Random Forest Experiment Report

Manikanta Kalyan Gokavarapu
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260
mgokavar@buffalo.edu

1 Random Forest Experiment Report

Random Forest is one of the Ensemble techniques. Ensemble means combining various models and these models are utilized to train the data and get a needed output. In ensemble techniques we have two types one is bagging and other one is boosting. Random forest is one of the techniques that uses bagging concept which is also known as Bootstrap aggregation. The concept of bagging is as follows, suppose if we have a dataset and we split the dataset using row sampling with replacement and create multiple bootstrapped datasets with different records and we feed these datasets into multiple models and the models gets trained on the given datasets and produces outputs. Then we use a voting classifier or a majority vote to get the final output, here as we are combining outputs from different models to produce a final output this is knows as aggregation. The flow of bagging concept is given in the below figure [2.1].

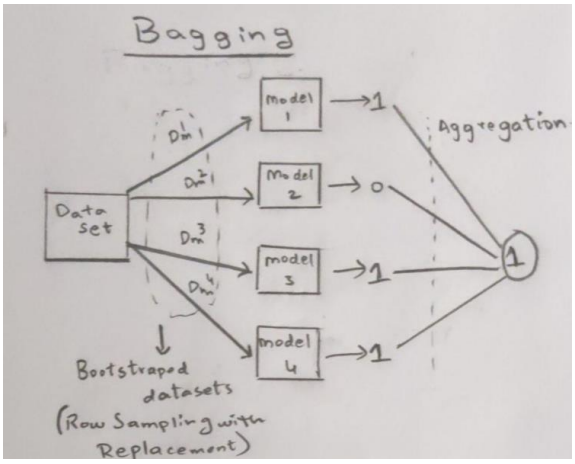

Figure [2.1]

Primarily, Random forests are made from decision trees. Although decision trees are easy to build, easy to use and easy to interpret they have one aspect that prevents them from being the ideal tool for predictive learning, namely inaccuracy. This is because they are accurate with the data used to create them, but they are not flexible when it comes to classifying new samples. Random forests combine the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy. In the above the figure 1.1 we use decision tree in place of models in random forests.

The Basic working of Random Forest is as follows, suppose if we have a dataset with ‘d’ records or rows and ‘m’ number of features or columns. Now we select some rows using row sampling with replacement (RS) and we select some columns or features known as feature sampling (FS) from the original dataset and we feed this data into the first decision tree. Similarly, we create multiple bootstrapped datasets using RS and FS and we will feed this data into different decision trees. The decision trees get trained using the input data and produces an output. Now in random forests we take the majority voting we consider the output that has maximum number of votes. The basic follow of random forest is given in the figure [2.2]

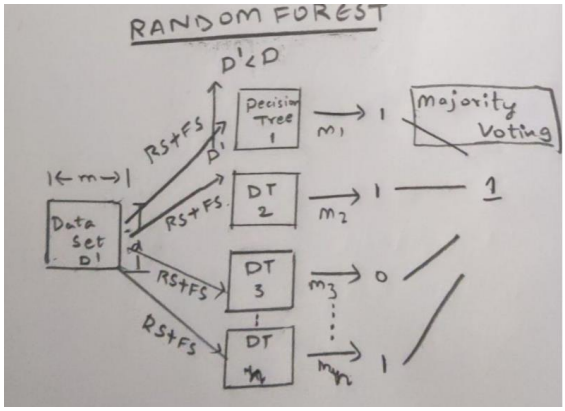

Figure [2.2]

A single decision tree has two properties low bias and high variance. Low bias indicates that if the decision tree is created to its complete depth, then the tree is properly trained with the given training data set. So, in this case the training error is less. High Variance indicates when we have new test data the trees are prone to give large number of errors. So, when we are creating a decision tree to its complete depth, we face the problem of overfitting. In random forest as we are using multiple decision trees and considering a majority voting from the outputs of each decision tree the high variance turns into low variance as each tree becomes an expert in the data that it is trained on. One more main advantage of random forest is even if there is a change in some records in the test data the random forest output it is not affected. It will still produce a low variance and high accuracy not only classification problems random forest can be used for regression problems as well. In regression problems we get a continuous value output from each decision tree and we consider the mean or median of the outputs based on the test data in the regression problems.

2 Data Set Description and Analysis.

For the Experiment purposes I have taken a dataset which contain 14 attributes. The attributes data is as follows.

