STAT 5703 Data Mining I - Winter 2018

FINAL PROJECT REPORT

Enrique Reveron 101066270

Michael Armanious 100978616

Alexander El-Hajj 100887389

Muneer Khan 100650973

April 17, 2018

Table of Contents

1. Introduction

This document contains the Final Course Project Report for STAT 4601/5703 Data Mining I - Winter 2018.

The main goal of the project is use different Data Mining methodologies in order to attempting to distinguish heart disease presence from absence, based on "Cleveland" Heart Disease Database.

The report include the following sections:

- Dataset
- Methodologies Implemented
- Main Summary/Interpretation of Findings
- Main Plots/Details Results

The complete R code along with its output is included as an Appendix as a separated file.

2. Dataset

The Dataset contains information related with heart disease diagnosis collected from "Cleveland Clinic Foundation". It could be download it from the following URL *http://archive.ics.uci.edu/ml/datasets/Heart+Disease.*

There are 303 observations on the following 13 variables (the last variable is the class identifier):

Variable	**Description**	**Type / Value**
Age	Age of the patient	Integer
Sex	Sex of the patient	0: female 1: male
CP	Chest pain type	1: Typical angina 2: Atypical angina 3: Non-anginal pain 4: Asymptomatic
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Integer

Chol	Cholesterol in mg/dl	Integer
Fbs	Fasting blood sugar > 120 mg/dl	0: false 1: true
Restecg	Resting electrocardiographic results	0: normal 1: having ST-T wave abnormality 2: showing left ventricular hypertrophy
Thalach	Maximum heart rate achieved	Integer
Exang	Exercise induced angina	0: no 1: yes
Oldpeak	ST depression induced by exercise relative to rest	numeric
Slope	The slope of the peak exercise ST segment	1: upsloping 2: flat 3: downsloping
Ca	Number of major vessels (0-3) colored by fluoroscopy	0-3
Thal	thalassemia	3: normal 6: fixed defect 7: reversible defect
Num	Diagnosis of heart disease (angiographic disease status)	0-4

Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (Num values 1,2,3,4) from absence (Num value 0), for this reason the following binary variable was created:

Variable	**Description**	**Type / Value**
Disease	Diagnosis of heart disease	0: healthy (Num = 0) 1: non-healthy (Num > 0)

The **Disease** variable is used as a predictor in the different methodologies implemented.

3. Data Mining Methodologies Implemented

The following include a list of the different Data Mining methodologies implemented in this project:

- Data Visualization
- Data Cleaning
- Data Splitting
- Data Preprocessing
- Cross-Validation
- Mining Association Rules (apriori)
- Logistic Regression
    - Quasibinomial (glm)
    - Binomial) (glm)
    - Binomial with Factor Variables (glm)
- Linear Model with Stepwise Feature Selection (glmStepAIC)
- Kmeans Clustering (eclust)
    - Using Principal Component Analysis (PCA)
    - All Data
- Support Vector Machine (SVM)
    - SVM Radial (svmRadial)
    - Linear SVM (svmLinear)
- Neural Networks (nnet)
- K-Nearest Neighbors (knn)
- Random Forest
    - Using Tuning Parameters (rf)
    - Boosted Tree (bstTree)
    - Boost with Tuning Parameters (gbm)
    - Stochastic Gradient Boost (gbm)
- Classification Tree (rpart)
- Fast-and-Frugal Decision Trees (FFTrees)
- Custom Fast-and-Frugal Decision Trees (FFTrees)
- Forest of Fast-and-Frugal Decision Trees (FFForest)

4. Methodology and Metrics Used to Compare Prediction Results

After cleaning the data (remove the NAs, less than 1% of the total data), we split the dataset into ⅔ for training and ⅓ for test set considering each value of the Num variable (0, 1, 2, 3, and 4).

In terms of data pre-processing, we consider to use the Standardized Data (Kmeans) and Centered and Scaled the data for the others (where apply). Where was possible we use cross-validation and different parameters in order to improve the accuracy of the models (tuning).

For each methodology we collect several parameters to be used to compare the results:

- Prediction Accuracy in Training Set
- Prediction Accuracy in Test Set
- Root Mean Square Error (RMSE)
- Residual Plots: plots of difference between predicted and class value
- Area under the Receiver Operating Characteristic (ROC) Curve
- Time Elapsed (sec): the computation time (seconds) that was necessary to build the model (train set)

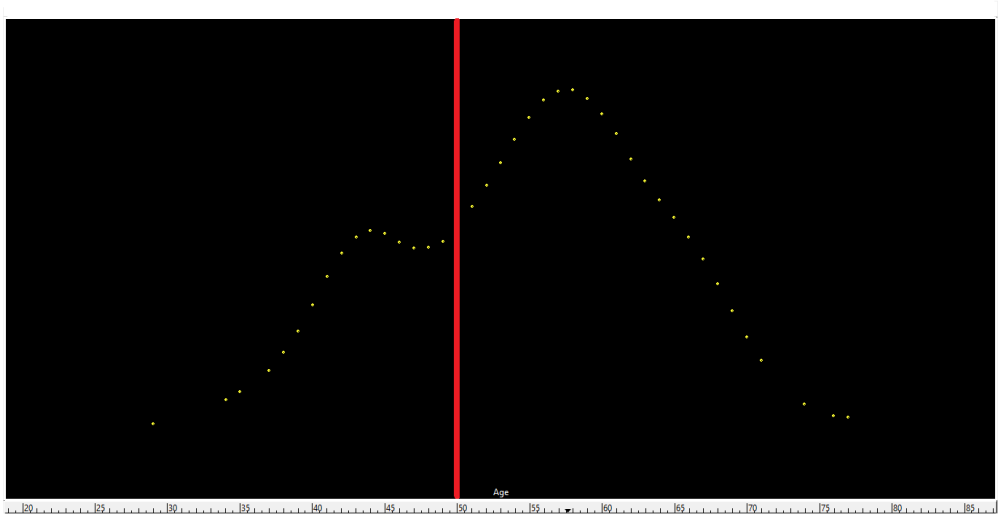
In the following section we will mention the most important findings for each Data Mining methodology implemented.

For details about each of them please refer to the Appendix.

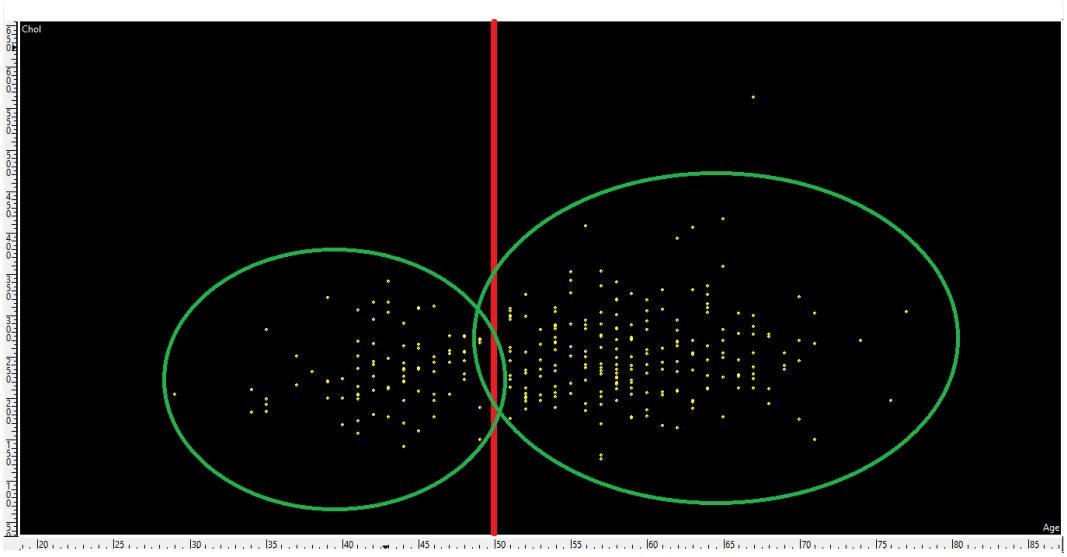
5. Main/Summary Interpretation of Findings

5.1 Data Visualization

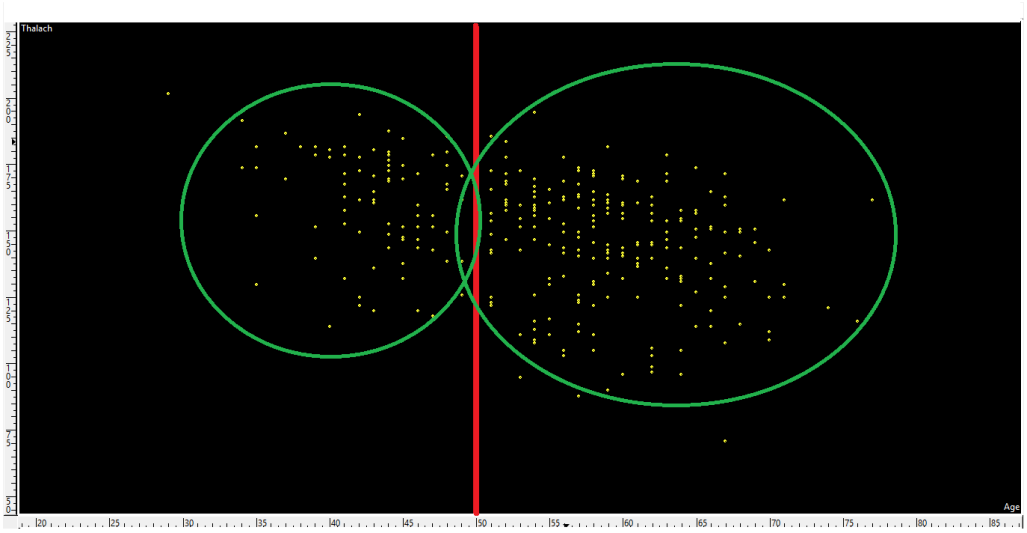
The majority of the individuals are over 50 years old. This proves to be an interesting threshold.



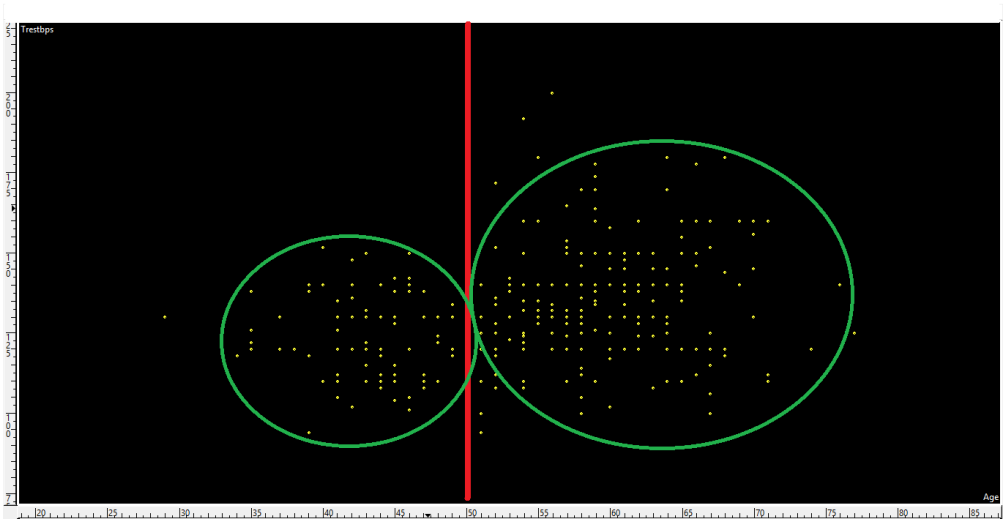
The figure below shows how after age 50, cholesterol levels increase



