

# Identification of Heart Disease from Common Indicators

Group 6

Megan Petralia, Chris Kujawa, Seth Hanson, Ryan Pykor

12/17/2020

## Abstract

Heart disease describes a range of conditions that affect the heart including blood vessel diseases, heart rhythm problems, and congenital heart defects, among others. Identification and especially early detection of heart disease is important for improving patient outcome. Using a dataset titled "Heart Disease UCI" Naive Bayes classifiers were developed using three subsets of the available predictors for identification of heart disease from common indicators.

Under the fullest model utilizing all predictors there was an observed AUC of 0.8939 and accuracy of 84.3%. Under a model utilizing only predictors which may be obtained from a regular physical with bloodwork there was an observed AUC of 0.8374 and accuracy of 73.0%. The most important predictors were found to be:

- Cp – Chest Paint
- Thal -Stress test blood flow observed w/Thallium
- Thalach - Maximum Heart Rate Achieved

The development of these models shows that prediction – including early prediction - of heart disease is possible from common indicators. Further investigation including medical professionals should be performed to determine the impact of misclassification when using models such as the types explored.

## Contents

Background & Project Goals .....	3
Dataset Investigation.....	4
Predictor Descriptions and Subsets.....	4
Correlation Matrix.....	5
Predictor Value Distributions.....	6
Model Development and Selection.....	8
Winning Models .....	8
Results .....	9
1. Can a model of $AUC \geq 0.90$ be created to predict heart disease? .....	9
1. Can a subset of the predictors be used to create a model of $AUC \geq 0.80$ ? .....	10
2. What are the most significant factors when predicting heart disease? .....	12
Conclusions .....	13
Appendix A– Detailed Model Results .....	14
Appendix B – All Data Investigation Results .....	18
Appendix C - Full Code .....	24
KNN Model.....	27
Decision Tree .....	31
Naive Bayes.....	39
Logistic Models.....	42

## Background & Project Goals

Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others (1). In 2019, there were 523 million cases of heart disease, a near doubling of the 271 million cases in 1990 (2). Over the same period, heart disease deaths have climbed from 12.1 million to 18.6 million deaths, now accounting for one-third of worldwide deaths (2). In addition to being the leading causes of death, heart disease is a major cause of disability and rising health care costs (2). It is estimated that by 2030, annual medical costs associated with heart disease are projected to rise to greater than \$818 billion (3). To combat these problems, identification and early detection of heart disease are key, as they have shown to improve patient outcome and reduce associated medical costs.

The dataset titled “Heart Disease UCI”, sourced from Kaggle.com’s library of datasets, contains 13 predictors and a single response indicating the presence of heart disease or not for 303 observations (patients). Using this dataset, the team sought out answers to the three following questions:

- 1. Can a model of AUC  $\geq 0.90$  be created to predict heart disease?**
- 2. Can a subset of the predictors be used to create a model of AUC  $\geq 0.80$ ?**
- 3. What are the most significant factors when predicting heart disease?**

To answer these questions, the team investigated the dataset to understand its features, their correlation with the response, and their value distributions. Upon full understanding of the dataset, multiple variants of classification models were created to predict the presence of heart disease, given all or a subset of the features. Finally, the various models were evaluated using distinguishing performance metrics the highest performing model was identified. This process is detailed in the following sections.

## Dataset Investigation

Data used for this analysis can be found on Kaggle or the UCI Machine Learning repository.

<https://www.kaggle.com/ronitf/heart-disease-uci>

### Predictor Descriptions and Subsets

Table 1: Predictors Described

Name	Predictor	Type	Values
age	1. Age	Continuous	
sex	2. Sex	Categorical	[0 = female , 1 = male]
cp	3. Chest Pain Type	Categorical	[0 = asymptomatic, 1 = atypical angina, 2 = pain without relation to angina, 3 = typical angina]
trestbps	4. Resting Blood Pressure (mm Hg)	Continuous	
chol	5. Serum Cholesterol (mg/dl)	Continuous	
fbs	6. Fasting Blood Sugar (mg/ml)	Categorical	[0 = <=120 mg/dl , 1 = >120 mg/dl]
restecg	7. Resting Electrocardiographic Results	Categorical	[0 = probable left ventricular hypertrophy, 1 = normal, 2 = abnormalities in T wave or ST segment]
thalach	8. Maximum Heart Rate Achieved	Continuous	
exang	9. Exercise Induced Angina	Categorical	[0 = no, 1 = yes]
oldpeak	10. Decrease of ST Segment During Exercise	Continuous	
slope	11. Slope of Peak Exercise ST Segment	Categorical	[0 = descending, 1 = flat, 2 = ascending]
ca	12. # Major Vessels colored by Fluoroscopy	Categorical	Low value and range integer
thal	13. Stress test blood flow observed w/Thallium	Categorical	[1 = fixed defect, 2 = normal, 3 = reversible defect]
Name	Response	Type	Values
target	1. Target; Heart Disease	Categorical	[0 = heart disease, 1 = no heart disease]

Table 1 describes the predictors in detail, including types and categorical value meanings. To answer our second question (*“Can a subset of the predictors be used to create a model of  $AUC \geq 0.80$ ?”*), the predictors were split into three subsets (See Table 2) based on the magnitude of medical testing required to measure them:

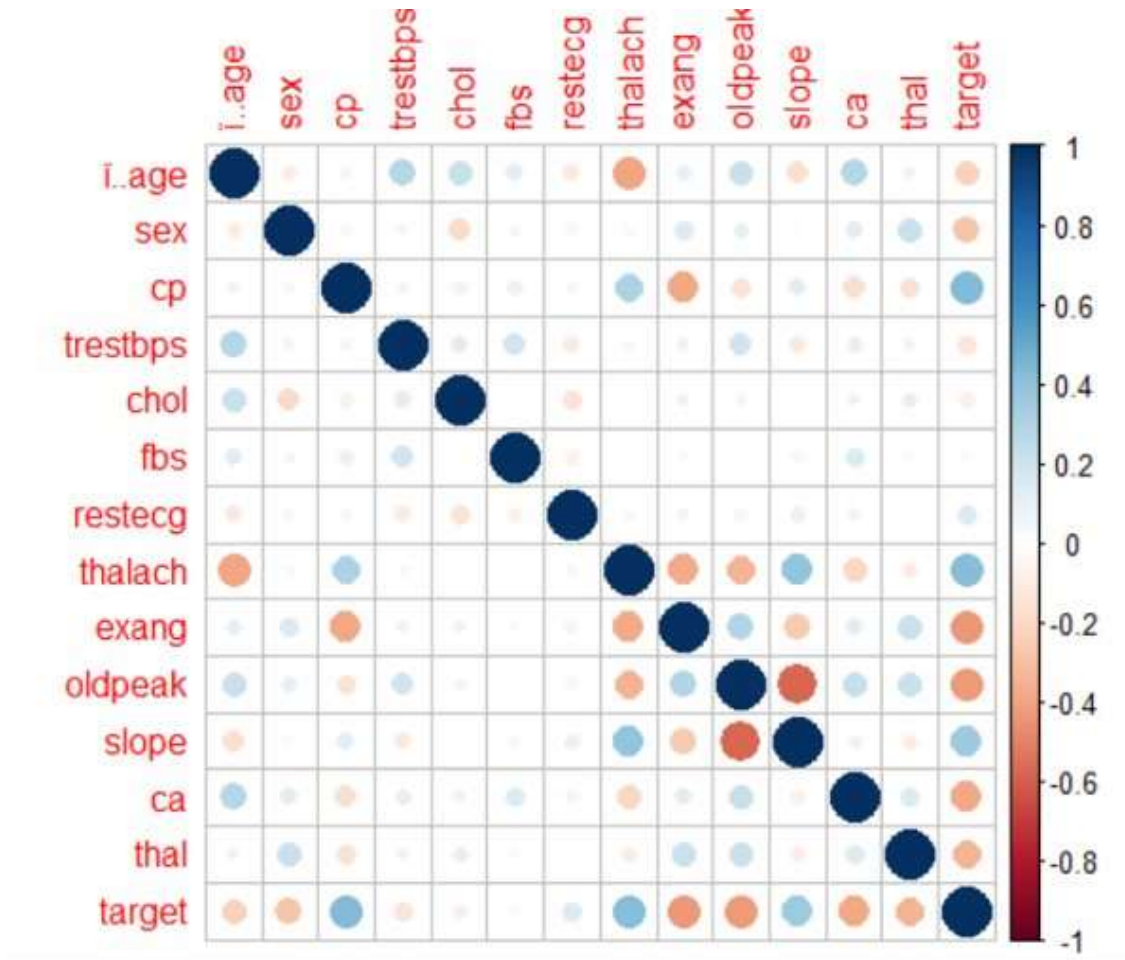
- **Early Warning**– Predictors (1-6) are innate or measurable at a routine annual physical and/or by bloodwork.
- **Classification** - Predictors (1-11) are a combination of the Early Warning subset and predictors measured by a stress test and/or electrocardiograph. This set requires looking for heart disease specifically.
- **Confirmation** - Predictors (1-13) encompasses all the available predictors. This subset includes additional predictors that require time consuming, invasive, and/or expensive radiological tests.

Table 2 – Predictor Subsets

Predictor
1. Age
2. Sex
3. Chest Pain Type
4. Resting Blood Pressure (mm Hg)
5. Serum Cholesterol (mg/dl)
6. Fasting Blood Sugar (mg/ml)
7. Resting Electrocardiographic Results
8. Maximum Heart Rate Achieved
9. Exercise Induced Angina
10. Decrease of ST Segment During Exercise
11. Slope of Peak Exercise ST Segment
12. # Major Vessels colored by Fluoroscopy
13. Stress test blood flow observed w/Thallium

## Correlation Matrix

Figure 1: Correlation Matrix



The correlation matrix in Figure 1 shows the relationships between all the features in the dataset. Correlation closer to -1 and 1 indicated a strong negative or positive correlation while, correlation near 0 indicates little to no relationship.

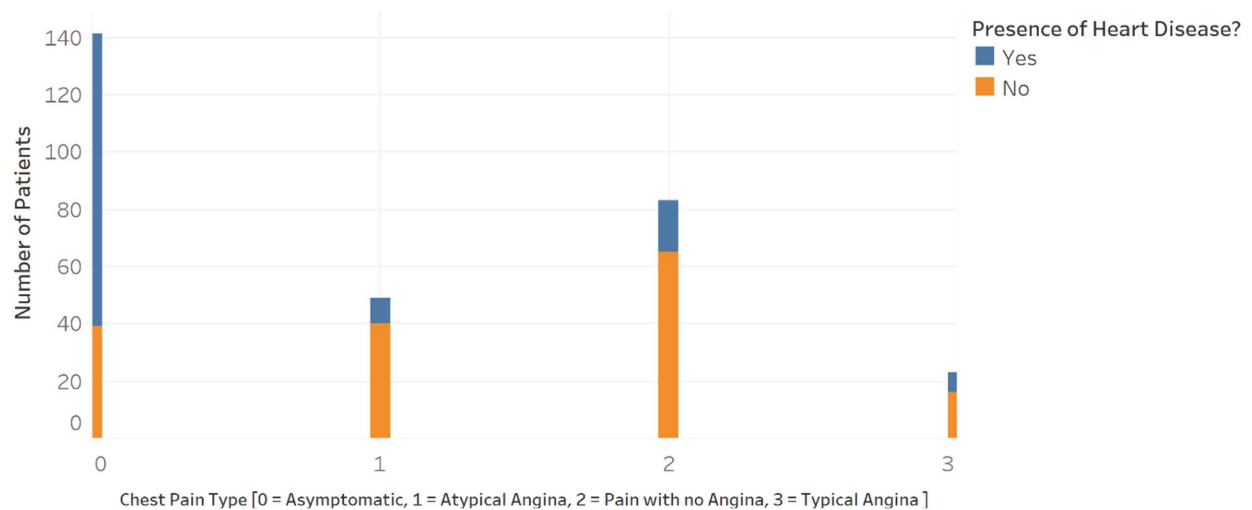
### Noteworthy Correlation Matrix Findings

- It is unexpected that chol (Cholesterol) has such a weak correlation with the response.
  - Value: [-0.0852]
- It is interesting that cp (chest pain type) has a relatively strong positive correlation with the response.
  - Value: [0.4338]
  - Target = 0 when heart disease is present and cp = 0 when asymptomatic, indicating that chest pain is not prevalent among patients with heart disease

## Predictor Value Distributions

Figure 2: Chest Pain vs. Presence of Heart Disease

Number of Patients With/Without Heart Disease by Chest Pain Type

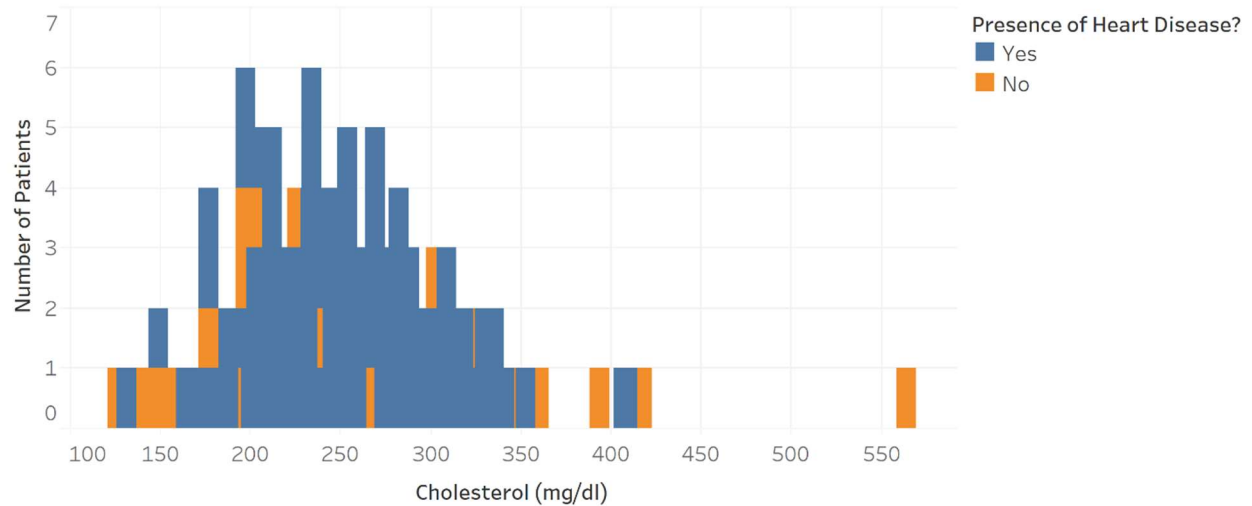


Observing Figure 2, the following conclusions can be made:

- Heart disease can, and does, present without chest pain
- 75% of patients with heart disease reported no chest pain
- 40% of patients without heart disease reported suffering from chest pain but were not diagnosed with angina (a type of chest pain caused by reduced blood flow to the heart).

