ISyE 6414: Regression Analysis
Project Report

Predicting the diagnosis of Coronary Artery Disease using Statistical Models

April 16, 2021

Submitted by -
Raj Desai (GTID: 903575149)
Sukanya R. Iyer (GTID: 903622016)
Tushar Vende (GTID: 903590888)

Contents

1. Introduction

1.1 Background

Heart disease is the leading cause of death in the United States, causing about 1 in 4 deaths. The term “heart disease” describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you are born with (congenital heart defects), among others. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves, or rhythm, also are considered forms of heart disease. In the United States, the most common type of heart disease is coronary artery disease (CAD), which can lead to heart attack.

CVDs are concertedly contributed by hypertension, diabetes, overweight and unhealthy lifestyles. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions. According to a news article, heart disease proves to be the leading cause of death for both women and men. These diseases take an economic toll, as well, costing our health care system $214 billion per year and causing $138 billion in lost productivity on the job.

A heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients. Early prediction of a possible heart disease will lead to better decision making for a clinical treatment. We have a data which classified if patients have heart disease or not according to features in it. We will try to use this data to create a model which tries predicting if a patient has this disease or not.

1.2 Causes of Heart Disease

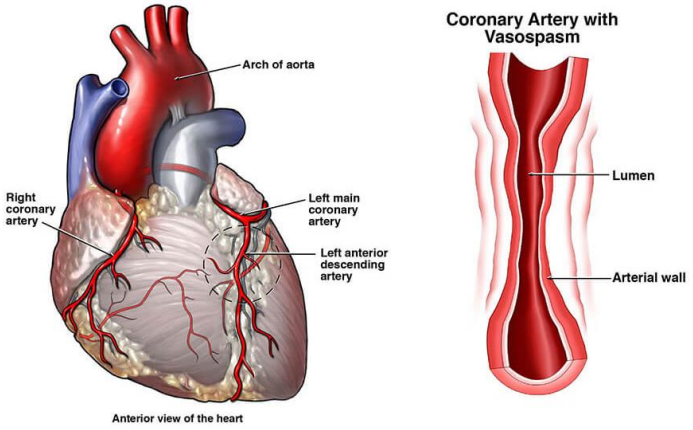
There are three main types of coronary heart disease: obstructive coronary artery disease, nonobstructive coronary artery disease, and coronary microvascular disease. Coronary artery disease affects the large arteries on the surface of the heart. Many people have both obstructive and nonobstructive forms of this disease. Coronary microvascular disease affects the tiny arteries in the heart muscle. Plaque build-up in the arteries is called atherosclerosis. When this build-up happens in the heart's arteries over many years, the arteries become narrower and harden, reducing oxygen-rich blood flow to the heart. The result is coronary artery disease. Small plaques can also develop in the small blood vessels in the heart, causing coronary microvascular disease.

Problems with how the heart’s blood vessels work can cause coronary heart disease. For example, the blood vessels may not respond to signals that the heart needs more oxygen-rich

blood. Normally, the blood vessels widen to allow more blood flow when a person is physically active or under stress. But if you have coronary heart disease, the size of these blood vessels may not change, or the blood vessels may even narrow. The cause of these problems may involve:

- Damage or injury to the walls of the arteries or tiny blood vessels from chronic inflammation, high blood pressure, or diabetes.
- Molecular changes that are part of the normal aging process. Molecular changes affect the way genes and proteins are controlled inside cells.

Figure 1.1: Numerical features and heart disease diagnosis



In nonobstructive coronary artery disease, damage to the inner walls of the coronary arteries can cause them to spasm (suddenly tighten). This is called vasospasm. The spasm causes the arteries to narrow temporarily and blocks blood flow to the heart.

1.3 Risk Factors

There are many risk factors for coronary heart disease. Your risk of coronary heart disease goes up with the number of risk factors you have and how serious they are. Some risk factors—such as high blood pressure and high blood cholesterol—can be changed through heart-healthy lifestyle changes. Other risk factors, such as sex, older age, family history and genetics, and race and ethnicity, cannot be changed.

Age: In men, the risk for the disease starts to increase around age 45. Before menopause, women have a lower risk of coronary heart disease than men. After around age 55, women's risk goes up.

Environment and occupation: Air pollution in the environment can put you at higher risk of coronary heart disease. The increase in risk may be higher in older adults, women, and people who have diabetes or obesity.

Gender: Coronary heart disease affects men and women. Obstructive coronary artery disease is more common in men. However, nonobstructive coronary artery disease is more common

in women. Since the nonobstructive type is harder to diagnose, women may not be diagnosed and treated as quickly as men.

Lifestyle Habits: Over time, unhealthy lifestyle habits increase your risk of coronary heart disease because they can lead to plaque buildup in the heart's blood vessels. Unhealthy lifestyle habits that are risk factors include the following: Being physically inactive, Not getting enough good quality sleep, Smoking tobacco or long-term exposure to secondhand smoke, Stress, Unhealthy eating patterns.

Other medical conditions: Other medical conditions that can raise your risk of developing coronary heart disease include:

- Atherosclerosis
- Autoimmune and inflammatory diseases
- Chronic kidney disease
- Congenital coronary artery defects
- Diabetes
- High blood LDL cholesterol (sometimes called "bad cholesterol")
- High blood pressure
- HIV/AIDS, especially among older adults. Part of the risk might be due to side effects of HIV treatments.
- Mental health conditions, including anxiety, depression, and posttraumatic stress disorder (PTSD)
- Metabolic syndrome
- Overweight and obesity
- Sleep disorders, such as sleep apnea or sleep deprivation and deficiency

1.4 Diagnosis Techniques

Diagnoses of coronary heart disease is done based on the symptoms, medical and family history, risk factors, and the results from tests and procedures.

- **Blood tests** to check the levels of cholesterol, triglycerides, sugar, lipoproteins, or proteins, such as C-reactive protein, that are a sign of inflammation.
- **Electrocardiogram (EKG or ECG)** to determine whether the heart's rhythm is steady or irregular. An EKG also records the strength and timing of electrical signals as they pass through the heart.
- **Coronary calcium scan** to measure the amount of calcium in the walls of your coronary arteries.
- **Stress tests** to check how your heart works during physical stress. During stress testing, you walk or run on a treadmill or pedal a stationary bike to make your heart work hard and beat fast. To detect reduced blood flow to your heart muscle, while you exercise you will be monitored by ECG and possibly echocardiogram or a CT scan.
- **Cardiac MRI (magnetic resonance imaging)** to detect tissue damage or problems with blood flow in the heart or coronary arteries. It can help your doctor diagnose coronary microvascular disease or nonobstructive or obstructive coronary artery disease. Cardiac MRI can help explain results from other imaging tests such as chest X-rays and CT scans.

- **Cardiac positron emission tomography (PET) scanning** to assess blood flow through the small coronary blood vessels and into the heart tissues. This is a type of nuclear heart scan that can diagnose coronary microvascular disease.
- **Coronary angiography** to show the insides of your coronary arteries. To get the dye into your coronary arteries, your doctor will use a procedure called cardiac catheterization. This procedure is often used if other tests show that you are likely to have coronary artery disease.
- **Coronary computed tomographic angiography** to show the insides of your coronary arteries rather than an invasive cardiac catheterization. It is a noninvasive imaging test using CT scanning.

2. Objectives of the Project

Objectives of the project are to:

I. Assess the efficacy of various simple clinical factors in predicting a possible occurrence of a heart disease in a patient.

II. Contrast various statistical models to decide the best model for predicting presence of a heart disease based on simple clinical factors.

3. Data Description

This Heart Disease Data Set dataset dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. The dataset present in the repository originally describes 76 features or attributes from 303 patients; however, published studies chose only 14 features that are relevant in predicting heart disease. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = disease and 1 = no disease.