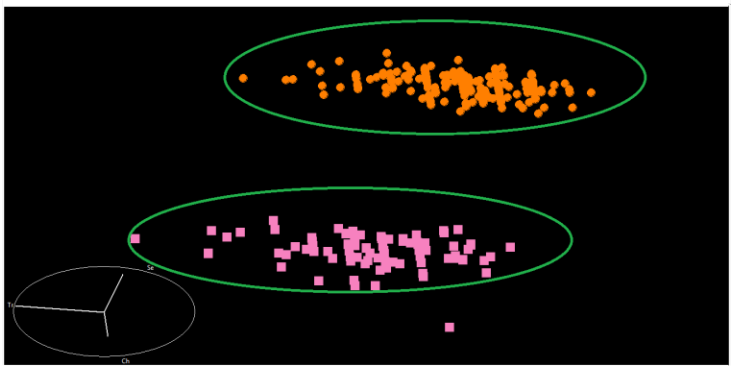
Heart rates go down after age 50



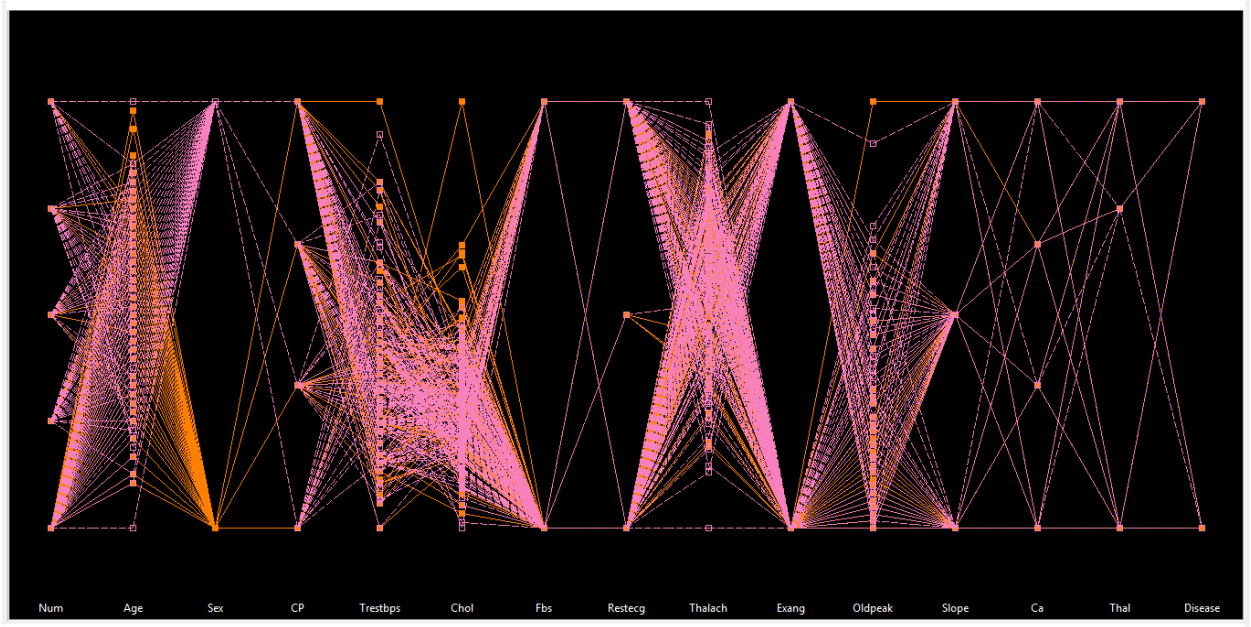
Blood pressure increases after age 50



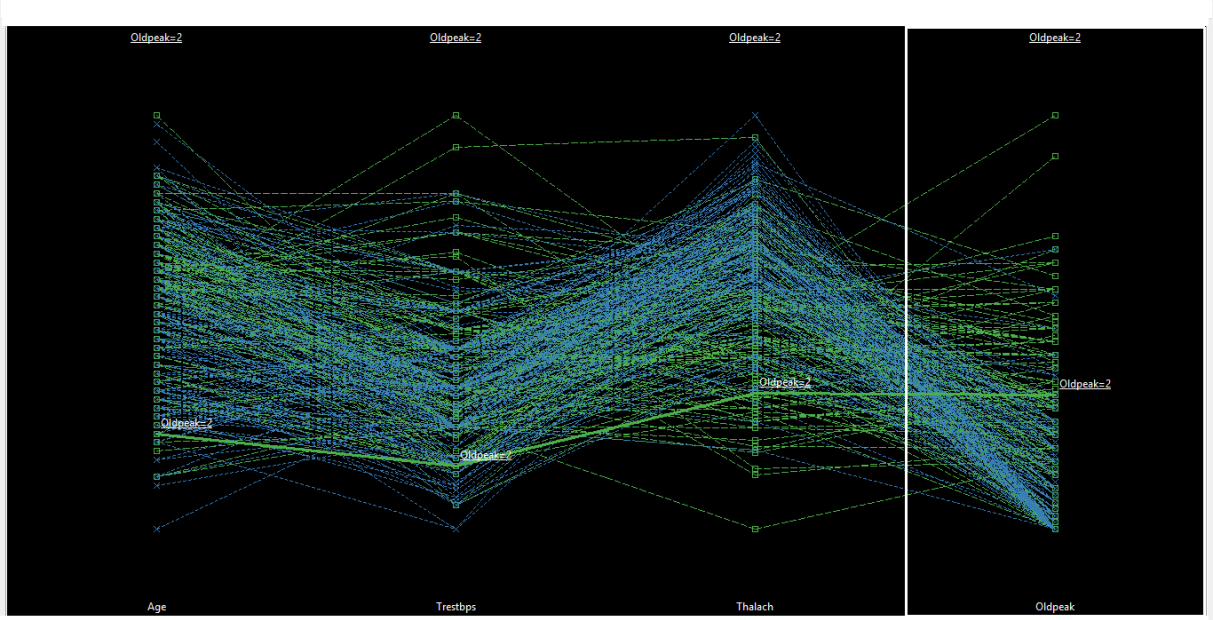
The graph below distinguishes gender, cholesterol and blood pressure.



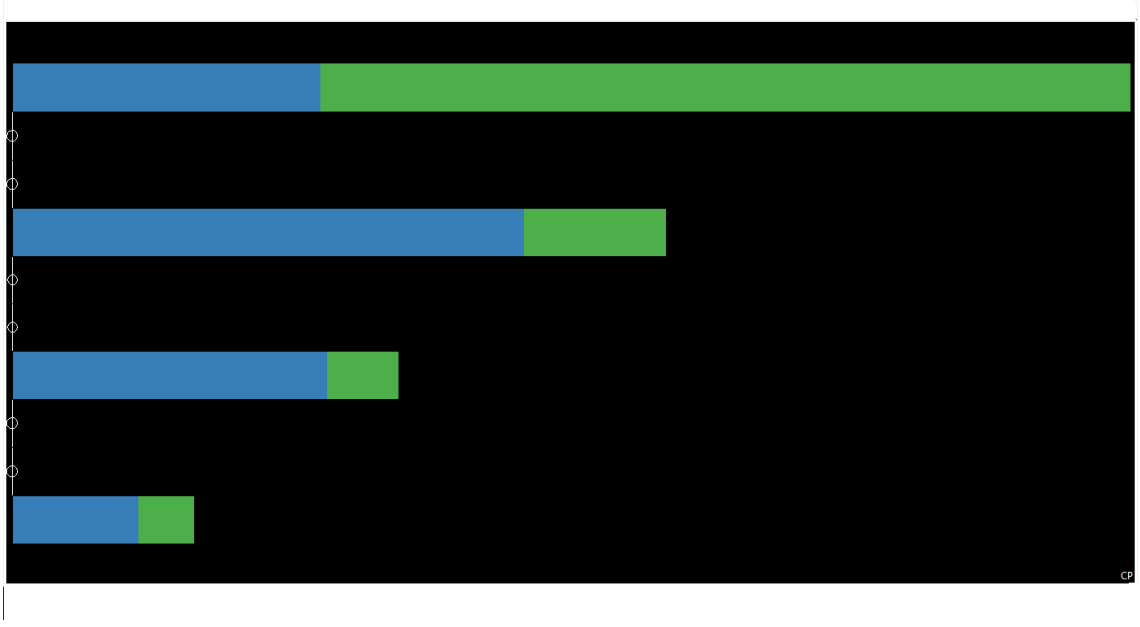
The figure below also helps show some distinction in cholesterol levels between genders. Males have some higher levels (orange)



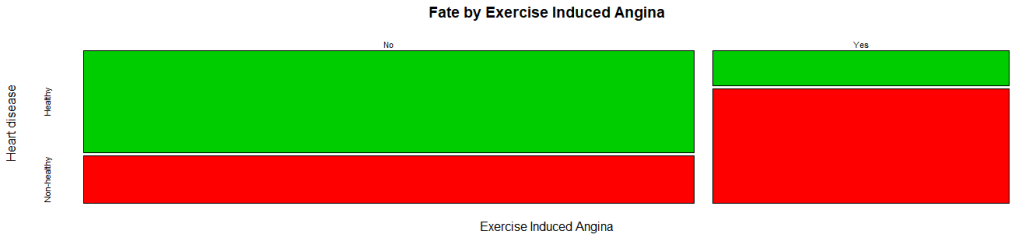
Individuals with heart disease are identifiable in the figure below. Generally, individuals without heart diseases are able to achieve higher max. heart rates vs. those with heart disease. Generally, individuals with heart disease have oldPeak (ST depression measurements) above 2.



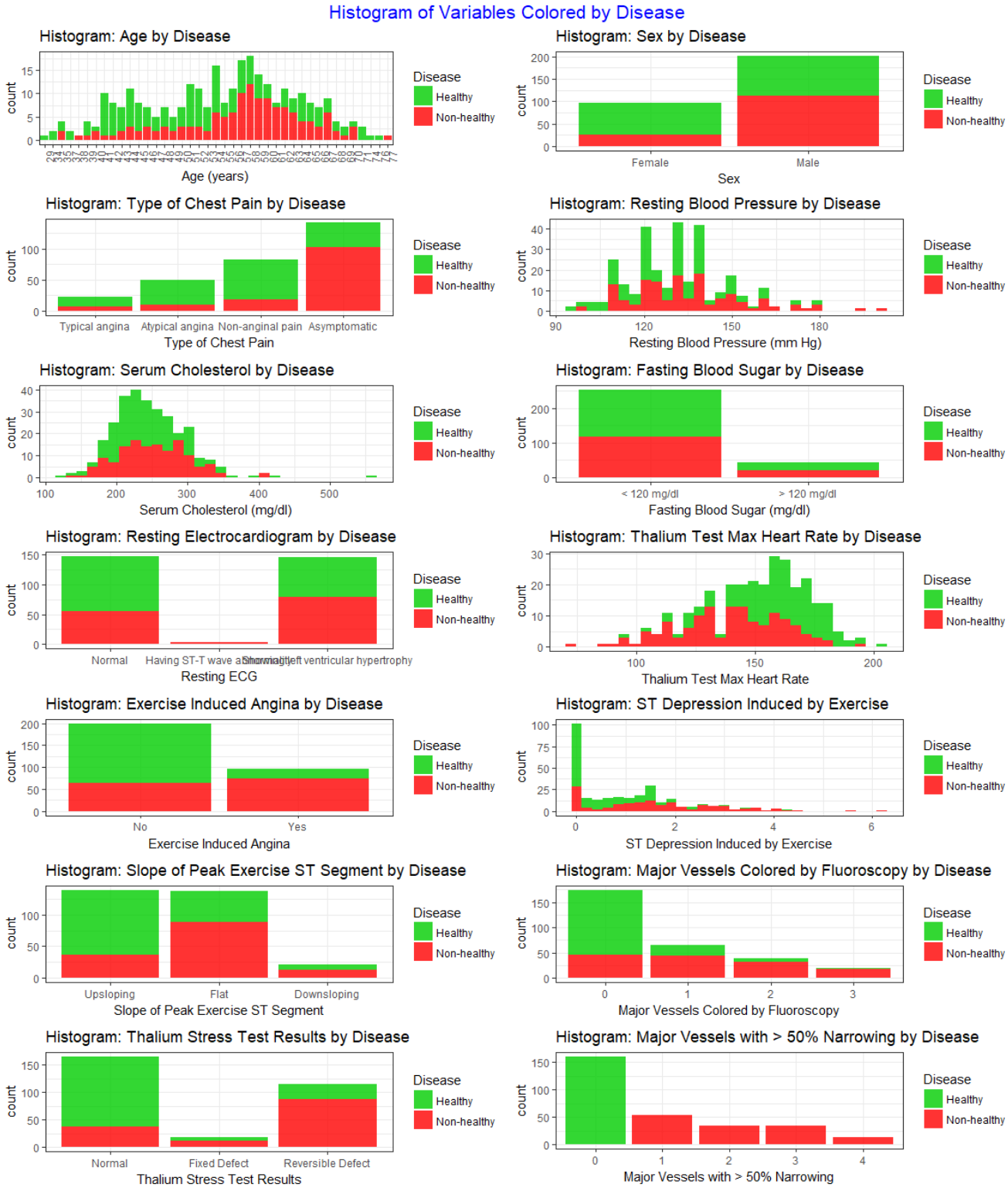
Individuals with heart disease (green in figure below) tend to not have chest pain symptoms (CP=4)



Exercise-induced angina plays a significant role as illustrated below



As expected, older individuals (50+) are more likely to have heart disease


Histogram of Variables Colored by Disease
Histogram: Age by Disease
count
Age (years)
Disease
Healthy
Non-healthy
Histogram: Sex by Disease
Female
Male
Sex
Histogram: Type of Chest Pain by Disease
Typical angina
Atypical angina
Non-anginal pain
Asymptomatic
Type of Chest Pain
Histogram: Resting Blood Pressure by Disease
Resting Blood Pressure (mm Hg)
Histogram: Serum Cholesterol by Disease
Serum Cholesterol (mg/dl)
Histogram: Fasting Blood Sugar by Disease
< 120 mg/dl
> 120 mg/dl
Fasting Blood Sugar (mg/dl)
Histogram: Resting Electrocardiogram by Disease
Normal
Resting ECG
Histogram: Thalium Test Max Heart Rate by Disease
Thalium Test Max Heart Rate
Histogram: Exercise Induced Angina by Disease
No
Yes
Exercise Induced Angina
Histogram: ST Depression Induced by Exercise
ST Depression Induced by Exercise
Histogram: Slope of Peak Exercise ST Segment by Disease
Upsloping
Flat
Downsloping
Slope of Peak Exercise ST Segment
Histogram: Major Vessels Colored by Fluoroscopy by Disease
Major Vessels Colored by Fluoroscopy
Histogram: Thalium Stress Test Results by Disease
Fixed Defect
Reversible Defect
Thalium Stress Test Results
Histogram: Major Vessels with > 50% Narrowing by Disease
Major Vessels with > 50% Narrowing