Figure 3: Cholesterol vs. Presence of Heart Disease

Number of Patients With/Without Heart Disease by Cholesterol



Using Figure 3 to take a closer look at the distribution of cholesterol in patients:

- 154 patients - 'high' cholesterol levels ( $\geq 240$ ) (4)
  - 78 have heart disease
  - 76 do not have heart disease
- 48 patients - 'good' cholesterol levels ( $< 200$ ) (4)
  - 20 have heart disease
  - 28 do not have heart disease
- 94 patients - 'borderline to moderately elevated' cholesterol levels (200–239) (4)
  - 38 have heart disease
  - 56 do not have heart disease



## Model Development and Selection

To produce models for comparison and selection to answer our questions, a parallel work stream method was applied, allowing each of the four members to analyze the data using a specific analytical technique:

- Logistic Regression – Seth
- Decision Tree – Chris
- K-Nearest Neighbor – Megan
- Naïve Bayes – Ryan

For each of the subsets previously identified (Early Warning, Classification, and Confirmation) each team member developed a model of their assigned type using a 70%/30% Training/Validation split of the data set. These models were then evaluated using the following metrics, in order of precedence in ranking:

1. AUC – as obtained from the ROC Curve
2. Accuracy – (#Correct Classifications/ #Total Classifications) as obtained from testing the model against the validation set

The results from model development are detailed in the table below:

Subset		Early Alarm Subset Models				Classification Subset Models				Confirmation Subset Models			
Model		Logistic	KNN	Naïve Bayes	Decision Tree	Logistic	KNN	Naïve Bayes	Decision Tree	Logistic	KNN	Naïve Bayes	Decision Tree
Threshold		0.65	k= 15	N/A	N/A	0.5	k = 20	N/A	N/A	0.55	k = 25	N/A	N/A
Threshold Selection		Manual	Manual	N/A	N/A	Manual	Manual	N/A	N/A	Manual	Manual	N/A	N/A
AUC		0.7542	0.7143	0.8374	0.7684	0.8087	0.6930	0.8449	0.7308	0.8884	0.6895	0.8939	0.8485
Accuracy		71.9%	65.2%	73.0%	74.2%	78.7%	62.9%	76.4%	71.9%	86.5%	62.8%	84.3%	80.9%
Training Set Ratio		70%	70%	70%	70%	70%	70%	70%	70%	70%	70%	70%	70%
Predictors	Selection Process	Backward	N/A	N/A	rpart	Backward	N/A	N/A	rpart	Backward	N/A	N/A	rpart
	1. age	yes	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes
	2. sex	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	3. cp	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	4. trestbps	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no
	5. chol	no	yes	yes	yes	no	yes	yes	yes	no	yes	yes	no
	6. fbs	no	yes	yes	no	no	yes	yes	yes	no	yes	yes	yes
	7. restecg					no	yes	yes	yes	no	yes	yes	yes
	8. thalach					yes	yes	yes	yes	yes	yes	yes	yes
	9. exang					yes	yes	yes	yes	yes	yes	yes	yes
	10. oldpeak					yes	yes	yes	yes	yes	yes	yes	yes
	11. slope					no	yes	yes	yes	no	yes	yes	yes
	12. ca									yes	yes	yes	yes
	13. thal									yes	yes	yes	yes

## Winning Models

From Tables 3 through 5, Naïve Bayes models outperformed all others on AUC and thus were selected as the winners of each predictor subset.

The generation R Code for all models as well as their results can be found in the appendices.



## Results

### 1. Can a model of AUC $\geq 0.90$ be created to predict heart disease?

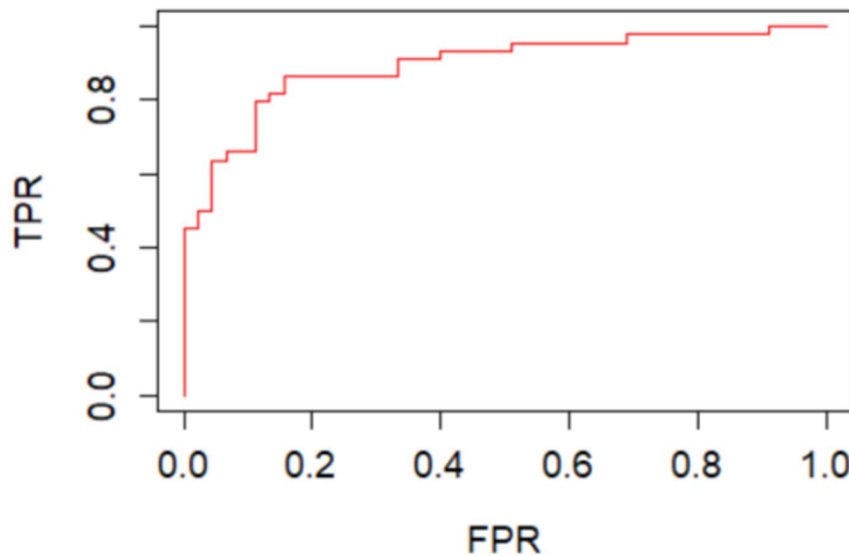
**No**, the highest AUC value was obtained from the Naïve Bayes model using the Confirmation data subset with an AUC value of 0.8939.

**Table 6: Naïve Bayes (Confirmation Subset) Confusion Matrix:**

Prediction	Actual	
	HD	No HD
HD	34	5
No HD	9	41

The confusion matrix (Table 6) shows the output of the classifications from the model. The Naïve Bayes model had a validation set accuracy of 83.4%. False Negatives occurred at a rate of **1.8 : 1** to False Positives.

**Figure 7: Naïve Bayes (Confirmation Subset) ROC Chart:  
ROC curve (n = 89)**



Overall, the Confirmation subset model is very accurate on predicting heart disease. While the target of  $\geq 0.90$  was not achieved, the “9” False Negatives is low, it is important still as each instance is a patient with heart disease without diagnosis. The “5” False Positives is also low. Meaning a person does not have heart disease but is diagnosed they do.

## 2. Can a subset of the predictors be used to create a model of AUC $\geq 0.80$ ?

**Yes**, the best performing models in both the Early Warning and Classification predictor subsets met the target of having an AUC  $\geq 0.80$ .

### Classification Subset Model

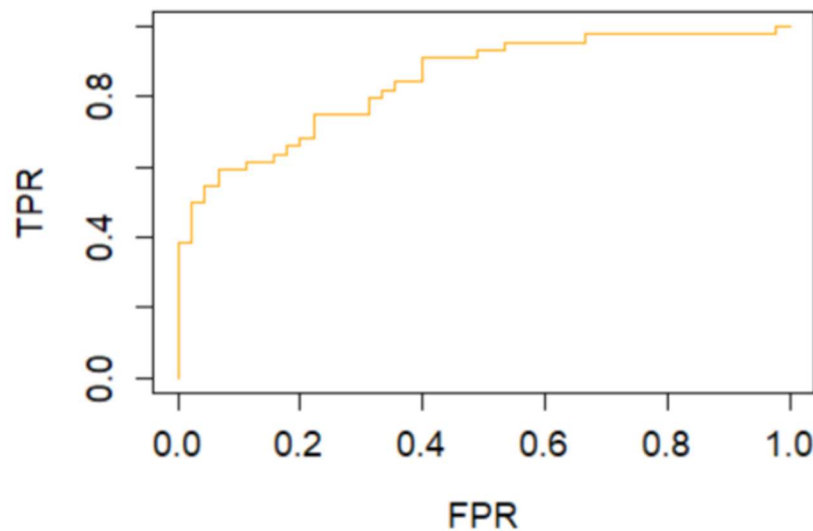
The selected best performing Classification model was the Naïve Bayes model with an AUC of 0.8449.

**Table 7: Naïve Bayes (Classification Subset) Confusion Matrix:**

Pred	Actual	
	HD	no HD
HD	30	8
No HD	13	38

The confusion matrix (Table 7) shows the output of the classifications from the model. The Naïve Bayes model had a validation set accuracy of 76.4%. False Negatives occurred at a rate of **1.63 :1** to False Positives. The ROC Curve for this model is shown below in Figure 2:

**Figure 8: Naïve Bayes (Classification Subset) ROC Chart**  
**ROC curve (n = 89)**



## Early Warning Subset Model

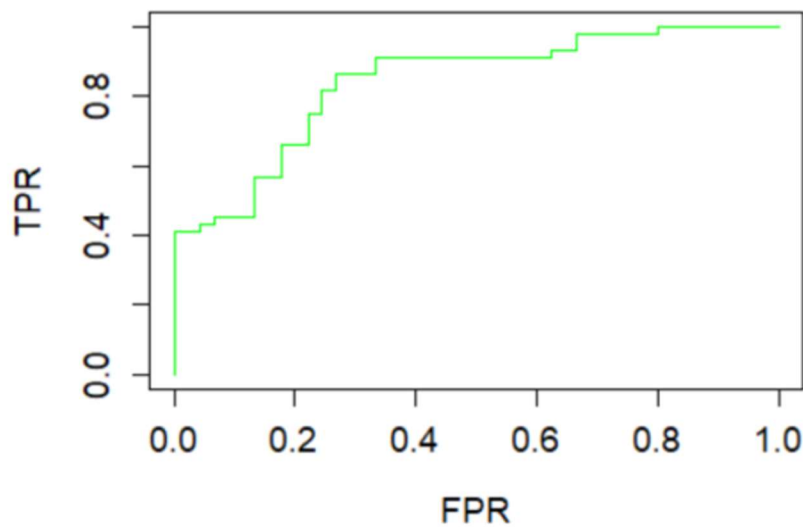
The selected best performing Early Warning model was the Naïve Bayes model with an AUC of 0.8374.

**Table 8: Naïve Bayes (Early Warning Subset) Confusion Matrix:**

Pred	Actual	
	HD	No HD
HD	27	8
No HD	16	38

The confusion matrix (Table 8) shows the output of the classifications from the model. The Naïve Bayes model had a validation set accuracy of 73.0%. False Negatives occurred at a rate of **2 : 1** to False Positives. The ROC Curve for this model is shown below in Figure 3:

**Figure 9: Naïve Bayes (Early Warning Subset) ROC Chart**  
**ROC curve (n = 89)**



### 3. What are the most significant factors when predicting heart disease?

Using the Rpart() function, the Importance Values for predictors from the Decision Tree models were discovered. Importance Value is the count of a variable being used in the primary and surrogate splits. The predictors with the top 3 Importance Values for each predictor subset are shown below in Tables 9 through 11:

**Table 9: Confirmation Predictor Subset, Top 3 Importance Value**

<u>Predictor</u>	<u>Importance Value</u>
thal	25
cp	19
thalach	15

**Table 10: Classification Predictor Subset, Top 3 Importance Value**

<u>Predictor</u>	<u>Importance Value</u>
cp	24
oldpeak	19
thalach	17

**Table 11: Early Warning Predictor Subset, Top 3 Importance Value**

<u>Predictor</u>	<u>Importance Value</u>
cp	42
age	27
chol	22

When taking the results of these three sets together, the top three predictors were identified by showing the most repeats in high Importance Value. These are:

1. Cp - Chest Pain Type
2. Thal - Stress test blood flow observed w/Thallium
3. Thalach - Maximum Heart Rate Achieved

## Conclusions

While a model having an AUC  $\geq 0.90$  could not be developed several models with AUC values  $\geq 0.80$  were developed using varying predictor subsets. This shows that it is possible to produce a reasonably effective classifier to predict heart disease in a patient from the common indicators available. We found that using a Naïve Bayes produced the highest AUC values, although the Logistic model had a higher validation set accuracy for the Confirmation subset models. The top three predictors were identified as cp, thal, and thalach.

A noteworthy finding is that using the Early Warning predictor subset a Naïve Bayes model was developed with an AUC of 0.8374. This AUC is only 6% worse than the winning model using the Confirmation predictor subset with an AUC 0.8939 despite using only predictors which can be measured by a routine physical with bloodwork. This shows that detection of heart disease on a routine – and possibly early - basis is possible.

Since heart disease is a potentially fatal condition, it is important to consider the potential outcomes of misclassification using these models. Further investigation including medical professionals should be performed to determine the risk vs. reward associated with minimizing false-positives vs. False- negatives classifications from the predictive models.

