Heart Disease Problem, Associated Risk Factors and Methods of Prediction

Data of 303 individuals is available. This data includes the following parameters which seems to have an impact on existence of heart disease

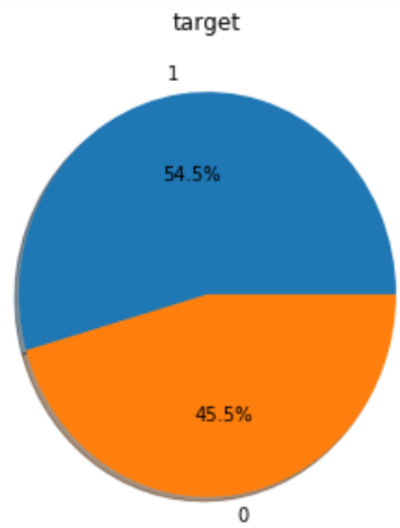
- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest pain type -- Value 0: typical angina -- Value 1: atypical angina -- Value 2: non-anginal pain -- Value 3: asymptomatic
- trestbps: resting blood pressure (in mm Hg on admission to the hospital) chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment -- Value 0: downsloping -- Value 1: upsloping -- Value 2: flat
- ca: number of major vessels (0-4) colored by fluoroscopy
- thal: 1 = normal; 2 = fixed defect; 3 = reversible defect

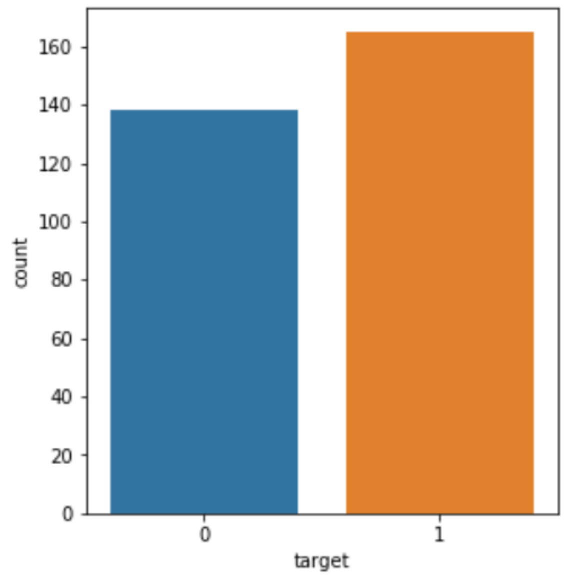
First, the data is analyzed and studied in order to understand the importance and effects of different variables.

By using a simple count function, below information can be obtained. This shows that the data set is spread well between two categories and we have enough data from both individuals with and without heart disease.

Target	Count
1 (individuals with heart disease)	165
0 (individuals without heart disease)	138

This information is also shown in below plots





In this data set we have some numerical parameters and some categorical parameters. We do not have any values such as strings that need to be converted to numbers since already all the categorical parameters are labeled.

