(beats/min). A normal heart rate considered to be 60 - 80 beats/min	37-220	beats/min(BPM)
GLUCOSE	Continuous	High Risk Factor for CHD	To measure the glucose levels in the blood of participants at exam	39-478	(mg/dL)

We performed visualization techniques to find patterns among the participant's risk factors. We did pre modelling explorations in determining the trends in coronary heart disease(CHD) occurrence.

As indicated from the graph below, Mean(AGE) & Mean (Systolic Blood Pressure) for heart patients are higher as compared to patients with no heart problem

Figure 2: Mean(AGE)

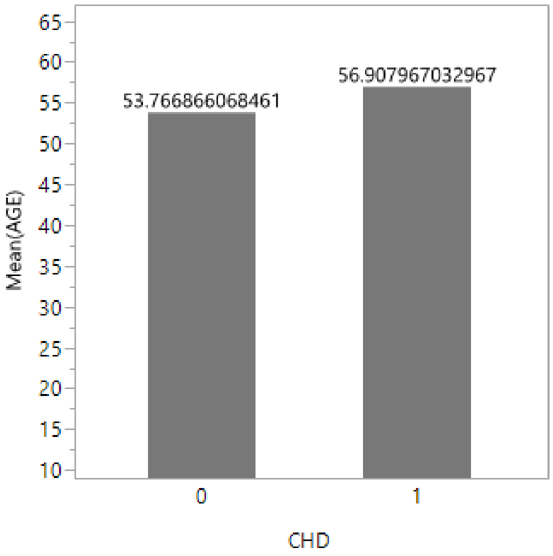
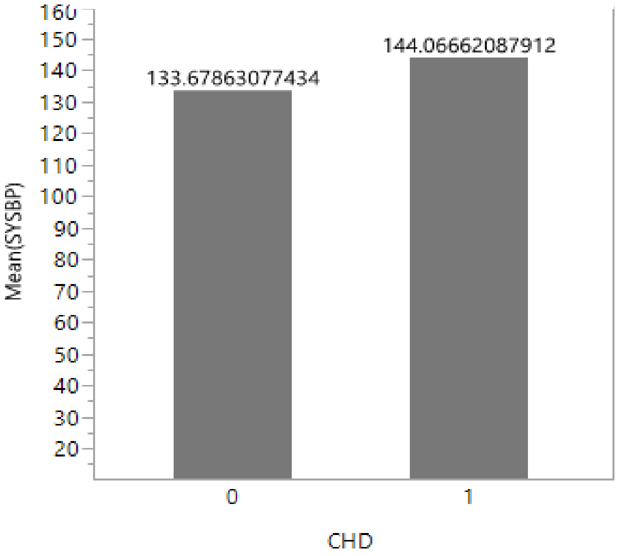


Figure 3: Mean(SYSBP)



CHD=1 are the People with coronary heart disease

CHD=0 are the People without coronary heart disease

The summary tables below give basic insights about how CHD is varying among participants with respect to risk factors. We observed most of participants attacked by CHD are male. We observed mean and median of total cholesterol, Age, systolic Blood pressure, diastolic blood is high for participants with CHD than the people who are not affected by CHD

Table 1.3 Variation in CHD

CHD	Median (SEX)	Median (TOTCHOL)	Median (AGE)	Median (SYSBP)	Median (DIABP)	Median (BMI)
0	2(Female)	238	53	130	81	25.15
1	1(Male)	243	58	140	84	26.255

Table 1.4 Variation in CHD

CHD	Mean (TOTCHOL)	Mean (AGE)	Mean (SYSBP)	Mean (DIABP)	Mean (BMI)	Mean (GLUCOSE)
0	238.1053694	53.8030411	133.5962818	82.06200998	25.57313139	82.47042053
1	249.1265985	57.30530691	143.5065537	85.61636829	26.68508951	86.77429668

SEMMA- Modifying Data

Data Preprocessing:

Missing Data Handling:

The risk factors [Independent Variables] which are empty or unmentioned for a participant record in Framingham heart study data set, are termed as Missing data. Major reasons of missing data in this scenario could be, a participant being unaware of some risk factors such as Total cholesterol level, Glucose levels. We assumed some participants might not be willing to share their medical history with respect to risk factors like Number of Cigarettes consumption per day, consuming BP (Blood Pressure) medicines or not.

To deal with this we calculated the biases involved in missing data by building models with missing data and without missing data. We observed a slight decrease in the accuracy of our best prediction without missing data rows. Therefore, we decided not to delete the missing data. We dealt with missing data by imputing, we found 565 entries are missing in BPMEDS (Whether or not patient is taking medicines to stabilize blood pressure), We imputed these values based on formulae using systolic blood pressure and diastolic blood pressure. We imputed BP Meds with 1(on medicine) when the participant is having High BP (140/90 mm Hg) or low BP (60/90 hg mm Hg.

Total cholesterol (which is one of the most important risk factors in heart diseases, this column has few (less than 200) observations missing. We imputed these values such that it will lead a prediction towards heart disease occurrence. We believed the cost involved in error of not predicting heart disease occurrence is more than cost of error in predicting heart disease occurrence. We imputed this value with median based upon the cost factor assumption.

We had missing values for number of cigarettes smoking per day when a participant is a smoker. We imputed with median value of number of cigarettes smoking per day by a smoker. We handled the missing data entries in glucose levels with mean imputation. However, we know that diabetes (High or low sugar levels) affect glucose levels of a person we were unclear about this concept. This lead us to take decision towards mean imputation. We are familiar with multivariate correlation imputation and multivariate normal imputation, but we decided to go with the risk factors involved in CHD based imputation rather than above methods.

Outlier handling:

Outlier is some observations which are not in similar range with the most of the participants in the Framingham heart study. A method was proposed by the team to detect outliers which is to use quantile range distances (Values farther than some quantile ranges from the tail quantile), we found 75 outliers in the Glucose levels of participants. However, these glucose levels are normal in case of high sugar levels, for diabetes patients. After an active discussion among team we decided not to remove them.

We determined impact of outliers by building models with and without outliers to predict ten-year CHD (Coronary Heart Disease). We noticed by removing outliers there is a slight decrease in our prediction. We believe to count on every possibility in increasing the successful prediction of CHD occurrence, as this will give the participants a heads up if they are in the high risk zone of CHD.

Principal Components:

Principal components are applied on data to reduce dimensionality and redundancy among the participant's risk factors. This works by determining the correlations between the different risk factors and change them to linearly uncorrelated elements by applying the dimensionality reduction algorithm. This works only on continuous numerical data like systolic BP, diastolic BP, cigarettes per day, Total cholesterol, BMI etc. With more number of principal components, we can reduce the data loss. We have applied this and found, with six principal components we are maintaining 80% data.

However, we did not use the principal components in predicting coronary heart disease. We believed it is important for us to explain the major risk factors involved in CHD, we were losing this capability by using PC's. we determined that TOTCHOL, AGE, SYSBP, CURSMOKE, DIABETES, BMI are some of the major factors in CHD. We came to a decision that as we have few important risk factors it is not required for us to reduce them even further and we wanted to explain the risk factors in normal medical measures.

SEMMA- Modeling

Selecting the best model:

Confusion Matrix:

Predicted values

Actual values

	0	1
0	0(True Negative)	1(False Negative)
1	1(False Positive)	0(True Positive)

The accuracy is the sum of the diagonals in the confusion matrix, the observations that were classified correctly, divided by the total number of observations in our test set (all the outcomes in the confusion matrix).

We select the model which has the highest accuracy based on the above calculation. However, we have few more considerations while selecting the model based on the wrong predictions (False Negative and False Positive). These are discussed below.

Penalty error matrix:

Predicted values

Actual values

	0	1
0	0	1*False Negative
1	2*False Positive	0

In our case selecting the best prediction model was a hurdle, we had two types of errors.

Error 1: predicting a participant will get a heart disease and in actual case he did not get (False Negative).

Error 2: predicting a participant will not get a heart disease and in actual case got a heart disease (False Positive).

After an active discussion within the team, we believed that error 2(predicting that participant won't get a heart disease but got a heart disease) type has high weightage than the error

1(predicting participant will get a heart disease but did not get a heart disease) type, so we decided to consider cost for each error case. We decided to give twice importance to the error 2 compared to the error 1, therefore we constructed a penalty matrix as shown above. We multiplied the confusion matrix with the penalty matrix, then we took sum of all the values in new matrix, these are observations classified wrongly multiplied with penalty matrix weightages, divided by the total number of observations in our test set. This value is called as the penalty error for the model**. (I think this has come out really well ☺. We will just make some minor changes to the verbatim specially the way the para starts)**

Based upon the above discussion, we selected model that has a highest accuracy and lowest penalty error among all the models we built.

Baseline Model:

To get an idea how successful the numerical model was, we used simple prediction methods to act as baseline method. The fact relied only few among all the participants had cardio vascular disease, our baseline model predicts that no participant in the study will get a heart disease. With this predicted values we have calculated the accuracy and penalty error by using the above stated methods.

In our case a simple baseline method predicting no one will get CHD has an accuracy of 72.908% and a penalty error of 54.183%. We will try beating this by achieving higher accuracy and lower penalty error than baseline method.

Logistic regression Model:

To identify the risk factors which are important in the predictive models we have performed one-way analysis on variables with respect to the target variable CHD. We initially identified Total cholesterol, Age, Diastolic BP, systolic BP, BMI, Smoking, glucose as the important variables in determining heart disease while comparing the means (Hypothesis Testing).

However, we were not sure of the less significant variables. so we decided to build models with all the Risk factors. We built different models by interchanging risk factors and selected the best among all the models. Our most significant model was built using the logistic regression function which has an accuracy of 74.40% and penalty error of 29.47%, which has higher accuracy and a very small penalty error than the baseline. Below table shows the accuracy and penalty error we got with the logistic model.

Table 4: Accuracies of Nominal Logistic Regression on testing data

Accuracy (predicting the correct outcomes)	**0.74355**
True Negative (rate of correctly predicting participants without heart disease)	**0.7554**
True Positive (rate of correctly predicting participants with heart disease)	**0.6456**
Penalty error (rate of wrong predictions, with weightage of penalty matrix)	**0.29469**

Training a Model:

Logistic regression predicts the probability of the outcome variable(CHD) being true with help of one or more risk factors (Independent Variables). We use threshold value t to determine participants getting heart disease, we selected the threshold value based on which errors are better.

