The University of New Mexico

CS 529 - Introduction to Machine Learning

Program 1 – Decision Trees

Submitted By:

Vasman Kaur

UNM ID:101892375

Submitted To:

Prof. Lydia Tapia

WHITE WINE DATASET:

The wine dataset determines the quality of white vihno verde wine samples. The data set is taken from the UCI machine learning repository [1].

The total attributes for this data set are 12 which are as follows:

1.) Fixed Acidity

2.) Volatile Acidity

3.) Citric Acidity

4.) Residual Sugar

5.) Chlorides

6.) Free Sulfur Dioxide

7.) Total Sulfur Dioxide

8.) Density

9.) pH

10.) Sulphates

11.) Alcohol

12.) Quality (score between 0 and 10) which is Output variable (based on sensory data)

The accuracy for training and testing dataset using Decision Tree Classifier when the parameters were set to default was:

Training Dataset: 1.00

Testing Dataset: 0.552

The accuracy for training and testing dataset using Random Forest when the parameters were set to default was:

Training Dataset: 1.00

Testing Dataset: 0.648

It is seen that there is a huge difference between the accuracy for training and testing data in both the models. Random forest however performs better than the Decision tree classifier model even with the default parameters. But, in both the cases there is overfitting which is the accuracy for training data is extremely high but for testing data it is rather low. Outliers are mainly one of the main reasons when it comes to overfitting. So, exploring the data a little further through the boxplots it is seen that there are many outliers for the attribute volatile acidity, citric acidity, chlorides, free sulfur dioxide, pH and sulphates and the number of outliers a lesser for fixed acidity, residual sugar, total sulfur dioxide and density and none for alcohol. One of the things that is a part of preprocessing the data which deals with outliers in machine learning is feature scaling. It is however not required for the trees because decision trees and random forests are scaling invariant. [2]

Below are box plots to better understand the wine data and the outliers.