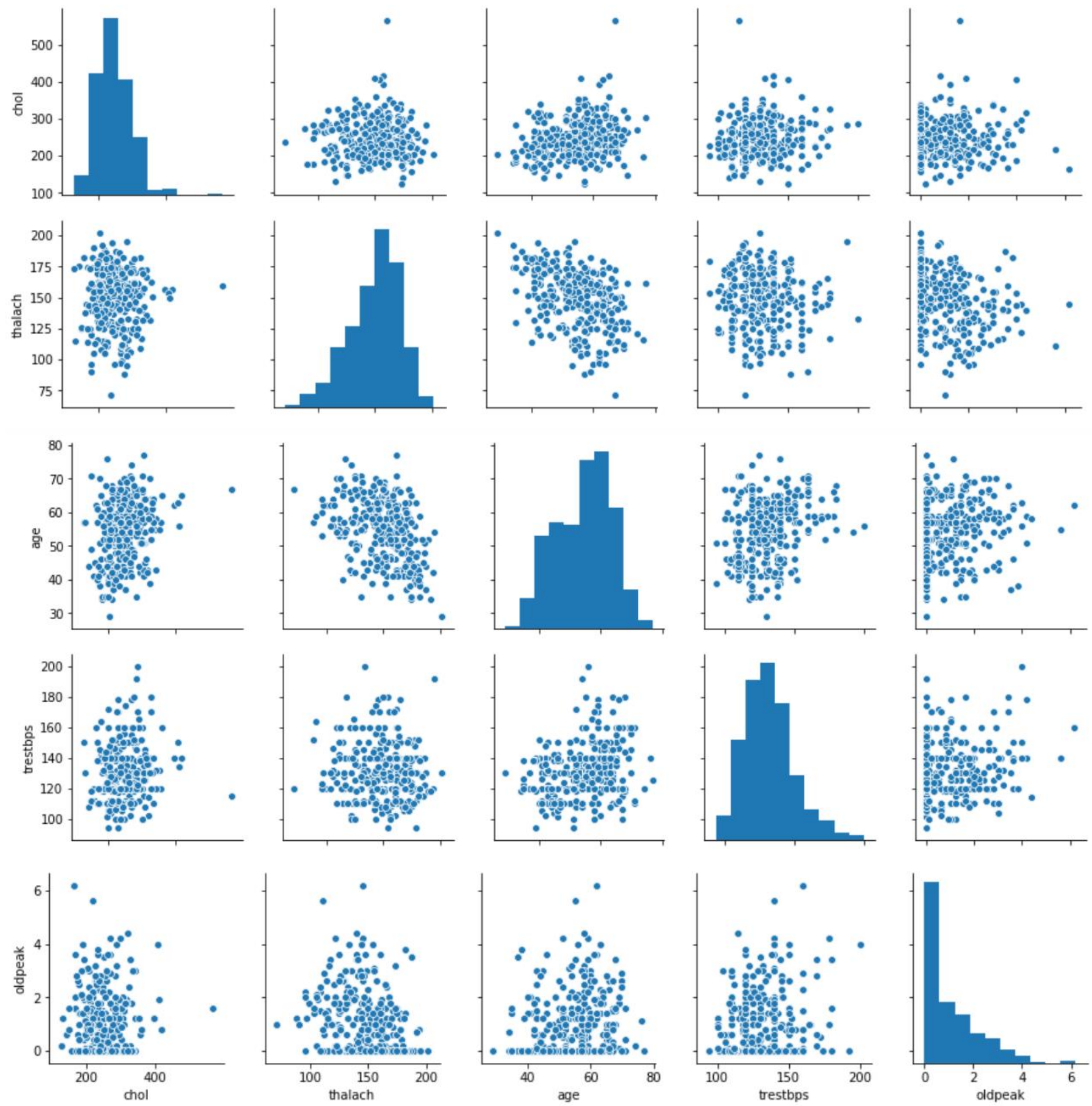
Numerical values: 'chol', 'thalach', 'age', 'trestbps', 'oldpeak'

Categorical values: 'sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal', 'target'

Below information describes the numerical values of the data set

	chol	thalach	age	trestbps	oldpeak
count	303.000000	303.000000	303.000000	303.000000	303.000000
mean	246.264026	149.646865	54.366337	131.623762	1.039604
std	51.830751	22.905161	9.082101	17.538143	1.161075
min	126.000000	71.000000	29.000000	94.000000	0.000000
25%	211.000000	133.500000	47.500000	120.000000	0.000000
50%	240.000000	153.000000	55.000000	130.000000	0.800000
75%	274.500000	166.000000	61.000000	140.000000	1.600000
max	564.000000	202.000000	77.000000	200.000000	6.200000