Feature	Description
Age	Age in years
Gender	Value 1 = male, Value 0 = female
Chest Pain Type (cp)	Value 1 = typical angina, Value 2 = atypical angina, Value 3 = non-anginal pain, Value 4 = asymptomatic
Resting Blood Pressure in mm Hg (trestbps)	In mm Hg on admission to hospital
Cholesterol (chol)	Serum cholesterol in mg/dl
Fasting Blood Sugar (fbs)	Fasting blood sugar >120 mg/dl, 1=true, 0=false
restecg	Resting electrocardiographic results, Value 0 = normal, Value 1 = ST wave abnormality, Value 2 = showing probable hypertrophy
Max Heart Rate (thalach)	Maximum heart rate achieved
Exercise Angina (exang)	Exercise induced angina, Value 1=yes, Value 0=no
Old Peak	ST depression induced by exercise relative to rest
Slope	Slope of peak exercise ST segment, 1 = upsloping, 2 = flat, 3 = down slopping
Number of Major Vessels (ca)	Number of major vessels
Thal (thalassemia)	Value 0 = normal, Value 1 = fixed defect, Value 2 = reversable defect
Target (Diagnosis)	Heart disease diagnosed, Value 1 = yes, Value 0 = no (target variable)

4. Exploratory Data Analysis

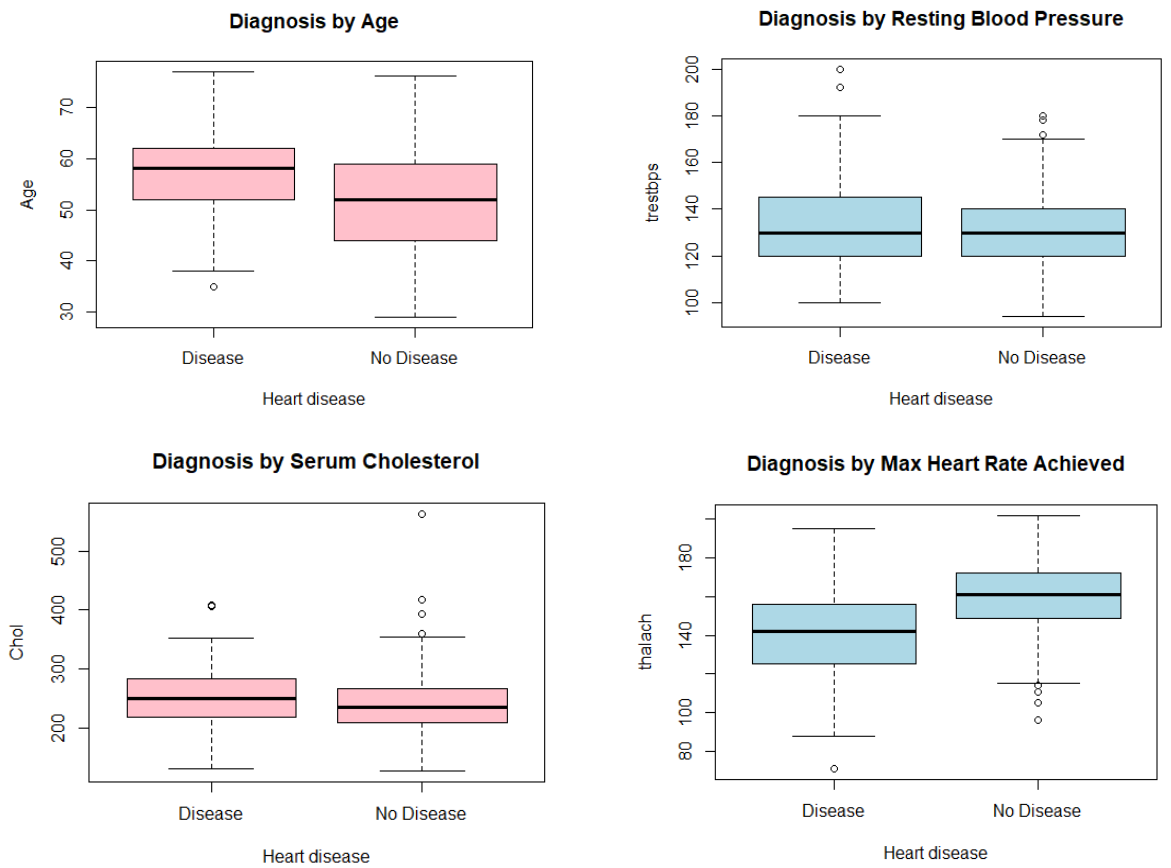
To understand the characteristics of the data, we did exploratory data analysis. EDA typically detects errors, finds appropriate data, checks assumptions, and determines the correlation among the explanatory variables. It enhances the insight into a given dataset and identifies anomalies. The dataset comprises 303 samples and each of them has 14 characteristics. There are 13 predictors and one target variable. We used data cleaning to check data type, data character mistakes, to deal with NA values, converted some variables to factors, removed the rows that had "?" for observations or any other anomaly, and reordered the variables within the data frame. After analysing the data, we split the 13 characteristics into two categories, 6 of them as numerical and rest 7 as categorical.

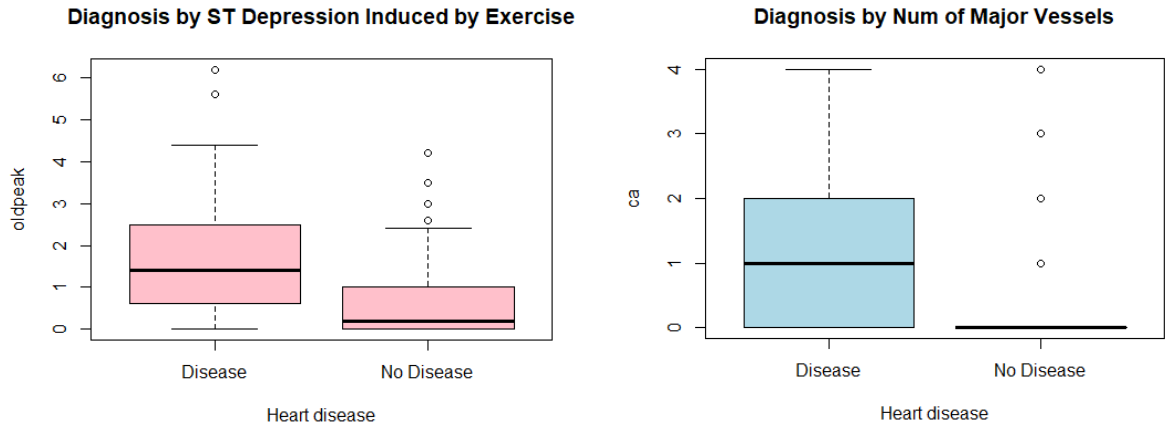
4.1 Exploratory Data Analysis of numerical features

For plotting the numerical features, we used boxplots as shown in Figure 1. These plots display the distribution of 6 numerical features for the presence and absence of the disease. The inferences that we can draw from the figure are such that what is the median age, resting blood pressure, cholesterol, maximum heart rate, depression induced by exercise and number of major vessels for patients having heart disease compared to patients not having a heart disease.

From the figure 4.1, we can see that samples with heart disease have a median age of 50 vs those without the disease have a median age of 57. Resting blood pressure, cholesterol, oldpeak (ST depression induced by exercise relative to rest) and number of major vessels are also higher for heart disease patients compared to controls.