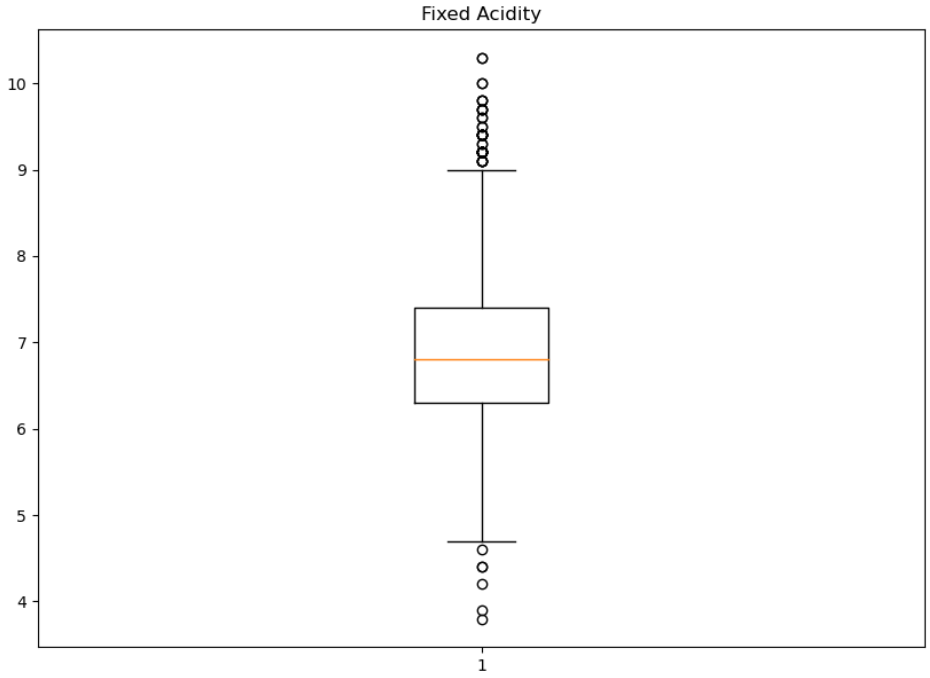

Fig 1. Fixed Acidity

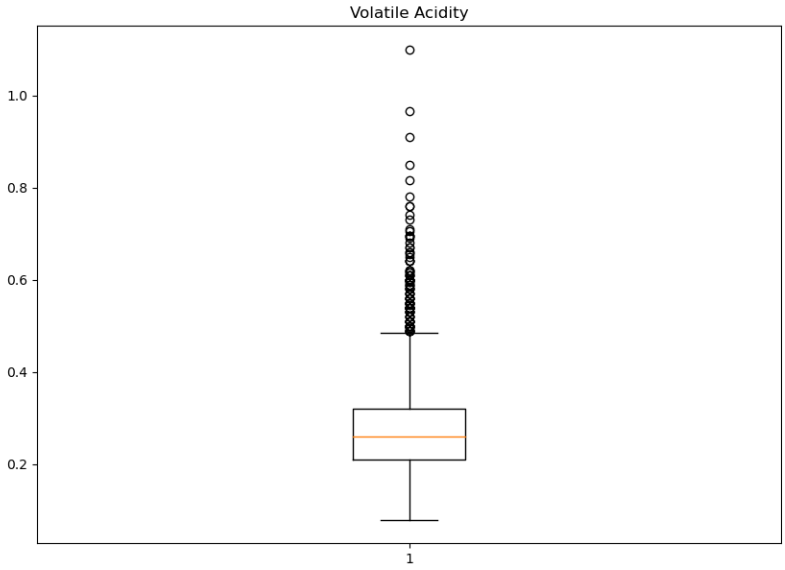

Fig 2. Volatile Acidity

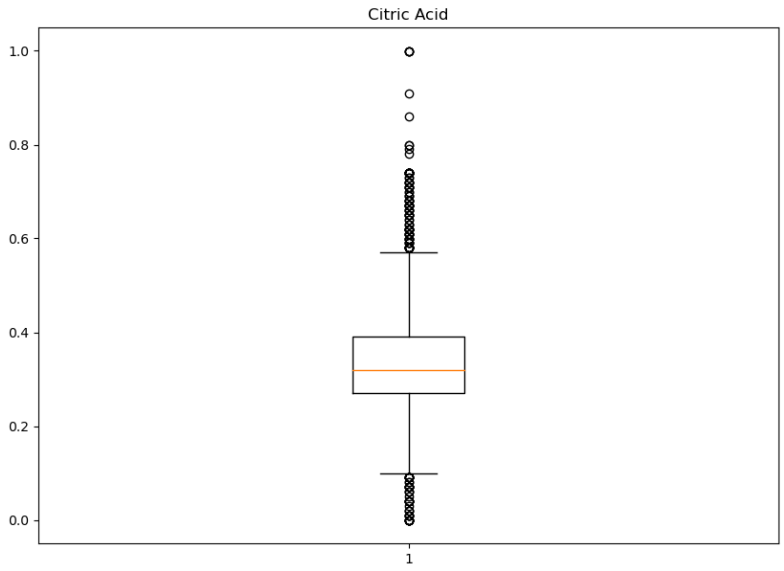