5.2 Data Cleaning

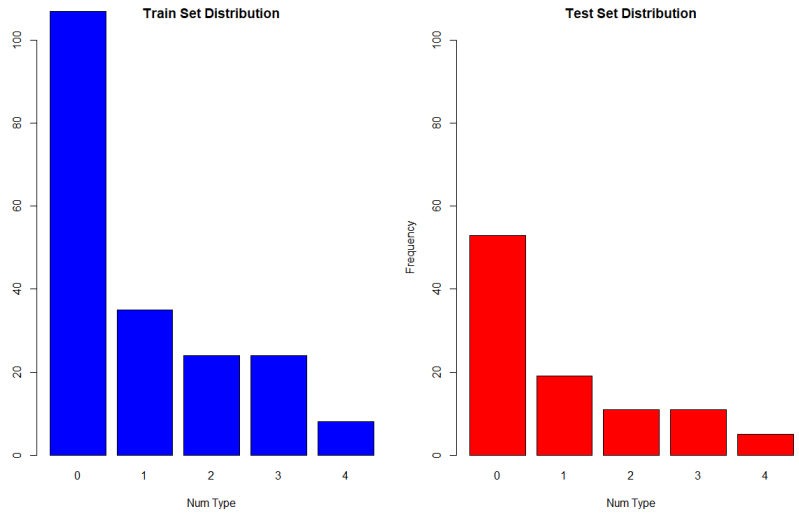
We found that the dataset contains 6 observations with NAs (related with Thal and Ca), because those observations was less than 2% of the data (6/303 = 1.98%), we considered to remove it from the analysis. The final dataset include 297 observations.

5.3 Data Preprocessing, Splitting and Cross-Validation

In terms of data preprocessing, mainly we use the following two methods:

- Standardized the data for Kmeans and
- Centered and Scaled the data for the other method (if apply)

We split the dataset into ⅔ for training and ⅓ for test set considering each value of the Num variable (0, 1, 2, 3, and 4), the following picture shows the distribution for each of the sets:



A cross-validation feature was implemented with the following parameters (if apply):

- 10 kfolds
- Repeated 10 times

5.3 Mining Association Rules (apriori)

In order to do association mining it would be best that we make all the variables discrete. Therefore, I created a new datafile doing just that with intervals of non-discrete variables assigned to a discrete value
.

Variable	**Value**
Age	29 - 40 corresponds to 1 41 - 53 corresponds to 2 54 - 65 corresponds to 3 66 - 77 corresponds to 4
Sex	0 corresponds to female 1 corresponds to male
Chest Pain Type (CP)	1 corresponds to typical angina 2 corresponds to atypical angina 3 corresponds to non-anginal pain 4 corresponds to asymptomatic
Resting blood pressure (in mm Hg on admission to the hospital) Trestbps	70 - 90 (Low) corresponds to 1 90 - 120 (Ideal) corresponds to 2 120 - 140 (Pre-High bp) corresponds to 3 140 - 200 (High bp) corresponds to 4
Cholesterol (Chol)	126 - 199 (Good) corresponds to 1 200 - 239 (Borderline) corresponds to 2 240+ (High) corresponds to 3
Fasting Blood Sugar > 120 mg/dl (FBS)	0 if false 1 if true
resting electrocardiographic results (Restecg)	0 if normal 1 if having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2 if showing probable or definite left ventricular hypertrophy by Estes' criteria
Max Heart Rate Achieved (Thalach range)	70 - 119 corresponds to 1 120 - 150 corresponds to 2 160 - 200 corresponds to 3 200+ corresponds to 4
Exercise induced Angina (Exang)	0 if no 1 if yes

ST depression induced by exercise relative to rest (Oldpeak)	0 - 1.9 corresponds to 1 2.0 - 3.9 corresponds to 2 4.0 - 6.2 corresponds to 3
The slope of the peak exercise ST segment (Slope)	1 if upsloping 2 if flat 3 if downsloping
Number of major vessels (0-3) colored by flouroscopy (Ca)	0 depending on color of fluoroscopy 1 2 3
Thalium stress test result (Thal)	3 if normal 6 if fixed defect 7 if reversible defect
Diagnosis of heart disease (angiographic disease status) (Disease)	0 if < 50% diameter narrowing 1 if > 50% diameter narrowing

After doing these changes to the dataset we can proceed with the association mining method. In this dataset we assume that Disease = 0 means the heart is okay and Disease = 1 means heart disease.

For our procedure we want to see the associations that lead to either Disease=0 or Disease=1
So we use R to compute our set of rules and specify the right hand side to be Disease = 0.
This gave us a set of 859 rules. We will show the top 10 sorted by support and also lift.

Sorting by count/support gives us this as our top 10 rules

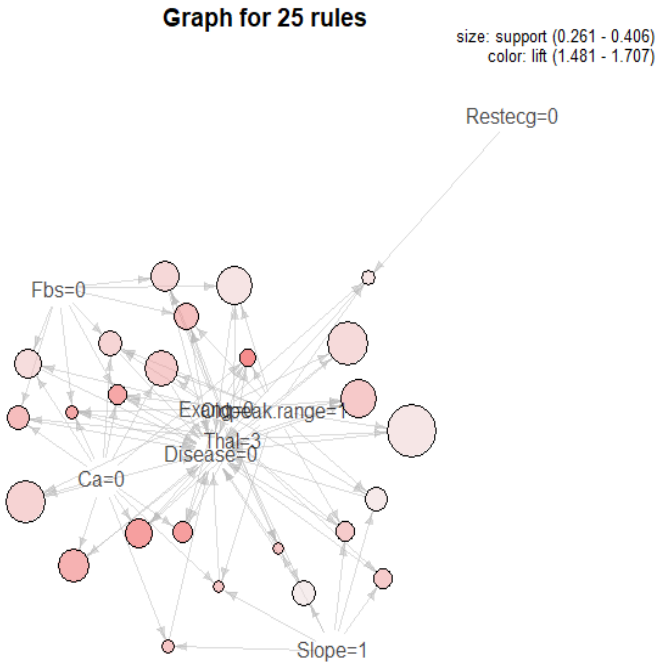
```
lhs                     rhs          support confidence  lift count
{Oldpeak.range=1,
 Thal=3}             => {Disease=0}  0.4059     0.8146 1.505   123
{Exang=0,
 Thal=3}             => {Disease=0}  0.3729     0.8370 1.546   113
{Exang=0,
 Ca=0}               => {Disease=0}  0.3696     0.8485 1.568   112
{Exang=0,
 Oldpeak.range=1,
 Thal=3}             => {Disease=0}  0.3564     0.8640 1.596   108
{Fbs=0,
 Oldpeak.range=1,
 Thal=3}             => {Disease=0}  0.3564     0.8182 1.512   108
{Exang=0,
 Oldpeak.range=1,
 Ca=0}               => {Disease=0}  0.3465     0.8607 1.590   105
{Ca=0,
 Thal=3}             => {Disease=0}  0.3366     0.8870 1.639   102
{Fbs=0,
 Exang=0,
 Thal=3}             => {Disease=0}  0.3267     0.8390 1.550    99
{Fbs=0,
 Exang=0,
 Ca=0}               => {Disease=0}  0.3234     0.8305 1.534    98
{Oldpeak.range=1,
 Ca=0,
 Thal=3}             => {Disease=0}  0.3201     0.9065 1.675    97
```

As we can see, the number one rules says that if a person's ST depression induced by exercise relative to rest is between 0 - 1.9 and their Thalium stress test result shows normal then 123 of the total amount do not have heart disease with high support, confidence and lift. The confidence shows that out of all those who had the combination of the previously mentioned predictors 81% of them do not have heart disease and all lifts >1.5, which makes those rules potentially useful for predicting the consequent in future data sets.

The next three are combinations of different predictors along with no exercise induced angina. In this top 10 the predictors are all combinations of:

- ST depression induced by exercise relative to rest is between 0 - 1.9
- Thalium stress test result shows normal
- No exercise induced angina
- 0 major vessels coloured by a fluoroscopy
- Fasting Blood Sugar is less than 120 mg/dl

Which clearly makes sense in this context.

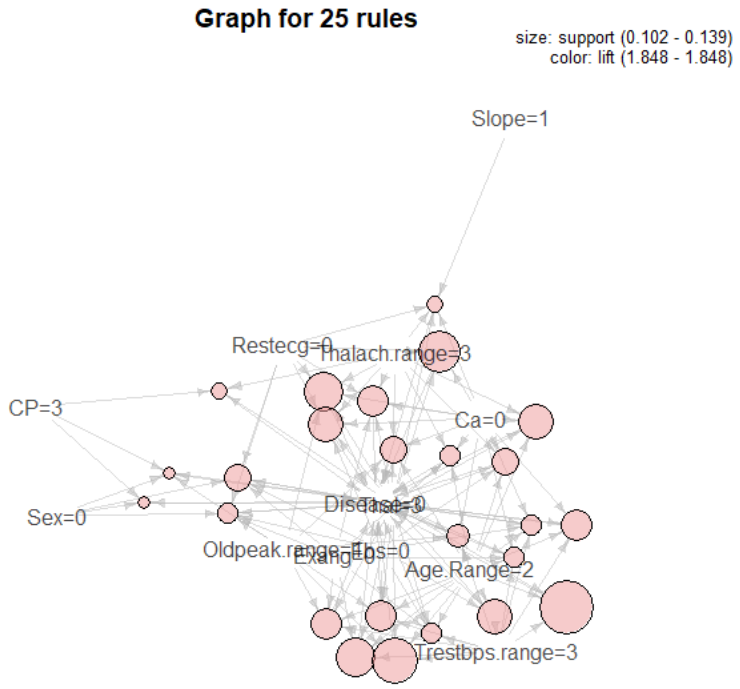


Here is a visualization of the top 25 rules sorted by support, all leading to no heart disease. The bigger the circle, the higher the support and the darker the colour, the higher the lift.

Now sorting by confidence or lift we get

lhs		rhs	support	confidence	lift	cou
{Sex=0,CP=3,Thal=3}	=>	{Disease=0}	0.1023102	1.0000000	1.847561	31
{CP=3,Thalach.range=3,Thal=3}	=>	{Disease=0}	0.1056106	1.0000000	1.847561	32
{Age.Range=2,Trestbps.range=3,Thal=3}	=>	{Disease=0}	0.1386139	1.0000000	1.847561	42
{Sex=0,CP=3,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.1023102	1.0000000	1.847561	31
{Age.Range=2,Thalach.range=3,Ca=0,Thal=3}	=>	{Disease=0}	0.1221122	1.0000000	1.847561	37
{Age.Range=2,Trestbps.range=3,Ca=0,Thal=3}	=>	{Disease=0}	0.1188119	1.0000000	1.847561	36
{Age.Range=2,Trestbps.range=3,Exang=0,Thal=3}	=>	{Disease=0}	0.1254125	1.0000000	1.847561	38
{Age.Range=2,Trestbps.range=3,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.1320132	1.0000000	1.847561	40
{Age.Range=2,Trestbps.range=3,Fbs=0,Thal=3}	=>	{Disease=0}	0.1221122	1.0000000	1.847561	37
{Restecg=0,Thalach.range=3,Ca=0,Thal=3}	=>	{Disease=0}	0.1287129	1.0000000	1.847561	39

This means that 100% of the people with these rules do not have heart disease (more than 10% of the dataset). The lift tells us that these associations are significant.