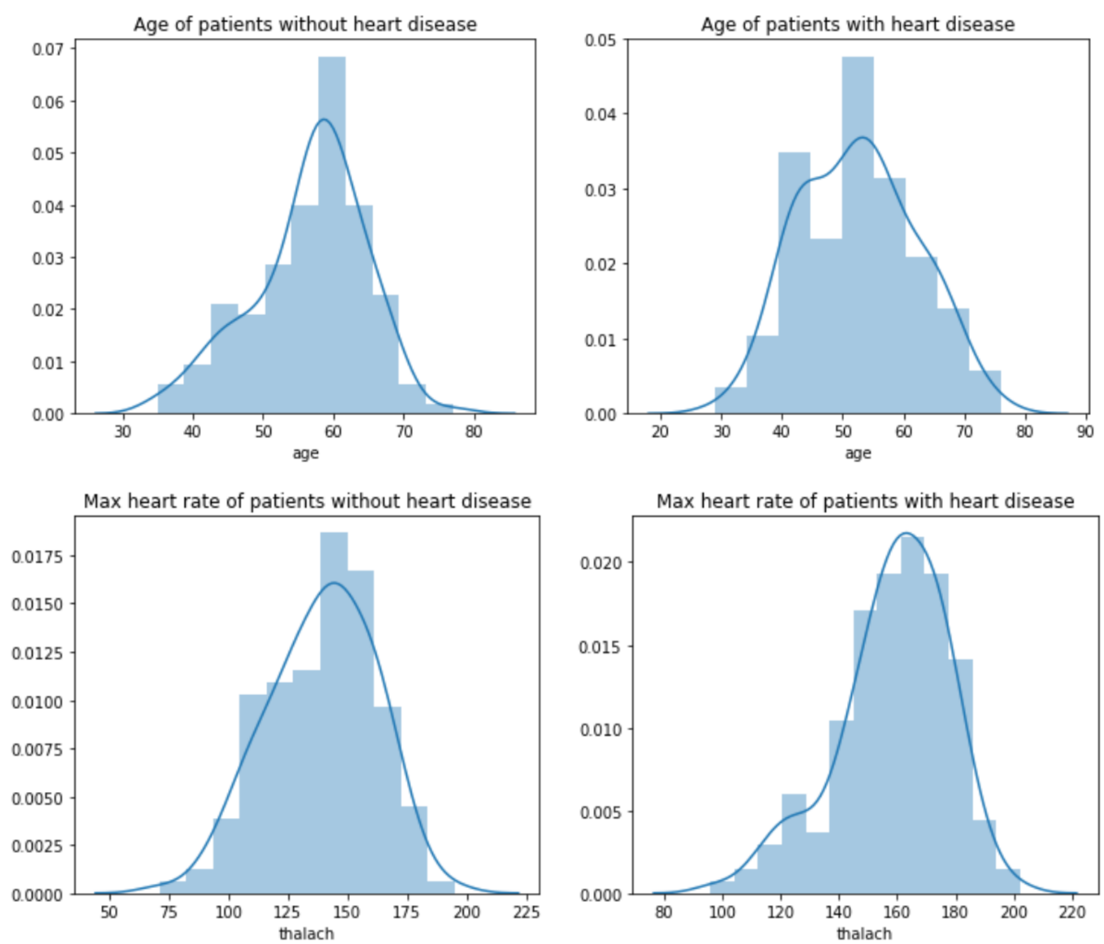
Below, the relationship between all the numerical data is shown



It can be seen that there is a negative relationship between age and maximum heart rate. There is no obvious or linear relationship between various parameters.

We can also plot the correlation of above parameters. Based on below, age and maximum heart rate seem to have a high impact on heart disease.

Age and maximum heart rate (thalach) are plotted below for patients with heart disease and patients without heart disease



Now, some studies are done on categorical parameters.