Fig 3. Citric acid

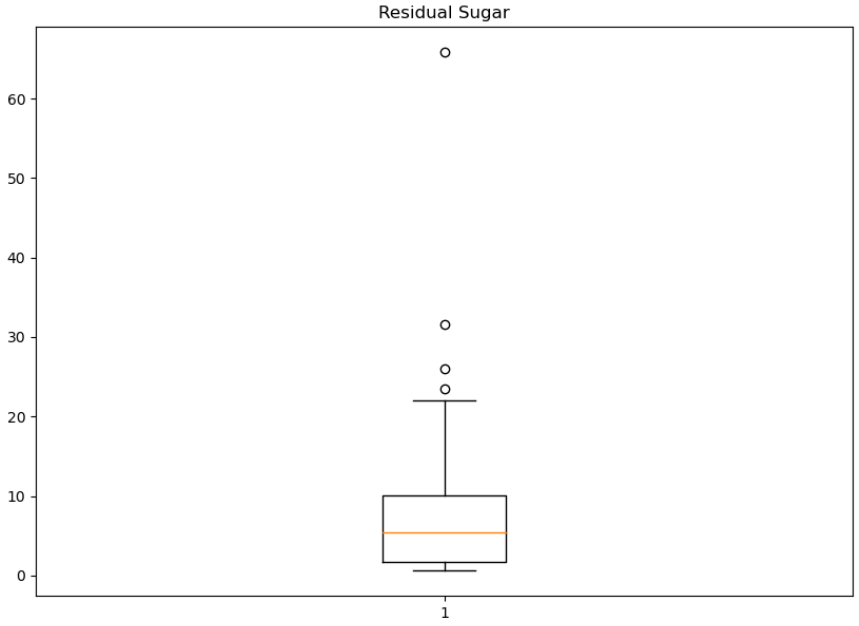

Fig 4. Residual Sugar

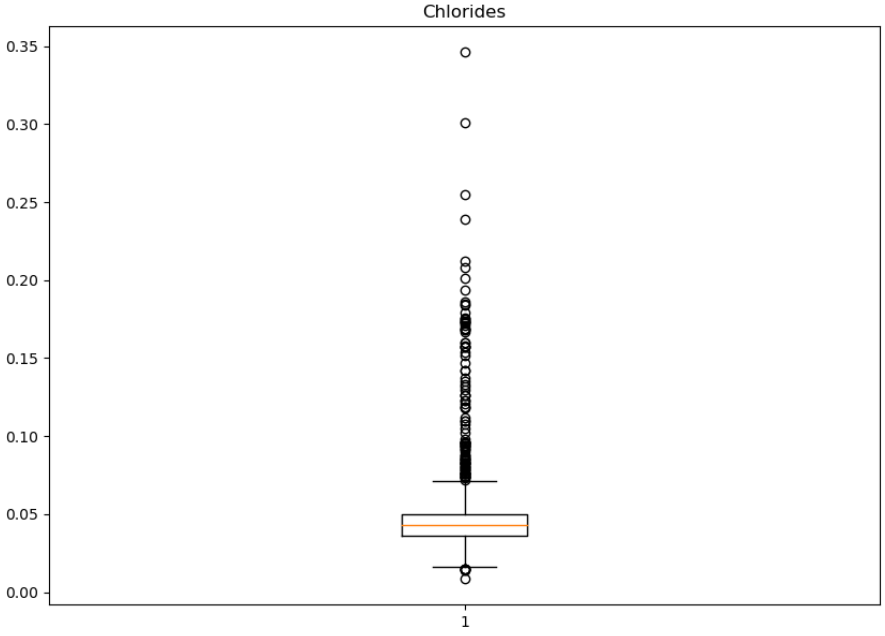

Fig 5. Chlorides

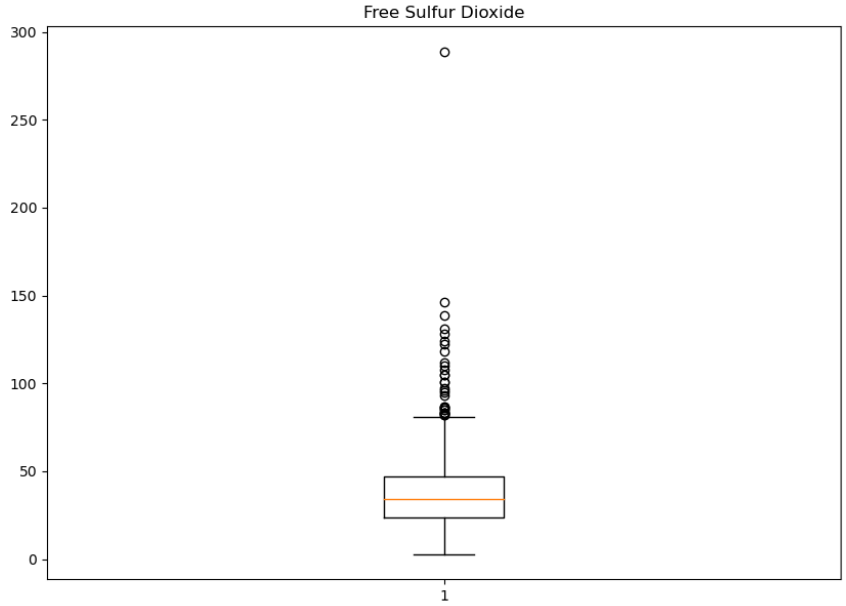