This is a visualization of the top 25 rules sorted by lift, all leading to no heart disease.
Next we will take a look at the rules that lead to heart disease sorted by support/count.

lhs		rhs	support	confidence	lift	count
{CP=4,Thal=7}	=>	{Disease=1}	0.2343234	0.9102564	1.984228	71
{CP=4,Exang=1}	=>	{Disease=1}	0.2310231	0.8750000	1.907374	70
{CP=4,Slope=2}	=>	{Disease=1}	0.2244224	0.8095238	1.764645	68
{Age.Range=3,CP=4}	=>	{Disease=1}	0.2112211	0.8000000	1.743885	64
{Sex=1,Exang=1}	=>	{Disease=1}	0.2046205	0.8051948	1.755209	62
{Slope=2,Thal=7}	=>	{Disease=1}	0.1980198	0.8571429	1.868448	60
{Sex=1,CP=4,Thal=7}	=>	{Disease=1}	0.1947195	0.8939394	1.948659	59
{CP=4,Fbs=0,Thal=7}	=>	{Disease=1}	0.1947195	0.9076923	1.978639	59
{CP=4,Fbs=0,Exang=1}	=>	{Disease=1}	0.1914191	0.8529412	1.859289	58
{Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1848185	0.8115942	1.769159	56
{Sex=1,CP=4,Exang=1}	=>	{Disease=1}	0.1848185	0.9032258	1.968902	56
{Exang=1,Slope=2}	=>	{Disease=1}	0.1815182	0.8593750	1.873314	55
{Exang=1,Thal=7}	=>	{Disease=1}	0.1749175	0.8983051	1.958176	53
{Fbs=0,Slope=2,Thal=7}	=>	{Disease=1}	0.1749175	0.8548387	1.863425	53
{Chol.range=3,Thal=7}	=>	{Disease=1}	0.1683168	0.8225806	1.793107	51
{Sex=1,CP=4,Slope=2}	=>	{Disease=1}	0.1683168	0.8793103	1.916770	51
{Sex=1,CP=4,Restecg=2}	=>	{Disease=1}	0.1650165	0.8620690	1.879186	50
{Sex=1,CP=4,Fbs=0,Thal=7}	=>	{Disease=1}	0.1650165	0.8928571	1.946300	50
{Restecg=2,Thal=7}	=>	{Disease=1}	0.1617162	0.8448276	1.841603	49
{CP=4,Exang=1,Thal=7}	=>	{Disease=1}	0.1617162	0.9423077	2.054095	49
{CP=4,Exang=1,Slope=2}	=>	{Disease=1}	0.1617162	0.9074074	1.978018	49
{Sex=1,Slope=2,Thal=7}	=>	{Disease=1}	0.1617162	0.8448276	1.841603	49
{Chol.range=3,Exang=1}	=>	{Disease=1}	0.1584158	0.8275862	1.804019	48
{Sex=1,CP=4,Thalach.range=2}	=>	{Disease=1}	0.1584158	0.8571429	1.868448	48
{Sex=1,CP=4,Fbs=0,Exang=1}	=>	{Disease=1}	0.1584158	0.8888889	1.937650	48

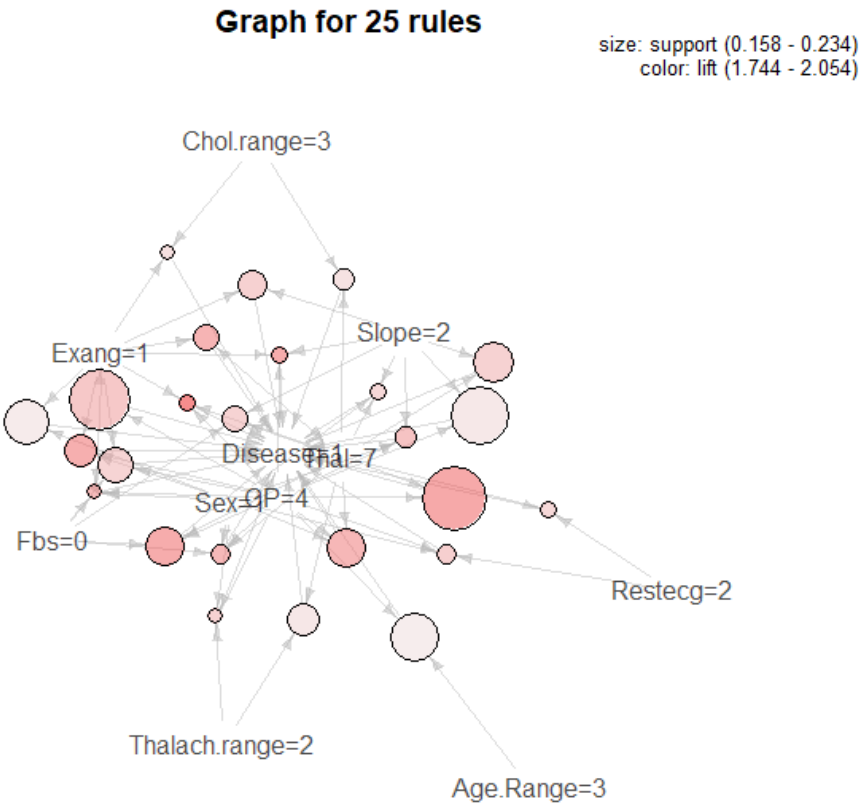
The rule with the highest support/count is an asymptotic chest pain type and a reversible defect result for the thalium stress test. This rule has a high lift and confidence as well. 91% of the people with this

combination have heart disease and all lifts are greater than 1.7 meaning these rules are potentially useful for predicting the consequent in future data sets.

Some of the predictors are combinations of:

- Asymptotic chest pain type
- reversible defect result for the thalium stress test
- Exercise induced angina
- Ages 54-65
- Male
- Fasting Blood sugar is less than 120 mg/dl
- Maximum heart rate achieved is from 120-150

The predictor that comes up most is an asymptotic chest pain type (CP=4). We can also see that females do not come up in the top 25 which could possibly suggest that females do not encounter heart disease as much and relatively healthier than males. A variable that both sets of rules have in common is having a fasting blood sugar less than 120mg/dl. This could suggest that fasting blood sugar is solely not a clear predictor for determining whether someone is healthy or not. However when it is combined with other predictors, it seems to give a high supported rule with high confidence and lift as well. It depends on the predictors that it is associated with in order to determine whether it possibly leads to heart disease or not.



Here is a visual of the top 25 rules by support.

Sorting by lift we get

lhs		rhs	support	confidence	lift	count
{CP=4,Thalach.range=2,Exang=1,Thal=7}	=>	{Disease=1}	0.1155116	1.0000000	2.179856	35
{CP=4,Thalach.range=2,Slope=2,Thal=7}	=>	{Disease=1}	0.1089109	1.0000000	2.179856	33
{CP=4,Restecg=2,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1056106	1.0000000	2.179856	32
{CP=4,Chol.range=3,Thal=7}	=>	{Disease=1}	0.1353135	0.9761905	2.127955	41
{CP=4,Restecg=2,Thal=7}	=>	{Disease=1}	0.1320132	0.9756098	2.126689	40
{CP=4,Fbs=0,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1287129	0.9750000	2.125360	39
{Sex=1,CP=4,Chol.range=3,Thal=7}	=>	{Disease=1}	0.1089109	0.9705882	2.115743	33
{CP=4,Chol.range=3,Fbs=0,Thal=7}	=>	{Disease=1}	0.1089109	0.9705882	2.115743	33
{CP=4,Exang=1,Slope=2,Thal=7}	=>	{Disease=1}	0.1056106	0.9696970	2.113800	32
{Sex=1,CP=4,Restecg=2,Thal=7}	=>	{Disease=1}	0.1056106	0.9696970	2.113800	32
{CP=4,Fbs=0,Restecg=2,Thal=7}	=>	{Disease=1}	0.1056106	0.9696970	2.113800	32
{Sex=1,CP=4,Ca=1}	=>	{Disease=1}	0.1023102	0.9687500	2.111736	31
{CP=4,Trestbps.range=4,Thal=7}	=>	{Disease=1}	0.1023102	0.9687500	2.111736	31
{Sex=1,CP=4,Fbs=0,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1023102	0.9687500	2.111736	31
{CP=4,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1551155	0.9591837	2.090882	47
{CP=4,Slope=2,Thal=7}	=>	{Disease=1}	0.1518152	0.9583333	2.089029	46
{CP=4,Fbs=0,Slope=2,Thal=7}	=>	{Disease=1}	0.1320132	0.9523810	2.076053	40
{Thalach.range=2,Exang=1,Thal=7}	=>	{Disease=1}	0.1221122	0.9487179	2.068069	37
{Sex=1,CP=4,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1221122	0.9487179	2.068069	37
{Exang=1,Slope=2,Thal=7}	=>	{Disease=1}	0.1188119	0.9473684	2.065127	36
{Sex=1,CP=4,Slope=2,Thal=7}	=>	{Disease=1}	0.1188119	0.9473684	2.065127	36
{Sex=1,CP=4,Thalach.range=2,Exang=1}	=>	{Disease=1}	0.1155116	0.9459459	2.062026	35
{Chol.range=3,Exang=1,Thal=7}	=>	{Disease=1}	0.1089109	0.9428571	2.055293	33
{Sex=1,CP=4,Chol.range=3,Exang=1}	=>	{Disease=1}	0.1089109	0.9428571	2.055293	33
{CP=4,Exang=1,Thal=7}	=>	{Disease=1}	0.1617162	0.9423077	2.054095	49

This shows a very high lift of greater than 2 for the top 25 rules which means the association is very significant. For the first two, we see that 100% of the people with asymptotic chest pain, Maximum heart rate achieved is from 120-150, exercise induced angina, and reversible defect result for the thalium stress test have heart disease (which is more than 10% of the dataset) and 100% of the people with asymptotic chest pain, Maximum heart rate achieved is from 120-150,a flat slope of the peak exercise ST segment, and reversible defect result for the thalium stress test have heart disease (again over 10% of the dataset).

Now we will validate this with a training and test set.

This is the top 10 rules of the training dataset that lead to no heart disease sorted by support.

lhs		rhs	support	confidence	lift	count
{Oldpeak.range=1,Ca=0}	=>	{Disease=0}	0.4088670	0.8058252	1.487114	83
{Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.3891626	0.8229167	1.518655	79
{Exang=0,Ca=0}	=>	{Disease=0}	0.3891626	0.8494624	1.567644	79
{Exang=0,Oldpeak.range=1,Ca=0}	=>	{Disease=0}	0.3645320	0.8705882	1.606631	74
{Exang=0,Thal=3}	=>	{Disease=0}	0.3596059	0.8488372	1.566490	73
{Exang=0,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.3448276	0.8860759	1.635213	70
{Fbs=0,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.3448276	0.8045977	1.484848	70
{Fbs=0,Exang=0,Ca=0}	=>	{Disease=0}	0.3448276	0.8333333	1.537879	70
{Fbs=0,Exang=0,Oldpeak.range=1,Ca=0}	=>	{Disease=0}	0.3300493	0.8589744	1.585198	67
{Ca=0,Thal=3}	=>	{Disease=0}	0.3251232	0.8684211	1.602632	66

This is the top 10 rules of the test dataset that lead to no heart disease sorted by support.

lhs		rhs	support	confidence	lift	count
{Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.44	0.8000000	1.481481	44
{Exang=0,Thal=3}	=>	{Disease=0}	0.40	0.8163265	1.511716	40
{Exang=0,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.38	0.8260870	1.529791	38
{Fbs=0,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.38	0.8444444	1.563786	38
{Slope=1}	=>	{Disease=0}	0.36	0.8000000	1.481481	36
{Oldpeak.range=1,Slope=1}	=>	{Disease=0}	0.36	0.8000000	1.481481	36
{Ca=0,Thal=3}	=>	{Disease=0}	0.36	0.9230769	1.709402	36
{Fbs=0,Exang=0,Thal=3}	=>	{Disease=0}	0.34	0.8500000	1.574074	34
{Exang=0,Ca=0}	=>	{Disease=0}	0.33	0.8461538	1.566952	33
{Oldpeak.range=1,Ca=0,Thal=3}	=>	{Disease=0}	0.33	0.9428571	1.746032	33

Sorting the training set by lift we get:

lhs		rhs	support	confidence	lift	count
{CP=3,Thalach.range=3,Oldpeak.range=1}	=>	{Disease=0}	0.1133005	1	1.845455	23
{CP=3,Trestbps.range=3,Thal=3}	=>	{Disease=0}	0.1083744	1	1.845455	22
{Restecg=0,Thalach.range=3,Ca=0}	=>	{Disease=0}	0.1379310	1	1.845455	28
{Age.Range=2,Trestbps.range=3,Thal=3}	=>	{Disease=0}	0.1428571	1	1.845455	29
{CP=3,Thalach.range=3,Exang=0,Oldpeak.range=1}	=>	{Disease=0}	0.1034483	1	1.845455	21
{CP=3,Trestbps.range=3,Exang=0,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{CP=3,Trestbps.range=3,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{Sex=0,Restecg=0,Exang=0,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{Age.Range=2,Trestbps.range=3,Thalach.range=3,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{Age.Range=2,Thalach.range=3,Ca=0,Thal=3}	=>	{Disease=0}	0.1182266	1	1.845455	24

The test set sorted by lift:

lhs		rhs	support	confidence	lift	count
{CP=3,Thalach.range=3,Oldpeak.range=1}	=>	{Disease=0}	0.1133005	1	1.845455	23
{CP=3,Trestbps.range=3,Thal=3}	=>	{Disease=0}	0.1083744	1	1.845455	22
{Restecg=0,Thalach.range=3,Ca=0}	=>	{Disease=0}	0.1379310	1	1.845455	28
{Age.Range=2,Trestbps.range=3,Thal=3}	=>	{Disease=0}	0.1428571	1	1.845455	29
{CP=3,Thalach.range=3,Exang=0,Oldpeak.range=1}	=>	{Disease=0}	0.1034483	1	1.845455	21
{CP=3,Trestbps.range=3,Exang=0,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{CP=3,Trestbps.range=3,Oldpeak.range=1,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{Sex=0,Restecg=0,Exang=0,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{Age.Range=2,Trestbps.range=3,Thalach.range=3,Thal=3}	=>	{Disease=0}	0.1034483	1	1.845455	21
{Age.Range=2,Thalach.range=3,Ca=0,Thal=3}	=>	{Disease=0}	0.1182266	1	1.845455	24

We can see that the tables are very similar and have many of the same predictor combinations which means that our association rules that lead to no heart disease are validated.