Error 1: While t is small, we predict more participants with heart disease. (More errors where we say heart disease, but in actual case participant will not get a heart disease).

Error 2: While t is large, we predict more participants without heart disease. (More errors where we say no heart disease, but in actual case participant will get a heart disease).

Probability(CHD=1)>=t, participants get a CHD (Coronary heart disease)

Probability(CHD=0) < t, participants will not get a CHD (Coronary heart disease)

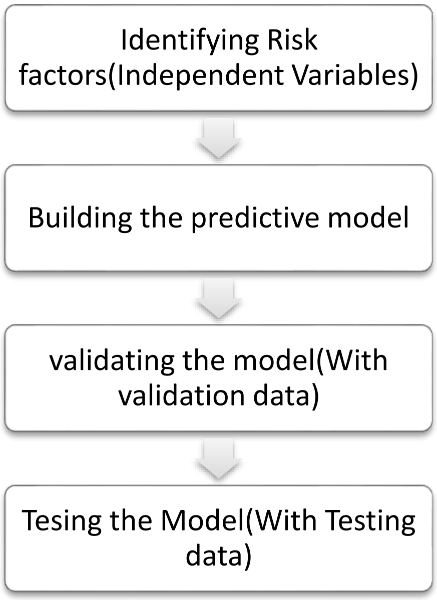
In this case, our aim is to successfully determine participants getting heart disease. So we need to make less error 2's. With help of ROC (Receiver optimistic curve) we chose a threshold where we make less error of type 2. However, by doing this our accuracy has significantly dropped. We decided to go with threshold value of 0.5, we achieved the highest accuracy and lowest penalty errors among all the models.

Validating the Model:

We have not chosen our model with training accuracy alone, rather we have used a validation data set to validate our model. Validation dataset has data different from training data and it is balanced (with target variable evenly distributed). Only when the model is performing as expected on the validation set, we will move the model for testing with real world data.

Testing the Model:

Once the model has passed through the training and validation, we perform predictions on the testing set which has data similar to real case scenario and unbalanced (Target variable being unevenly distributed). We measured the accuracies for the testing set and selected the model which performs with high significance on all the three (training, validation, testing) datasets. **Figure ( )**


Identifying Risk factors(Independent Variables)
Building the predictive model
validating the model(With validation data)
Tesing the Model(With Testing data)

Data Analysis Tools

SAS JMP:

We used SAS Jmp pro 12.1.0 as the tool to explore, analyze and to build predictive models. summary tables in Jmp guided us in initial data exploration techniques to observe patterns. Univariate (one variable) platform in columns section helped us in outlier analysis (Quantile range outlier method) and missing data handling. We used missing data patterns in tables section to deal with missing data. Multivariate (more than two variables) platform helped us to observe the relations among the continuous variables, Observing the multivariate outliers, creating the principal components (uncorrelated linear data). Most of the times we relayed on bivariate (using two variables) contingency platform to explore trends in variables.

Fit model option in Analyze helped us to implement stepwise regression, Nominal logistic regression to predict the Nominal variable CHD (Coronary Heart Disease) using the independent variables. in addition, we built predictive models using various modelling techniques available in Jmp. Such as, Partition in creating classification trees (Decision trees), Neural networks in predicting CHD. We used Multivariate technique called discriminant Analysis which uses covariance's in predicting the target variable CHD.

Conclusion

The Framingham Heart study helps in identifying the high risk factors for predicting the probability of getting a coronary heart disease. As per our analysis, we have identified BMI, Cholesterol, Systolic BP, Hypertension are major factors in Heart disease. These factors can be reduced to normal by doing physical exercise on a daily basis. By identifying the high risk factors at early stages, we can create awareness among the people to reduce these factors to the normal state. By predicting CHD at an early stage we can reduce the death rate from this disease.

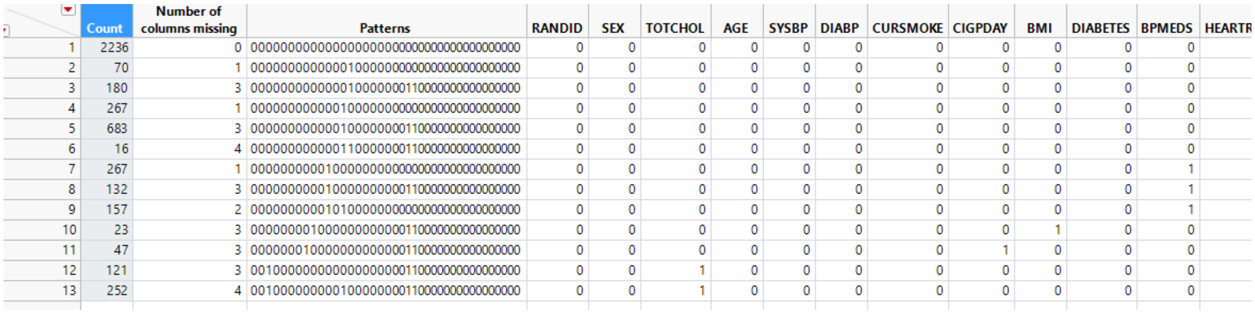
Recommendation

APPENDIX

A. Missing Data Pattern

In Framingham dataset there was no missing data in all the rows. In our dataset there was no missing data in all the rows, so we checked for missing data in every column and found that few rows are missing for the columns which we have considered as important risk factors to determine the population at risk for CHD disease.

(Jmp -> Tables -> Missing Data Pattern)

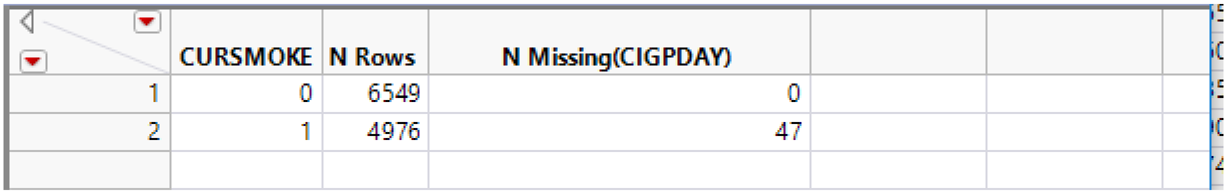


	Count	Number of columns missing	Patterns	RANDID	SEX	TOTCHOL	AGE	SYSBP	DIABP	CURSMOKE	CIGPDAY	BMI	DIABETES	BPMEDS	HEARTR
1	2236	0	0000000000000000000000000000000000000000	0	0	0	0	0	0	0	0	0	0	0	
2	70	1	0000000000000100000000000000000000000000	0	0	0	0	0	0	0	0	0	0	0	
3	180	3	0000000000000100000000110000000000000000	0	0	0	0	0	0	0	0	0	0	0	
4	267	1	0000000000001000000000000000000000000000	0	0	0	0	0	0	0	0	0	0	0	
5	683	3	0000000000001000000000110000000000000000	0	0	0	0	0	0	0	0	0	0	0	
6	16	4	0000000000001100000000110000000000000000	0	0	0	0	0	0	0	0	0	0	0	
7	267	1	0000000000100000000000000000000000000000	0	0	0	0	0	0	0	0	0	0	1	
8	132	3	0000000000100000000000110000000000000000	0	0	0	0	0	0	0	0	0	0	1	
9	157	2	0000000000101000000000000000000000000000	0	0	0	0	0	0	0	0	0	0	1	
10	23	3	0000000010000000000000110000000000000000	0	0	0	0	0	0	0	0	1	0	0	
11	47	3	0000000100000000000000110000000000000000	0	0	0	0	0	0	0	1	0	0	0	
12	121	3	0010000000000000000000110000000000000000	0	0	1	0	0	0	0	0	0	0	0	
13	252	4	0010000000001000000000110000000000000000	0	0	1	0	0	0	0	0	0	0	0	

Methodology followed for missing data pattern:

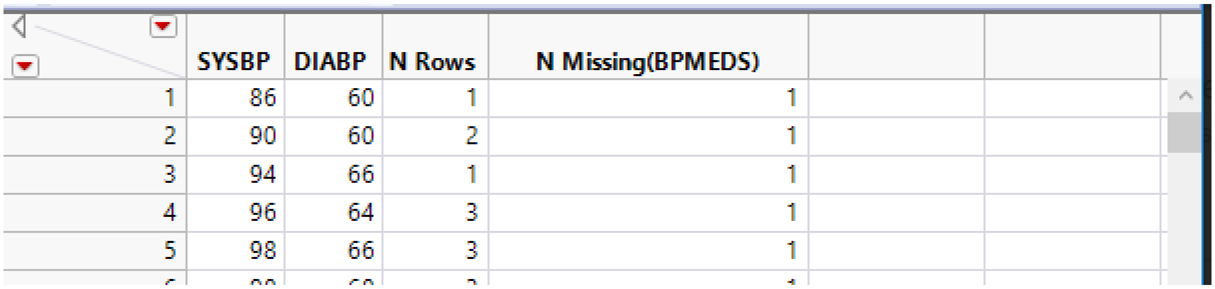
a) We have determined that TOTCHOL is an important risk factor for determining CHD as WebMD studies prove that high cholesterol builds up a wall in the arteries which causes heart diseases. Numerically, it is the sum of HDL (High Density Lipoprotein) and LDL (Low Density Lipoprotein). We have these 2 columns, HDL and LDL along with TOTCHOL, but these 2 columns are having more than 80% of the data missing, which we can get from TOTCHOL hence we have deleted these 2 columns.
TOTCHOL column is also having 200 missing values, we imputed these 200 values with the median based upon the cost factor assumption which is explained above in the visualization section.

b) CIGSPDAY is the number of cigarettes a person smoked each day. We found that this column is having 47 entries missing for the current smoker, so we have imputed them through the median of persons who smoked cigarettes.