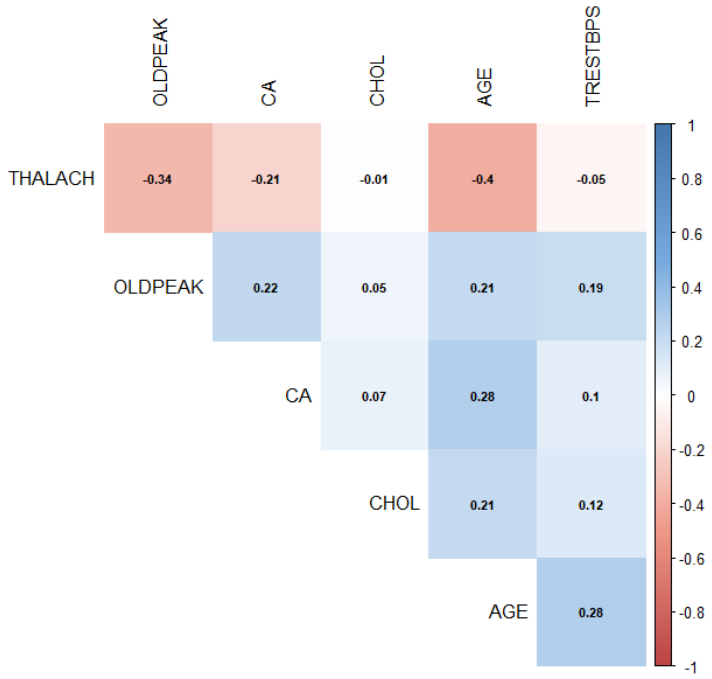
Figure 4.1: Numerical features and heart disease diagnosis





We also found the correlation between the numerical features. From the figure 4.2 we can see the correlation between the 6 numerical features. The scale of the correlation matrix is from negative one to positive one. Here, we find that all features except thalach, are positively correlated to each other. This makes sense, since “thalach” is lower for diseased person and high for non-diseased person, as seen in the boxplots above. Further, age has the higher correlation with other features, 36% with ca, 29% with resting blood pressure and 20% with oldpeak and cholesterol.

Figure 4.2: Correlation matrix for Numerical features

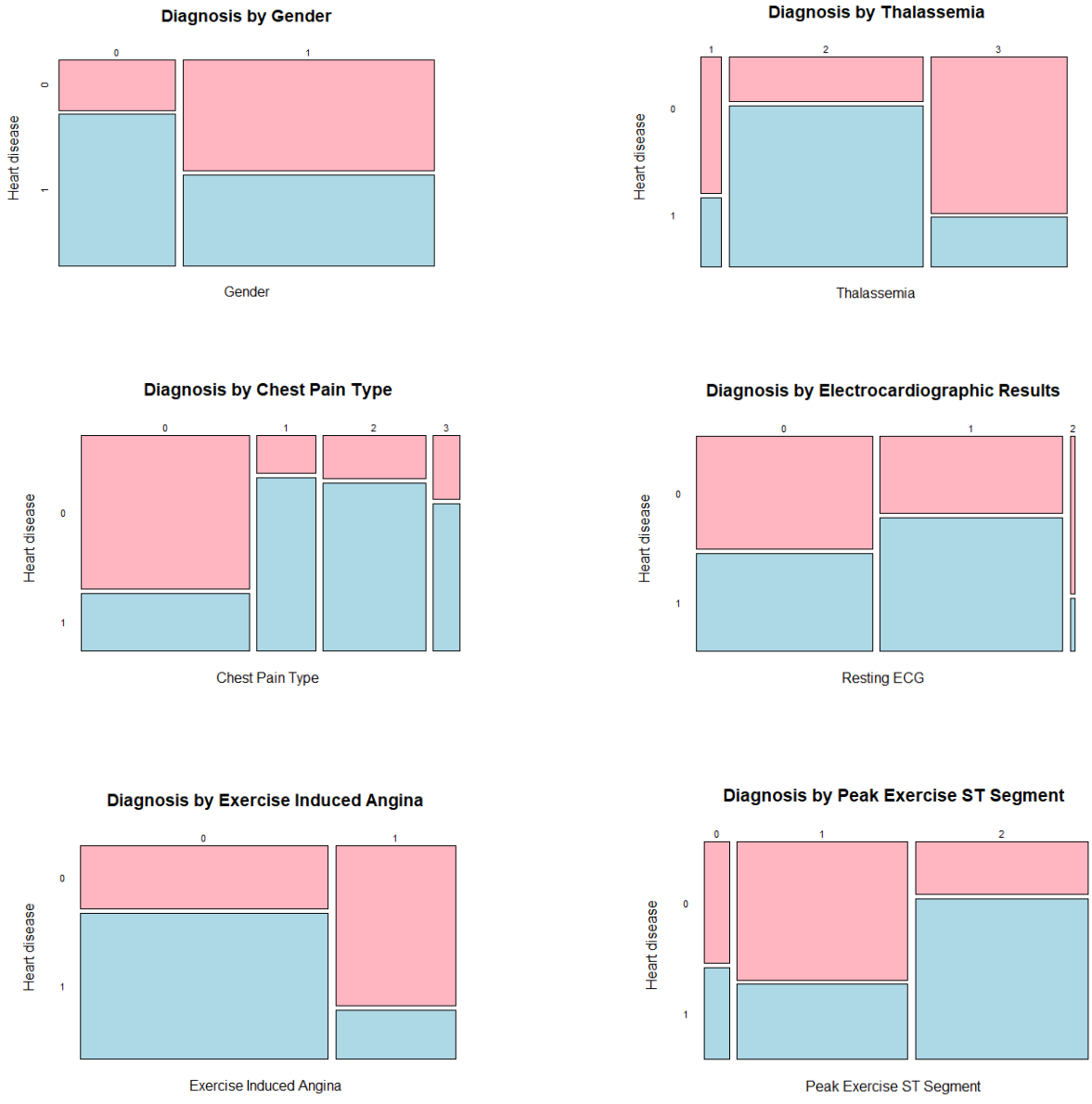


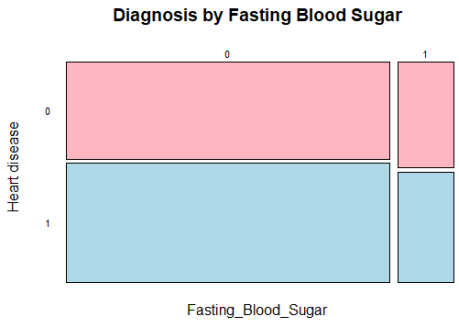
4.2 Exploratory Data Analysis of categorical features

We used mosaic plots to study the relationship between the target variable and the categorical features. Mosaic plots help in visualizing the statistical association between two variables. The surfaces of the rectangular fields that are available for a combination of features are proportional to the number of observations that have this combination of features. Here the width along the x axis is basically proportional to the percentage of distribution of the categorical feature, the length along y axis is proportional to percentage diagnosed with the disease.

After understanding the plot, we can make inferences such as how gender plays a role in resulting in heart disease. We notice that males are more likely to be diagnosed with heart diseases. Other categorical features have with increased prevalence of heart disease are asymptomatic angina chest pain (relative to typical angina chest pain, atypical angina pain, or non-angina pain), Presence of exercise induced angina, Lower fasting blood sugar, Flat or down-sloping peak exercise ST segment, Male, Higher thalassemia score.