## Appendix A– Detailed Model Results

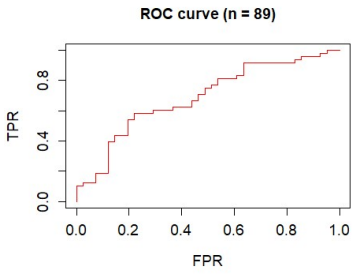
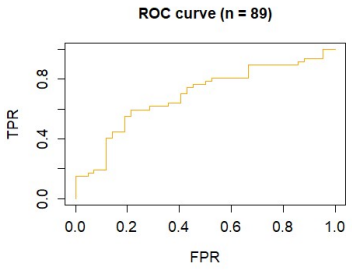
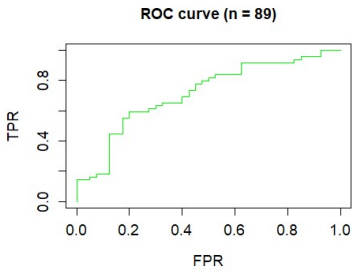
### K-Nearest Neighbor

*#KNN model*

*#Manually tested different K values for the best accuracy*

```
pred_knn <- knn(train= heart[trainset,], test=heart[validset,],  
cl=heart[trainset,]$target, k=15)
```

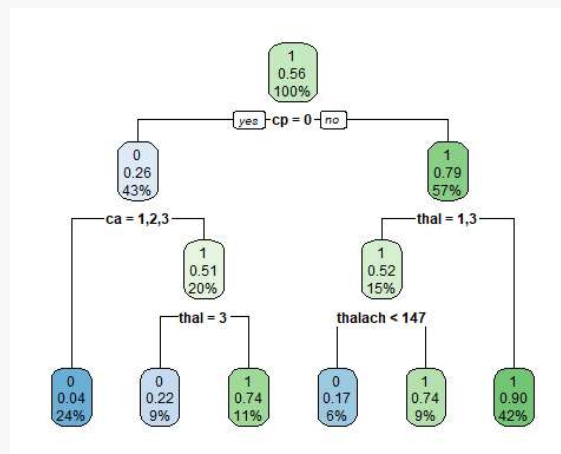
AUC value and ROC curve for each subset

Confirmation	Classification	Early Warning
0.6895325	0.6930091	0.7142857
		

## Decision Tree

```
#Create decision tree with selected variables from Confirmation Subset
tree_hrt_conf <- rpart(target ~ .-chol -trestbps, data = heart,method = 'class', subset = trainset)
#plot tree
rpart.plot(tree_hrt_conf)
```

For the full confirmation model, the tree output is as shown below. The tree can be interpreted as a series of “If this, go to this next” statements. Starting at the top-most node with a new observation the state of “cp” for that observation would be checked. If it was 0, the next check would be if ca = 1,2, or 3. If cp was not 0, the next check would be if thal = 1, or 3. These checks are continued until the bottom-most nodes are reached. The bottom-most nodes indicate the default classification from the probability, the probability of an observation NOT having heart disease, and the % of total training set that met the conditions to be placed at that node.



## AUC value and ROC curve for each subset

Confirmation	Classification	Early Warning
0.6895325	0.7308081	0.768453

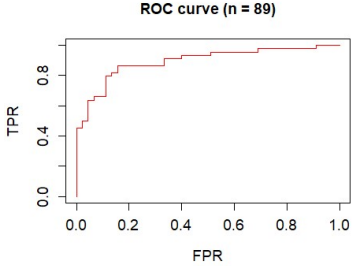
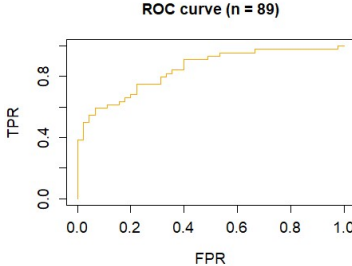
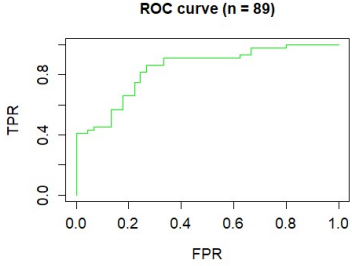


## Naïve Bayes

```
#Naive Bayes Model
```

```
nb_full <- naiveBayes(target~., data = heart, subset = trainset)
```

AUC value and ROC curve for each subset

Confirmation	Classification	Early Warning
0.8939	0.8449	0.8374
		

## Logistic Regression

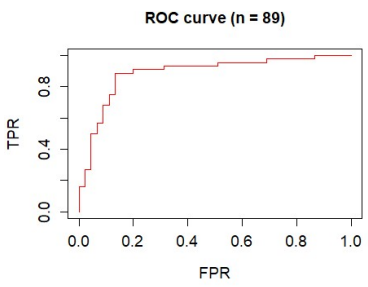
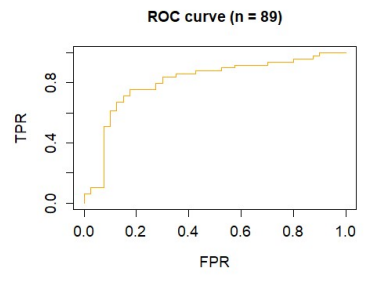
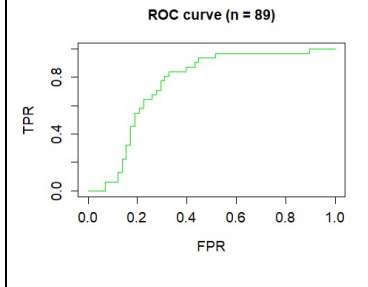
*#Model Creation, Predictors eliminated via backwards selection w/AIC*

```
md.log.conf <- glm(target ~ sex + cp + exang + thalach + oldpeak + ca + thal,
data = heart, subset = trainset, family = "binomial")
```

Logistic Regression		Early Alarm	Classification	Confirmation
Predictors	1. age	yes	no	no
	2. sex	yes	yes	yes
	3. cp	yes	yes	yes
	4. trestbps	yes	no	no
	5. chol	no	no	no
	6. fbs	no	no	no
	7. restecg		no	no
	8. thalach		yes	yes
	9. exang		yes	yes
	10. oldpeak		yes	yes
	11. slope		no	no
	12. ca			yes
	13. thal			yes

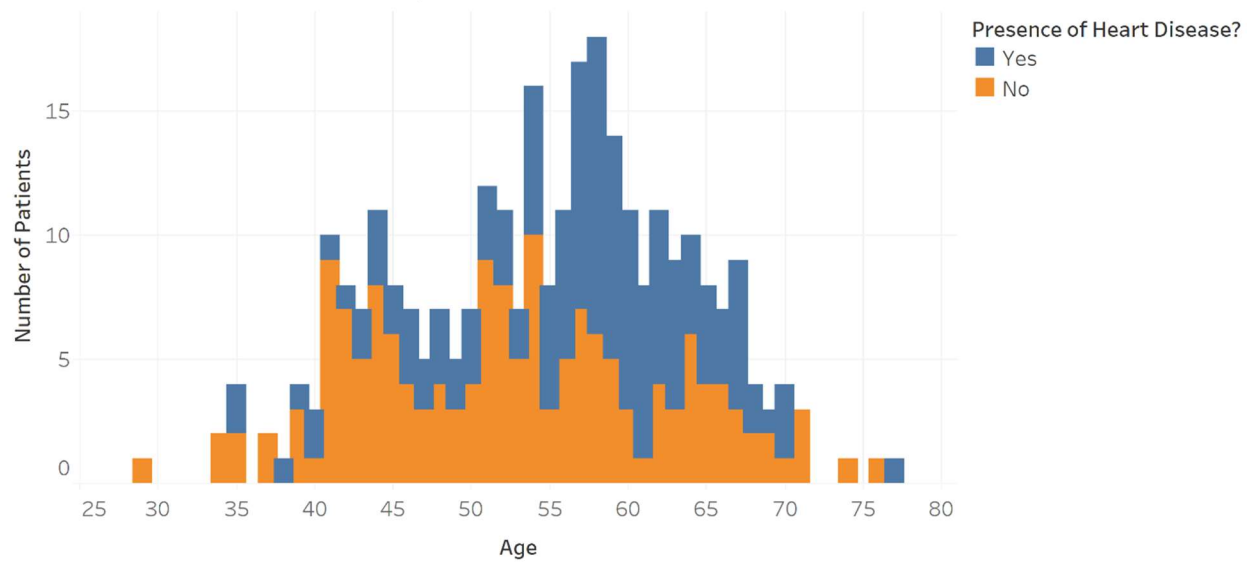
To create the optimal logistic model for each subset, backwards selection was used. The specific predictors selected for each subset are shown on the left.

AUC value and ROC curve for each subset

Confirmation	Classification	Early Warning
0.8884	0.8087	0.7542
		

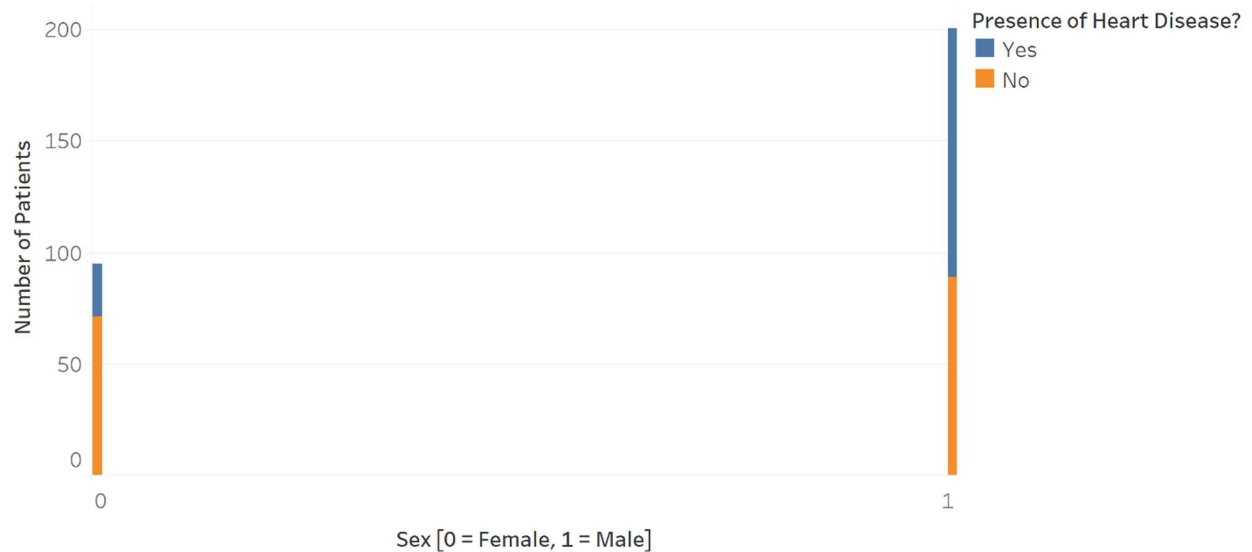
## Appendix B – All Data Investigation Results

Number of Patients With/Without Heart Disease by Age



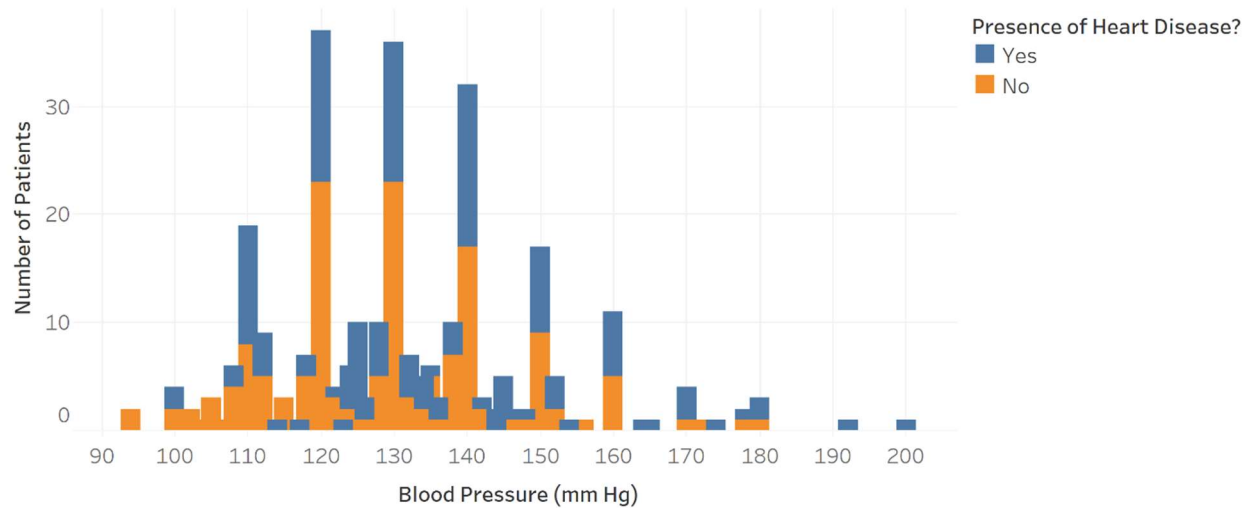
Observing the above figure, heart disease is most prevalent in patients aged between 50-65.

Number of Patients With/Without Heart Disease by Sex



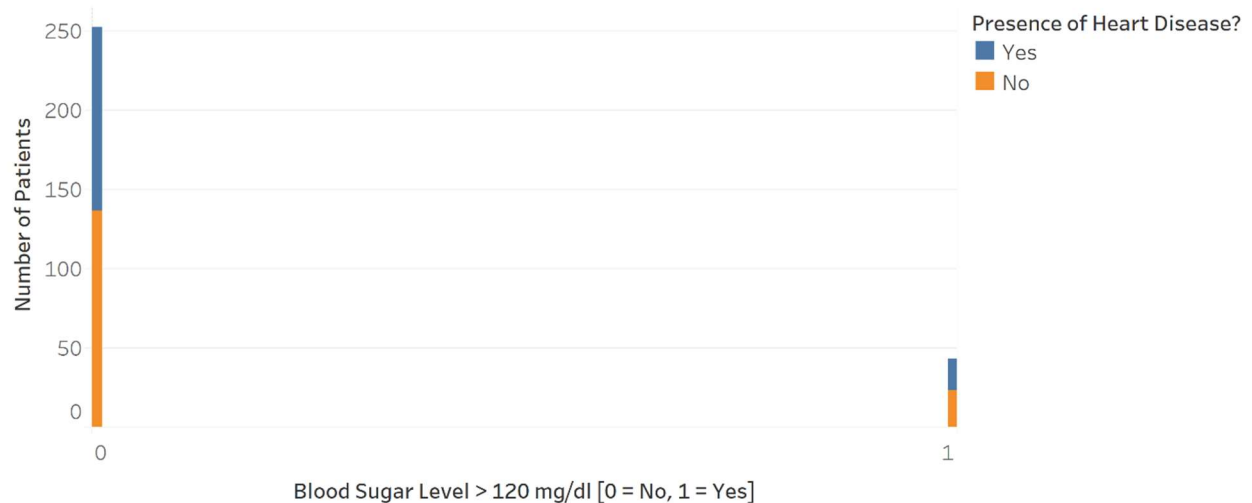
Observing the above figure, heart disease is more prevalent in male patients.