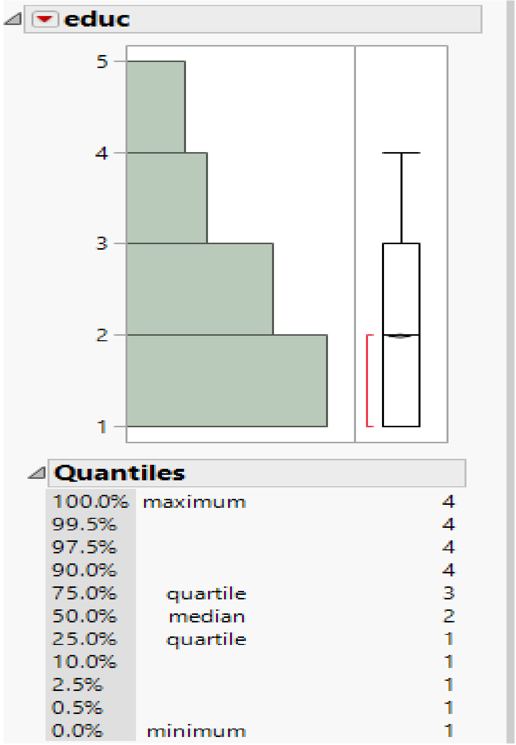
	CURSMOKE	N Rows	N Missing(CIGPDAY)		
1	0	6549	0		
2	1	4976	47		

c) BPMEDS indicates the number of participants taking anti-hypertensive medication, it is having 556 missing values. We can get this information from the columns of SYSBP and DIABP indicates the blood pressure measurement of High BP (140/90 mm Hg) or low BP (60/90 mm Hg). So we have imputed the values for BPMEDS based on the unit values of SYSBP and DIABP.



	SYSBP	DIABP	N Rows	N Missing(BPMEDS)
1	86	60	1	1
2	90	60	2	1
3	94	66	1	1
4	96	64	3	1
5	98	66	3	1

d) educ is a demographical factor which indicates the level of education of participants at exam, It has266 missing values, We have imputed this column with the maximum value due of time constraint.



Quantiles

100.0%	maximum	4
99.5%		4
97.5%		4
90.0%		4
75.0%	quartile	3
50.0%	median	2
25.0%	quartile	1
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

B. Uncritical variable to determine target variable

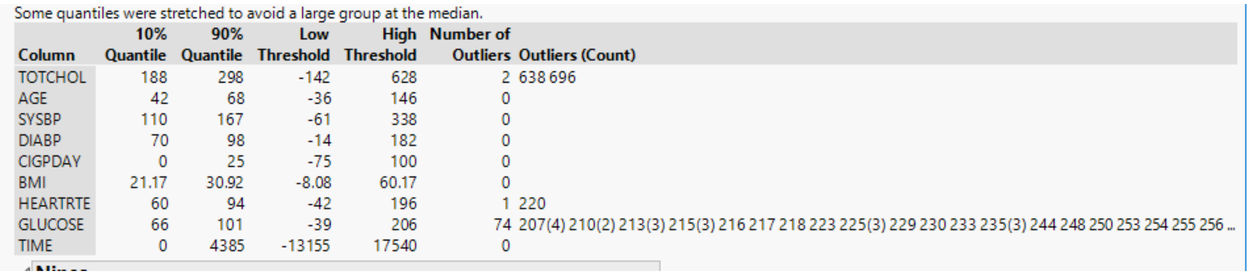
Variable	Description	Reason for deletion
RANDID	Unique Identification number of an individual	This number is uniquely assigned to a participant but our assumption is, if someone misses their medical report, they will assign a new RANDID. Moreover, most of the health care will have their unique ID file based on the current year so this variable will not help us as a risk factor.
HDLC and LDLC	High density Lipoprotein and Low density lipoprotein	These columns will help in determining the cholesterol level of an individual. But these columns are having more than 80% of missing data and we can recover this data with the variable TOTCHOL so we have deleted these 2 columns.
DEATH	Death of participant because of any cause	We have made use of this variable because we are predicting CHD (Target Variable) which occurs before the death of the participant. However, we could make use of "death" as an another Target variable which is out of scope for our project.
ANGINA	Individuals with have Angina Pectoris	Angina Pectoris is a pre-existing condition of individuals who suffered from CHD and it reveals the same information as CHD column so we decided to consider CHD and remove the ANGINA

C. Outlier Detection

We checked outliers to remove inconsistency from the data and found 74 outliers in glucose,but glucose is having unit value of 39-478(mg/DL). These high values of glucose indicate the high level of glucose in participants which is an important factor for determining the diabetic person.

Hence, we cannot remove these values by treating as outliers and we decided to keep these values.

(Jmp -> Cols -> Modelling Utilities -. Explore Outliers -> Quantile Range Outliers)

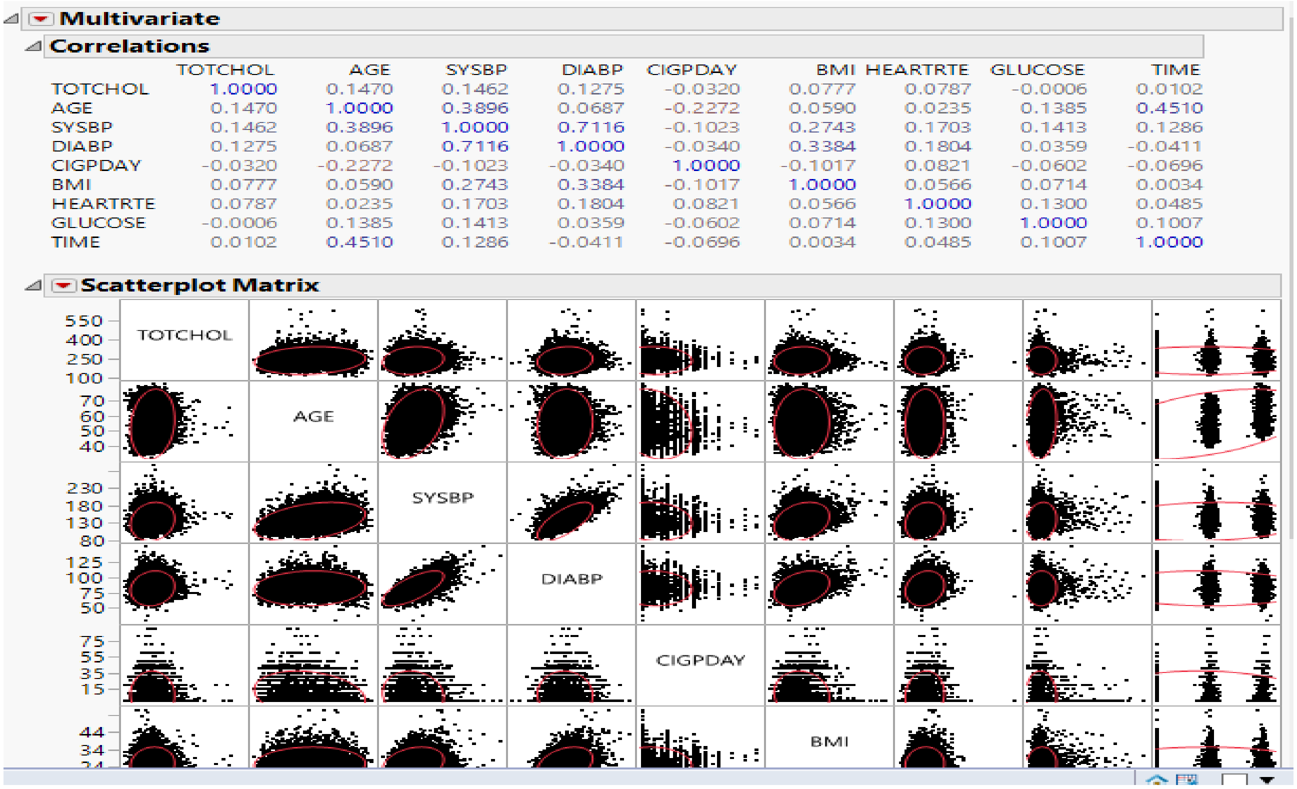


Some quantiles were stretched to avoid a large group at the median.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
TOTCHOL	188	298	-142	628	2	638 696
AGE	42	68	-36	146	0	
SYSBP	110	167	-61	338	0	
DIABP	70	98	-14	182	0	
CIGPDAY	0	25	-75	100	0	
BMI	21.17	30.92	-8.08	60.17	0	
HEARTRTE	60	94	-42	196	1	220
GLUCOSE	66	101	-39	206	74	207(4) 210(2) 213(3) 215(3) 216 217 218 223 225(3) 229 230 233 235(3) 244 248 250 253 254 255 256 ...
TIME	0	4385	-13155	17540	0	

D. Checking Correlations

We checked correlation using multivariate techniques to explore how many variables are related to each other. We tried to find the dependencies, outliers and clusters using scatterplot matrix and we found that there were no correlations between the continuous variables, and each variables is an independent variable and behaves as a unique risk factors for determining CHD.