The top 10 rules in the training set sorted by support that lead to heart disease are shown below:

lhs		rhs	support	confidence	lift	count
{Sex=1,CP=4}	=>	{Disease=1}	0.2758621	0.8000000	1.746237	56
{CP=4,Thal=7}	=>	{Disease=1}	0.2512315	0.9272727	2.024047	51
{CP=4,Exang=1}	=>	{Disease=1}	0.2266010	0.9200000	2.008172	46
{CP=4,Slope=2}	=>	{Disease=1}	0.2216749	0.8181818	1.785924	45
{CP=4,Restecg=2}	=>	{Disease=1}	0.2216749	0.8181818	1.785924	45
{CP=4,Fbs=0,Thal=7}	=>	{Disease=1}	0.2118227	0.9347826	2.040439	43
{Slope=2,Thal=7}	=>	{Disease=1}	0.2019704	0.8367347	1.826421	41
{Sex=1,CP=4,Thal=7}	=>	{Disease=1}	0.2019704	0.9111111	1.988769	41
{Sex=1,Exang=1}	=>	{Disease=1}	0.1871921	0.8085106	1.764814	38
{Restecg=2,Thal=7}	=>	{Disease=1}	0.1871921	0.8260870	1.803179	38

For the test set, we have:

lhs		rhs	support	confidence	lift	count
{Exang=1}	=>	{Disease=1}	0.32	0.8421053	1.830664	32
{CP=4,Exang=1}	=>	{Disease=1}	0.29	0.9062500	1.970109	29
{Sex=1,CP=4}	=>	{Disease=1}	0.29	0.8285714	1.801242	29
{Fbs=0,Exang=1}	=>	{Disease=1}	0.28	0.8235294	1.790281	28
{CP=4,Fbs=0,Exang=1}	=>	{Disease=1}	0.26	0.8965517	1.949025	26
{Sex=1,CP=4,Fbs=0}	=>	{Disease=1}	0.26	0.8125000	1.766304	26
{Sex=1,Exang=1}	=>	{Disease=1}	0.25	0.8620690	1.874063	25
{CP=4,Thal=7}	=>	{Disease=1}	0.25	0.9259259	2.012882	25
{CP=4,Slope=2}	=>	{Disease=1}	0.25	0.8333333	1.811594	25
{CP=4,Chol.range=3}	=>	{Disease=1}	0.24	0.8000000	1.739130	24

Sorting the training set by lift we have:

lhs		rhs	support	confidence	lift	count
{Oldpeak.range=2,Thal=7}	=>	{Disease=1}	0.1034483	1.0000000	2.182796	21
{CP=4,Trestbps.range=4,Thal=7}	=>	{Disease=1}	0.1133005	1.0000000	2.182796	23
{CP=4,Restecg=2,Thal=7}	=>	{Disease=1}	0.1527094	1.0000000	2.182796	31
{CP=4,Thalach.range=2,Exang=1,Thal=7}	=>	{Disease=1}	0.1182266	1.0000000	2.182796	24
{CP=4,Restecg=2,Slope=2,Thal=7}	=>	{Disease=1}	0.1083744	1.0000000	2.182796	22
{CP=4,Thalach.range=2,Slope=2,Thal=7}	=>	{Disease=1}	0.1182266	1.0000000	2.182796	24
{CP=4,Restecg=2,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1182266	1.0000000	2.182796	24
{Sex=1,CP=4,Restecg=2,Thal=7}	=>	{Disease=1}	0.1182266	1.0000000	2.182796	24
{CP=4,Fbs=0,Restecg=2,Thal=7}	=>	{Disease=1}	0.1231527	1.0000000	2.182796	25
{CP=4,Chol.range=3,Thalach.range=2,Thal=7}	=>	{Disease=1}	0.1034483	1.0000000	2.182796	21

For the test set we have:

lhs		rhs	support	confidence	lift	count
{Thalach.range=1}	=>	{Disease=1}	0.12	1	2.173913	12
{Thalach.range=1,Slope=2}	=>	{Disease=1}	0.11	1	2.173913	11
{CP=4,Thalach.range=1}	=>	{Disease=1}	0.10	1	2.173913	10
{Sex=1,Thalach.range=1}	=>	{Disease=1}	0.10	1	2.173913	10
{Fbs=0,Thalach.range=1}	=>	{Disease=1}	0.12	1	2.173913	12
{CP=4,Ca=2}	=>	{Disease=1}	0.11	1	2.173913	11
{Exang=1,Oldpeak.range=2}	=>	{Disease=1}	0.14	1	2.173913	14
{Oldpeak.range=2,Thal=7}	=>	{Disease=1}	0.14	1	2.173913	14
{Thalach.range=2,Oldpeak.range=2}	=>	{Disease=1}	0.12	1	2.173913	12
{Age.Range=3,Oldpeak.range=2}	=>	{Disease=1}	0.12	1	2.173913	12

Note: many of the top rules (more than 10) sorted by lift are tied so this is why the training and test may not seem as identical as the ones sorted by support.

We conclude that our findings are validated since the training and test sets are very similar.

5.4 Logistic Regression (glm) and Linear Model with Stepwise Feature Selection (glmStepAIC)

Because we are trying to do prediction using a binary variable we start trying to implement a Logistic Regression. We consider three different options:

- Quasibinomial (glm)
- Binomial (glm)
- Binomial with Factor Variables (glm)

We also conducted a Linear Model Stepwise Feature Selection (glmStepAIC) to validate the possibility of reducing the number of variables. According to the results; Sex, CP, Trestbps, Chol, Thalach, Oldpeak, Ca, and Thal (8 variables) are significant to the model. One unexplained result is that Trestbps is included while being insignificant [p=0.14>0.1]. One explanation of this result is that the Stepwise Feature

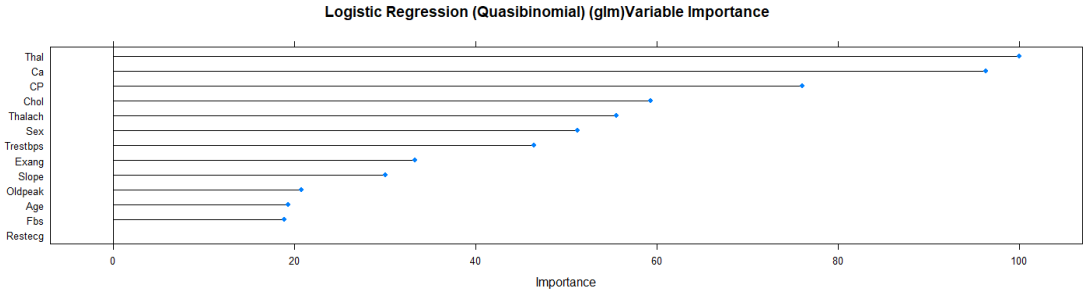
Selection selects the best model based on Akaike Information Criteria (AIC), not p-values.The most significant variables are Thal, CP and ca, which are significant at the 5% level.

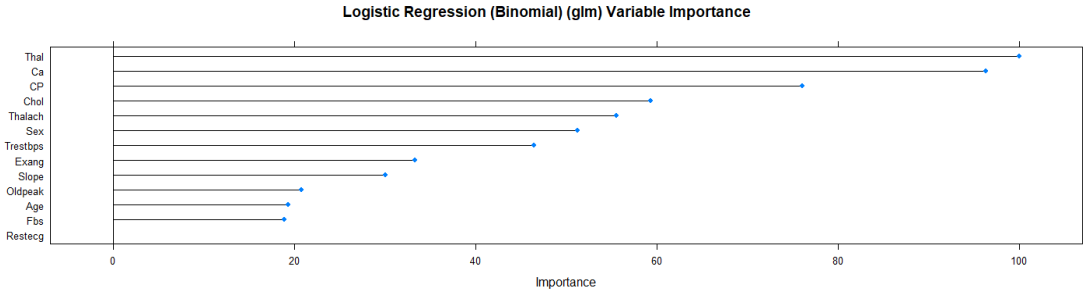
The results are shown in the following table:

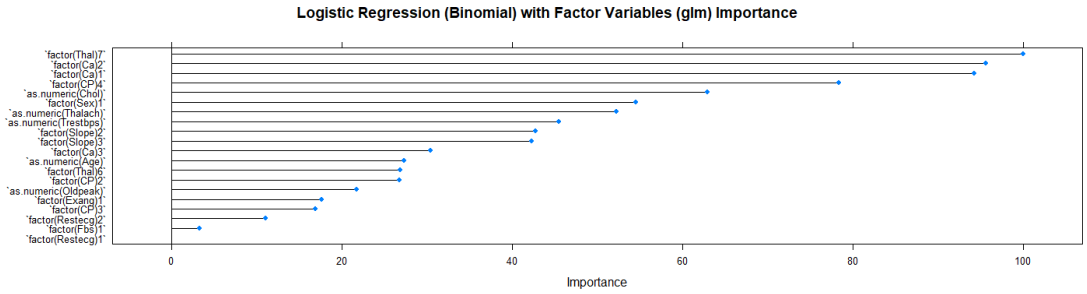
Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
Logistic Regression (Quasibinomial) (glm)	0.8013	0.8182	0.4264	0.9036	1.3
Logistic Regression (Binomial) (glm)	0.8167	0.8182	0.4264	0.9036	1.3
Logistic Regression (Binomial) with Factor Variables (glm)	0.8081	0.8081	0.4381	0.8720	0.86
Linear Model with Stepwise Feature Selection (glmStepAIC)	0.8080	0.8182	0.4264	0.8925	5

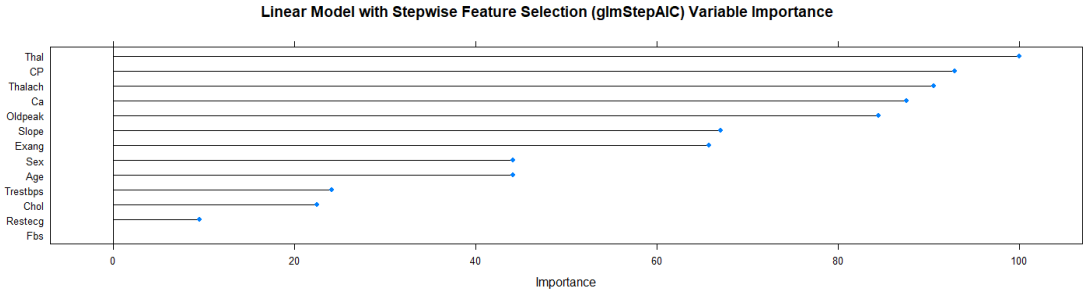
We can notice that using Factor variables did not improve the results and we got 81.82% accuracy in the test set for the others.

The most important variables (>80) were: (*Thal* , *Ca*) for Logistics and (Thal, CP, Thalach, OldPeak, Slope) for Stepwise Feature Selection:



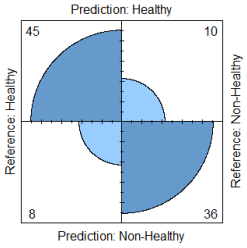




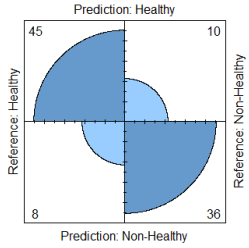


The Confusion Matrices are the following:

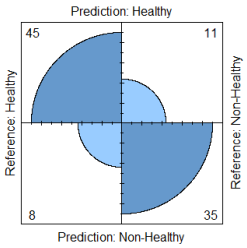
Logistic Regression (Quasibinomial) (glm)



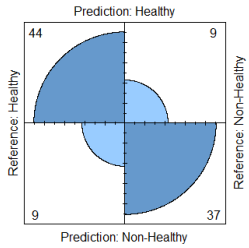
Logistic Regression (Binomial) (glm)



Logistic Regression (Binomial) with Factor Variables (glm)



Linear Model with Stepwise Feature Selection (glmStepAIC)

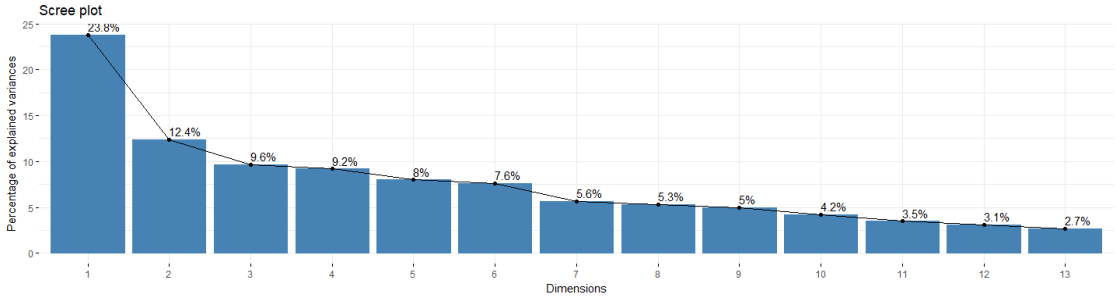


5.5 Kmeans Clustering (eclust)

We also try Kmeans Clustering to do Classification. We consider two different options:

- Using Principal Component Analysis (PCA)
- All Data

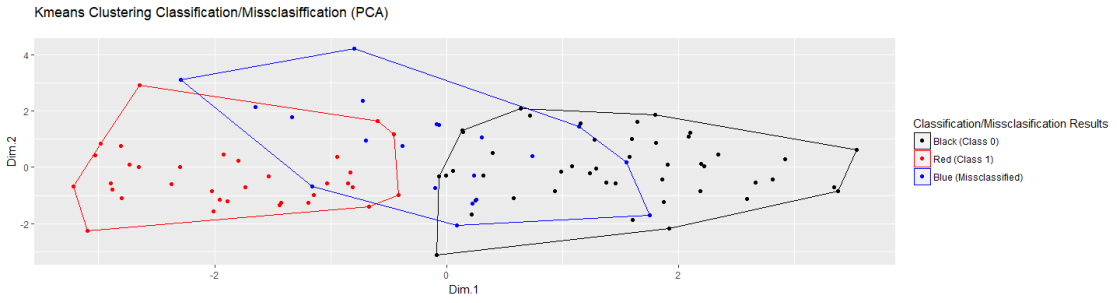
We consider 8 Principal Components (81.54% of cumulative variance) and compare the results with the Cluster using the total variables (13). We notice that we did not get any improvement (we got the same results). The next picture shows the scree plot:

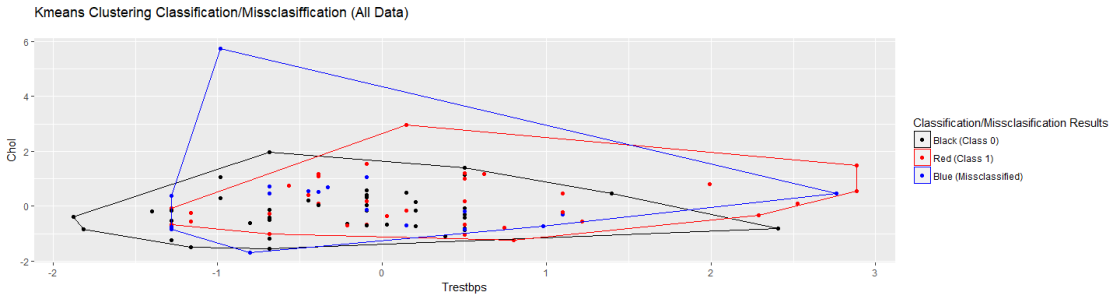


The results are shown in the following table:

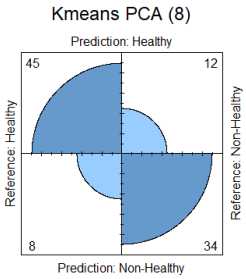
Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
Kmeans PCA (8)	0.8333	0.7980	0.4495	0.7941	3.7
Kmeans (13)	0.8333	0.8182	0.4264	0.8158	4.5

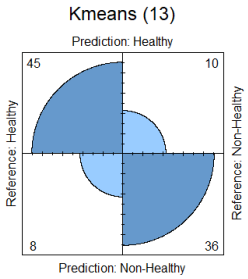
Those are some examples of Cluster Plots:





The Confusion Matrix are the following:





5.6 Support Vector Machine (SVM)

Support Vector Machine (SVM) shows to be a very powerful tool in our Research Project (Steganalysis). We try two different kernels:

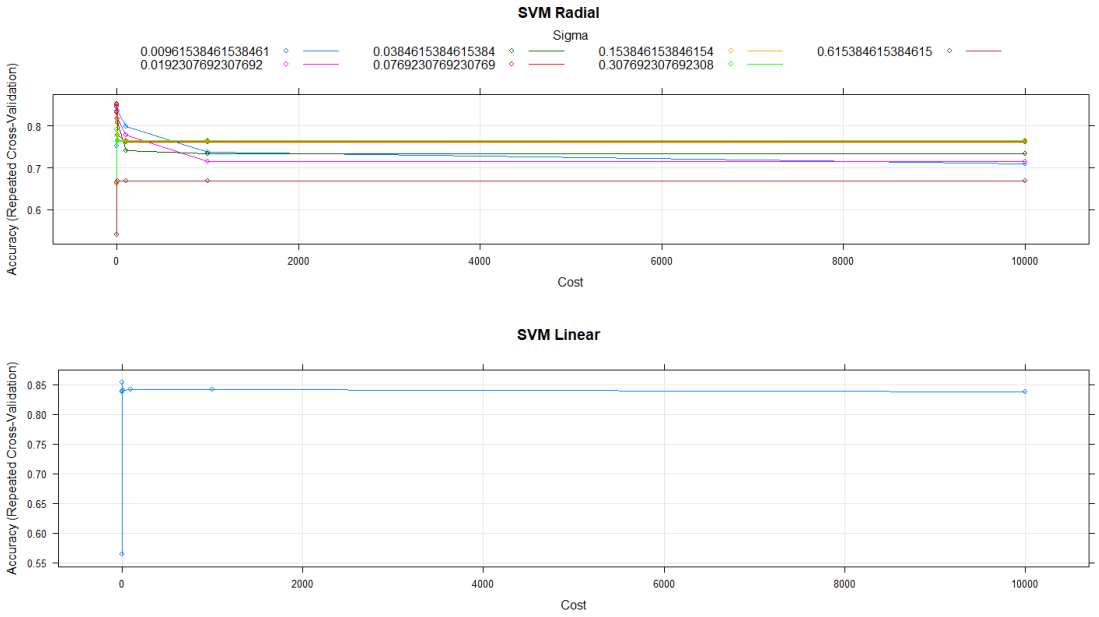
- SVM Radial (svmRadial)
- Linear SVM (svmLinear)

We notice that using a Linear Kernel we got the best results and a big improvement in the Accuracy in Training Set. In terms of computational cost the Linear Kernel Method was more costly.

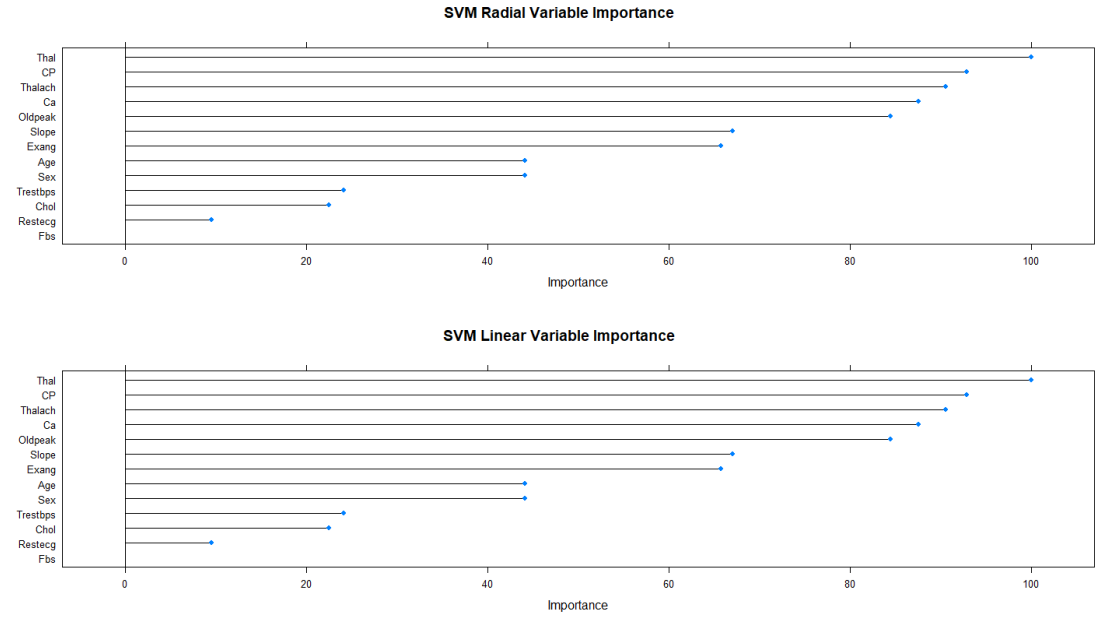
The results are shown in the following table:

Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
SVM Radial	0.8535	0.7980	0.4495		43
SVM Linear	0.8586	0.8182	0.4264		140

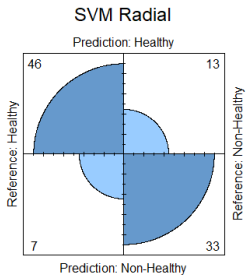
In terms of Accuracy, we got the best results using the following kernel parameters:

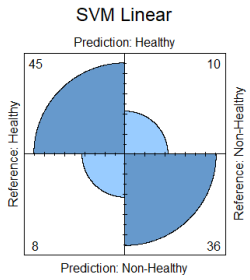


Regarding the variable importance (> 80), we got the same results (*Thal*, *CP*, *Thalach*, *Ca* and *OldPeak*):



The Confusion Matrices are the following:





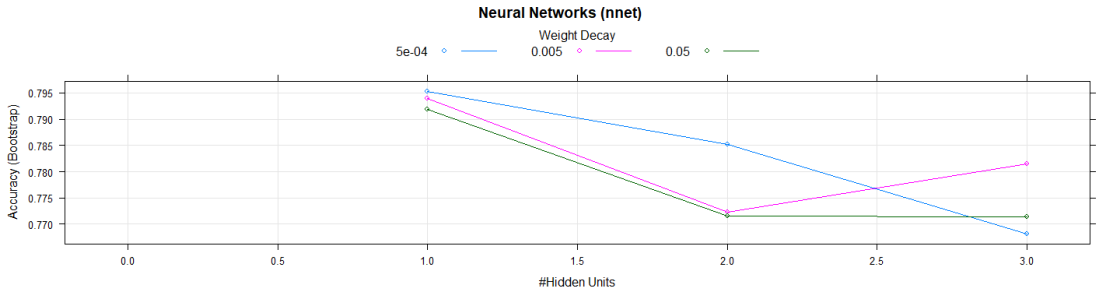
5.7 Neural Networks (nnet) and K-Nearest Neighbors (knn)

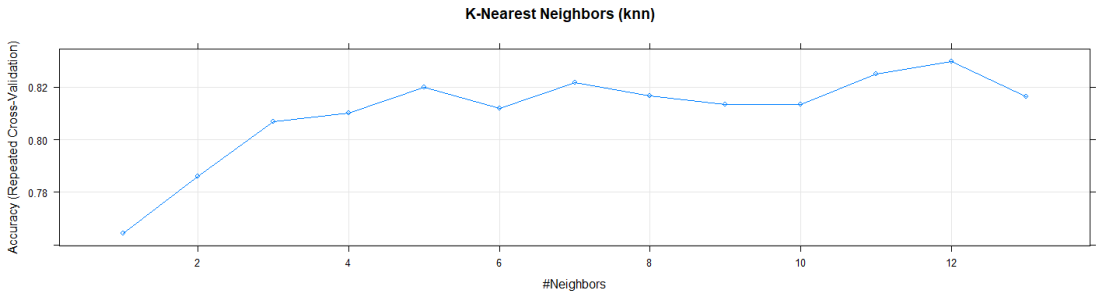
We also try to use Neural Networks (nnet) and K-Nearest Neighbors (knn). K-Nearest Neighbors shows to have a good improvement (83.84%) with a very low computational cost.