Number of Patients With/Without Heart Disease by Resting Blood Pressure



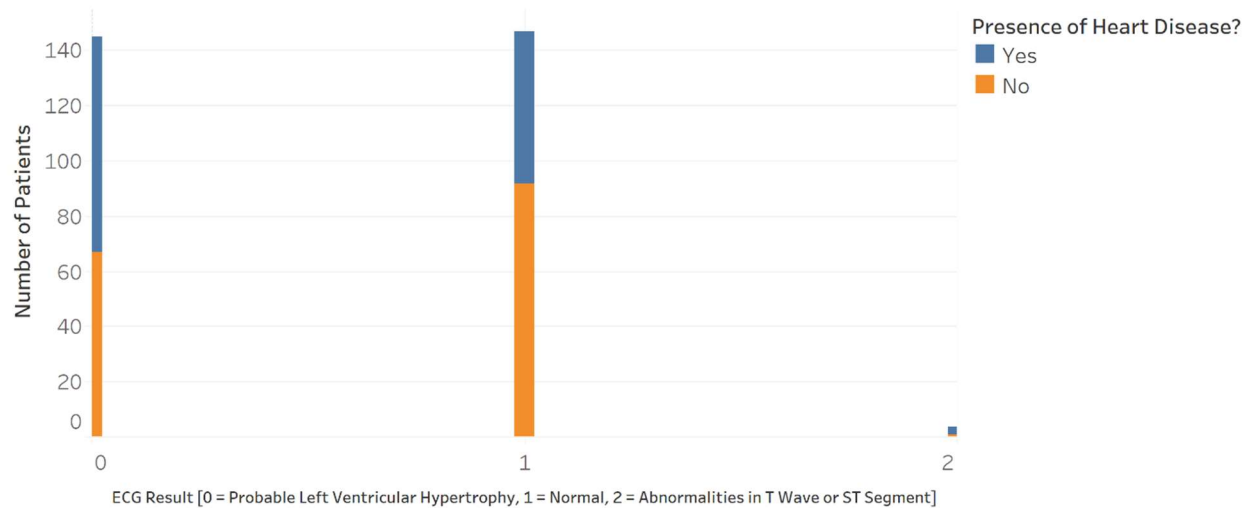
Observing the above figure, very high blood pressure can indicate the presence of heart disease.

Number of Patients With/Without Heart Disease by Blood Sugar



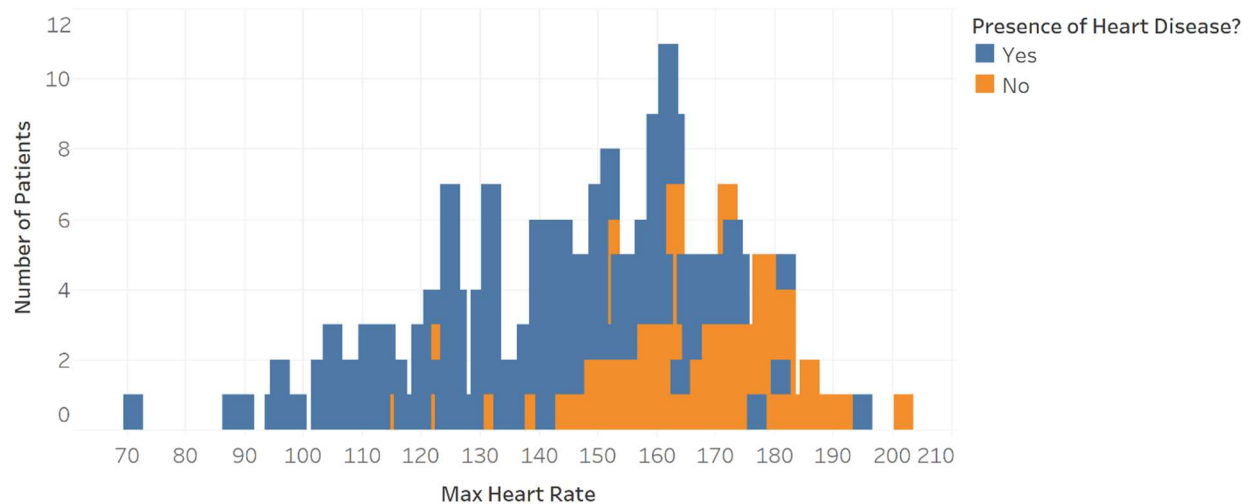
Observing the above figure, blood sugar doesn't seem to be a valuable indicator of heart disease as the distribution is split evenly given both blood sugar cases.

## Number of Patients With/Without Heart Disease by Electrocardiogram Results on Rest



Observing the above figure, probable hypertrophy does not seem to be very indicative of heart disease. Abnormalities in the T wave of ST segment seem to be very indicative of the presence of heart disease, as a large percentage of those cases were diagnosed with heart disease.

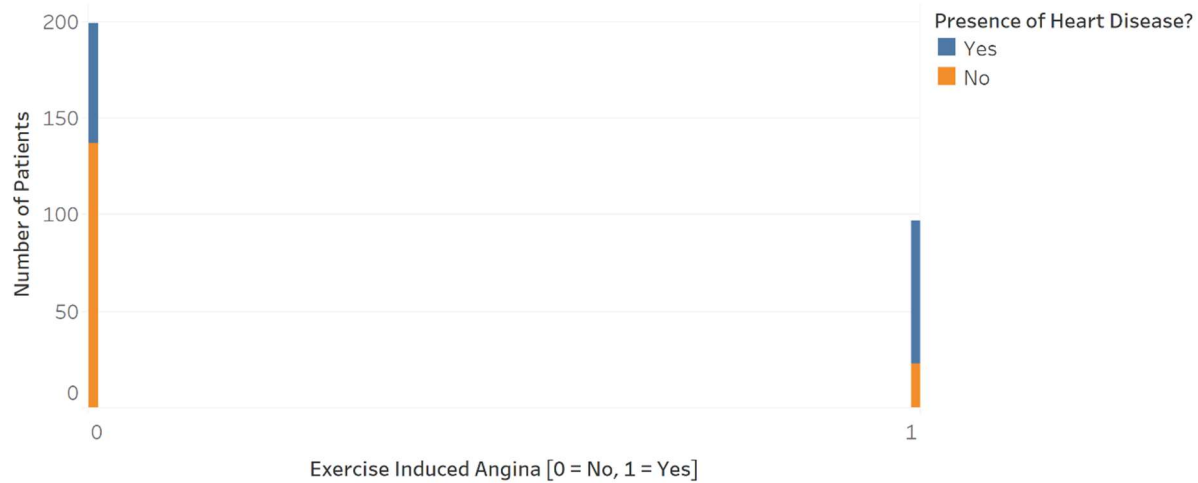
## Number of Patients With/Without Heart Disease by Max Heart Rate During Stress Test



Observing the data for max heart rate during stress test, the following conclusion can be made:

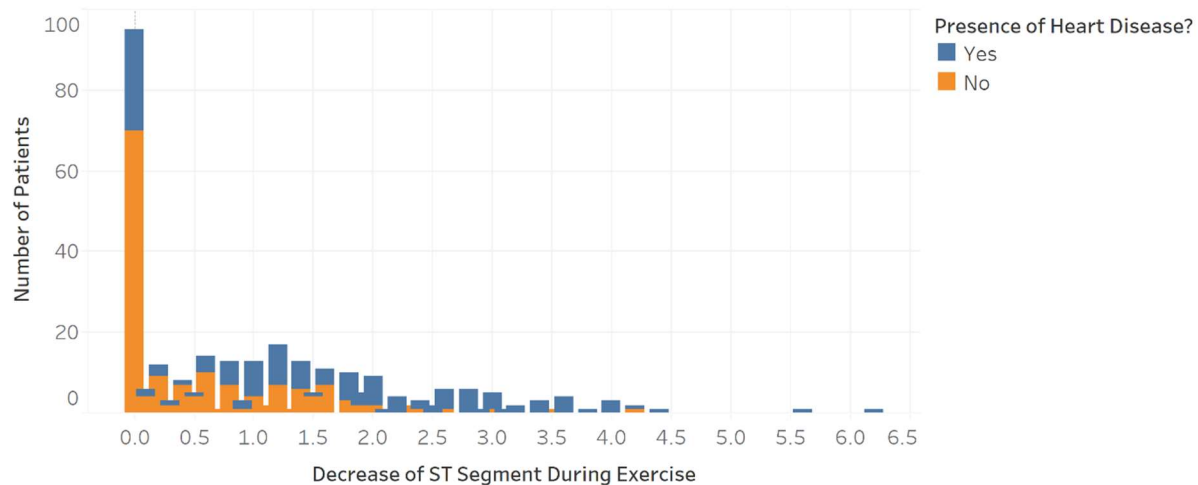
- Higher max heart rate is correlated with no presence of heart disease
- Lower max heart is linked with the presence of heart disease
- These takeaways seem logical because as age increases, max heart rate decreases

Number of Patients With/Without Heart Disease by Presence of Angina During Exercise



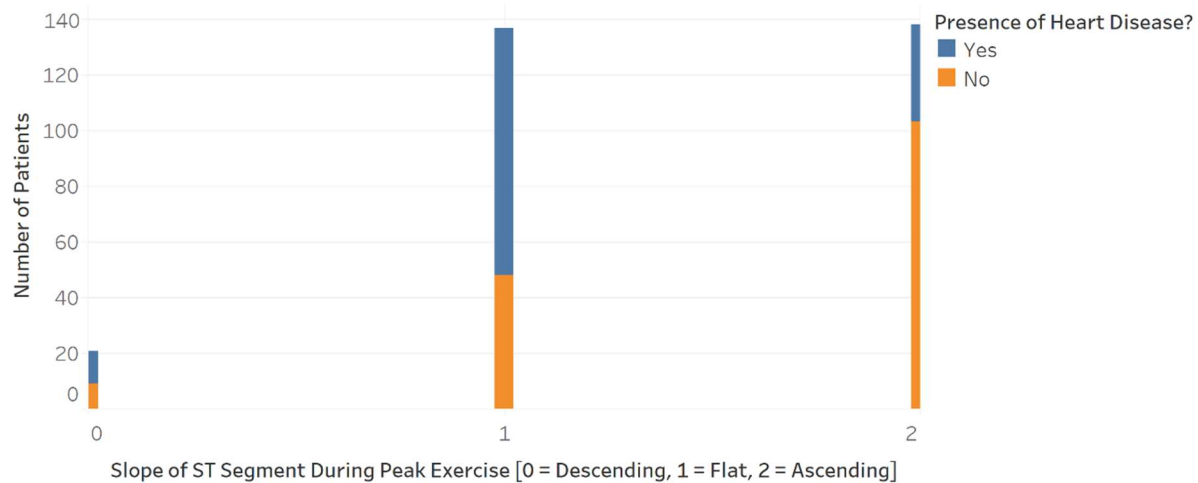
Observing the above figure, exercise induced angina is a good indicator of heart disease given the large percentage of patients diagnosed with heart disease who present with exercise induced angina.

Number of Patients With/Without Heart Disease by Decrease of ST Segment During Exercise



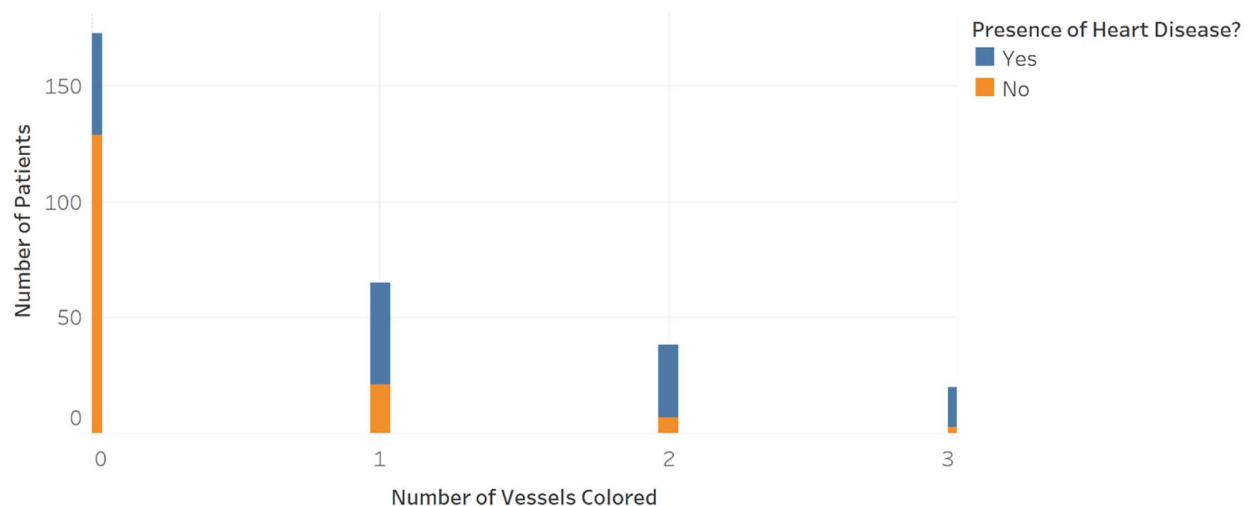
Observing the above figure, no displacement in the ST segment is a good indicator of the absence of heart disease. Displacements above 1.5 units appear to be highly correlated with the presence of heart disease.

Number of Patients With/Without Heart Disease by Slope of ST Segment During Peak Exercise



Observing the above figure, heart disease was most prevalent where the Slope of ST Segment during Peak Exercise was 1 (the slope was flat).