Multivariate

Correlations

	TOTCHOL	AGE	SYSBP	DIABP	CIGPDAY	BMI	HEARTRTE	GLUCOSE	TIME
TOTCHOL	1.0000	0.1470	0.1462	0.1275	-0.0320	0.0777	0.0787	-0.0006	0.0102
AGE	0.1470	1.0000	0.3896	0.0687	-0.2272	0.0590	0.0235	0.1385	0.4510
SYSBP	0.1462	0.3896	1.0000	0.7116	-0.1023	0.2743	0.1703	0.1413	0.1286
DIABP	0.1275	0.0687	0.7116	1.0000	-0.0340	0.3384	0.1804	0.0359	-0.0411
CIGPDAY	-0.0320	-0.2272	-0.1023	-0.0340	1.0000	-0.1017	0.0821	-0.0602	-0.0696
BMI	0.0777	0.0590	0.2743	0.3384	-0.1017	1.0000	0.0566	0.0714	0.0034
HEARTRTE	0.0787	0.0235	0.1703	0.1804	0.0821	0.0566	1.0000	0.1300	0.0485
GLUCOSE	-0.0006	0.1385	0.1413	0.0359	-0.0602	0.0714	0.1300	1.0000	0.1007
TIME	0.0102	0.4510	0.1286	-0.0411	-0.0696	0.0034	0.0485	0.1007	1.0000

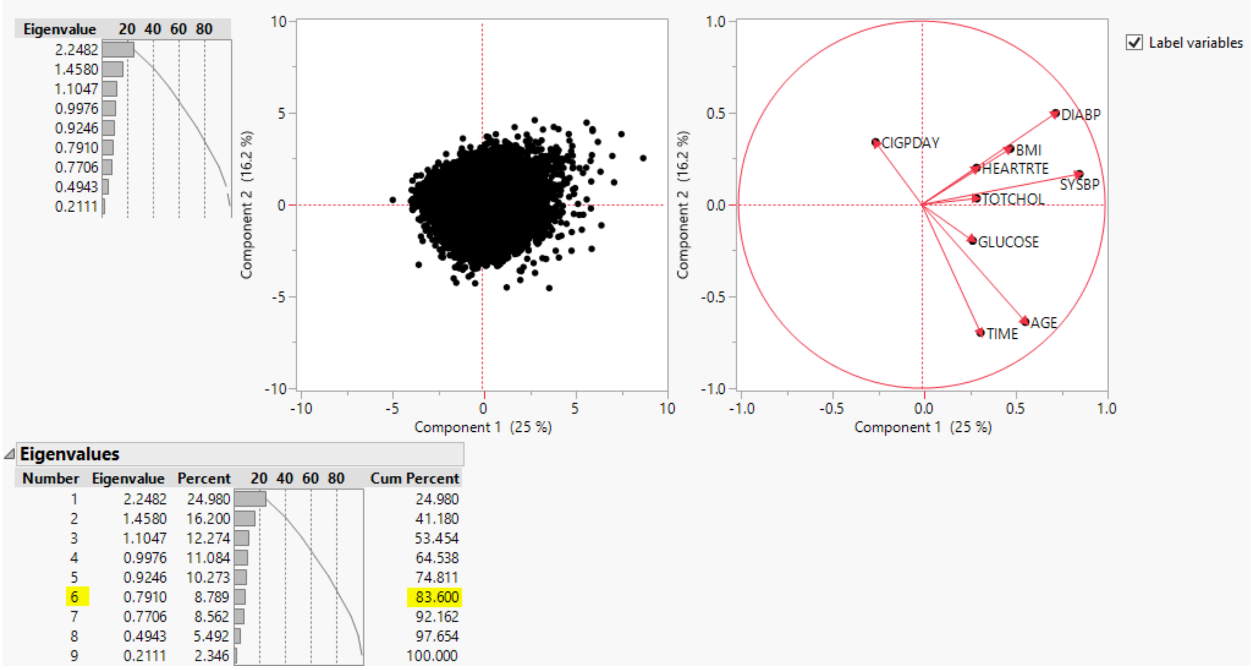
Scatterplot Matrix

E. Principal Component Analysis(PCA)

While constructing the predictive models for our project, we encountered the need for reducing the size of the available data both in terms of variables and observations so that we can minimize some associated cost with any variable. For this purpose, we have used principal component analysis, where all the principal components are completely uncorrelated and linear function of the original variables.

(Jmp – Analyze -Multivariate Methods – Principal Components)

After analysis for PCA, we found that first six principal components account for over 83.6% of total variance.



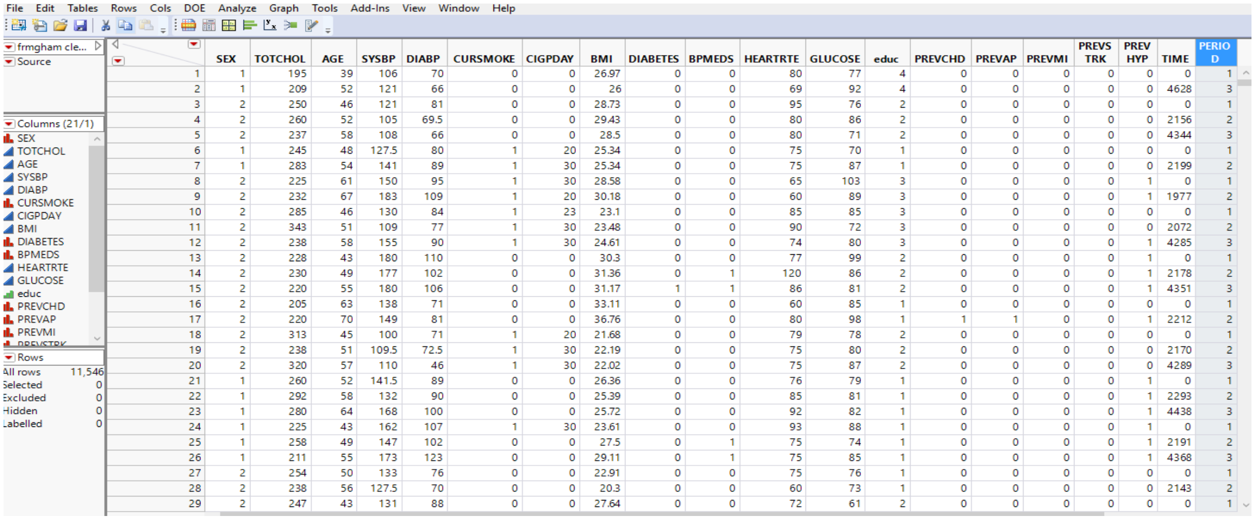
Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.2482	24.980		24.980
2	1.4580	16.200		41.180
3	1.1047	12.274		53.454
4	0.9976	11.084		64.538
5	0.9246	10.273		74.811
6	0.7910	8.789		83.600
7	0.7706	8.562		92.162
8	0.4943	5.492		97.654
9	0.2111	2.346		100.000

But we did not apply PCA for our data because doing PCA we can lose the high risk factors which are important for a participant to know that on what risk factor they should control to reduce the chances for coronary heart disease.

F. Generating the Sample

After preprocessing the data, we have sample data of table with exactly 11,546 observations with no empty rows and columns.



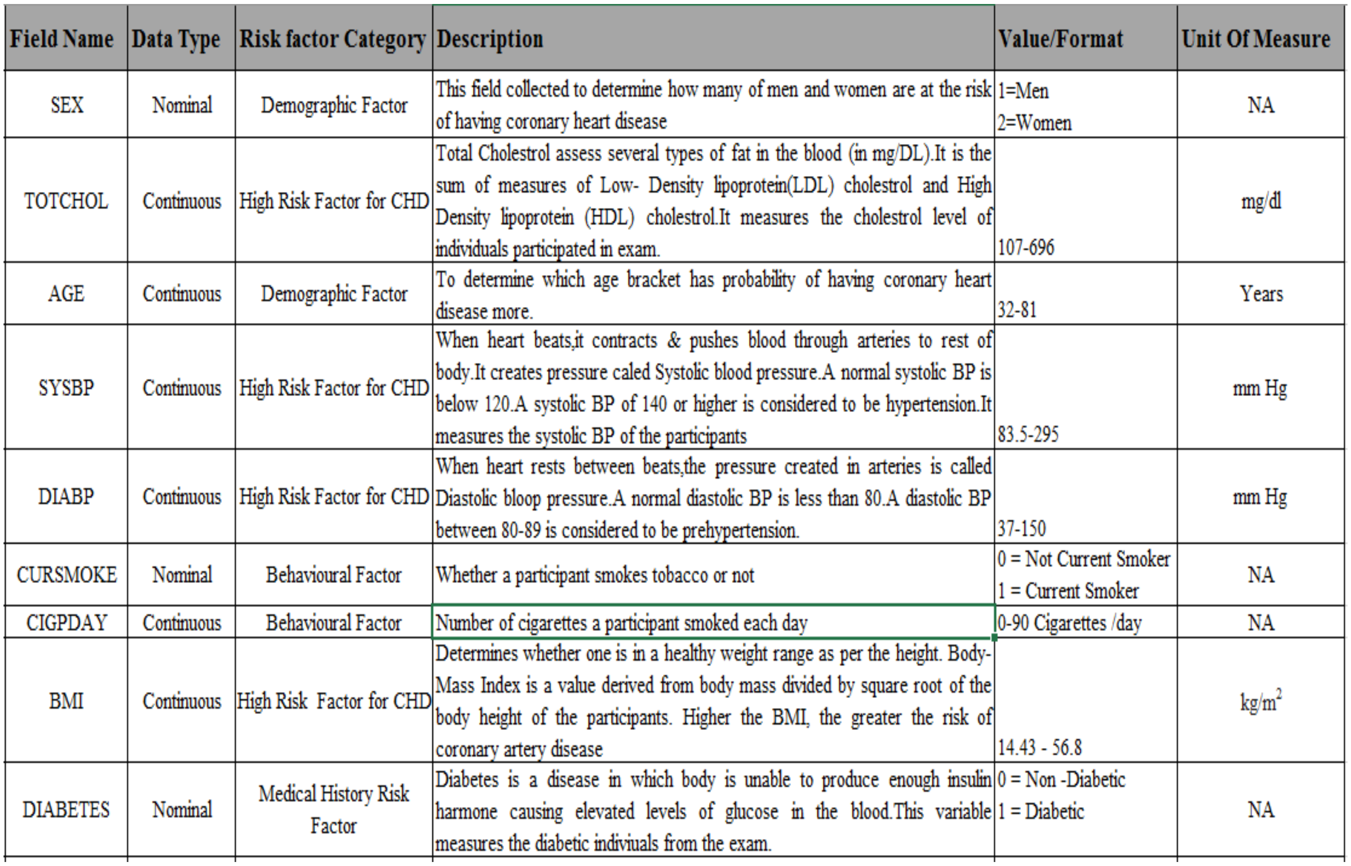
All rows 11,546
Selected 0
Excluded 0
Hidden 0
Labelled 0