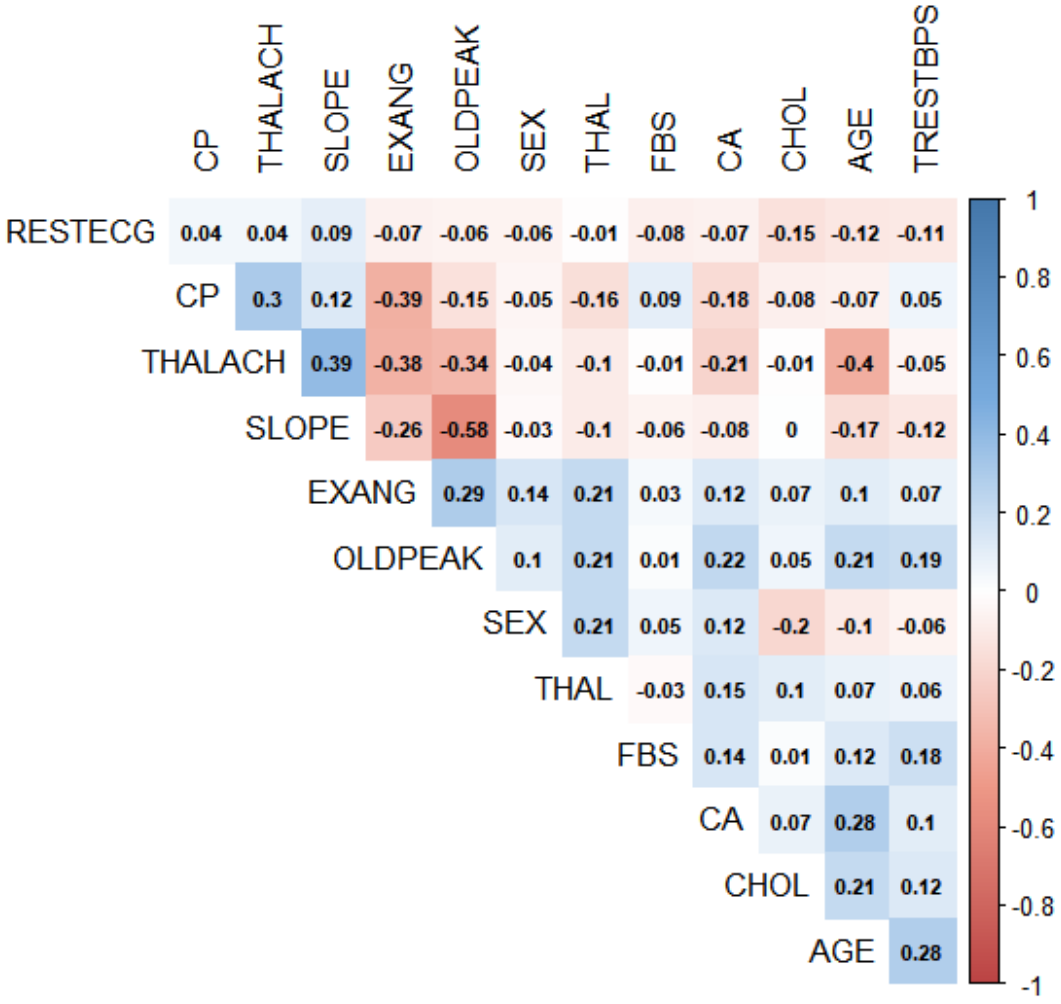
Figure 4.3: Categorical predictors and heart disease diagnosis





We further explored the correlation of all features as shown in the plot below. From the correlation matrix, we find that some features are strongly correlated, in this way, we do not need to put all variables into the prediction models.

Figure 4.4: Correlation matrix for both categorical & numerical features



5. Model Selection

Model selection is a process of selecting a model from a set of candidate models. In this step, we contrast various candidate models based on variety of criterion and select a model which best represents the true pattern in the data. Model selection also involves variable selection in which the candidate models consist of different combination of variables from the total variables set.

The best approach is to divide the dataset in a training and a testing set and fit a model using training dataset. A good model should perform well with the training dataset. In practice, a validation dataset is also used to determine which variables to be added in the model.

5.1 Elementary Model Selection Methods

A basic method that can be used for variable selection is by creating a logit model between the response variable and each of the predictor variable individually and checking if coefficient of that predictor variable is significantly different from 0. At significant level = 5%, we drop the predictor variable which has a p value of less than 5% in the individual model. This method suggests marginal association of CHOL, FBS, RESTECG is significantly very low, thus they can be dropped to make a simpler model with 9 predictor variables.

We also conducted a drop1 test to find with variables can be dropped considering significance level = 5%. The drop1 test suggests we can drop AGE, TRESTBPS, CHOL, FBS, RESTECG, THALACH, SLOPE to make a simpler model with only 7 predictor variables.

5.2 Model Selection Methodology

In general, when we have a lot of predictor variables, it becomes difficult to find a good model. If there are p variables including interactions, total 2^p models can be constructed. Thus, if p is large total number of candidate models will be very high.

To implement model selection from 2^p candidate models, we need:

- A criteria to access the models
- A search strategy

5.2.1 Model Assessment Criterion

1. Sum Square Error of the predictions on the testing dataset (Continuous response variable)
2. Mean Error of the predictions on the testing dataset (Discrete response variable)
3. $R^2_{adjusted}$
4. Mallow's Cp
5. Information Criteria: AIC or BIC
6. Cross Validation: LOOCV or GCV

5.2.2 Search Strategies

5.2.2.1 Best Subset Algorithms

Generally, there are two main reasons for using Subset Selection Technique. We improve prediction accuracy by removing some predictors and achieve a balance between bias and variance. Also, we get proper interpretation of the big picture by sacrificing some small details. Best Subset Model Selection compares all possible models using a specified set of predictors and displays the best-fitting models that contain one predictor, two predictors, and so on. The result is a number of models and their summary statistics. Out of these models, we select one best performing model. Sometimes the results do not point to one best model and manual judgement is required. Best subset provides more information by including more models, but it can be more complex to choose one. Because Best Subset assesses all possible models, large models may take a long time to process.

There is total 13 predictor variables in our dataset. To use best subset strategy, we will have to search through at least 2^{13} = 8192 candidate model. And if interaction between variables is considered, this number will shoot up very high. Considering the computational complexity of this strategy, we choose not to use it directly.

5.2.2.2 Stepwise Search

Stepwise Search is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. The result of this process is a single regression model, which makes it nice and simple. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Stepwise does not assess all models but constructs a model by adding or removing one predictor at a time.

We can use model assessment criteria like Information Criteria (AIC) and Mallows Cp to evaluate the model. These are an unbiased estimate of the model prediction error MSE. The lower these criteria, the better the model.

5.3 Model Selection Methods

5.3.1 Best Subset Algorithm using Mallow's Cp

To use the best subset algorithm, we will have to search through very high number of candidate models, which will get computationally complex. So, we used leaps() library in R which performs an exhaustive search for the best subsets of the variables for predicting the response, using an efficient branch-and-bound algorithm.

This strategy using leaps() gives Mallow's Cp values of the candidate models. The model with the minimum Cp value is selected as the best model. The Model selected with this strategy included 9 predictor variables. The predictor variables are dropped by this strategy are:- AGE, CHOL, FBS, RESTEG

5.3.2 Stepwise Search using AIC

The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model. To implement the search strategy, we first created a full model, *logitmodel* using all the parameters. Using Step() function from the MASS library in R, we implemented the search strategy with AIC criteria. The model selected by this strategy included 9 variables. The predictor variables are dropped by this strategy are:- AGE, CHOL, FBS, RESTEG

5.3.3 Penalized Least Square methods

When two or more predictor variables are highly co-related to one another, there will be a multi-collinearity in the model. When point estimations and predictions are drawn from the model having high collinearity, results achieved will not be good in the sense of larger standard error, weaker power in hypothesis testing, and wider confidence and prediction intervals.

Shrinkage reduction is one of the methods to overcome the collinearity. Shrinkage is the reduction in the effects of sampling variation. Shrinkage methods are based on subtracting a penalty from the risk, and the penalty is a function of a decay parameter(λ). A shrinkage method solves the following optimisation problem.

$$\min_{\beta} \left|\left|Y - X\beta\right|\right| + \lambda \sum_{j=1}^{p} J(|\beta_j|)$$

Where $J(|\beta_j|)$ = penalty function and λ ≥ 0 is the decay parameter.

The larger the penalising term, the more our model will be simple and inverted. The "shrinkage penalty" becomes small when the coefficients of our estimators are close to 0.

Two special cases of the shrinkage methods are (i) Ridge Regression (ii) LASSO (Least Absolute Shrinkage and Selection Operator).

5.3.3.1 Ridge Regression

Ridge regression improves prediction error by shrinking the sum of the squares of the regression coefficients to be less than a fixed value. Hence,

The Ridge Regression estimator is $\hat{\beta}^{ridge} = \min_{\beta} ||Y - X\beta|| + \lambda \sum_{j=1}^{p} {\beta_j}^2$

For our data, first the decay parameter was calculated using cross validation. We used the cv.glmnet and plot functions from “glmnet” and “plotmo” libraries of R. log(λ) observed to be -3.81 which gives the λ values of 0.022. We used glmnet to fit the ridge regression using alpha = 0 and nfold = 5. The plot obtained showed that parameters such as thalch, chol, trestbps, age, restecg, FBS can be dropped off the model.