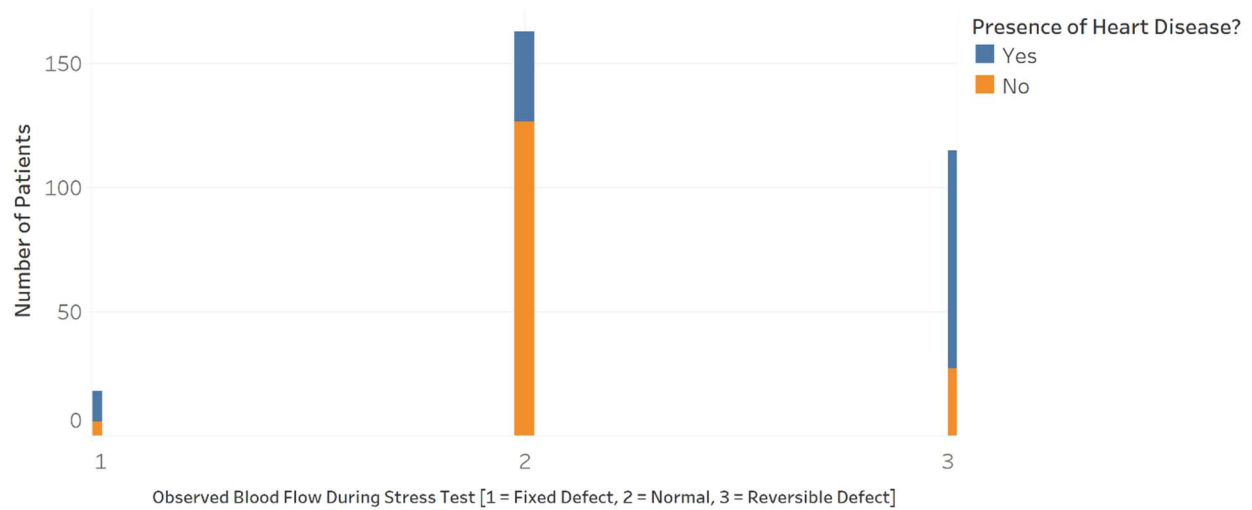
Number of Patients With/Without Heart Disease by Number of Major Blood Vessels Colored by Fluoroscopy



Observing the above figure, heart disease has a positive correlation with the Number of Major Blood Vessels Colored by Fluoroscopy. The absence of colored vessels does not however indicate that there is no heart disease.



## Number of Patients With/Without Heart Disease by Observed Blood Flow During Stress Test



Observing the above figure, heart disease is most prevalent where a defect exists whether it has or has not been resolved.

## Appendix C - Full Code

Call libraries and read in dataset

```
library(readr)
library(MASS)
library(class)
library(rpart)
library(rpart.plot)
library(e1071)

rm(list = ls())
heart <- read_csv("heart.csv")

## Parsed with column specification:
## cols(
##   age = col_double(),
##   sex = col_double(),
##   cp = col_double(),
##   trestbps = col_double(),
##   chol = col_double(),
##   fbs = col_double(),
##   restecg = col_double(),
##   thalach = col_double(),
##   exang = col_double(),
##   oldpeak = col_double(),
##   slope = col_double(),
##   ca = col_double(),
##   thal = col_double(),
##   target = col_double()
## )
```

Quick look at data

```
summary(heart)
```

##	age	sex	cp	trestbps
##	Min. :29.00	Min. :0.0000	Min. :0.0000	Min. : 94.0
##	1st Qu.:48.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:120.0
##	Median :56.00	Median :1.0000	Median :1.0000	Median :130.0
##	Mean :54.52	Mean :0.6791	Mean :0.9595	Mean :131.6
##	3rd Qu.:61.00	3rd Qu.:1.0000	3rd Qu.:2.0000	3rd Qu.:140.0
##	Max. :77.00	Max. :1.0000	Max. :3.0000	Max. :200.0
##	chol	fbs	restecg	thalach
##	Min. :126.0	Min. :0.0000	Min. :0.0000	Min. : 71.0
##	1st Qu.:211.0	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:133.0
##	Median :242.5	Median :0.0000	Median :1.0000	Median :152.5
##	Mean :247.2	Mean :0.1453	Mean :0.5236	Mean :149.6
##	3rd Qu.:275.2	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:166.0
##	Max. :564.0	Max. :1.0000	Max. :2.0000	Max. :202.0
##	exang	oldpeak	slope	ca

```
## Min. :0.0000 Min. :0.000 Min. :0.000 Min. :0.0000
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:1.000 1st Qu.:0.0000
## Median :0.0000 Median :0.800 Median :1.000 Median :0.0000
## Mean :0.3277 Mean :1.059 Mean :1.395 Mean :0.6791
## 3rd Qu.:1.0000 3rd Qu.:1.650 3rd Qu.:2.000 3rd Qu.:1.0000
## Max. :1.0000 Max. :6.200 Max. :2.000 Max. :3.0000
##      thal      target
## Min. :1.000 Min. :0.0000
## 1st Qu.:2.000 1st Qu.:0.0000
## Median :2.000 Median :1.0000
## Mean :2.328 Mean :0.5405
## 3rd Qu.:3.000 3rd Qu.:1.0000
## Max. :3.000 Max. :1.0000
```

`head(heart)`

```
## # A tibble: 6 x 14
##   age  sex  cp trestbps chol  fbs restecg thalach exang oldpeak slo
##   <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>
## 1    63    1    3    145    233    1     0    150     0    2.3
## 2    37    1    2    130    250    0     1    187     0    3.5
## 3    41    0    1    130    204    0     0    172     0    1.4
## 4    56    1    1    120    236    0     1    178     0    0.8
## 5    57    0    0    120    354    0     1    163     1    0.6
## 6    57    1    0    140    192    0     1    148     0    0.4
## # ... with 3 more variables: ca <dbl>, thal <dbl>, target <dbl>
```

Looking at the data as above, some variables are categorical. To better assess the categorical data, we will factor each categorical variable.

- CP for instance, is labeled as either 1, 2 or 3. While these categories are labeled with numbers, they should be treated categorically.
  - Note that this is chest pain type, not chest pain strength. The difference between pain labeled 1 and 3 is the same as the difference between pain labeled 2 and 3

```
#Converting to categorical type
heart$sex <- as.factor(heart$sex)
heart$cp <- as.factor(heart$cp)
heart$fbs <- as.factor(heart$fbs)
```

```

heart$restecg <- as.factor(heart$restecg)
heart$exang <- as.factor(heart$exang)
heart$slope <- as.factor(heart$slope)
heart$ca <- as.factor(heart$ca)
heart$thal <- as.factor(heart$thal)
heart$target <- as.factor(heart$target)

```

Before beginning our analysis, we will define functions that will be useful later.

```

ROC_func <- function(df, label_colnum, score_colnum, add_on = F, color = "black"){
  # Sort by score (high to low)
  df <- df[order(-df[,score_colnum]),]
  rownames(df) <- NULL # Reset the row number to 1,2,3,...
  n <- nrow(df)
  # Total # of positive and negative cases in the data set
  P <- sum(df[,label_colnum] == 1)
  N <- sum(df[,label_colnum] == 0)

  # Vectors to hold the coordinates of points on the ROC curve
  TPR <- c(0,vector(mode="numeric", length=n))
  FPR <- c(0,vector(mode="numeric", length=n))

  # Calculate the coordinates from one point to the next
  AUC = 0
  for(k in 1:n){
    if(df[k,label_colnum] == 1){
      TPR[k+1] = TPR[k] + 1/P
      FPR[k+1] = FPR[k]
    } else{
      TPR[k+1] = TPR[k]
      FPR[k+1] = FPR[k] + 1/N
      AUC = AUC + TPR[k+1]*(1/N)
    }
  }

  # Plot the ROC curve
  if(add_on){
    points(FPR, TPR, main=paste0("ROC curve", " (n = ", n, ")"), type = 'l', col=color, cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.2)
  } else{
    plot(FPR, TPR, main=paste0("ROC curve", " (n = ", n, ")"), type = 'l', col=color, cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.2)
  }
  return(AUC)
}

#This sets the train/validation subsets. The split is set as 70/30
resetSets <- function(){

```

```

trainset <- sample(1:nrow(heart), round(nrow(heart)*0.7))
validset <- setdiff(1:nrow(heart),trainset)
}

#Functions for logistic models
log_score <- function(model) {
  scores <- predict(model, heart[validset,])
  scores <- exp(scores)/(1+exp(scores))
  return(scores)
}

log_class <- function(model) {
  scores <- predict(model, heart[validset,])
  scores <- exp(scores)/(1+exp(scores))
  prediction <- ifelse(scores>threshold,1,0)
  return(prediction)
}

resetSets()

```

## KNN Model

### Confirmation Model

```

#subsets as s
earlywarning <- 6 #earlywarning
classification <- 11 #Classification
confirmation <- 13 #Confirmation

s = confirmation

#This sets predictor subsets: Early Alarm, Classification, Confirmation
traindata <- heart[trainset, c(1:s,14) ]
testdata <- heart[validset, c(1:s,14)]
hrt <- heart[, c(1:s,14)]

#Normalizing the data makes the accuracy jump higher in calculations below
normalize <- function(x){return ((x - min(x)) / (max(x) - min(x))) }

#Normalizing only non-factored/non-categorical variables
heart_n <- heart
heart_n$age <- normalize(heart$age)
heart_n$chol <- normalize(heart$chol)
heart_n$trestbps <- normalize(heart$trestbps)
heart_n$thalach <- normalize(heart$thalach)
heart_n$oldpeak <- normalize(heart$oldpeak)
n <- length(heart_n)-1

#KNN model

```

```

#Manually tested different K values for the best accuracy
pred_knn <- knn(train= heart[trainset,], test=heart[validset,], cl=heart[trainset,]$target, k=25)

#Confusion Matrix table
table( predictions = pred_knn, target = heart[validset,]$target)

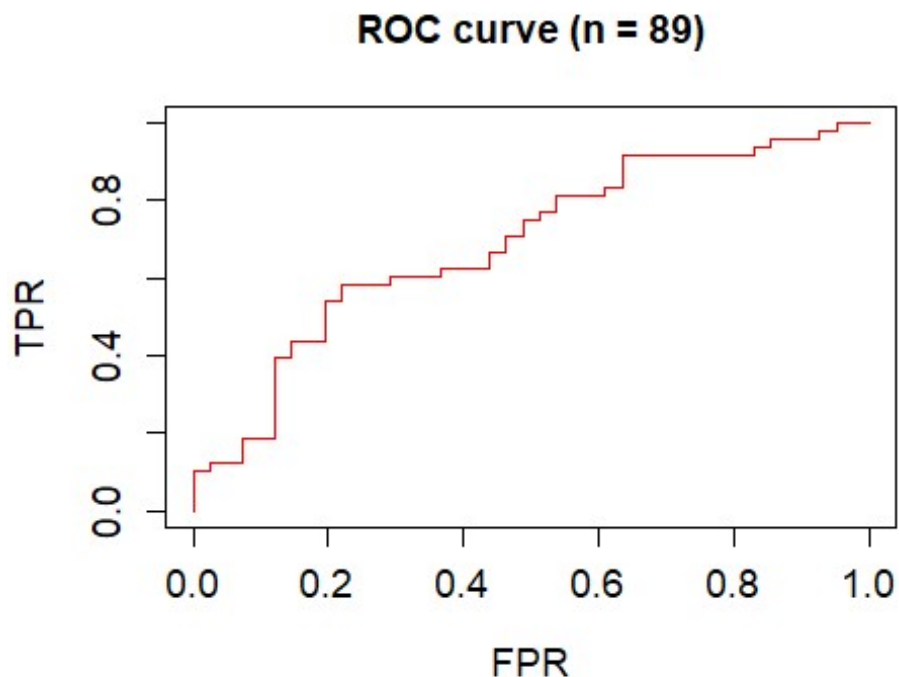
##           target
## predictions  0  1
##           0 26 15
##           1 19 29

#Error Rate
( sum(ifelse(pred_knn == heart[validset,]$target, 1, 0))/ length(heart[validset,]$target))

## [1] 0.6179775

#create an AUC curve
knn_pred_truelabel <- data.frame(pred_knn, heart[validset,]$target)
heart.knn.AUC <- ROC_func(knn_pred_truelabel, 1, 2 , color = 'red')

```



```

(heart.knn.AUC)
## [1] 0.6895325

```

**Classification Model**

```

s =classification

#This sets predictor subsets: Early Alarm, Classification, Confirmation
traindata <- heart[trainset, c(1:s,14) ]
testdata <- heart[validset, c(1:s,14)]
hrt <- heart[, c(1:s,14)]

#Normalizing the non-factored/numerical variables
heart_n <- heart
heart_n$age <- normalize(heart$age)
heart_n$chol <- normalize(heart$chol)
heart_n$trestbps <- normalize(heart$trestbps)
heart_n$thalach <- normalize(heart$thalach)
heart_n$oldpeak <- normalize(heart$oldpeak)
n <- length(heart_n)-1

#KNN model
#Manually tested different K values for the best accuracy
pred_knn <- knn(train= heart[trainset,], test=heart[validset,], cl=heart[trainset,]$target, k=20)

#Confusion Matrix table
table( predictions = pred_knn, target = heart[validset,]$target)

##           target
## predictions  0  1
##           0 27 15
##           1 18 29

#Error Rate
( sum(ifelse(pred_knn == heart[validset,]$target, 1, 0))/ length(heart[validset,]$target))

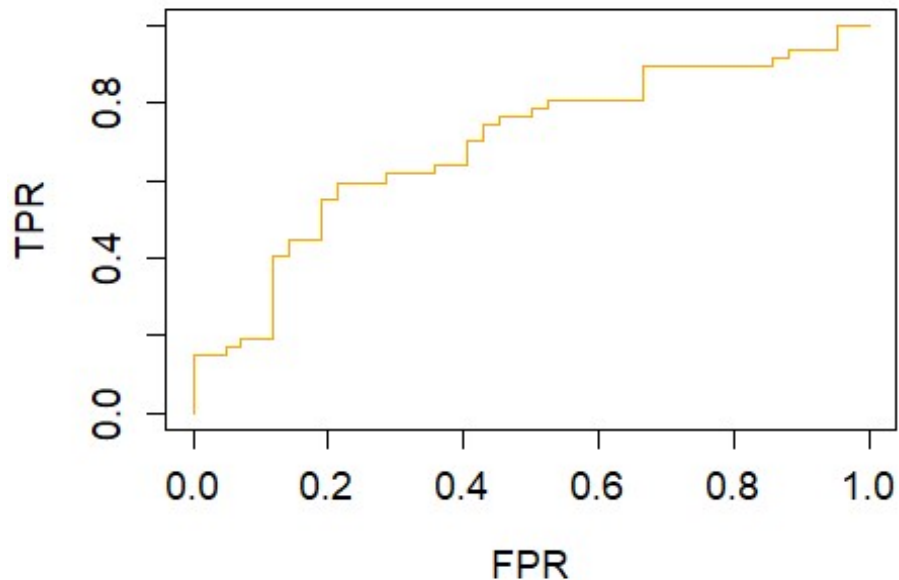
## [1] 0.6292135

#create an AUC curve
knn_pred_truelabel <- data.frame(pred_knn, heart[validset,]$target)
heart.knn.AUC <- ROC_func(knn_pred_truelabel, 1, 2 , color = 'orange')