	SEX	TOTCHOL	AGE	SYSBP	DIABP	CURSMOKE	CIGPDAY	BMI	DIABETES	BPMEDS	HEARTRTE	GLUCOSE	educ	PREVCHD	PREVAP	PREVMI	PREVSTRK	PREVHYP	TIME	PERIOD
1	1	195	39	106	70	0	0	26.97	0	0	80	77	4	0	0	0	0	0	0	1
2	1	209	52	121	66	0	0	26	0	0	69	92	4	0	0	0	0	0	4628	3
3	2	250	46	121	81	0	0	28.73	0	0	95	76	2	0	0	0	0	0	0	1
4	2	260	52	105	69.5	0	0	29.43	0	0	80	86	2	0	0	0	0	0	2156	2
5	2	237	58	108	66	0	0	28.5	0	0	80	71	2	0	0	0	0	0	4344	3
6	1	245	48	127.5	80	1	20	25.34	0	0	75	70	1	0	0	0	0	0	0	1
7	1	283	54	141	89	1	30	25.34	0	0	75	87	1	0	0	0	0	0	2199	2
8	2	225	61	150	95	1	30	28.58	0	0	65	103	3	0	0	0	0	1	0	1
9	2	232	67	183	109	1	20	30.18	0	0	60	89	3	0	0	0	0	1	1977	2
10	2	285	46	130	84	1	23	23.1	0	0	85	85	3	0	0	0	0	0	0	1
11	2	343	51	109	77	1	30	23.48	0	0	90	72	3	0	0	0	0	0	2072	2
12	2	238	58	155	90	1	30	24.61	0	0	74	80	3	0	0	0	0	1	4285	3
13	2	228	43	180	110	0	0	30.3	0	0	77	99	2	0	0	0	0	1	0	1
14	2	230	49	177	102	0	0	31.36	0	1	120	86	2	0	0	0	0	1	2178	2
15	2	220	55	180	106	0	0	31.17	1	1	86	81	2	0	0	0	0	1	4351	3
16	2	205	63	138	71	0	0	33.11	0	0	60	85	1	0	0	0	0	0	0	1
17	2	220	70	149	81	0	0	36.76	0	0	80	98	1	1	1	0	0	1	2212	2
18	2	313	45	100	71	1	20	21.68	0	0	79	78	2	0	0	0	0	0	0	1
19	2	238	51	109.5	72.5	1	30	22.19	0	0	75	80	2	0	0	0	0	0	2170	2
20	2	320	57	110	46	1	30	22.02	0	0	75	87	2	0	0	0	0	0	4289	3
21	1	260	52	141.5	89	0	0	26.36	0	0	76	79	1	0	0	0	0	1	0	1
22	1	292	58	132	90	0	0	25.39	0	0	85	81	1	0	0	0	0	1	2293	2
23	1	280	64	168	100	0	0	25.72	0	0	92	82	1	0	0	0	0	1	4438	3
24	1	225	43	162	107	1	30	23.61	0	0	93	88	1	0	0	0	0	1	0	1
25	1	258	49	147	102	0	0	27.5	0	1	75	74	1	0	0	0	0	1	2191	2
26	1	211	55	173	123	0	0	29.11	0	1	75	85	1	0	0	0	0	1	4368	3
27	2	254	50	133	76	0	0	22.91	0	0	75	76	1	0	0	0	0	0	0	1
28	2	238	56	127.5	70	0	0	20.3	0	0	60	73	1	0	0	0	0	0	2143	2
29	2	247	43	131	88	0	0	27.64	0	0	72	61	2	0	0	0	0	0	0	1

G. Defining Data Dictionary

Data Dictionary is a repository which contains the description & category of data objects used in the model for the benefit of data analyst and business users who want to understand the Framingham heart study risk factors. We have created our own data dictionary for easy reference.

Data
Dictionary_Farmingh



Field Name	Data Type	Risk factor Category	Description	Value/Format	Unit Of Measure
SEX	Nominal	Demographic Factor	This field collected to determine how many of men and women are at the risk of having coronary heart disease	1=Men 2=Women	NA
TOTCHOL	Continuous	High Risk Factor for CHD	Total Cholestrol assess several types of fat in the blood (in mg/DL).It is the sum of measures of Low- Density lipoprotein(LDL) cholestrol and High Density lipoprotein (HDL) cholestrol.It measures the cholestrol level of individuals participated in exam.	107-696	mg/dl
AGE	Continuous	Demographic Factor	To determine which age bracket has probability of having coronary heart disease more.	32-81	Years
SYSBP	Continuous	High Risk Factor for CHD	When heart beats,it contracts & pushes blood through arteries to rest of body.It creates pressure caled Systolic blood pressure.A normal systolic BP is below 120.A systolic BP of 140 or higher is considered to be hypertension.It measures the systolic BP of the participants	83.5-295	mm Hg
DIABP	Continuous	High Risk Factor for CHD	When heart rests between beats,the pressure created in arteries is called Diastolic bloop pressure.A normal diastolic BP is less than 80.A diastolic BP between 80-89 is considered to be prehypertension.	37-150	mm Hg
CURSMOKE	Nominal	Behavioural Factor	Whether a participant smokes tobacco or not	0 = Not Current Smoker 1 = Current Smoker	NA
CIGPDAY	Continuous	Behavioural Factor	Number of cigarettes a participant smoked each day	0-90 Cigarettes /day	NA
BMI	Continuous	High Risk Factor for CHD	Determines whether one is in a healthy weight range as per the height. Body-Mass Index is a value derived from body mass divided by square root of the body height of the participants. Higher the BMI, the greater the risk of coronary artery disease	14.43 - 56.8	kg/m^2
DIABETES	Nominal	Medical History Risk Factor	Diabetes is a disease in which body is unable to produce enough insulin harmone causing elevated levels of glucose in the blood.This variable measures the diabetic indiviuals from the exam.	0 = Non -Diabetic 1 = Diabetic	NA

H. Data Modelling

We have performed one-way analysis on the risk factors with respect to the target variables, by performing compare means (Hypothesis Test) test we tried to identify some of the important factors. Most likely the variables having Probability>|t| as <.0001 are the most significant variables in determining the heart disease. However, we have built models with all the variables to determine differences in the accuracies. Below is the table showing mean difference test of risk factors with respect to CHD.

Table Comparing means of 1(having heart disease) with 0(not having heart disease)

Term	Difference	Confidence	t Ratio	DF	Probability> |t|	Probability > t	Probability < t
Total Cholesterol	11.0212	0.95	11.85352	11544	<.0001	<.0001	1
AGE	3.50227	0.95	17.73318	5584.344	<.0001	<.0001	1
Systolic BP	9.9103	0.95	20.02608	5049.282	<.0001	<.0001	1
Diastolic BP	3.55436	0.95	13.94887	5083.546	<.0001	<.0001	1
CIGPDAy	0.43182	0.95	1.652864	5356.849	0.0984	0.0492	0.9508
BMI	1.11196	0.95	12.53507	5189.09	<.0001	<.0001	1
Heart rate	0.07636	0.95	0.288002	5422.18	0.7734	0.3867	0.6133
Glucose	4.30388	0.95	7.207877	4104.625	<.0001	<.0001	1

Data Models are the abstraction of reality; it is required to do data modelling in analytics to predict the future outcome. To do this h Jmp has provided us a Validation approach where we estimate the model parameters and using the other assess part to predict the ability of the model.

Whenever we are trying to fit models, it is advisable to partition the data into three parts:
Training dataset - To build the model
Validation dataset - To estimate how well our model has been trained
Test dataset - To estimate how well the model fits in the real world
We have 11,546 rows after cleaning the data having 8418 with 0's of CHD (Participants who will not get Coronary heart disease) and 3128 with 1's of CHD (Participants who will get Coronary heart disease).

We have partitioned the dataset based on the target variable value being 1(participants who has CHD, 3128 rows).

Our Training set has 2800 records with half of them having a heart disease and the rest are who doesn't have a heart disease. We built our models using this dataset.

Our validation set has 1800 records with half of them are having a heart disease and the rest are who doesn't have a heart disease. We validated our model using this dataset. We have used balanced datasets for training and validating the data set.

Our testing set has 738 records with participants having a heart disease and 6018 records who doesn't have a heart disease. we expect real world data to be similar to test data, having this in mind we did not use a balanced dataset for testing purpose as we don’t know what type of data we might get from the real world.

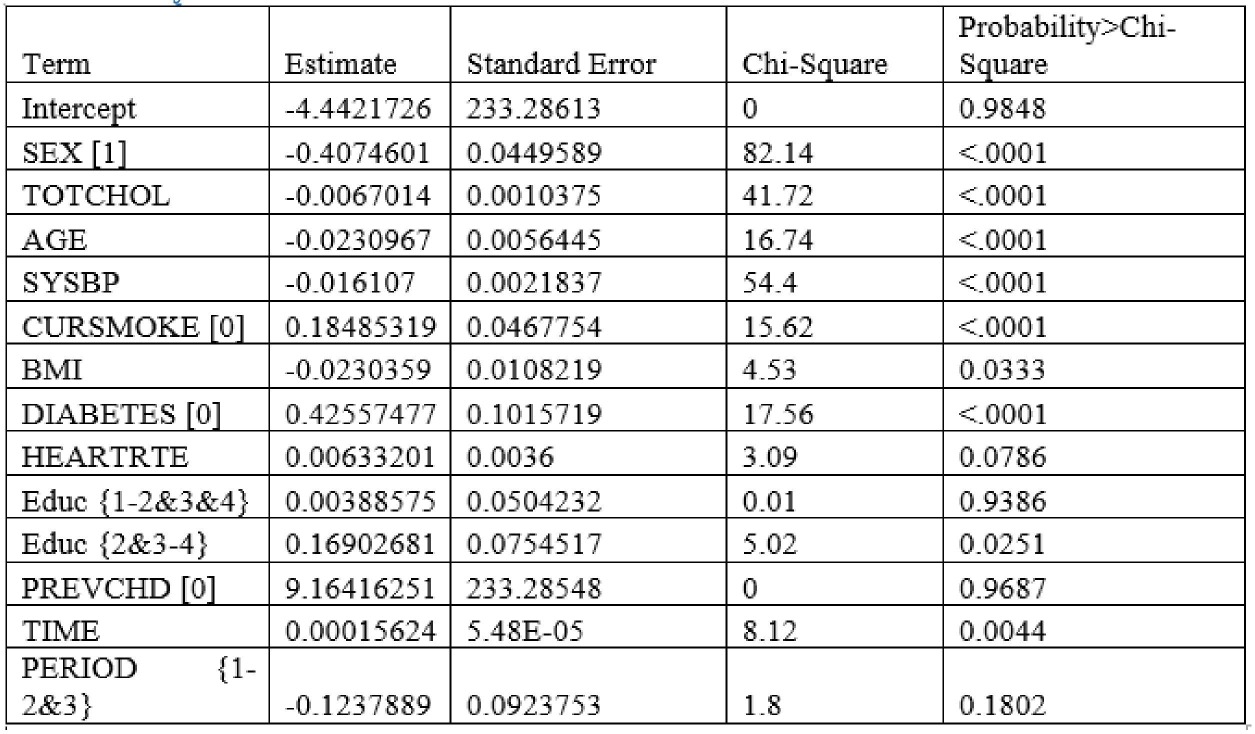
Our team has proposed the idea to perform modelling on each method to predict the probability that how many individuals will get CHD based on the risk factors involved and then select the best fit model based on high accuracy and low penalty error.