Figure 5.1: Binomial deviance v/s Log(λ) to find the best λ

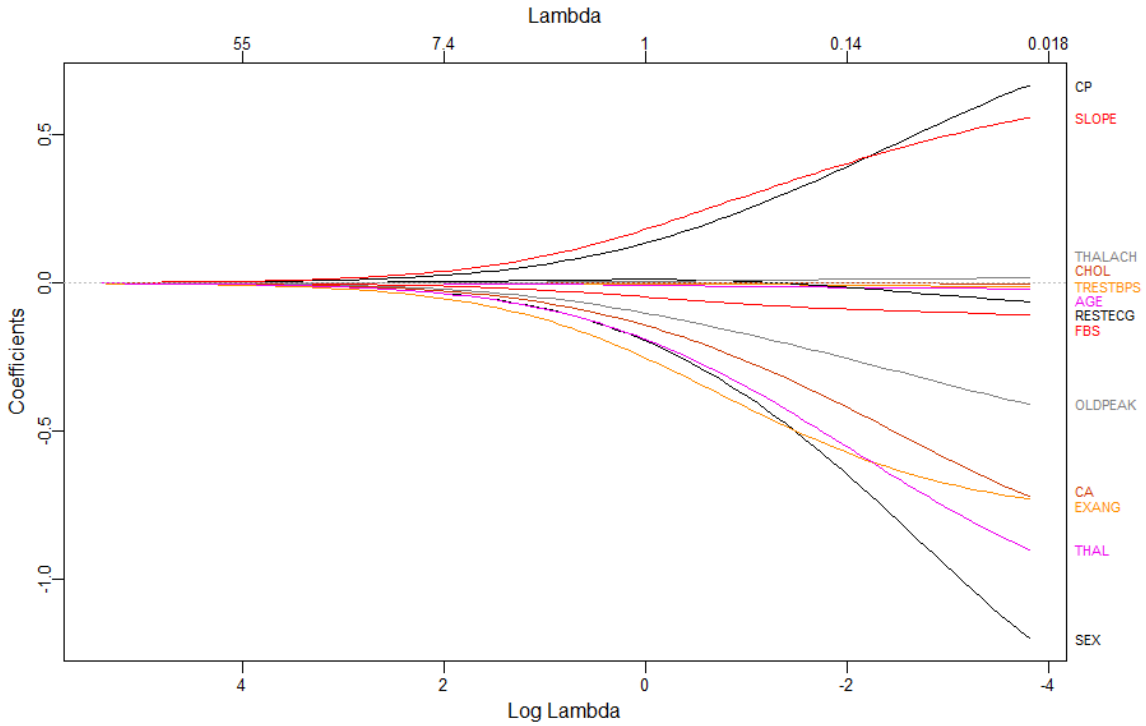
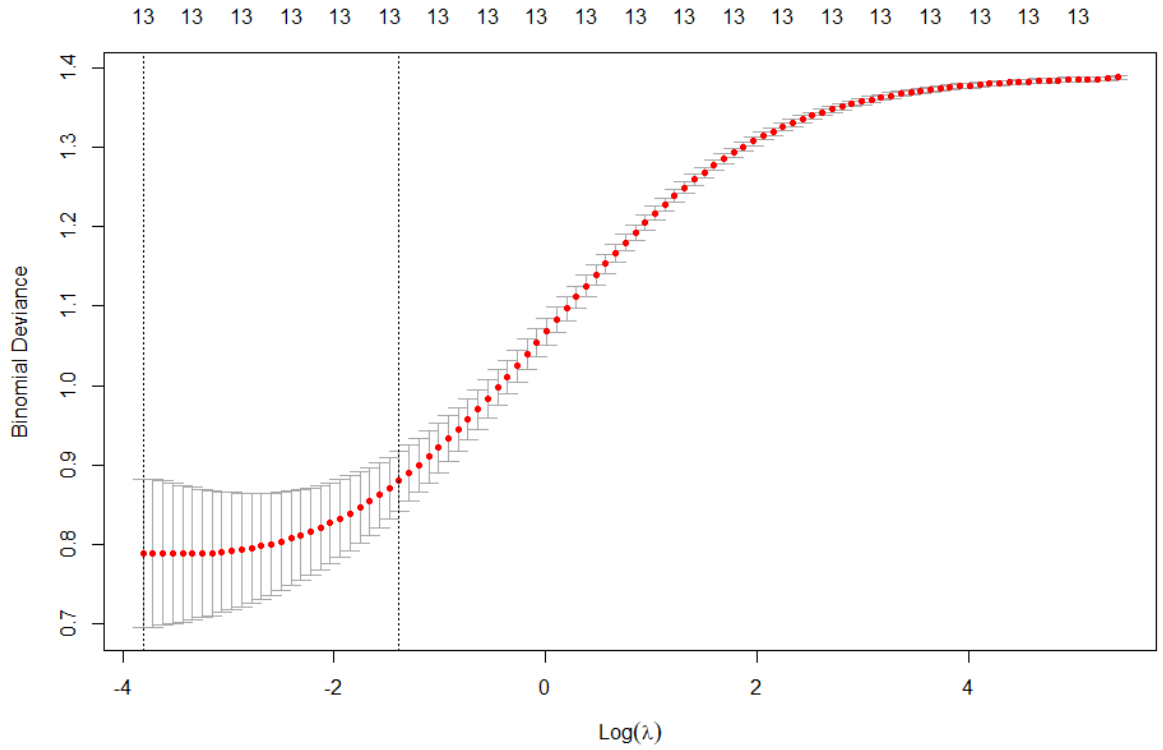


Figure 5.2: Coefficient's v/s λ and Log(λ) for selection of the parameters from Ridge classifier fit



5.3.3.2 LASSO (Least Absolute Selection and Shrinkage Operator)

LASSO achieves the reduction in prediction error by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value.

The big difference with the Ridge estimator is that with LASSO, the value of the coefficients may be so penalized that it can be zero. This will therefore act as a means of selecting the variables of our model. Ridge regression does not perform covariate selection and therefore does not help to make the model more interpretable. Lasso achieves this goal by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to zero, effectively excluding them.

LASSO estimator is $\hat{\beta}^{lasso} = \min_{\beta} ||Y - X\beta|| + \lambda \sum_{j=1}^{p} |\beta_j|$

The essential difference can be notices that absolute value function is applies in case of LASSO compared to square function in case of Ridge regression for the estimators β_j's.

Just like ridge regression, we calculated decay parameter λ using cross validation. log(λ) observed to be -4.2 which gives the λ values of 0.01498. We used glmnet to fit the lasso but using alpha = 1 (instead of 0) and nfold = 5. The plot obtained showed that parameters such as thalch, chol, trestbps, age, restecg, FBS can be dropped off the model. Here, the decision obtained for dropping the predictor variable is quite same as obtained from ridge regression.