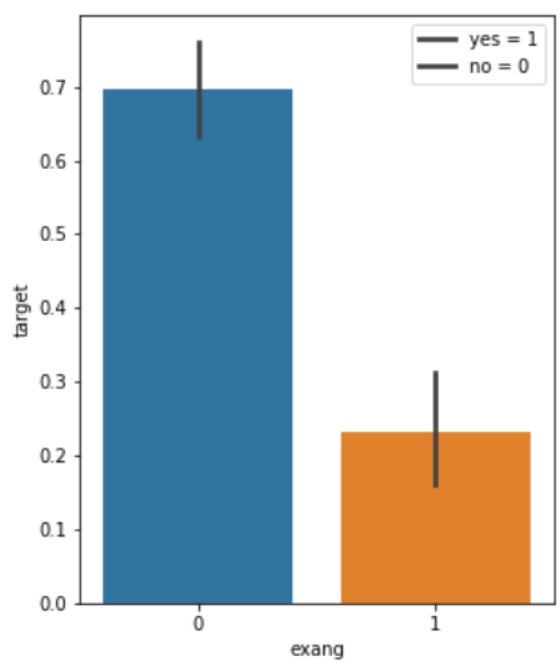
Below table shows the division of patients based on gender

Gender	With Disease	Without Disease
Female	72	24
Male	93	114

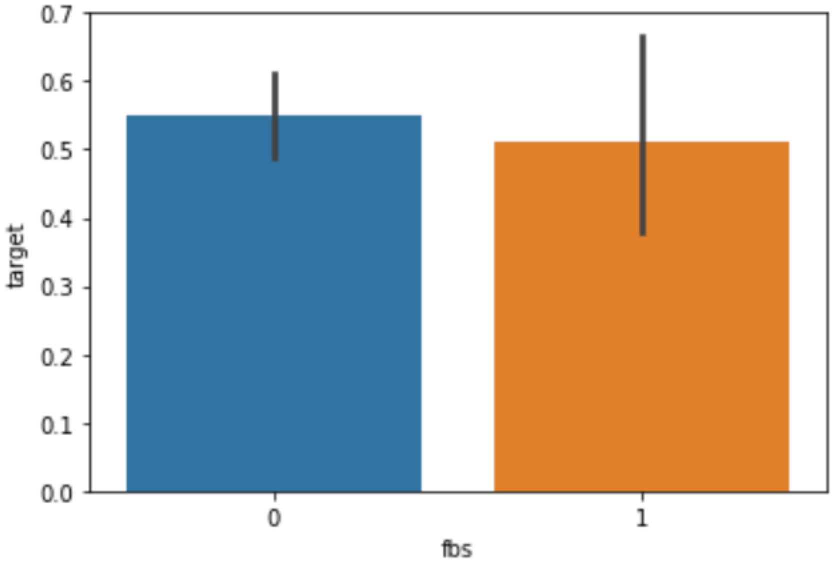
Below data shows the average target number based on the number of major vessels colored by fluoroscopy (ca). We cannot see an obvious trend.

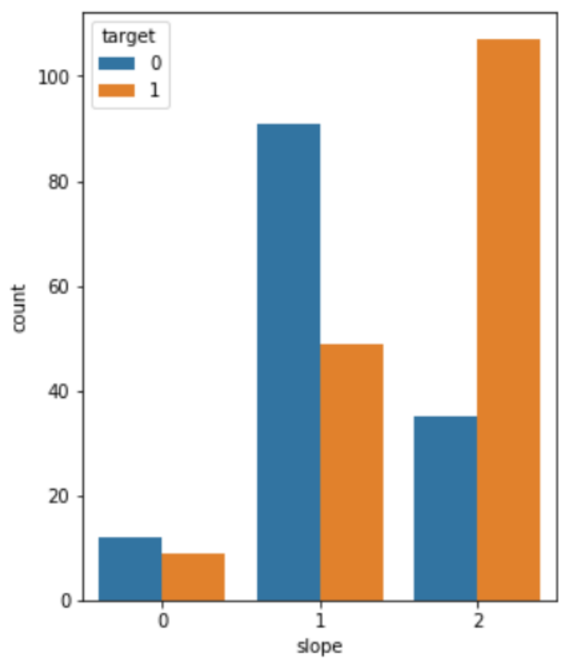
ca	target
0	0.742857
1	0.323077
2	0.184211
3	0.150000
4	0.800000