Modelling Methods Implemented:

1. **Logistic Regression** – Our target variable is CHD which is of nominal category so we will go for nominal logistic regression using stepwise method to choose one of the response levels as a smooth function of the x-factor. We used the forward selection method with no variables in the model and used step by step process to determine the model that have strongest statistical significant variables based on AIC value.

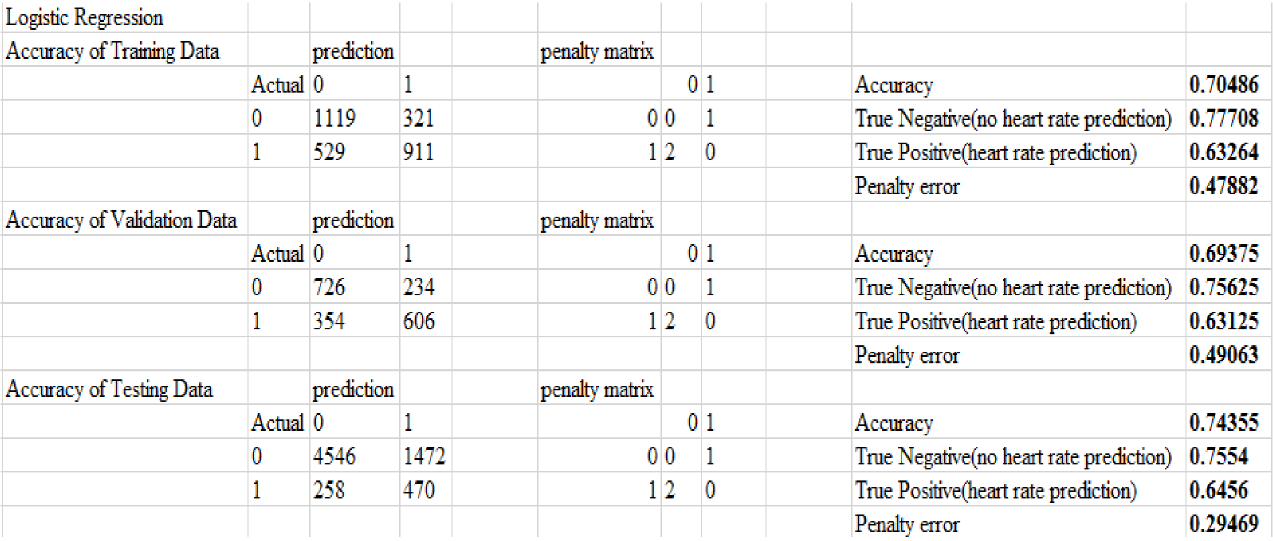
   After running the model, we found our best model with step 13 because any further addition of the variables using step method increased AIC value.

Below table shows the significance of the variables in the logistic regression model with the best accuracy of 74.40%.



Term	Estimate	Standard Error	Chi-Square	Probability>Chi-Square
Intercept	-4.4421726	233.28613	0	0.9848
SEX [1]	-0.4074601	0.0449589	82.14	<.0001
TOTCHOL	-0.0067014	0.0010375	41.72	<.0001
AGE	-0.0230967	0.0056445	16.74	<.0001
SYSBP	-0.016107	0.0021837	54.4	<.0001
CURSMOKE [0]	0.18485319	0.0467754	15.62	<.0001
BMI	-0.0230359	0.0108219	4.53	0.0333
DIABETES [0]	0.42557477	0.1015719	17.56	<.0001
HEARTRTE	0.00633201	0.0036	3.09	0.0786
Educ {1-2&3&4}	0.00388575	0.0504232	0.01	0.9386
Educ {2&3-4}	0.16902681	0.0754517	5.02	0.0251
PREVCHD [0]	9.16416251	233.28548	0	0.9687
TIME	0.00015624	5.48E-05	8.12	0.0044
PERIOD {1-2&3}	-0.1237889	0.0923753	1.8	0.1802

Accuracy and Penalty Error Analysis of the best logistic model



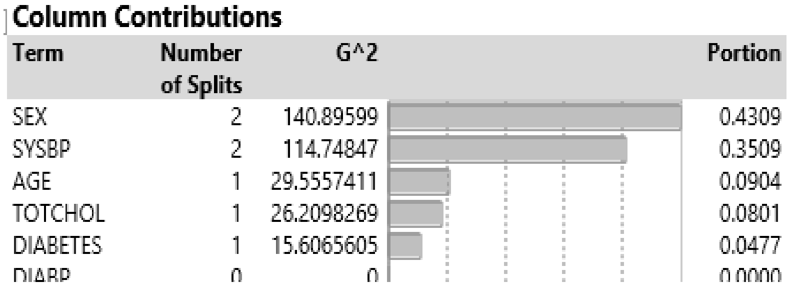
Logistic Regression										
Accuracy of Training Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.70486**
	0	1119	321		0	0	1		True Negative(no heart rate prediction)	**0.77708**
	1	529	911		1	2	0		True Positive(heart rate prediction)	**0.63264**
									Penalty error	**0.47882**
Accuracy of Validation Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.69375**
	0	726	234		0	0	1		True Negative(no heart rate prediction)	**0.75625**
	1	354	606		1	2	0		True Positive(heart rate prediction)	**0.63125**
									Penalty error	**0.49063**
Accuracy of Testing Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.74355**
	0	4546	1472		0	0	1		True Negative(no heart rate prediction)	**0.7554**
	1	258	470		1	2	0		True Positive(heart rate prediction)	**0.6456**
									Penalty error	**0.29469**

2 **Decision Tree**: Decision trees or classification trees provide a better visual representation of a Model with a flow chart structure. Decision trees split the independent variables based on the target variable by putting decisions in different splits.

Jmp provides partition in modelling options which recursively splits the data to predict a response using decision trees.

Flow in SAS JMP→ (Jmp -Analyze- Modelling -Partition).

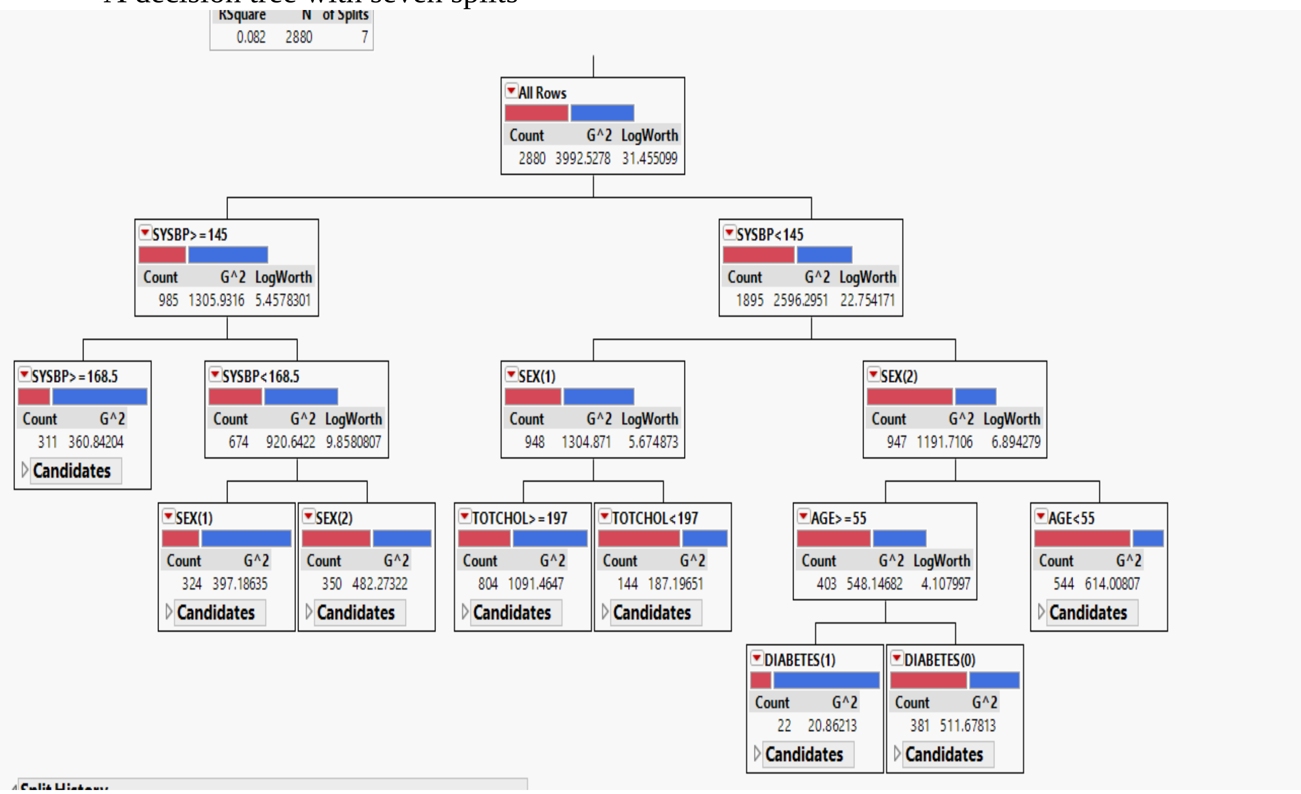
By building this decision trees we recognized risk factors like Systolic BP, Sex, Total cholesterol, Age and diabetes are the most important factors leads to CHD. We can observe the same in the below decision tree.