Figure 5.3: Binomial deviance v/s Log(λ) to find the best λ

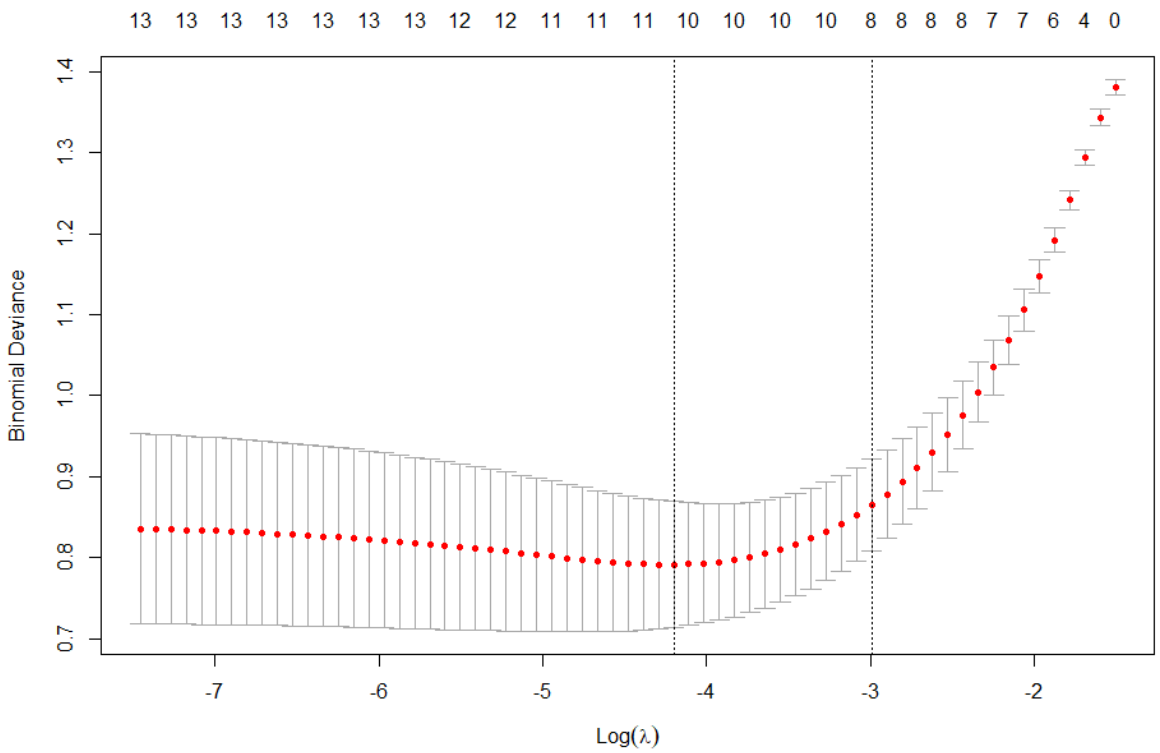
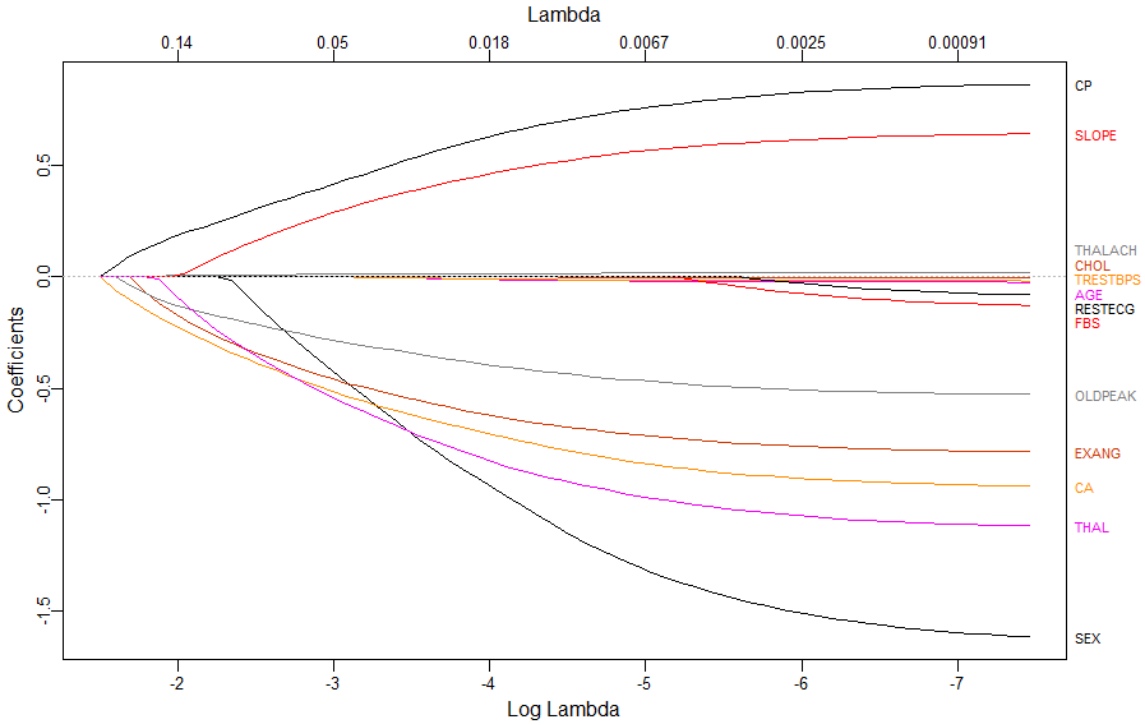


Figure 5.4: Coefficient's v/s λ and Log(λ) for selection of the parameters from LASSO classifier fit



5.3.4 Comparison of Different Models

We use AIC criterion to compare the result of different methods of model selection. Following table compares AIC values of the models obtained from various model selection methods.

Model Selection Technique	**AIC value**
Full Model	181.94
Stepwise Regression	175.93
Ridge Regression	180.13
LASSO	180.13

The model obtained from stepwise regression includes 9 variables out of 13, except AGE, CHOL, FBS, RESTEG. We will consider the 9 remaining parameters for model formulation and prepare models using various techniques and check the prediction accuracy of these models.

6. Model Fitting:

Based on the model's selection methodologies used in the previous session, we selected 2 models to predict heart disease possibility based on given predictor variables. The selected models are:

- **Model 1:** Full model consisting of 13 predictor variables
- **Model 2:** A smaller model with 9 predictor variables (AGE, CHOL, FBS, RESTEG dropped)

We divided the input dataset in a training and a testing dataset. As a commonly followed practice, we randomly placed 25% observations in testing dataset (*datatest*) and remaining 75% observations in training dataset (*datatrain*). Training dataset is used to fit the parameters of the model whereas testing dataset is used to assess quality of the model.

6.1 logistic Regression:

6.1.1 Model 1

```
logitmodel1<-lm(TARGET~.,family=binomial(link=logit),data=datatrain)
testpred1 <- predict(logitmodel1, datatest, type="response")
Ytrue <- t(datatest[14])
```

The prediction testpred1 generated by the logitmodel1 consists of the probability(response = 1), but the actual response (*Ytrue*) associated with the testing dataset is a Binary variable. Thus, we need to convert the predicted probability values to a binary prediction. This is achieved by setting a cutoff value for the probability. If the probability is above the cutoff value, predicted response = 1, otherwise predicted response = 0.

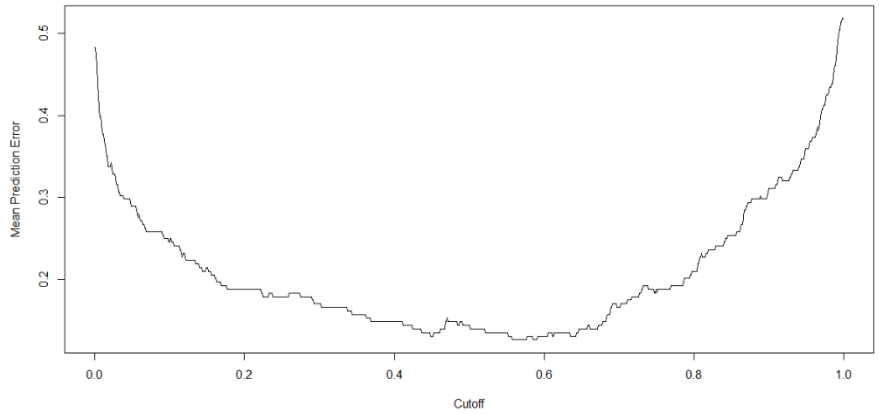
To select a cutoff value, following methods are used:

1) Central cutoff value: Setting cutoff value = 0.5
This is a commonly used method. This method is not very accurate as it ignores the actual distribution of the response variable in the observed data.

2) Cutoff value as the proportion of Y = 1 in the training data:
This method is better than the first method as it sets a realistic cutoff on probability of Y being 1 based on the training data.