Below plot illustrates the data for exang (exercise induced angina (1 = yes; 0 = no))



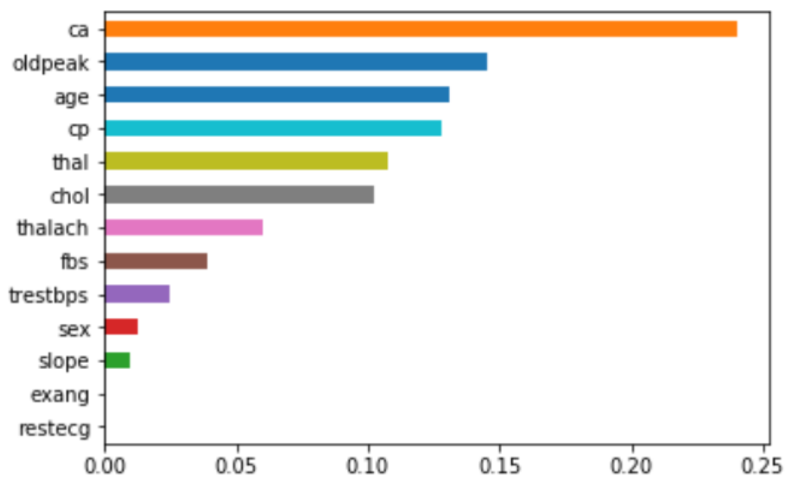
Lastly, the slope parameter (the slope of the peak exercise ST segment) and fbs ((fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)) are depicted below





The data is splitted into train and test to be used for machine learning algorithms. Different methods are used with the objective of optimizing accuracy, confusion matrix results and test and training errors. Logistic regression, decision tree and random forest seem to have acceptable errors. However, since the objective is to have the model understandable for people without knowledge of programming and statistics, random forest and decision tree results are reported since these algorithms are easy to be used by general public.

A decision tree algorithm is used to train the machine learning algorithm for heart disease data. Below plot shows the feature importance