- Age - age in years
- Sex -1=male, 0=female
- Cp - chest pain type
- Trestbps - resting blood pressure (in mm Hg on admission to the hospital)
- Chol - serum cholesterol in mg/dl
- Fbs - (fasting blood sugar > 120 mg/dl) (1=true; 0=false)
- Restecg – resting electrocardiographic results
- Thalach – maximum heart rate achieved.
- Exang – exercise induced angina (1=Yes; 0=No)
- Oldpeak – ST depression induced by exercise
- Slope – the slope of the peak exercise ST segment
- Ca – number of major vessels (0-3) colored by flourosopy
- Thal – 3 = normal; 6= fixed defect; 7= reversable defect
- Target - have disease or not (1=yes, 0=no)

The first five rows of the dataset are given in the figure [2.3]

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure [2.3]

And the attribute information such as count, mean, std, min of the dataset are given in the figure [2.4].

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000

thal	target
303.000000	303.000000
2.313531	0.544554
0.612277	0.498835
0.000000	0.000000
2.000000	0.000000
2.000000	1.000000
3.000000	1.000000
3.000000	1.000000

Figure [2.4]

So, As I want to build a Random Forest classification model to classify the people who will have heart disease or not (target) using some input features. First, I need to choose the input parameters or features to classify. Based on the correlation matrix output I have chosen all the features as all the features are co-related with the target attribute to classify whether a person has heart disease or not. The correlation matrix output is given the below figure [2.5]

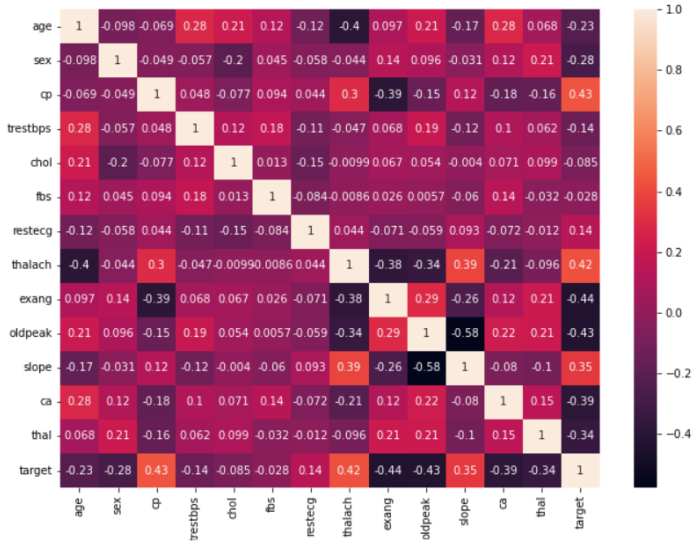

Figure [2.5]

I have plotted the box plots for some highly correlated features like chest pain, thalach (maximum heart rate achieved), Restecg (resting electro cardiography results), slope to understand the relation between input features and target variables better. The plot is given in the figure 2.6.

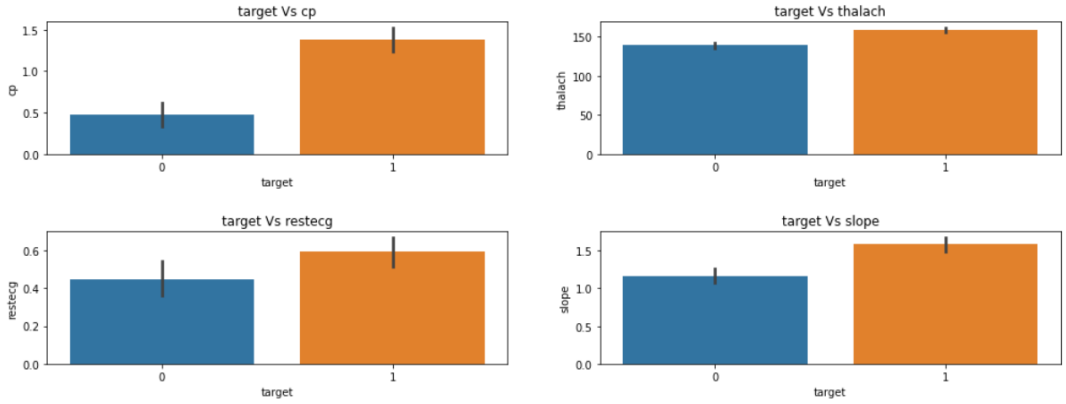

Figure [2.6]

'Thalach' feature is highly correlated with target variable out of all the other features. The histogram distribution of thalach is given in the figure [2.7] below.