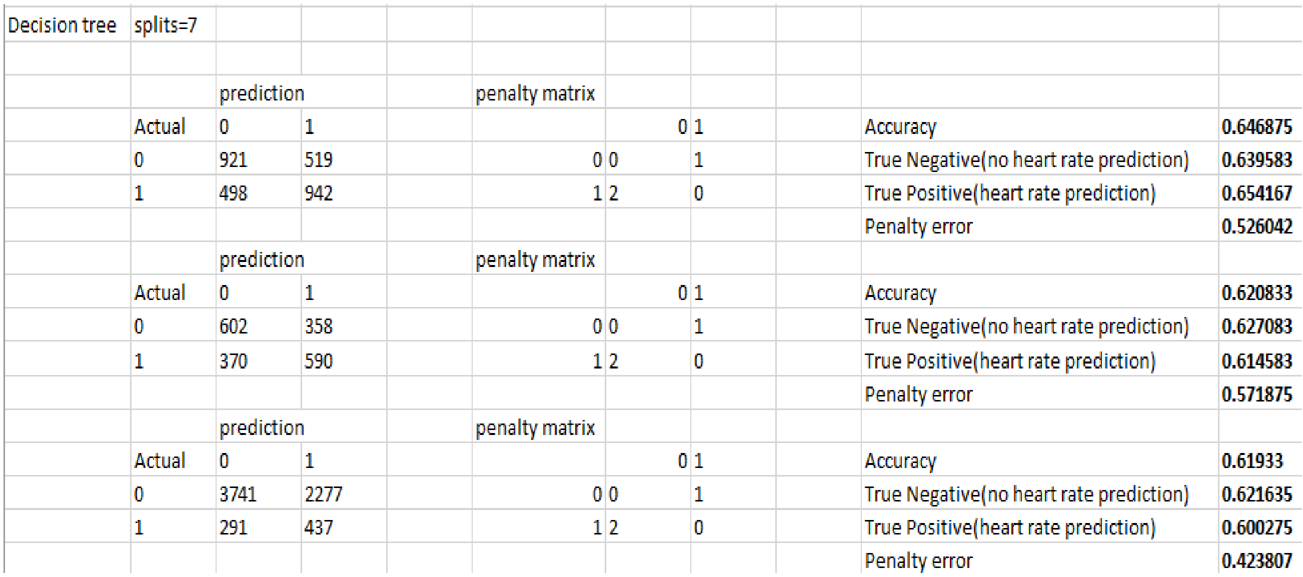
Column Contributions

Term	Number of Splits	G^2	Portion
SEX	2	140.89599	0.4309
SYSBP	2	114.74847	0.3509
AGE	1	29.5557411	0.0904
TOTCHOL	1	26.2098269	0.0801
DIABETES	1	15.6065605	0.0477
DIABP	0	0	0.0000

A decision tree with seven splits



Accuracy and Penalty Error Analysis of the best Decision Tree model



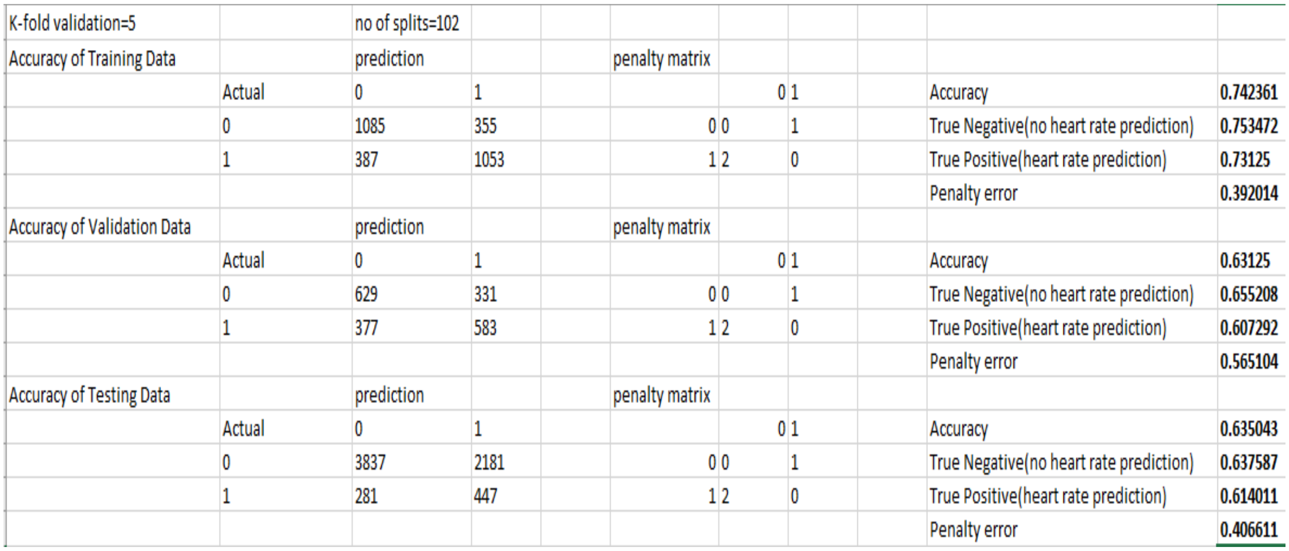
Decision tree	splits=7									
		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.646875
	0	921	519		0	0	1		True Negative(no heart rate prediction)	0.639583
	1	498	942		1	2	0		True Positive(heart rate prediction)	0.654167
									Penalty error	0.526042
		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.620833
	0	602	358		0	0	1		True Negative(no heart rate prediction)	0.627083
	1	370	590		1	2	0		True Positive(heart rate prediction)	0.614583
									Penalty error	0.571875
		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.61933
	0	3741	2277		0	0	1		True Negative(no heart rate prediction)	0.621635
	1	291	437		1	2	0		True Positive(heart rate prediction)	0.600275
									Penalty error	0.423807

3. K-fold cross validation:

In decision trees the number of splits might affect the accuracy, If the number of splits are less the model might be too simple, if the number of splits is more the model might over fit the data. We could select the model based on the testing set accuracy. But the testing set should always be used to measure performance rather than using it to select the model.

This why we used cross validation which splits our data into K folds based on K value given, from these folds' model will use one fold to test the model by building the model with all the other folds. This process goes by building models with different folds and testing model with a different fold every time. By doing this we can pick the model which has highest out of sample accuracy.

(Jmp – Analyze – Modelling -Partition – K Fold cross validation)
Accuracy and Penalty Error Analysis of the best Decision Tree model with K-fold cross validation



K-fold validation=5		no of splits=102								
Accuracy of Training Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.742361**
	0	1085	355		0	0	1		True Negative(no heart rate prediction)	**0.753472**
	1	387	1053		1	2	0		True Positive(heart rate prediction)	**0.73125**
									Penalty error	**0.392014**
Accuracy of Validation Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.63125**
	0	629	331		0	0	1		True Negative(no heart rate prediction)	**0.655208**
	1	377	583		1	2	0		True Positive(heart rate prediction)	**0.607292**
									Penalty error	**0.565104**
Accuracy of Testing Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.635043**
	0	3837	2181		0	0	1		True Negative(no heart rate prediction)	**0.637587**
	1	281	447		1	2	0		True Positive(heart rate prediction)	**0.614011**
									Penalty error	**0.406611**

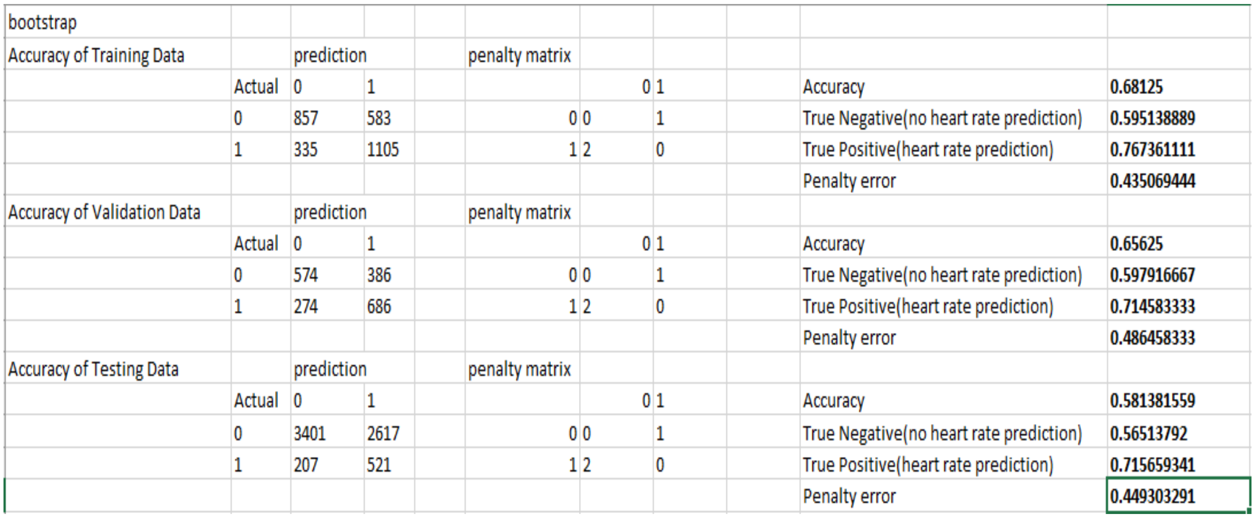
4. Bootstrap Forest:

This method was designed to improve accuracy of decision trees. It works by building a large number of decision trees which makes a forest with trees. However, it can't be visualized as a decision tree which makes it less interpretable. Each tree in the forest is given votes based on the outcome and we select the outcome with the majority of votes.

Each tree in the bootstrap is built on randomly selected data from available data to build training model.