```



ROC curve (n = 89)



```
(heart.knn.AUC)
```

```
## [1] 0.6930091
```

### Early Warning Model

```
s = earlywarning
```

```
#This sets the predictor subsets: Early Alarm, Classification, Confirmation
```

```
traindata <- heart[trainset, c(1:s,14) ]
```

```
testdata <- heart[validset, c(1:s,14)]
```

```
hrt <- heart[, c(1:s,14)]
```

```
#Only normalizing some non-factored/numerical variables
```

```
heart_n <- heart
```

```
heart_n$age <- normalize(heart$age)
```

```
heart_n$chol <- normalize(heart$chol)
```

```
heart_n$trestbps <- normalize(heart$trestbps)
```

```
heart_n$thalach <- normalize(heart$thalach)
```

```
heart_n$oldpeak <- normalize(heart$oldpeak)
```

```
n <- length(heart_n)-1
```

```
#KNN model
```

```
#Manually tested different K values for the best accuracy
```

```
pred_knn <- knn(train= heart[trainset,], test=heart[validset,], cl=heart[trainset,]$target, k=15)
```

```

#Confusion Matrix table
table( predictions = pred_knn, target = heart[validset,]$target)

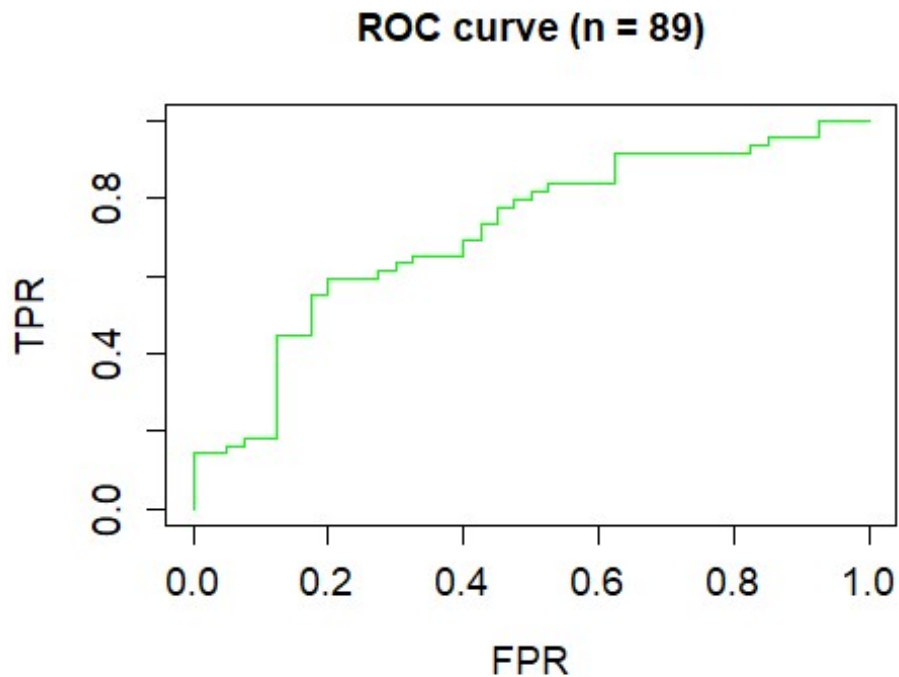
##           target
## predictions  0  1
##           0 27 13
##           1 18 31

#Error Rate
( sum(ifelse(pred_knn == heart[validset,]$target, 1, 0))/ length(heart[validset,]$target))

## [1] 0.6516854

#create an AUC curve
knn_pred_truelabel <- data.frame(pred_knn, heart[validset,]$target)
heart.knn.AUC <- ROC_func(knn_pred_truelabel, 1, 2, color = 'green')

```



```

(heart.knn.AUC)
## [1] 0.7142857

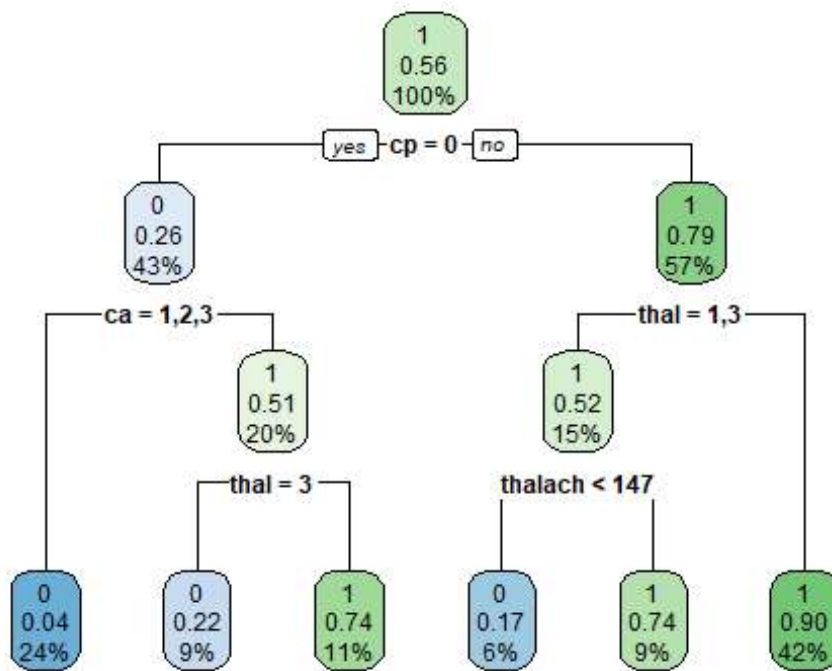
```

## Decision Tree

### Confirmation Model

```
#Create decision tree with selected variables from Confirmation Subset
tree_hrt_conf <- rpart(target ~ .-chol -trestbps, data = heart,method = 'class', subset = trainset)
```

```
#plot tree
rpart.plot(tree_hrt_conf)
```



Confirmation model confusion matrix and calculated accuracy

```
#Create predictive model based on "validset"
t_pred_conf = predict(tree_hrt_conf,heart[validset,],type="class")
#Create a confusion matrix and find accuracy
(confMat <- table(heart[validset,]$target,t_pred_conf))

##      t_pred_conf
##      0  1
## 0 34 11
## 1  6 38

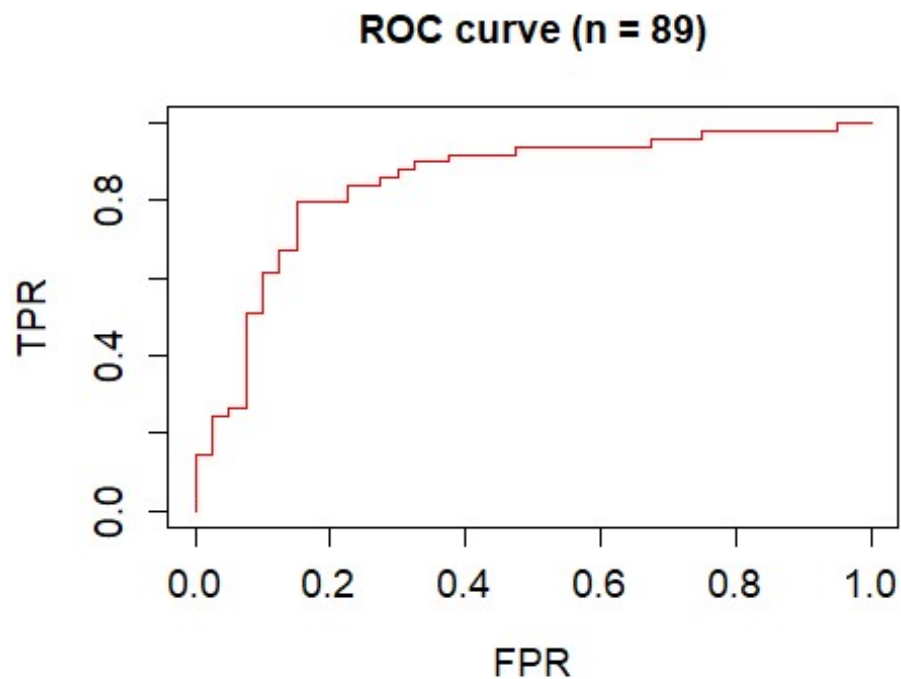
(accuracy <- sum(diag(confMat))/sum(confMat))

## [1] 0.8089888
```

ROC Curve and associated AUC for Confirmation Model

```
dt_pred_truelabel_conf <- data.frame(t_pred_conf, heart[validset,]$target)
heart.dt.AUC_conf <- ROC_func(dt_pred_truelabel_conf, 1, 2, color = 'red')
```

```
## Warning in Ops.factor(df[, score_colnum]): '-' not meaningful for factors
```

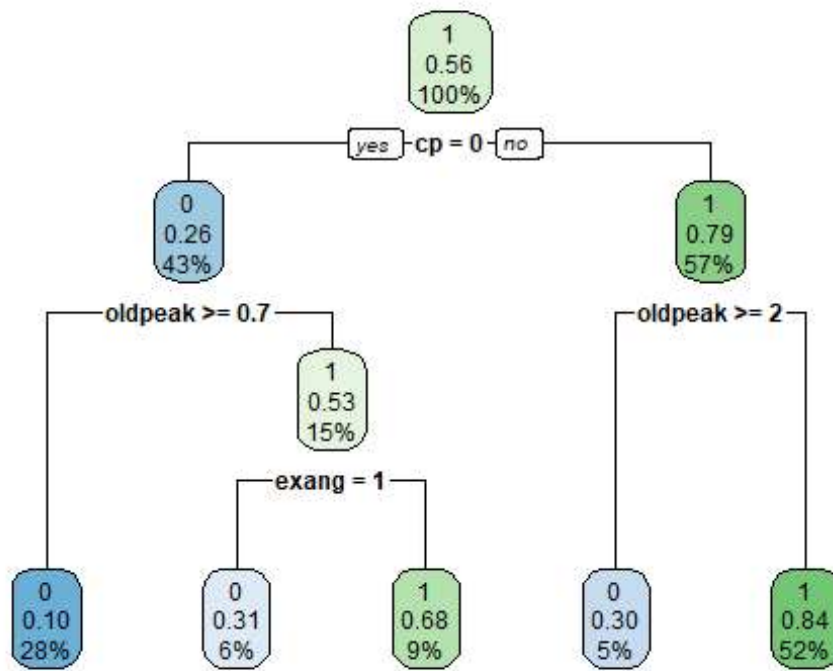


```
(heart.dt.AUC_conf)
```

```
## [1] 0.8484694
```

## Classification Model

```
#Create decision tree with selected variables from Classification Subset  
tree_hrt_class <- rpart(target ~age+sex+cp+trestbps+chol+fbs+restecg+thalach+  
exang+oldpeak+slope, data = heart,method = 'class', subset = trainset)  
#tree_hrt_class  
rpart.plot(tree_hrt_class)
```



```
#summary(tree_hrt_class)
```

Classification model confusion matrix and calculated accuracy

```
#Create predictive model based on "validset"
t_pred_class = predict(tree_hrt_class,heart[validset,],type="class")
#Create a confusion matrix and find accuracy
(confMat <- table(heart[validset,]$target,t_pred_class))

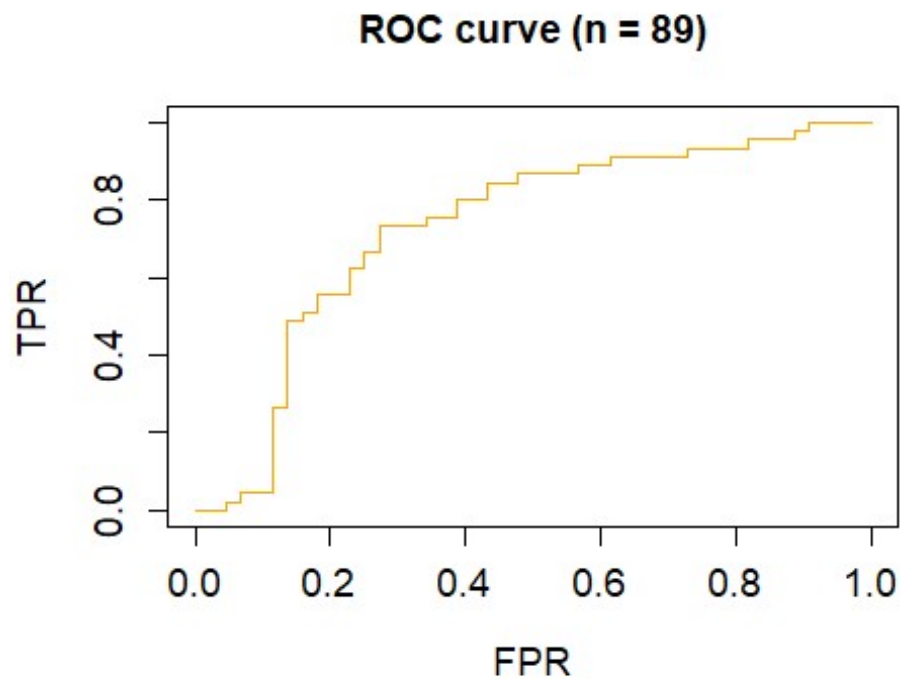
##      t_pred_class
##      0  1
##  0 32 13
##  1 12 32

(accuracy <- sum(diag(confMat))/sum(confMat))

## [1] 0.7191011
```