The results are shown in the following table:

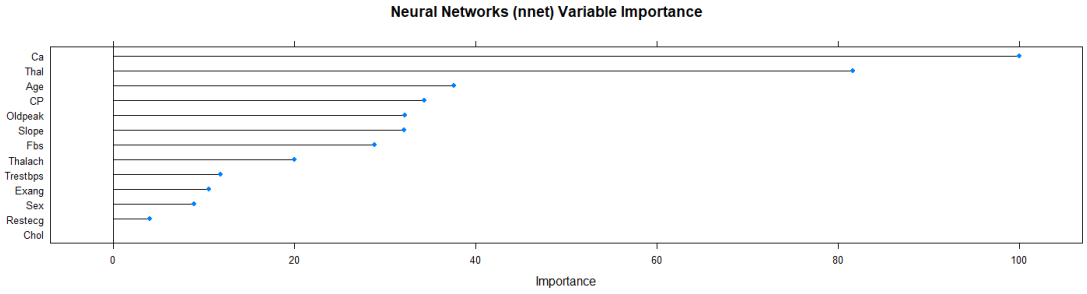
Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
Neural Networks (nnet)	0.7953	0.7778	0.4714	0.8376	9.3
K-Nearest Neighbors (knn)	0.8299	0.8384	0.4020	0.9100	6.5

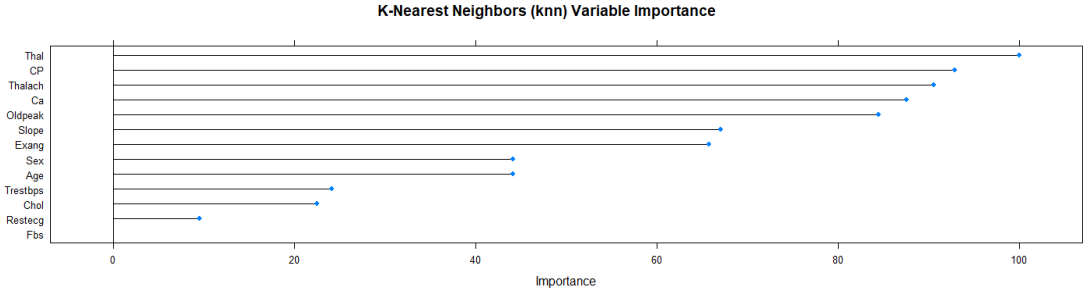
In terms of Accuracy, we got the best results using the following kernel parameters:



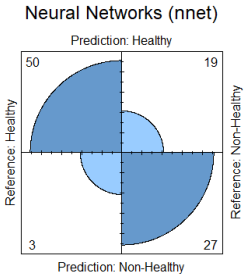


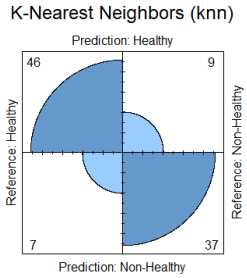
Regarding the variables importance (>80), the most important for Neural Network was (*Ca, Thal)*, and (*Thal, CP,Thalach,Ca, OldPeak)* for K-Nearest Neighbors.





The Confusion Matrices are the following:





5.8 Random Forest

Using Random Forest we try four different scenarios:

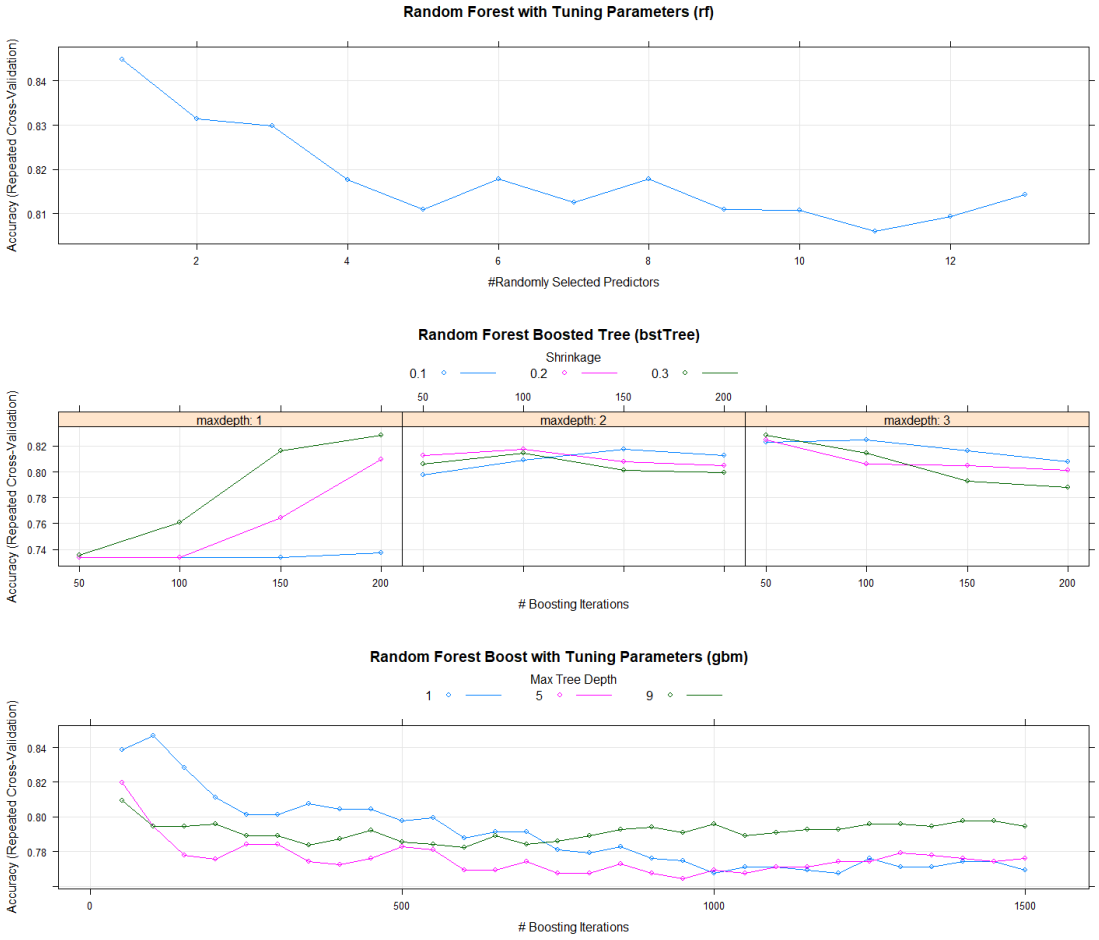
- Using Tuning Parameters (rf)
- Boosted Tree (bstTree)
- Boost with Tuning Parameters (gbm)
- Stochastic Gradient Boost (gbm)

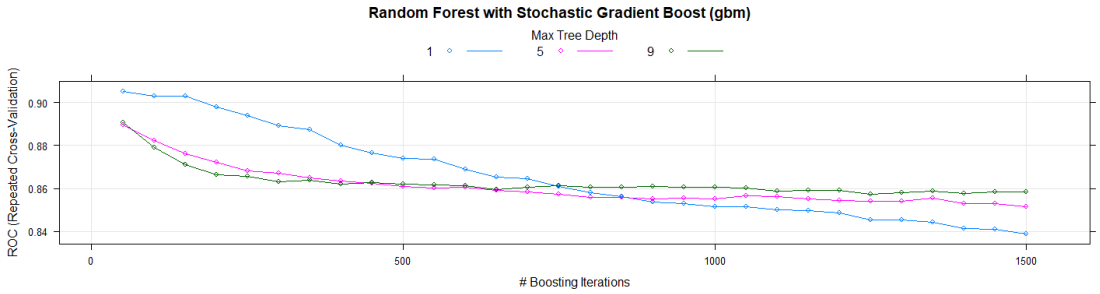
The best results was using Tuning Parameters (rf) (83.84%) with a ROC equal to 91.82%. Boosted Tree was very computationally costly and did not improve the results.

The results are shown in the following table:

Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
Random Forest with Tuning Parameters (rf)	0.8449	0.8384	0.4020	0.9182	53
Random Forest Boosted Tree (bstTree)	0.8283	0.7778	0.4714		499
Random Forest Boost with Tuning Parameters (gbm)	0.8469	0.8384	0.4020	0.8897	31
Random Forest with Stochastic Gradient Boost (gbm)	0.9053	0.8182	0.4264	0.8819	102

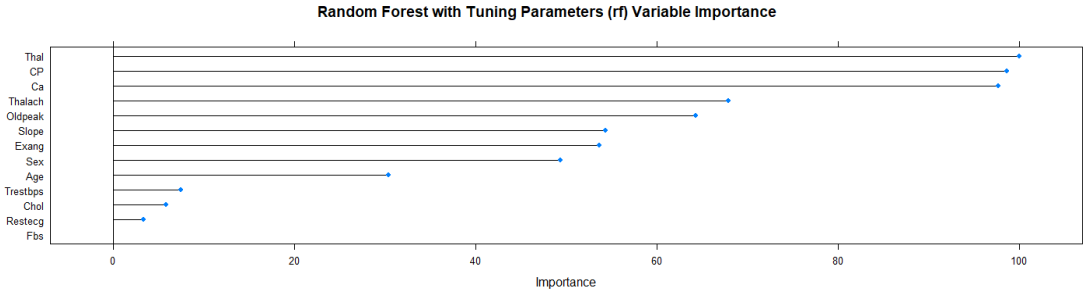
In terms of Accuracy, we got the best results using the following parameters:

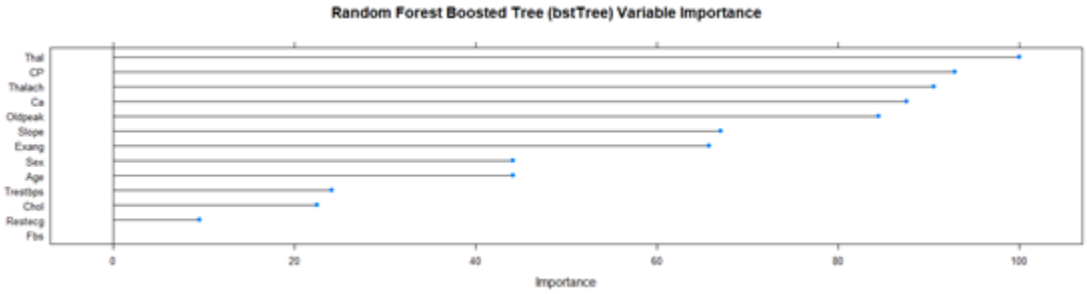


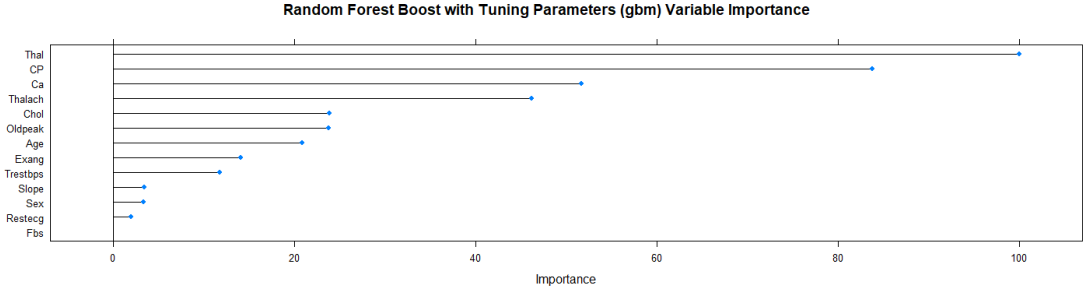


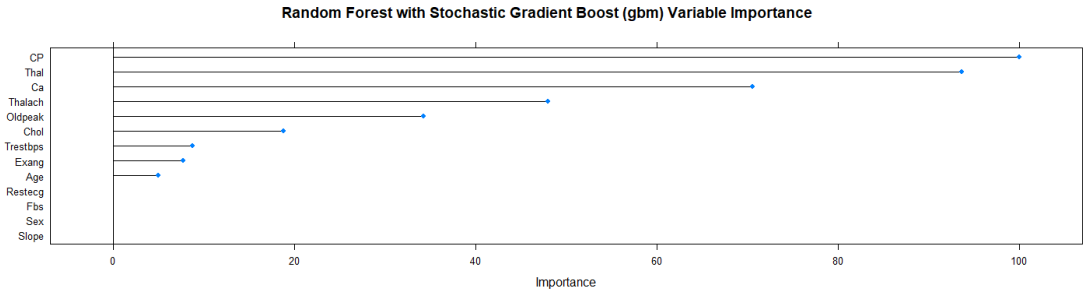
Regarding the variables importance (>80) we got the following results:

- Tuning Parameters (rf): (*Thal, CP, Ca*)
- Boosted Tree (bstTree): (*Thal, CP,Thalach, Ca, OldPeak)* and
- *(Thal, CP)* for the others



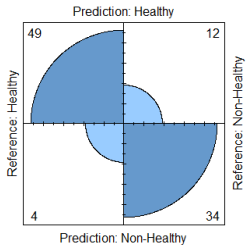




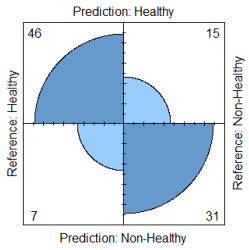


The Confusion Matrices are the following:

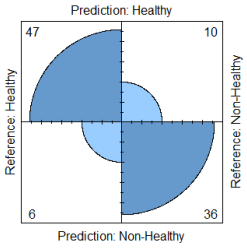
Random Forest with Tuning Parameters (rf)



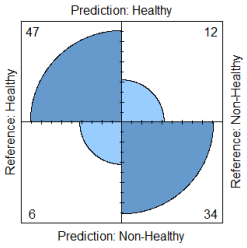
Random Forest Boosted Tree (bstTree)



Random Forest Boost with Tuning Parameters (gbm)



Random Forest with Stochastic Gradient Boost (gbm)



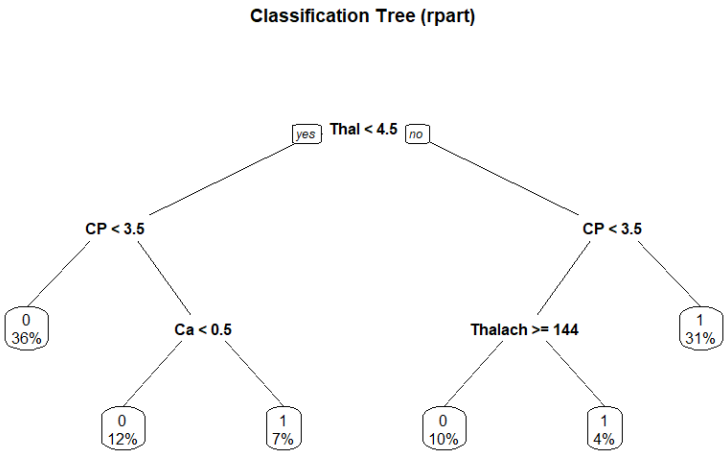
5.9 Classification Tree (rpart)

Using a Classification Tree (rpart) we got a very low prediction accuracy (78%). The following table shown the results:

Method	Prediction	Prediction	RMSE	ROC	Time

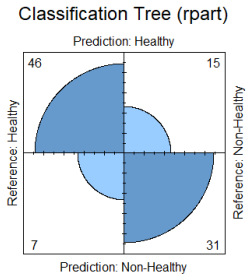
	Accuracy in Training Set	Accuracy in Test Set	Test		Elapsed
Classification Tree (rpart)		0.7778	0.4714	0.8117	102

The following figure shows the model:



According with this the main variables used to do the classification are Thal, CP, Ca and Thalach.