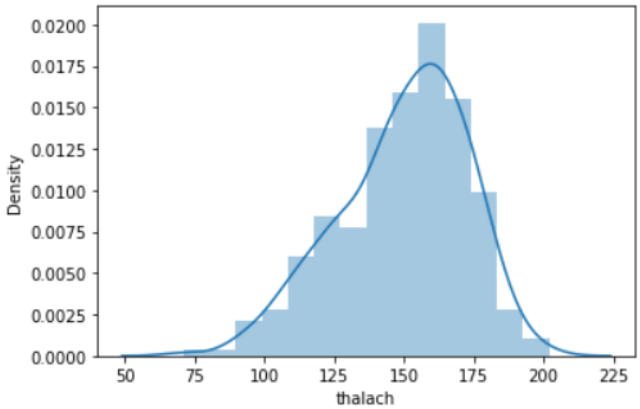

Figure [2.7]

The pair plots of all the features are plotted in below figure 2.8. From the pair plots we can infer that patients who are most likely to not suffer from the disease have a slightly greater blood pressure than the patients who have heart diseases.

Figure [2.8]

3 Experiment and Results

I have considered variable Y which holds the target variable (Having heart disease or not) and variable X which holds all the features except the target variable. I have selected all features because all the features are affecting the target variable either positively or negatively which we can observe in the data analysis above. I have split the data into 75% as train data and rest 25% as test data. I have given n_estimators parameter as 100. This parameter indicates the number of decision trees to be constructed before maximum voting this will help in improving the performance of the model. The test results are in below figure [2.9] plotted via confusion matrix and from the confusion matrix we can infer that 25 patients were predicted that they will not have heart disease, the prediction was correct (True-negative). 40 patients were predicted that they will have heart disease, the prediction was correct (True-positive). 8 patients were predicted that they will have heart disease, but the prediction was wrong (False-positive). 3 patients were predicted that they will not have heart disease, but the prediction was wrong (False-Negative). As False-Negative and False-Postive values are very less our model is performing good and has good accuracy of 85.5%.

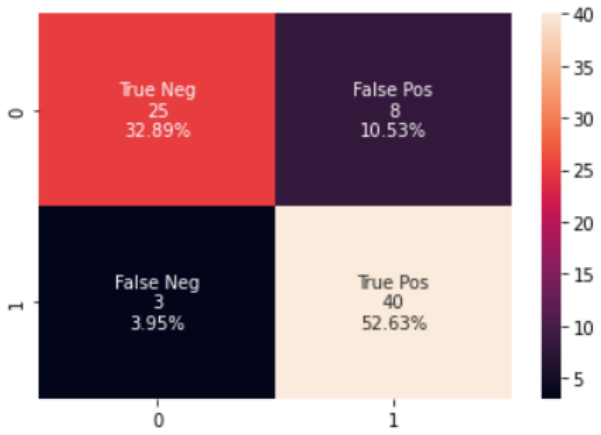

Figure [2.9]

The ROC curve for the random forest model is given in the below figure [2.10]

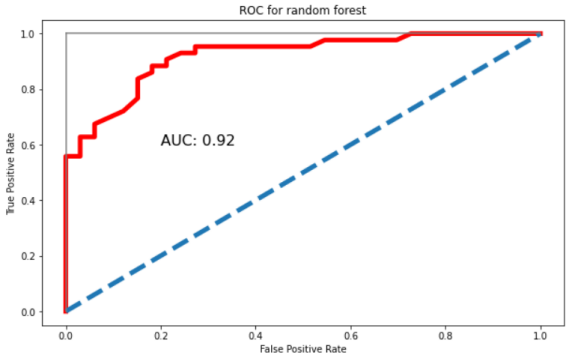

Figure [2.10]

The overall accuracy or classifier score of the model is **85.5%.** And remaining results of model like Roc_Auc score, precision score, recall score, F1 score are given in the below figure [2.11].

	precision	recall	f1-score	support
0	0.89	0.76	0.82	33
1	0.83	0.93	0.88	43
accuracy			0.86	76
macro avg	0.86	0.84	0.85	76
weighted avg	0.86	0.86	0.85	76

Figure [2.11]

References

[1] Charan kakararaparthi, Heart_Diseases, Prediction of Heart Disease by Effective Machine Learning techniques.

[2] Chengyou Chen, Ensemble Learning, Lecture Slides, 10/11/2022.

[3] James Thorn, A summary of the Basic Machine Learning models, 15/02/2021.