3) Best cutoff value from the training data:
In this method, we vary the cutoff value between (0,1) in discrete steps of 0.001 and for each value measure the mean prediction error on the training dataset. The cutoff value which minimizes the mean prediction error is finalized. The variation in the mean prediction error on training dataset with the cutoff value is represented in the graph below.

Fig. 6.1 Variation in mean prediction error v/s cut-off value to determine the best cut-off value

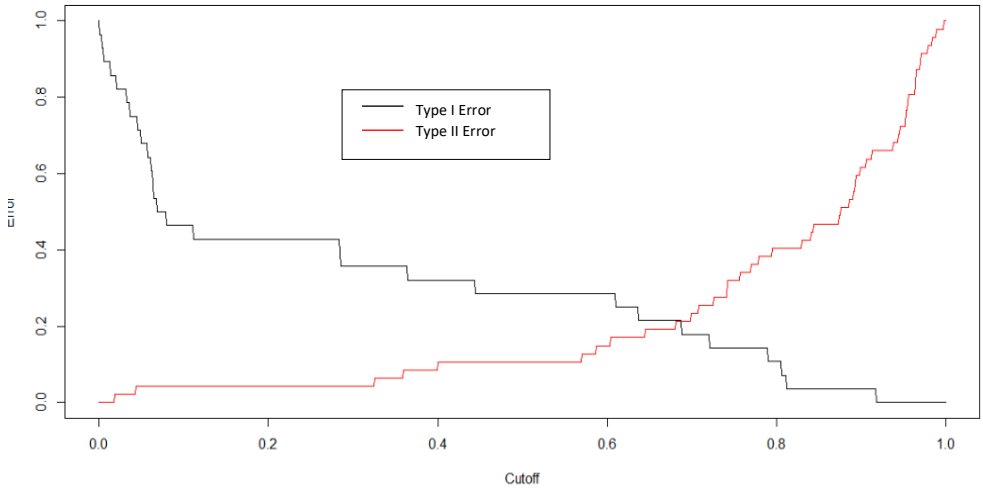


4) Using a validation data to choose the cutoff value:

In this method, we treat the test data as a validation data. A penalty of "1" is assigned for mis-classifying *Ytrue* = 0 and a penalty of "2.5" is assigned for mis-classifying *Ytrue* = 1. We assign a larger penalty on mis-classifying *Ytrue* = 1, because we want to ensure that if a person actually has a heart disease, he must not be classified as healthy. Thus, we try to minimize the Type II error. In an effort to reduce the Type II error, Type I error gets increased, but we are more concerned about minimizing the misclassification of actually affected people.

Variation in the Type I and Type II error with the Cutoff value is demonstrated in the figure below. As the cutoff value is increased, Type I error reduces but Type II error increases. We find the best cutoff value by assigning penalties on the two errors and selecting the cutoff value which gives the minimum weighted error.

Fig. 6.2 Variation in type-I and type-II error v/s cut-off value to determine the best cut-off Value



6.1.2 Model 2

```
logitmodel2 <- step(logitmodel1)
testpred2 <- predict(logitmodel2, datatest, type="response")
```

The main reason behind using a smaller model is it reduces computational complexity by dropping insignificant variables. To find the probability cutoff value for the predicted response variable, same methodology as that of Model 1 is used.

The results obtained for the two models using various cutoff values are tabulated below:

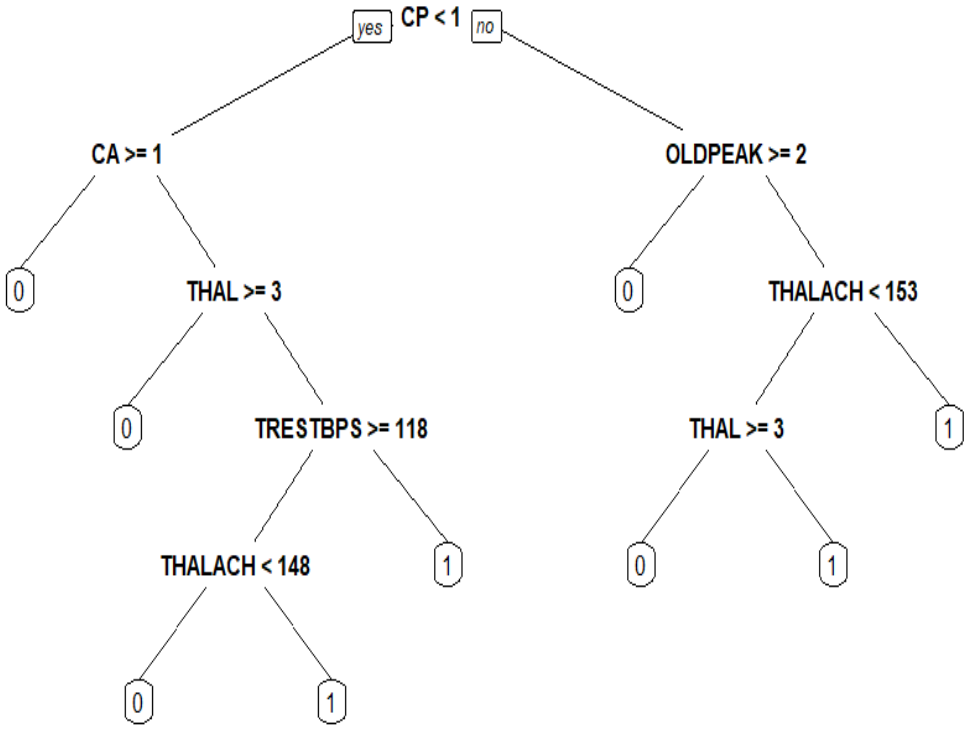
Model	Cutoff Selection Method	Cutoff Value	Mean Error	Type II Error
Model 1	Central Cutoff value	0.5	0.1733	0.0667
	Proportion of Y=1 in the training data	0.5175	0.1733	0.0667
	Best Cutoff from the training data	0.557	0.1733	0.0667
	Cutoff value using a validation data	0.285	0.16	0.0267
Model 2	Central Cutoff value	0.5	0.1733	0.0667
	Proportion of Y=1 in the training data	0.5175	0.1733	0.0667
	Best Cutoff from the training data	0.514	0.1733	0.0667
	Cutoff value using a validation data	0.38	0.1733	0.0533

Based on the results in the above table, we conclude that the best logistic regression model is Model 1 with probability cutoff value = 0.285, selected using a validation dataset.

6.2 Decision Tree:

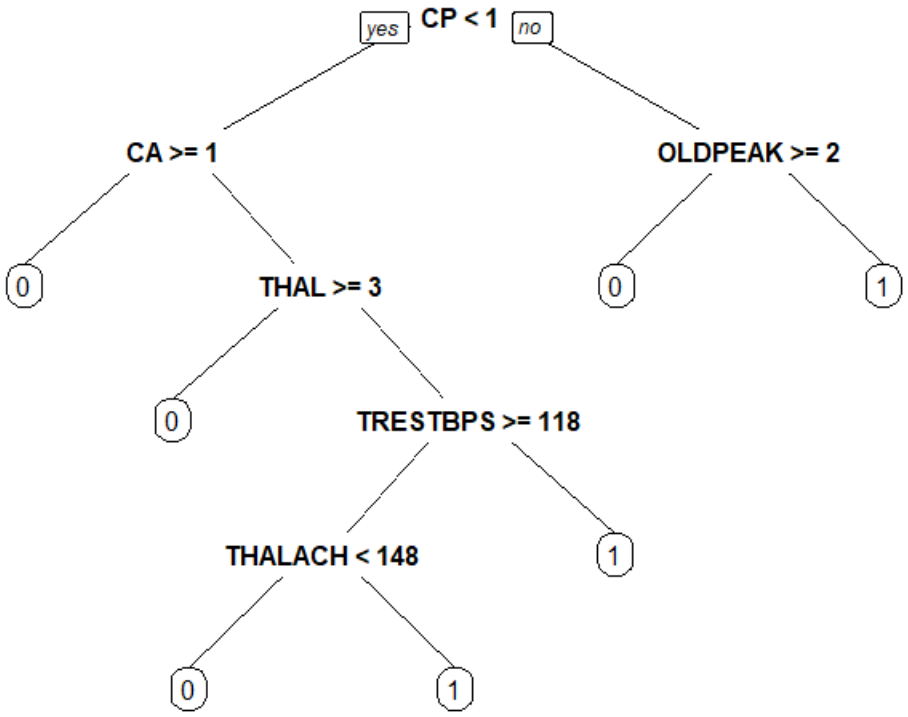
Decision Trees are versatile Machine Learning algorithm that can perform both classification and regression tasks. They are very powerful algorithms, capable of fitting complex datasets. It is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. The main benefit of decision tree is that they are simple to understand and interpret. We used "rpart" package of the R. The decision tree and confusion matrix obtained are show below. It can be seen that prediction error achieved is 18.67% for Model 1 and 21.33% for Model 2.