ROC Curve and associated AUC for Classification Model

```
dt_pred_truelabel_class <- data.frame(t_pred_class, heart[validset,]$target)
heart.dt.AUC_class <- ROC_func(dt_pred_truelabel_class, 1, 2, color = 'orange')
## Warning in Ops.factor(df[, score_colnum]): '-' not meaningful for factors
```

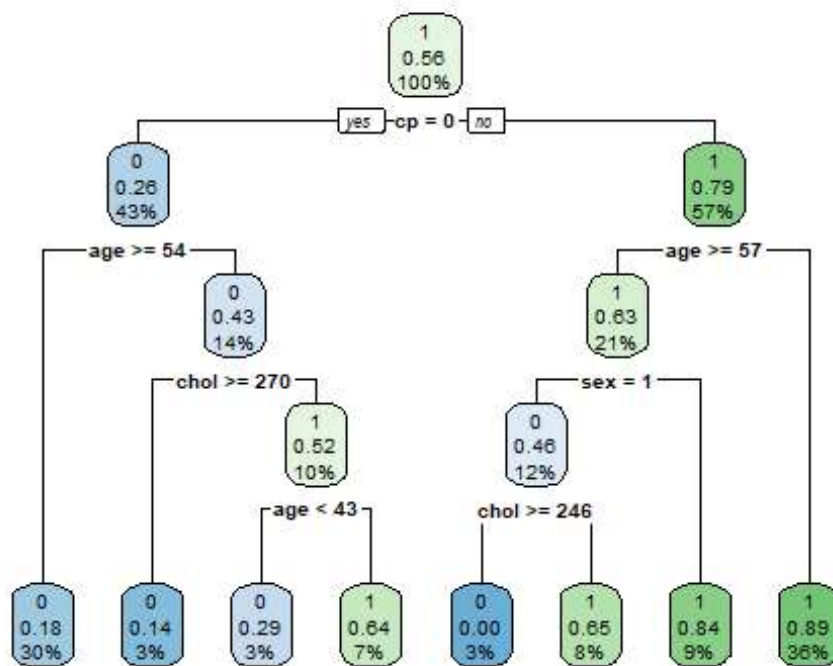


```
(heart.dt.AUC_class)
```

```
## [1] 0.7308081
```

### Early Warning Model

```
#Create decision tree with selected variables from Early Warning Subset  
tree_hrt_early <- rpart(target ~age+sex+cp+chol, data = heart,method = 'class',  
  subset = trainset)  
#tree_flight  
rpart.plot(tree_hrt_early)
```



```
#summary(tree_hrt_early)
```

Early Warning model confusion matrix and calculated accuracy

```
#Create predictive model based on "validset"
t_pred_early = predict(tree_hrt_early,heart[validset,],type="class")
#Create a confusion matrix and find accuracy
(confMat <- table(heart[validset,]$target,t_pred_early))

##      t_pred_early
##      0  1
## 0 34 11
## 1 12 32

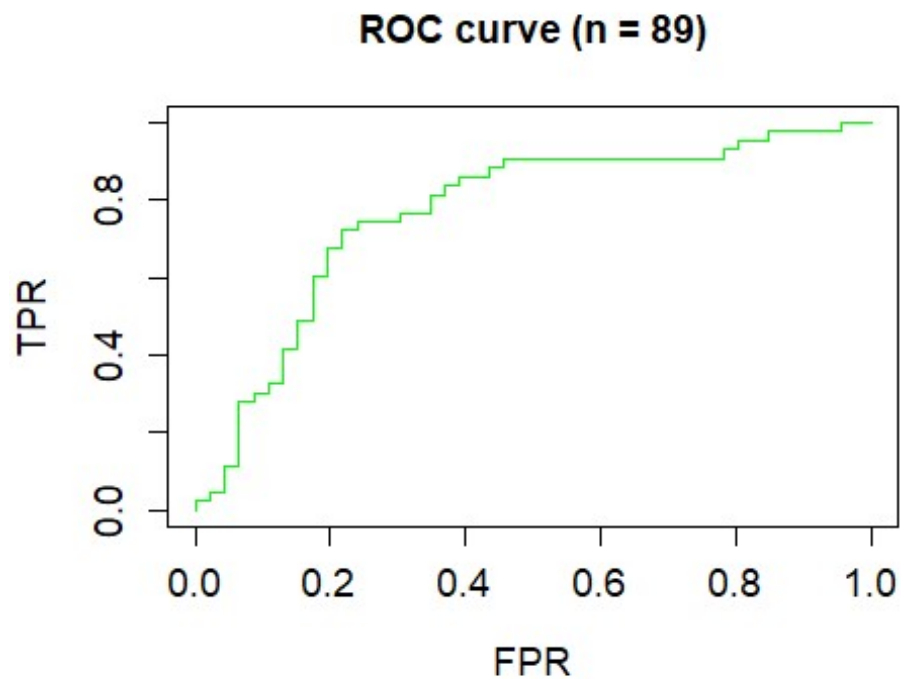
(accuracy <- sum(diag(confMat))/sum(confMat))

## [1] 0.741573
```

Early Model ROC

```
dt_pred_truelabel_early <- data.frame(t_pred_early, heart[validset,]$target)
heart.dt.AUC_early <- ROC_func(dt_pred_truelabel_early, 1, 2, color = 'green'
)
```





```
(heart.dt.AUC_early)
```

```
## [1] 0.768453
```

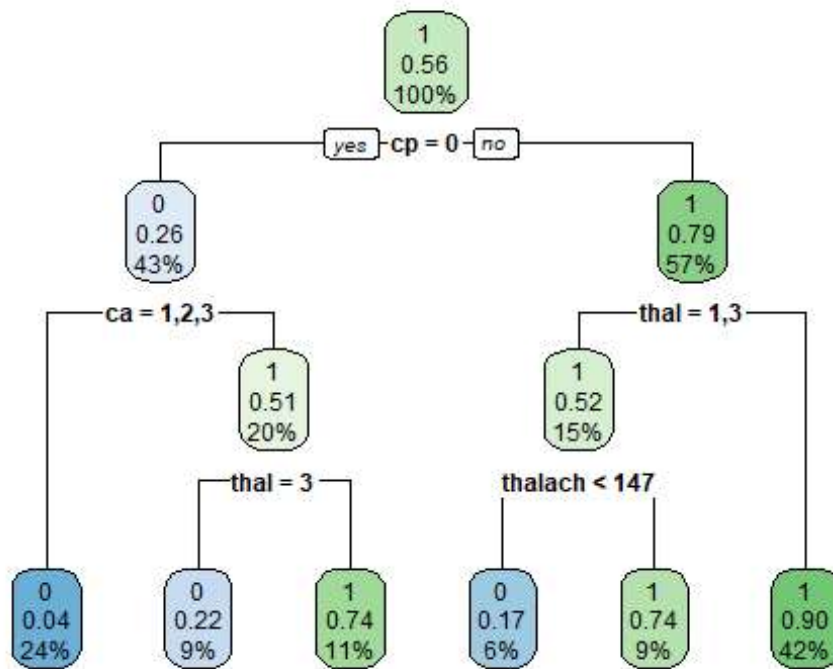
Confirmation Model

```
#Use rpart library to create decision tree with all predictors
```

```
tree_hrt <- rpart(target ~ ., data = heart, method = 'class', subset = trains  
et)
```

```
#tree_hrt
```

```
rpart.plot(tree_hrt)
```



```
#summary(tree_hrt)
```

Full model confusion matrix and calculated accuracy

```
#Create predictive model based on "validset"
t_pred = predict(tree_hrt,heart[validset,],type="class")
#Create a confusion matrix and find accuracy
(confMat <- table(heart[validset,]$target,t_pred))

##      t_pred
##      0  1
##  0 34 11
##  1  6 38

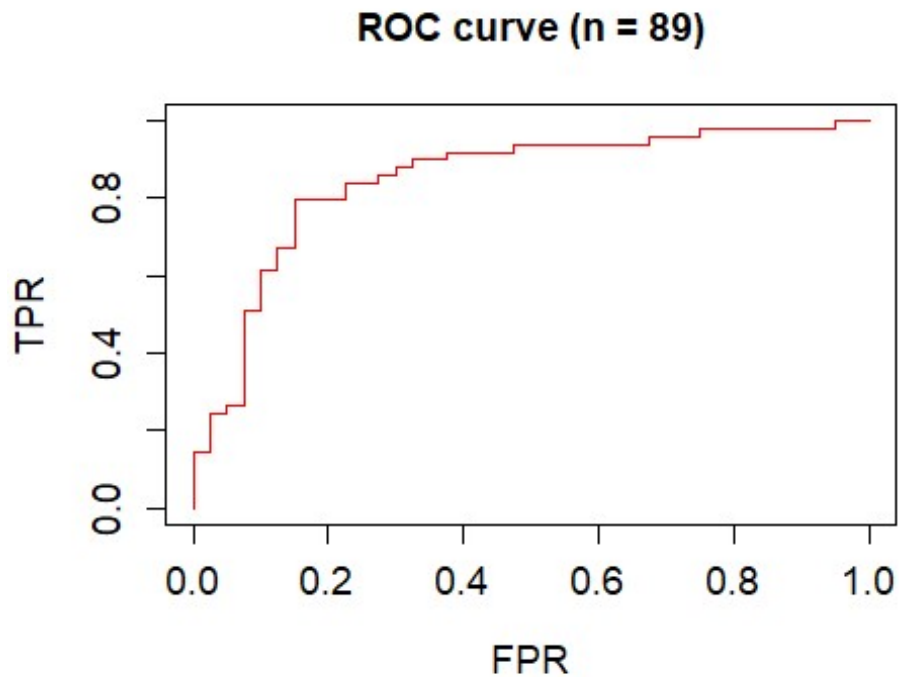
(accuracy <- sum(diag(confMat))/sum(confMat))

## [1] 0.8089888
```

ROC Curve and associated AUC for Full Model/Confirmation

```
dt_pred_truelabel <- data.frame(t_pred, heart[validset,]$target)
heart.dt.AUC <- ROC_func(dt_pred_truelabel, 1, 2, color = 'red')

## Warning in Ops.factor(df[, score_colnum]): '-' not meaningful for factors
```



```
(heart.dt.AUC)
## [1] 0.8484694
```

## Naive Bayes

### Confirmation Model

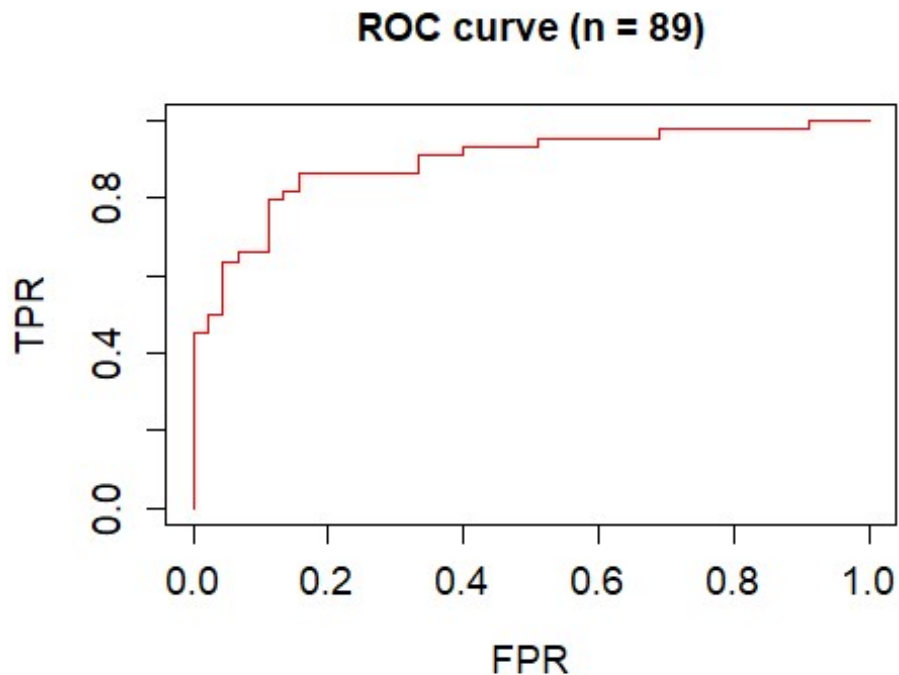
```
#Naive Bayes Model
nb_full <- naiveBayes(target~., data = heart, subset = trainset)

#Calculate predictions
pred1 <- predict(nb_full, heart[validset,])

#Model accuracy
table(pred1, heart[validset,]$target,dnn = c('Pred','Actual'))

##      Actual
## Pred  0  1
##    0 37  6
##    1  8 38

pred1_raw <- predict(nb_full, heart[validset,],type='raw')
pred1_df <- data.frame(score = pred1_raw[, '1'], true.class = ifelse(heart[validset,]$target == '1',1,0))
ROC_func(pred1_df,2,1, color = 'red')
```



```
## [1] 0.8939394
```