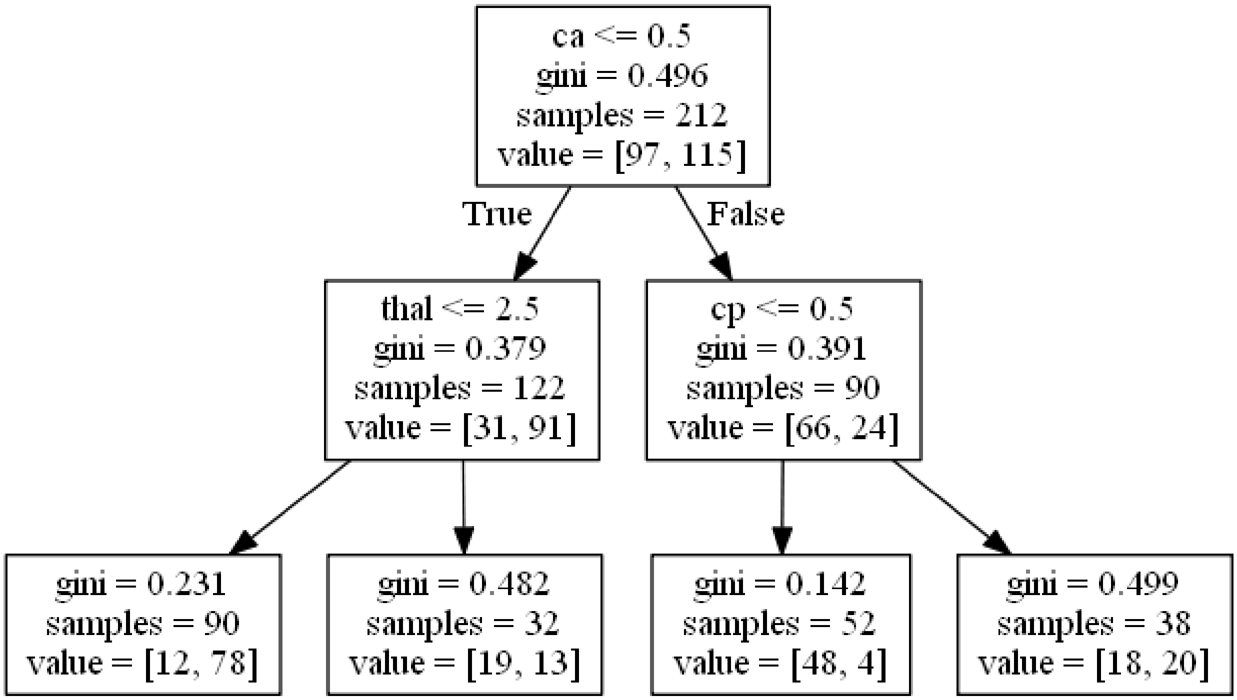


This plot closely corresponds to the studies that were done on the data at the beginning of the report.

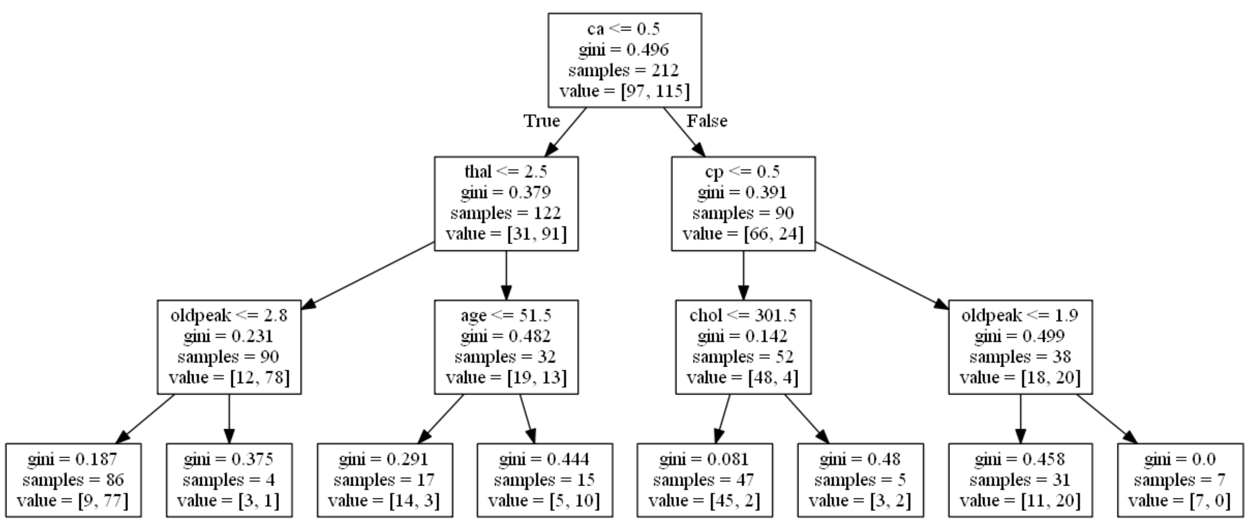
The depth of the tree is increased to find the optimal value. Based on below errors, maximum depth of 3 or 2 seem to be optimized.

```
max_depth = 1
Train accuracy: 74.06%
Test accuracy: 72.53%
max_depth = 2
Train accuracy: 77.83%
Test accuracy: 84.62%
max_depth = 3
Train accuracy: 84.43%
Test accuracy: 81.32%
max_depth = 4
Train accuracy: 87.74%
Test accuracy: 73.63%
max_depth = 5
Train accuracy: 90.57%
Test accuracy: 65.93%
max_depth = 6
Train accuracy: 93.40%
Test accuracy: 64.84%
max_depth = 7
Train accuracy: 97.17%
Test accuracy: 72.53%
```

Below decision tree is obtained as a result of setting the maximum depth to 2



Below decision tree shows the results for a tree with maximum depth of 3. In this model gini index is lower than previous one which will lead to an easier decision making at the last step.