(Jmp – Analyze – Modelling -Partition – K Fold cross validation)

Accuracy and Penalty Error Analysis of the best Decision Tree model



bootstrap										
Accuracy of Training Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.68125
	0	857	583		0	0	1		True Negative(no heart rate prediction)	0.595138889
	1	335	1105		1	2	0		True Positive(heart rate prediction)	0.767361111
									Penalty error	0.435069444
Accuracy of Validation Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.65625
	0	574	386		0	0	1		True Negative(no heart rate prediction)	0.597916667
	1	274	686		1	2	0		True Positive(heart rate prediction)	0.714583333
									Penalty error	0.486458333
Accuracy of Testing Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.581381559
	0	3401	2617		0	0	1		True Negative(no heart rate prediction)	0.56513792
	1	207	521		1	2	0		True Positive(heart rate prediction)	0.715659341
									Penalty error	0.449303291

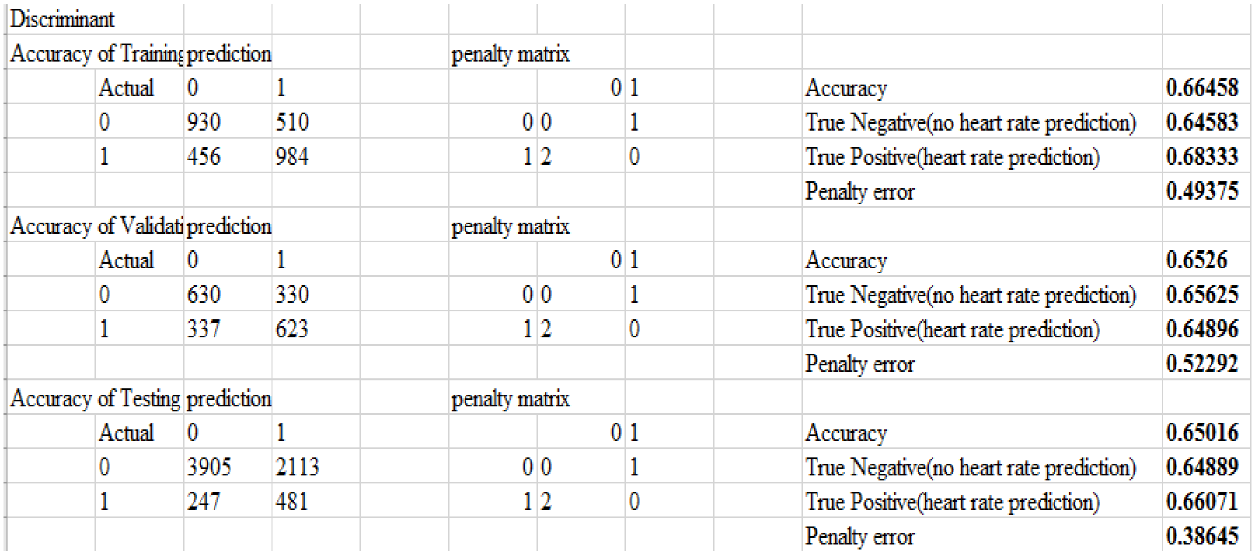
5. Discriminant Analysis:

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. In our example we have two groups of participant's, one is with coronary heart disease and the other is without coronary heart disease. We implemented this method to identify what variables are determining the participant to get a heart disease.

We used Jmp multivariate platform to implement this algorithm on Framingham Heart Study.

(Jmp – Analyze – Multivariate Methods-Discriminant)

Accuracy and Penalty Error Analysis of the best Decision Tree model

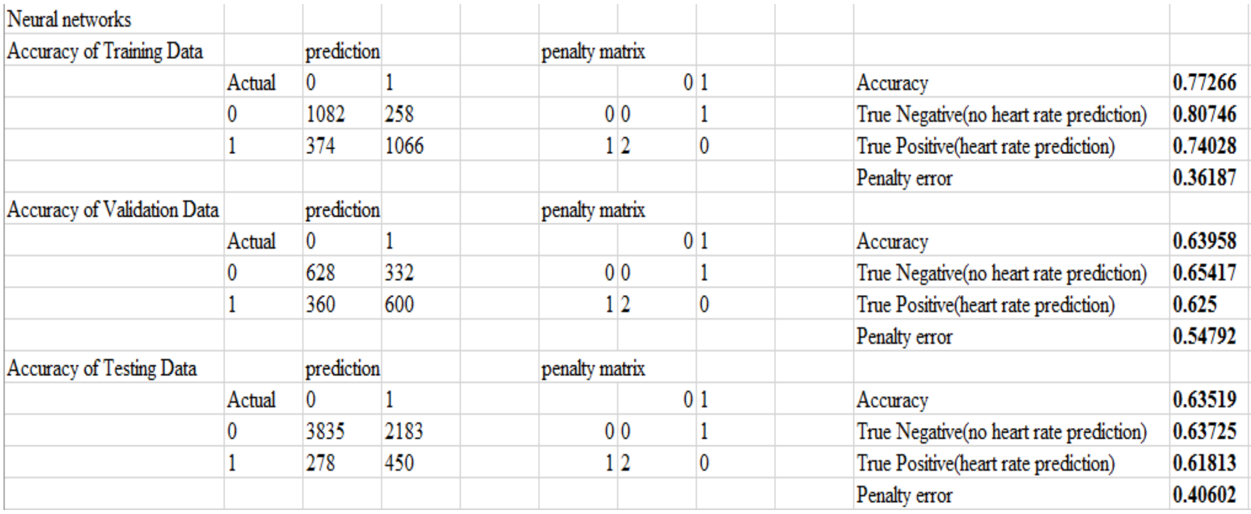


Discriminant										
Accuracy of Training		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.66458
	0	930	510		0	0	1		True Negative(no heart rate prediction)	0.64583
	1	456	984		1	2	0		True Positive(heart rate prediction)	0.68333
									Penalty error	0.49375
Accuracy of Validati		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.6526
	0	630	330		0	0	1		True Negative(no heart rate prediction)	0.65625
	1	337	623		1	2	0		True Positive(heart rate prediction)	0.64896
									Penalty error	0.52292
Accuracy of Testing		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	0.65016
	0	3905	2113		0	0	1		True Negative(no heart rate prediction)	0.64889
	1	247	481		1	2	0		True Positive(heart rate prediction)	0.66071
									Penalty error	0.38645

6. Neural Networks:

A neural network is a model which is built by highly interconnecting independent variables to predict or estimate a dependent variable, this model is inspired by biological neural networks (central nervous system of human beings, primarily Brain).

In our case we used predictions from different models (Logistic Regression, Decision trees, Random Forest, Discriminant Analysis) as input to the neural networks. Our Target variable is CHD, which states whether or not a participant has Heart Disease.
(**Jmp – Analyze – Modeling-Neural)**



Neural networks										
Accuracy of Training Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.77266**
	0	1082	258		0	0	1		True Negative(no heart rate prediction)	**0.80746**
	1	374	1066		1	2	0		True Positive(heart rate prediction)	**0.74028**
									Penalty error	**0.36187**
Accuracy of Validation Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.63958**
	0	628	332		0	0	1		True Negative(no heart rate prediction)	**0.65417**
	1	360	600		1	2	0		True Positive(heart rate prediction)	**0.625**
									Penalty error	**0.54792**
Accuracy of Testing Data		prediction			penalty matrix					
	Actual	0	1			0	1		Accuracy	**0.63519**
	0	3835	2183		0	0	1		True Negative(no heart rate prediction)	**0.63725**
	1	278	450		1	2	0		True Positive(heart rate prediction)	**0.61813**
									Penalty error	**0.40602**

Below table shows the accuracies of all the models and the respective penalty errors.

	Logistic	Decision Trees	K Fold Cross Validation	Boot Strap	Discriminant	Neural Networks
Accuracy - Training Data	70.50%	64.69%	74.24%	68.13%	66.46%	77.27%
True Negative	77.70%	63.96%	75.35%	59.51%	64.58%	80.75%
True Positive	63.30%	65.42%	73.13%	76.74%	68.33%	74.03%
Penalty error	47.88%	52.60%	39.20%	43.51%	49.38%	36.19%
Accuracy - Validation Data	69.40%	62.08%	63.13%	65.63%	65.26%	63.96%
True Negative	75.60%	62.71%	65.52%	59.79%	65.63%	65.42%
True Positive	63.10%	61.46%	60.73%	71.46%	64.90%	62.50%
Penalty error	49.06%	57.19%	56.51%	48.65%	52.29%	54.79%
Accuracy - Testing Data	74.40%	61.93%	63.50%	58.14%	65.02%	63.52%
True Negative	75.50%	62.16%	63.76%	56.51%	64.89%	63.73%
True Positive	64.60%	60.03%	61.40%	71.57%	66.07%	61.81%
Penalty error	29.47%	42.38%	40.66%	44.93%	38.65%	40.60%

References

1. History of the Framingham Heart Study, *Framingham Heart Study*.
2. The Framingham Heart Study: The Town That Changed America's Heart, *Framingham Heart Study*.
3. Score Sheet Can Estimate Individual's Risk for Developing Heart Disease, *Framingham Heart Study*
4. Gina Wei, The Framingham Heart Study (FHS), *National Heart and Blood institute*.
5. Christian Nordqvist, C. (2016), what is cardiovascular disease? What causes cardiovascular disease? medical news today. Heart Disease: Scope and Impact, Heart Disease Facts, *The Heart Foundation*.
6. World Health Federation-Cardio Vascular Disease Risk Factors. *World Health Federation*.
7. Mendis S, Puska P, Norrving B, C. (2011), Global Atlas on Cardiovascular Disease Prevention and Control. *World Health Organization (in collaboration with the World Heart Federation and World Stroke Organization), Geneva.*
8. Cardiovascular Disease (10-year risk), *Framingham Heart Study*.
9. Framingham Heart Study Longitudinal Data Documentation, *Framingham Heart Study.*
10. Correlation and Multivariate Techniques, *Jmp.*
11. Principal components for modeling, *Abbott analytics*.
12. Validation, *Jmp.(http://www.jmp.com/support/help/Validation_2.shtml).*
13. Partition Models, Jmp, *http://www.jmp.com/support/help/Partition_Models.shtml*
14. Discriminant Function Analysis, *University of Arlington.*

:

- National Heart, Lung and Blood Institute

END