### Classification Model

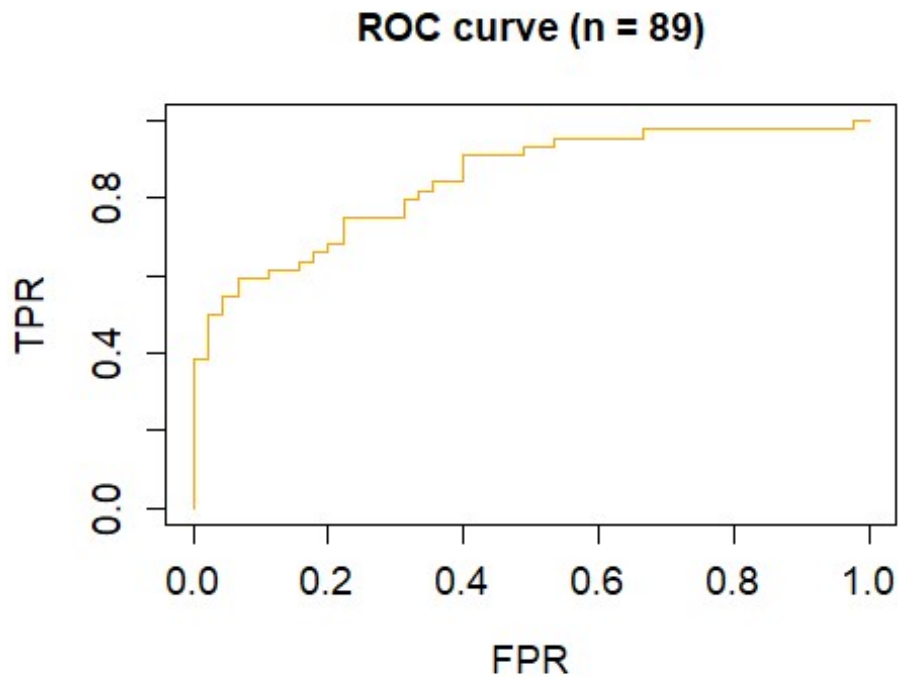
```
#Create model
nb_classification <- naiveBayes(target~ age + sex + cp + trestbps + chol + fb
s + restecg + thalach + exang + oldpeak + slope, data = heart, subset = train
set)

#calculate predictions
pred3 <- predict(nb_classification, heart[validset,])

#Model Accuracy
table(pred3, heart[validset,]$target, dnn = c('Pred', 'Actual'))

##      Actual
## Pred  0  1
##    0 30  9
##    1 15 35

pred3_raw <- predict(nb_classification, heart[validset,], type='raw')
pred3_df <- data.frame(score = pred3_raw[, '1'], true.class = ifelse(heart[val
idset, 'target'] == '1', 1, 0))
ROC_func(pred3_df, 2, 1, color = 'orange') #removed add_on = T
```



```
## [1] 0.8449495
```

### Early Warning Model

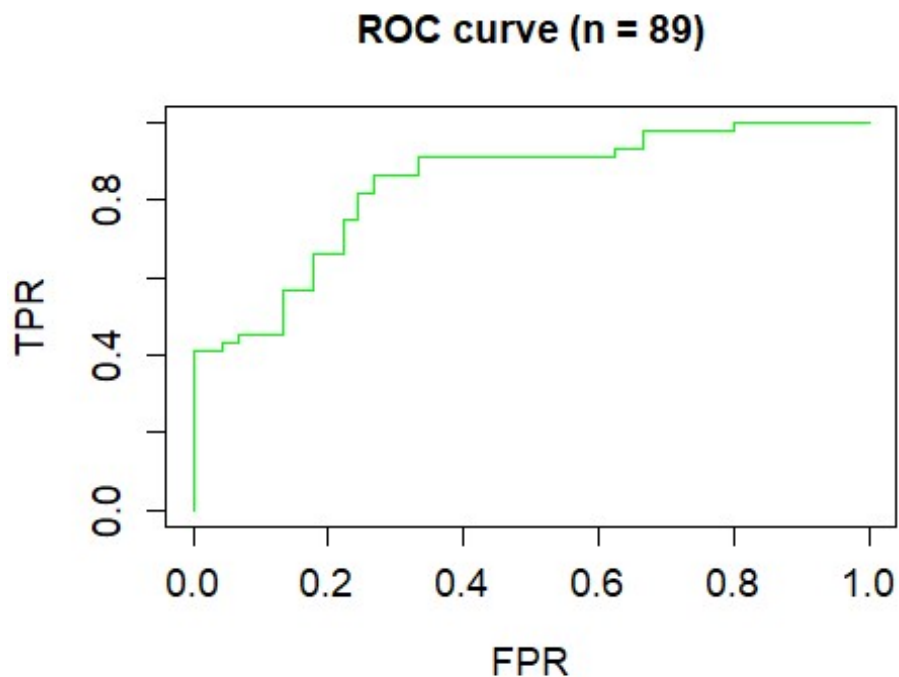
```
#Create model
nb_early_alarm <- naiveBayes(target~ age + sex + cp + trestbps + chol + fbs,
data = heart, subset = trainset)

#calculate predictions
pred2 <- predict(nb_early_alarm, heart[validset,])

#Model Accuracy
table(pred2, heart[validset,]$target,dnn = c('Pred','Actual'))

##      Actual
## Pred  0  1
##    0 37 16
##    1  8 28

pred2_raw <- predict(nb_early_alarm, heart[validset,],type='raw')
pred2_df <- data.frame(score = pred2_raw[, '1'], true.class = ifelse(heart[validset, 'target'] == '1', 1, 0))
ROC_func(pred2_df, 2, 1, color = 'green')
```



```
## [1] 0.8393939
```

## Logistic Models

### Confirmation Model

*#Model Creation, Predictors eliminated via backwards selection w/AIC & p-values*

```
md.log.conf <- glm(target ~ sex + cp + exang + thalach + oldpeak + ca + thal,
  data = heart, subset = trainset, family = "binomial")
```

*#Threshold manually selected to maximize AUC*

```
threshold <- 0.55
```

```
md.log.conf.class <- log_class(md.log.conf)
```

```
md.log.conf.score <- log_score(md.log.conf)
```

*#Model accuracy and AUC output*

```
table(md.log.conf.class, heart[validset,]$target,dnn = c('Pred', 'Actual'))
```

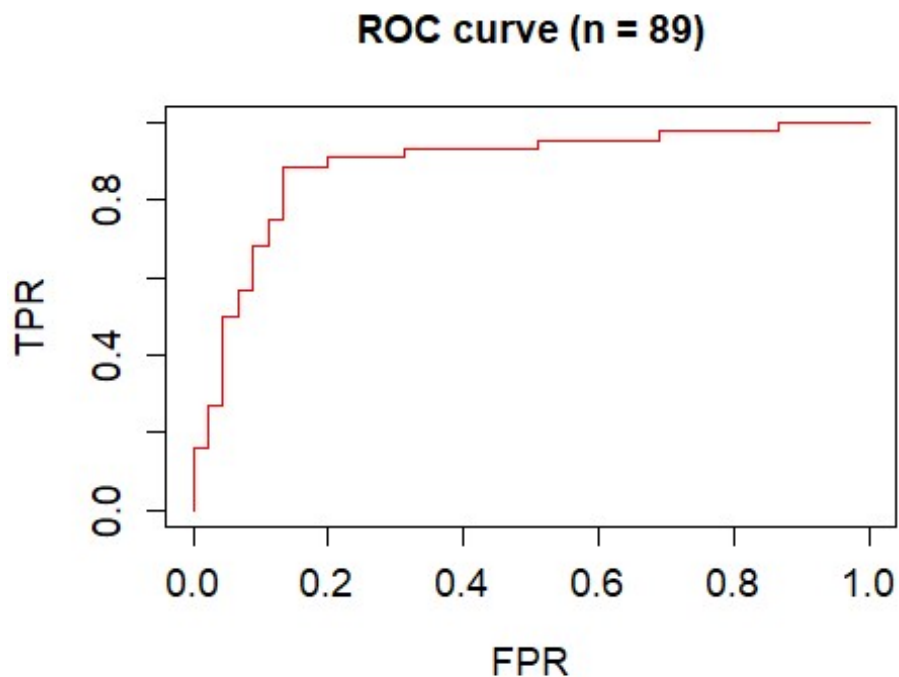
```
##      Actual
```

```
## Pred  0  1
```

```
##    0 39  6
```

```
##    1  6 38
```

```
(md.log.conf.AUC <- ROC_func(data.frame(md.log.conf.class,true.label = heart[
  validset,'target']), 1, 2 , color = 'red'))
```



```
## [1] 0.8883838
```

Classification Model

```
#Model Creation, Predictors eliminated via backwards selection w/AIC & p-values
```

```
md.log.clas <- glm(target ~ sex + cp + thalach + exang + oldpeak, data = heart,
  subset = trainset, family = "binomial")
```

```
#Threshold manually selected to maximize AUC
```

```
threshold <- 0.5
```

```
md.log.clas.class <- log_class(md.log.clas)
```

```
md.log.clas.score <- log_score(md.log.clas)
```

```
#Model accuracy and AUC output
```

```
table(md.log.clas.class, heart[validset,]$target,dnn = c('Pred','Actual'))
```

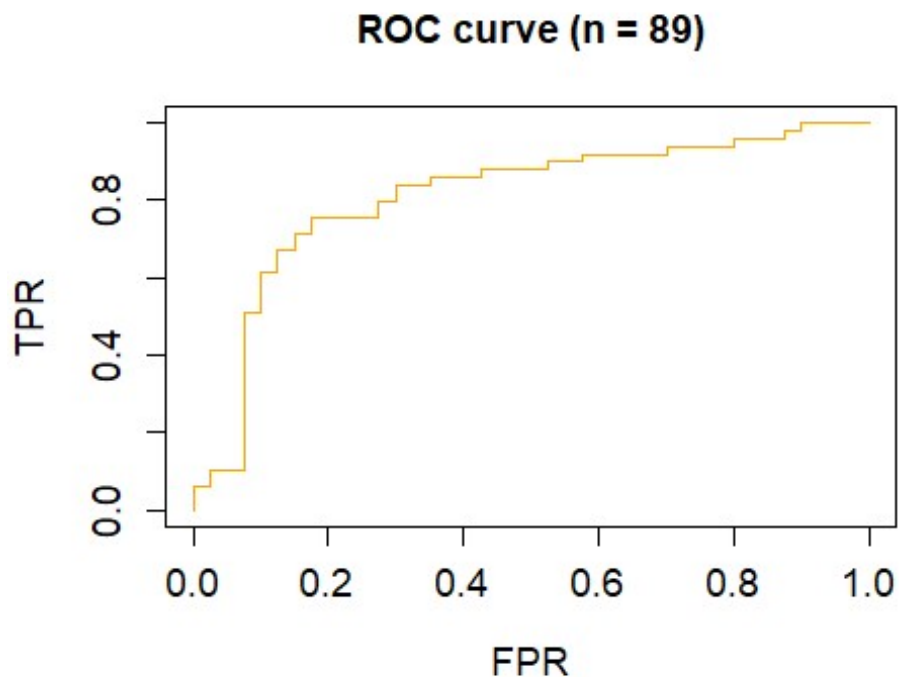
```
##      Actual
```

```
## Pred  0  1
```

```
##      0 33  7
```

```
##      1 12 37
```

```
(md.log.clas.AUC <- ROC_func(data.frame(md.log.clas.class,true.label = heart[
  validset,'target']), 1, 2 , color = 'orange'))
```



```
## [1] 0.8086735
```

Early Warning Model

```
#Model Creation, Predictors eliminated via backwards selection w/AIC & p-values
```

```
md.log.earl <- glm(target ~ age + sex + cp + trestbps, data = heart, subset =  
trainset, family = "binomial")
```

```
#Threshold manually selected to maximize AUC
```

```
threshold <- 0.65
```

```
md.log.earl.class <- log_class(md.log.earl)
```

```
md.log.earl.score <- log_score(md.log.earl)
```

```
#Model accuracy and AUC output
```

```
table(md.log.earl.class, heart[validset,]$target,dnn = c('Pred','Actual'))
```

```
##      Actual
```

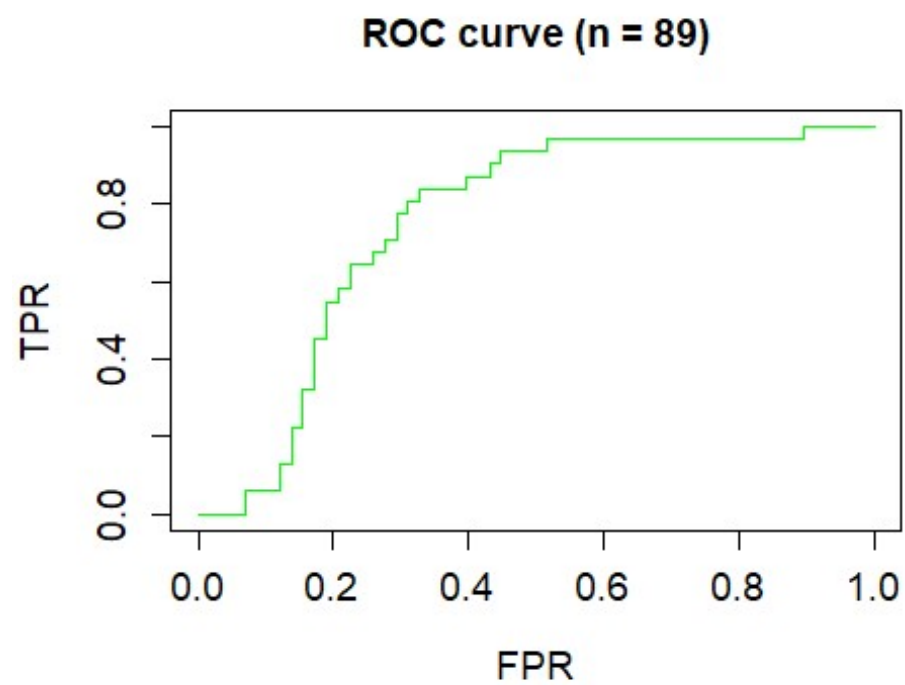
```
## Pred  0  1
```

```
##      0 39 19
```

```
##      1  6 25
```

```
(md.log.earl.AUC <- ROC_func(data.frame(md.log.earl.class,true.label = heart[  
validset,'target']), 1, 2 , color = 'green'))
```





```
## [1] 0.7541713
```