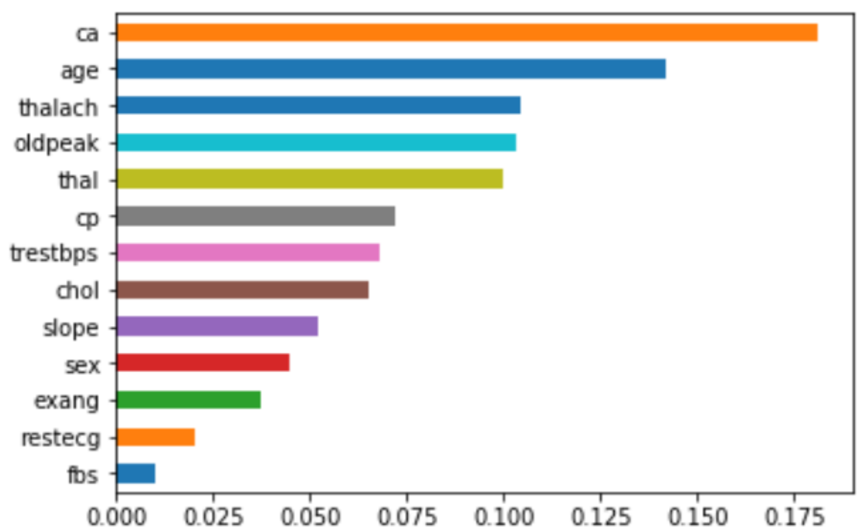


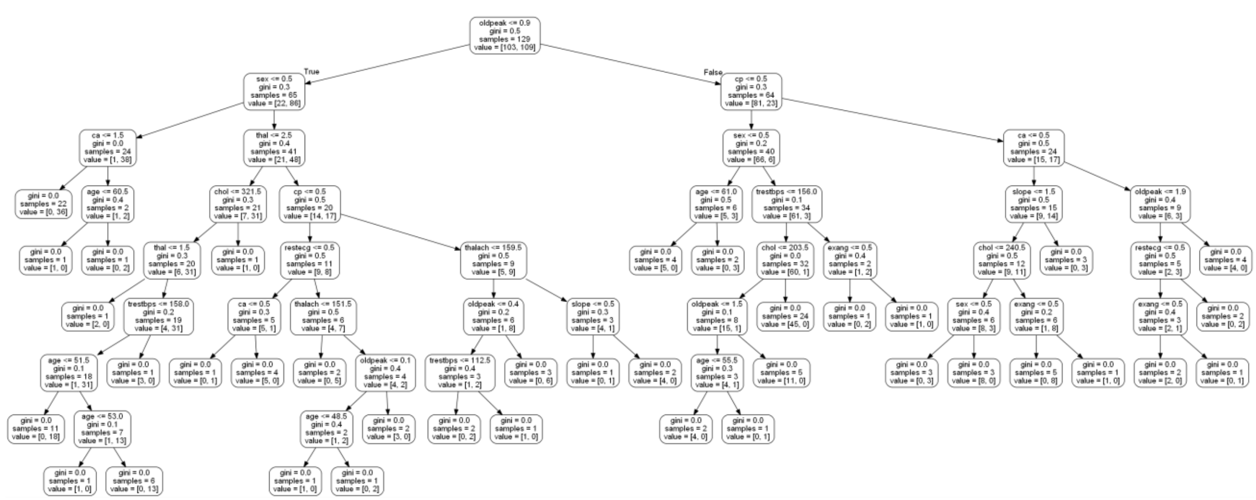
Confusion matrix of above model is shown below. When optimizing the model, the emphasis was on getting a lower number for false negatives. It is very important to diagnose the heart disease correctly. Therefore, we do not want to predict no heart disease for somebody who actually suffers from heart disease. This case is more important than predicting somebody has heart disease while they do not have. There is less harm in the second case.