Fig. 6.3 Prediction tree for Model 1



Confusion matrix for Prediction Tree		
	0	1
0	21	7
1	9	38

Fig. 6.4 Prediction tree for Model 2



Confusion matrix for Prediction Tree		
	0	**1**
0	19	9
1	5	42

6.3 K nearest neighbors (KNN):

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point and predict the label from these.

Class and gmodels packages of R were used to implement KNN algorithm for our data. 61.4% of accuracy is achieved with k=4 for model-1 and 66.7% of accuracy is achieved with k=15 for Model 2.

For Model 1:

```
Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  75


                 | pred5
datatest$TARGET |         0 |         1 | Row Total |
----------------|-----------|-----------|-----------|
              0 |        17 |        11 |        28 |
                |     0.607 |     0.393 |     0.373 |
                |     0.531 |     0.256 |           |
                |     0.227 |     0.147 |           |
----------------|-----------|-----------|-----------|
              1 |        15 |        32 |        47 |
                |     0.319 |     0.681 |     0.627 |
                |     0.469 |     0.744 |           |
                |     0.200 |     0.427 |           |
----------------|-----------|-----------|-----------|
   Column Total |        32 |        43 |        75 |
                |     0.427 |     0.573 |           |
----------------|-----------|-----------|-----------|
```

For Model 2:

```
Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  75


                 | predred4
datatest$TARGET |         0 |         1 | Row Total |
----------------|-----------|-----------|-----------|
              0 |        16 |        12 |        28 |
                |     0.571 |     0.429 |     0.373 |
                |     0.552 |     0.261 |           |
                |     0.213 |     0.160 |           |
----------------|-----------|-----------|-----------|
              1 |        13 |        34 |        47 |
                |     0.277 |     0.723 |     0.627 |
                |     0.448 |     0.739 |           |
                |     0.173 |     0.453 |           |
----------------|-----------|-----------|-----------|
   Column Total |        29 |        46 |        75 |
                |     0.387 |     0.613 |           |
----------------|-----------|-----------|-----------|
```

7. Model Fitting: Results Summary

We compared the 4 model fitting methods based on Mean Error and Type II error. It was found that Model 1 with logistic regression yields minimum value for both the Mean and Type II error. We also tried few machine learning methodologies like decision tree and K-nearest neighbor. The decision tree method provided descent results, but the K-nearest neighbor method provided slightly higher values for the Mean and Type II error. Overall, we recommend using Model 1 with logistic regression to achieve the best results.

The results for the fitted models are tabulated below:

Model	Model fitting method	Mean Error	Type II Error
Model 1	Logistic Regression	0.16	0.0267
	Decision Tree	0.187	0.067
	K-nearest neighbor with k = 4	0.347	0.2
Model 2	Logistic Regression	0.1733	0.0533
	Decision Tree	0.213	0.12
	K-nearest neighbor with k = 15	0.333	0.2

8. Conclusion and Future Directions

8.1 Statistical Conclusions

We performed a thorough data analysis and statistical modelling to determine features which can best predict the diagnosis of heart disease. First, we performed exploratory data analysis to assess all features of our dataset. Clearly, based on the p-value of the t-test, we found that patients with heart disease show significant difference in age, resting blood pressure, thalach, number of major vessels and oldpeak (ST depression induced by exercise) compared to controls. Next, we found that some features were positively correlated, and some were negative correlated to each other. However, the correlation values were moderate and not very high.

We carried out model selection using best subset algorithm, stepwise search and penalized least square methods and found that model with 9 predictor variables sex, cp, trestbps, thalach, exang, oldpeak, slope, ca, thal having best AIC value of 175.93.

We performed logistic regression, decision tree and k nearest neighbour methods on both Model 1 and Model 2 to fit our data. We compared mean error and type II error for all models and found that logistic regression performed on Model 1 provides the best result with mean error of 16% and type II error of 2.67%.

8.2 Clinical and Biological Conclusions

Following features are the best predictors of heart disease:

a) Thalassemia (thal), which is a genetic blood disorder of having less hemoglobin and can be determined if a patient is anemic if he/she experiences weakness, fatigue or has pale-yellowish skin.
b) Based on gender, men are found to be at higher risk at an average age of 50.
c) Chest pain is the first symptom that can be easily diagnosed in routine check-up.
d) Maximum heart rate can be determined with blood pressure measurements.
e) If a greater number of major vessels are present with more than 50 % narrowing, this can lead to heart attack.
f) Abnormality in ECG is one of the first tests performed if a patient is suspected to have heart disease. ECG indicates about oldpeak and slope.

Therefore, if a person comes for a routine check up to the doctor, these basic clinical measurements can be taken, and the patient can be indicated about his or her risk of getting

the heart disease based on our statistical predictive models. Most of our selected features include **basic clinical measurements** which can be taken in a **routine visit to the doctor**. Our models can serve as an **initial screening test for risk prediction of heart disease** and can be provided to hospitals or general physicians. With this information, both the doctor and the patient can be benefitted. The doctor can start the treatment well in time and the patient can make necessary lifestyle modifications well in time before the disease worsens. Risk assessment from our models will also help asymptomatic patients to take necessary preventive action. These **preventive measures** can help to prevent several deaths due to heart disease.

8.3 Future Directions

For future directions, based on the literature, it is also very important that data about patient demographics must be available. Studies have shown that individuals from different ethnic groups have different risk heart disease. Genetic studies have shown that risk of heart disease is higher in African Americans than Europeans or Asians. So, having this kind of data is useful to assign priors to our models based on the ethnicity of patients. Also, testing more interaction terms, such as interactions between these clinical measurements, heart disease and environment must be included in the model to assess all possibilities. Gene environment interaction studies have been getting a lot of attention. One way to do this is to include interaction terms. The other way is to stratify the patients based on their environment and then fit the model for each group and compare their differences. For example, create 4 groups of individuals as underweight (BMI < 20), normal (BMI 20-25), overweight (BMI > 25) and obese (BMI > 30) and for each group, fit the models separately and compare their outcomes. Next, non-linear models can be tested to see if the prediction models and prediction accuracy can be improved. Finally, with studies like these, ethical concerns must be kept in mind. Patient consent is highly important, which means risk prediction assessment and any information disclosure from the study or analyses is only done with complete consent from the patient. Our findings provide a basis for predicting the diagnosis of heart disease based on simple clinical measurement and this can be further improved upon following the future directions, with larger datasets and further polishing the predictive models.

9. Learning outcomes from the project:

- We express our gratitude towards Dr. Mei for providing us this opportunity of working on an independent project work.
- We learned many new concepts of Regression Analysis in this course and the project helped us in implementing those concepts on a real dataset, hence we were able to visualize the concepts properly. The examples from class and assignments were helpful to think about the problem with perspective of data analysis.
- We also tried some of the new technics like decision tree and K nearest neighbors and studies some literature on our own. We tried some of the materials from the optional topics like LASSO and ridge regression.
- We also learned the importance of working in a team and co-operation. We get to express and explain our idea to other members and have a healthy discussion on it. Everyone in our team was from diverse backgrounds and hence, we were able to know the concepts of other fields in a nice way by doing this project and interacting.