The Confusion Matrix is the following:



5.10 Fast-and-Frugal Decision Trees (FFTrees) (also Custom)

A fast-and-frugal tree (FFT) is a set of hierarchical rules for making decisions based on very little information (usually 4 or fewer). We use the FFT to predict which cues are the best predictors of heart disease risk. Specifically, it is a decision tree where each node has exactly two branches, where one (or in the cast of the final node, both) branches is an exit branch.

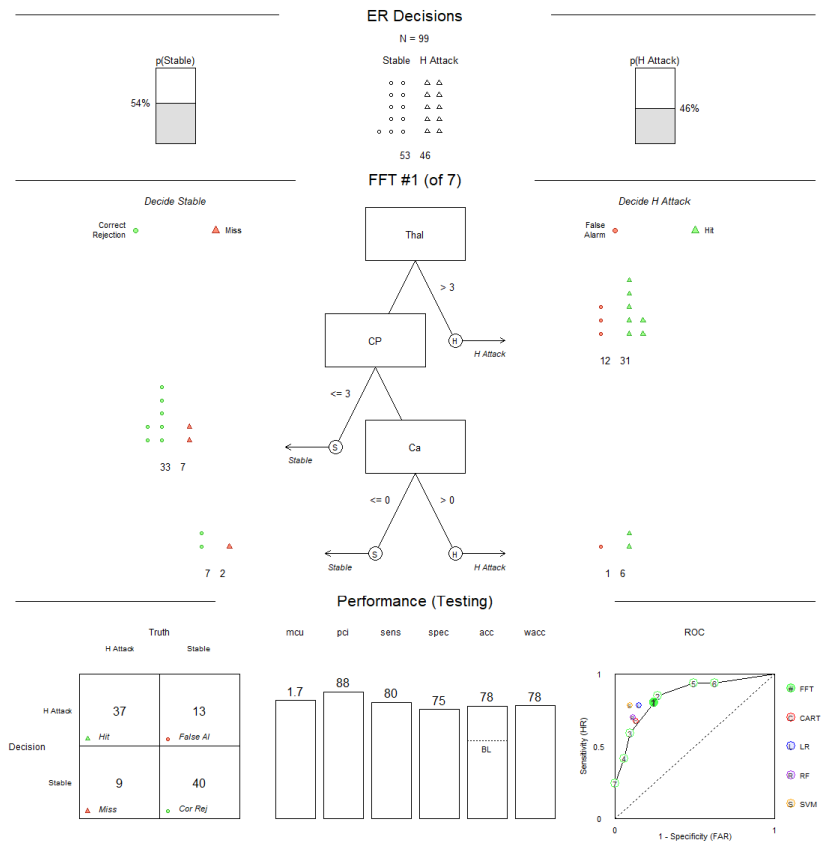
The following table show the results:

Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
Fast-and-Frugal Decision Trees (FFTrees)	0.8283	0.7778	0.4714	0.7914	96
Custom Fast-and-Frugal Decision Trees (FFTrees)	0.8283	0.7778	0.4714	0.7914	96

FFTs are simple, convenient decision strategies that use minimal information to make decisions. In our report, we conducted fast and frugal trees since this method rarely over-fits data and is easy to interpret and implement in real-world decision tasks, such as for detecting depression and increasing decision making abilities in emergency rooms.

We note that using the default settings that thal, cp, and ca are used as indicators to predict heart disease in the FFTree plot.

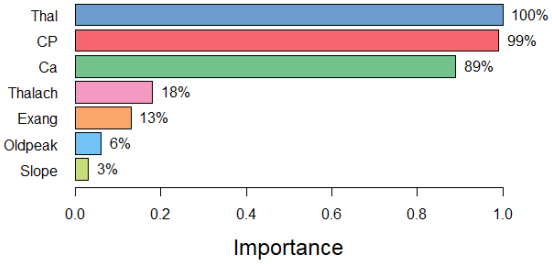
```
##
## [1] If Thal > 3, predict H Attack.
## [2] If CP <= 3, predict Stable.
## [3] If Ca <= 0, predict Stable, otherwise, predict H Attack.
```

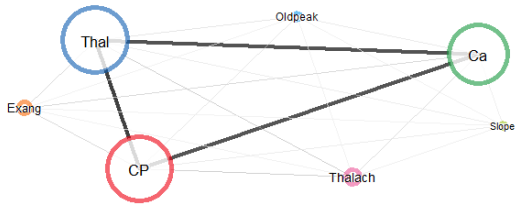


- On the top row, we observed that there were 99 patients (cases) in the test data, where 46 patients were having heart attacks (46%), and 53 patients were not (54%).
- In the middle row, the tree makes decisions for each of the patients using easy–to–understand icon arrays. For example, you see that 43 patients suspected of having heart attacks were (virtually) sent to the CCU after the first question, where 12 were not having heart attacks (false–alarms), and 31 were having heart attacks (hits).
- The bottom row of the plot indicates the aggregate summary statistics for the tree. On the bottom row, you have a 2 x 2 confusion matrix, which shows you a summary of how well the tree was able to classify patients, levels indicating overall summary statistics, and an ROC curve which compares the accuracy of the tree to other algorithms such as logistic regression (LR) and random forests (RF). Here, where the fast-and-frugal tree is represented by the green circle "1", you can see that the fast-and-frugal tree had a higher sensitivity than logistic regression and random forests, but at a cost of a lower specificity.

5.11 Forest of Fast-and-Frugal Decision Trees (FFForest)

Trying a Forest of Fast-and-Frugal Decision Trees (FFForest), we see that the three cues CP, Thal, and Ca occur the most often in the forest and thus appear to be the most important three cues in the dataset.





In the above table, we observe that Thal, CP, and ca are of the most importance to predict heart disease based on the FFForest plot. The plot indicates that at n=100 (number of trees to create), thal is of 100% importance, chest pain (CP) is of 99% importance, and ca is of 89% importance. These are important indicators to note and are valid attributes in understanding the causes of heart disease.

6. Main Plots/Detail Results

6.1 All Prediction Models Results

Method	Prediction Accuracy in Training Set	Prediction Accuracy in Test Set	RMSE Test	ROC	Time Elapsed
Logistic Regression (Quasibinomial) (glm)	0.8132	0.8182	0.4264	0.9036	1.3
Logistic Regression (Binomial) (glm)	0.8021	0.8182	0.4264	0.9036	1.4
Logistic Regression (Binomial) with Factor Variables (glm)	0.8009	0.8081	0.4381	0.8720	1.2
Linear Model with Stepwise Feature Selection (glmStepAIC)	0.8076	0.8182	0.4264	0.8925	5.3
Kmeans PCA (8)	0.8333	0.7980	0.4495	0.7941	5.8
Kmeans (13)	0.8333	0.8182	0.4264	0.8158	4.5
SVM Radial	0.8535	0.7980	0.4495		58
SVM Linear	0.8586	0.8182	0.4264		169
Neural Networks (nnet)	0.7953	0.7778	0.4714	0.8376	9.5
K-Nearest Neighbors (knn)	0.8299	0.8485	0.3892	0.9100	6.9
Random Forest with Tuning Parameters (rf)	0.8449	0.8384	0.4020	0.9182	52
Random Forest Boosted Tree (bstTree)	0.8283	0.7778	0.4714		425

Random Forest Boost with Tuning Parameters (gbm)	0.8469	0.8384	0.4020	0.8897	29
Random Forest with Stochastic Gradient Boost (gbm)	0.9053	0.8182	0.4264	0.8819	96
Classification Tree (rpart)		0.7778	0.4714	0.8117	96
Fast-and-Frugal Decision Trees (FFTrees)	0.8283	0.7778	0.4714	0.7914	96
Custom Fast-and-Frugal Decision Trees (FFTrees)	0.8283	0.7778	0.4714	0.7914	96
Forest of Fast-and-Frugal Decision Trees (FFForest)	0.8283	0.7980	0.4495	0.8273	127

Our main objective is to observe the results from the testing data, from this we can determine how accurate our model is in predicting heart disease. In the Results table above, we observe that the best method of predicting heart disease is k-nearest neighbours with an accuracy of 0.8485 (or 84.85%) for the test set.

RMSE indicates the absolute fit of the model to the data and shows how close the observed data points are to the model's predicted values. Therefore, the smaller the RMSE, the better the method is. ROC is the relationship between sensitivity [The fraction of people with the disease that the test correctly identifies as positive] and specificity [The fraction of people without the disease that the test correctly identifies as negative]. Therefore, the higher the ROC, the stronger the relationship and the better the predictor. When observing the table above, K nearest neighbours also has the smallest RMSE [0.3892] and largest ROC [0.9100], therefore, this confirms that the k nearest neighbours method is the best and most accurate in predicting heart disease.

From observing the results table, we found identical results using the random forest boosted tree method, the FFTree, and the custom FFTree for the accuracy of the test set and RMSE. The similar results between the two FFTree's show that even when adding parameters to optimize for a sensitivity weight of .99 and maximize for accuracy, the results will remain similar to the original FFTree. Thus, the original FFTree is the most optimal option for our data.

We used excel to calculate the averages for each test and with these results we conclude that, on average, the training set produces better results than the test set [training set = 0.831647, test set = 0.806978]. On average, the RMSE, ROC, and time elapsed was 0.438622, 0.856053, and 71.10556 seconds, respectively.

7. Optimizing Costs

The fan algorithms [ifan and dfan] can be used to try to minimize costs, if goal= "cost" and/or goal.chase = "cost". We can specify two types of costs; cost.cues, the cost of using a cue to classify a case; and cost.outcomes, the cost of different outcomes.

cost.cues is a data frame with two columns, one column giving the names of cues with costs, and one column giving the actual costs. cost.outcome is a vector of length 4 indicating the cost of hits, false

alarms, misses, and correct rejections respectively. In our report we classify a false alarm with a cost of $500 and a miss with a cost of $1000.

Our first FFT was built with the goal of maximizing balanced accuracy and ignoring these costs:

When examining the outcome of the best performing tree, we observe that bacc = 0.8145073 and cost = 248.8519.

Our second FFT, is built to respect these costs. We observe a slightly lower balanced accuracy of bacc = 0.7317518 and a much lower cost of cost = 151.5152.

Therefore, we can suggest that the minimum cost to classify a case of heart disease is $151.5152.

8. Conclusions

In this project we compared the heart disease performance prediction of different supervised and unsupervised data mining methodologies including: linear regression models (logistic, stepwise feature selection), non-linear regression models (support vector machines, neural network, classification and regression trees, boosting and random forest) and clustering (kmeans).

From our findings, we can sufficiently suggest which attributes are better indicators of heart disease and specifically, which combination of these indicators and at what levels result in a patients more at more risk than others to get a heart disease. By setting a parameter of 80% accuracy we were able to eliminate attributes that were not significant indicators of heart disease and minimize the thirteen attribute list down to three. Almost all the models considered the attributes ‘thalassemia (thal), number of major vessels coloured by fluoroscopy (ca), and chest pain (CP) the most important variables for predicting heart disease. This can be seen visually when observing the FFForest plot; thal has an importance of 100%, CP has an importance of 99%, and ca has an importance of 89%, while all other attributes were under 20%. Only Thal, Ca, and CP are significant at the 5% level, which corresponds with our results in the Fast and Frugal Trees

Generally, all considered models have similar performances but Random Forest with Tuning Parameters (rf) and K-Nearest Neighbor (knn) seem to be slightly better, (83.84%) ROC (>91%). Overall, we can conclude that the best method of predicting heart disease is k-nearest neighbours with an accuracy, RMSE, and ROC of 0.8485 (or 84.85%), 0.3892, and 0.9100, respectively, for the test set.

When implementing this machine learning capability into the medical field, this study can be used by medical personnel to better predict future patients probability of having a heart disease. By analyzing which attributes and their values, medical personnel can accurately predict the patient's likelihood of having heart disease. This new technology will make doctors and nurses lives much easier and advance evidence based decision making, which in turn can result in cost optimization and cost saving over time.