	Predicted NO	Predicted YES
Actual NO	28	13
Actual YES	4	46

In this case we have only missed diagnosis of 4 heart disease patients which is not perfect but it is ideal.

Below feature importance is plotted based on **random forest** training. Also, the full plot is shown below





Below values summarize the performance of the random forest machine learning method used for this data set. First error is related to the complete random forest and second one is related to a sample tree

```
Train accuracy: 100.00%
Test accuracy: 94.51%
Train accuracy: 85.85%
Test accuracy: 95.60%
```

Confusion matrix is shown below

	Predicted NO	Predicted YES
Actual NO	38	3
Actual YES	1	49

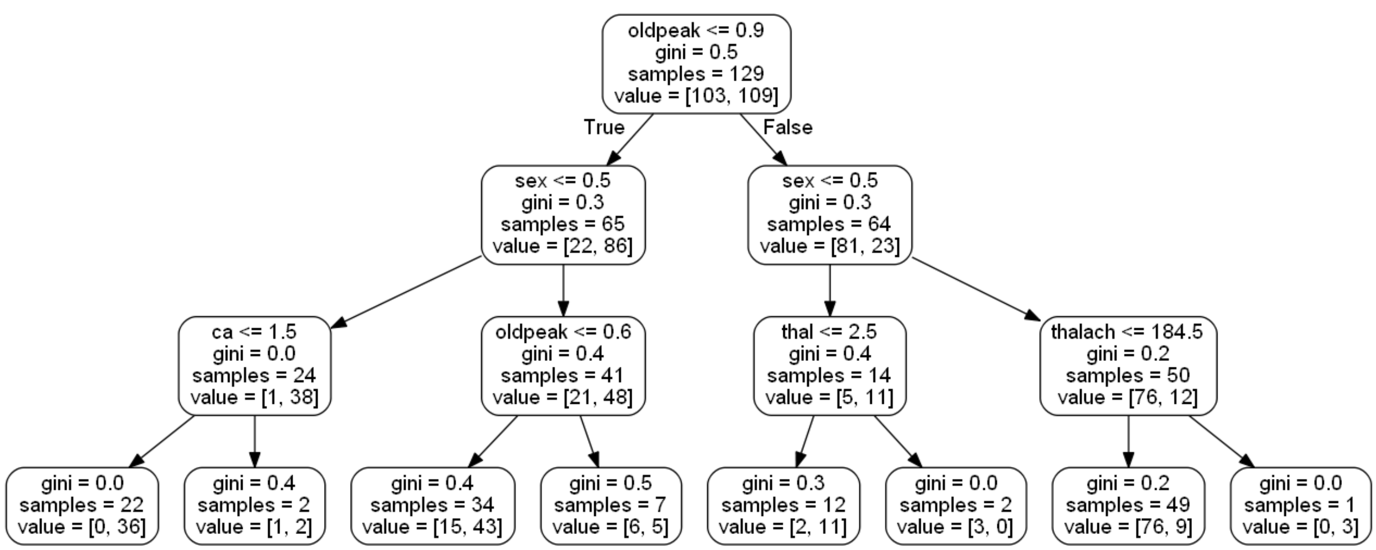
In this case we have only missed diagnosis of 1 heart disease patient which is not perfect but it is very ideal.

```
             precision    recall  f1-score   support

    Healthy       0.97      0.93      0.95        41
       Sick       0.94      0.98      0.96        50

avg / total       0.96      0.96      0.96        91
```

Below is a tree from random forest than can be used by anyone to predict if an individual is suffering from heart disease



Resources

- McGill, Statistical Machine Learning course slides and example codes
- Github and stackoverflow example codes