Fig 6. Free sulfur Dioxide

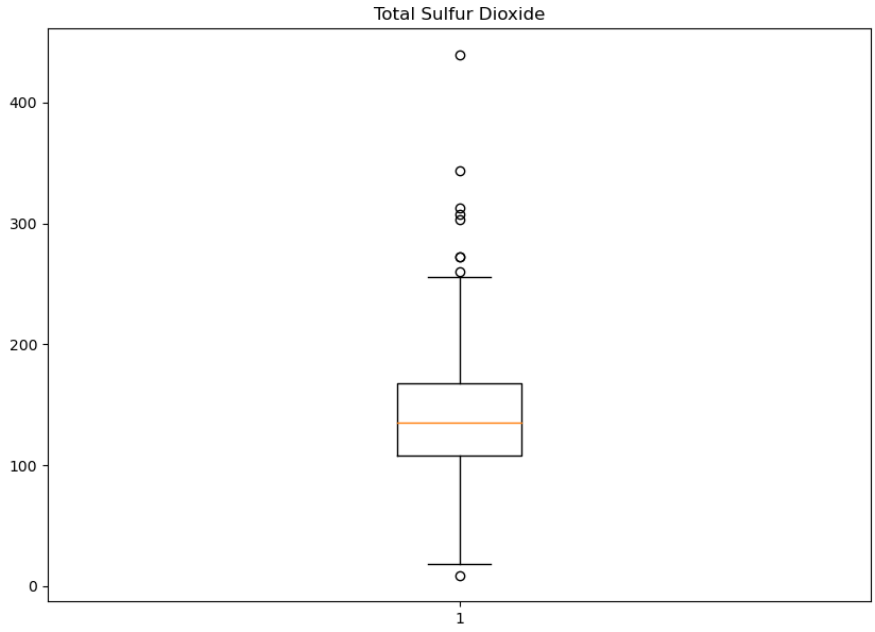

Fig 7. Total Sulfur Dioxide

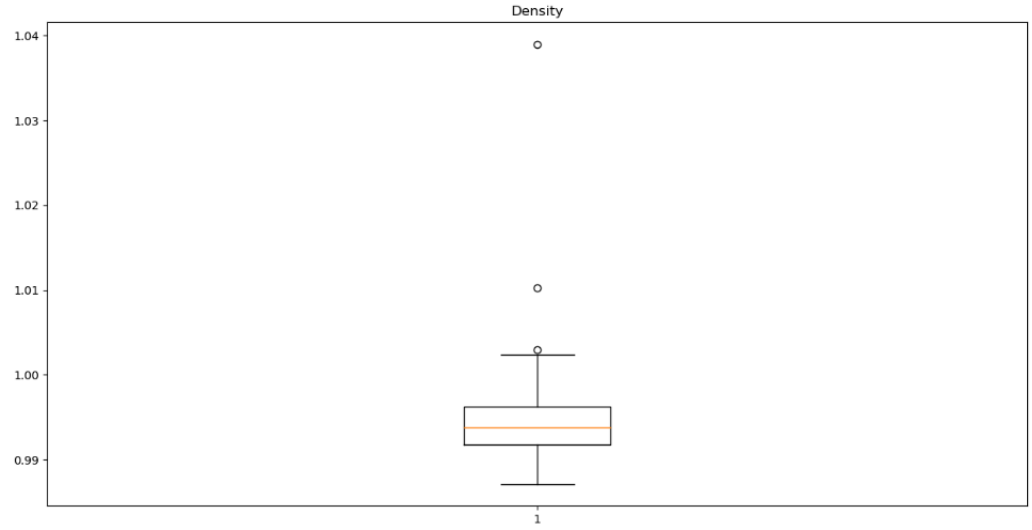

Fig 8. Density

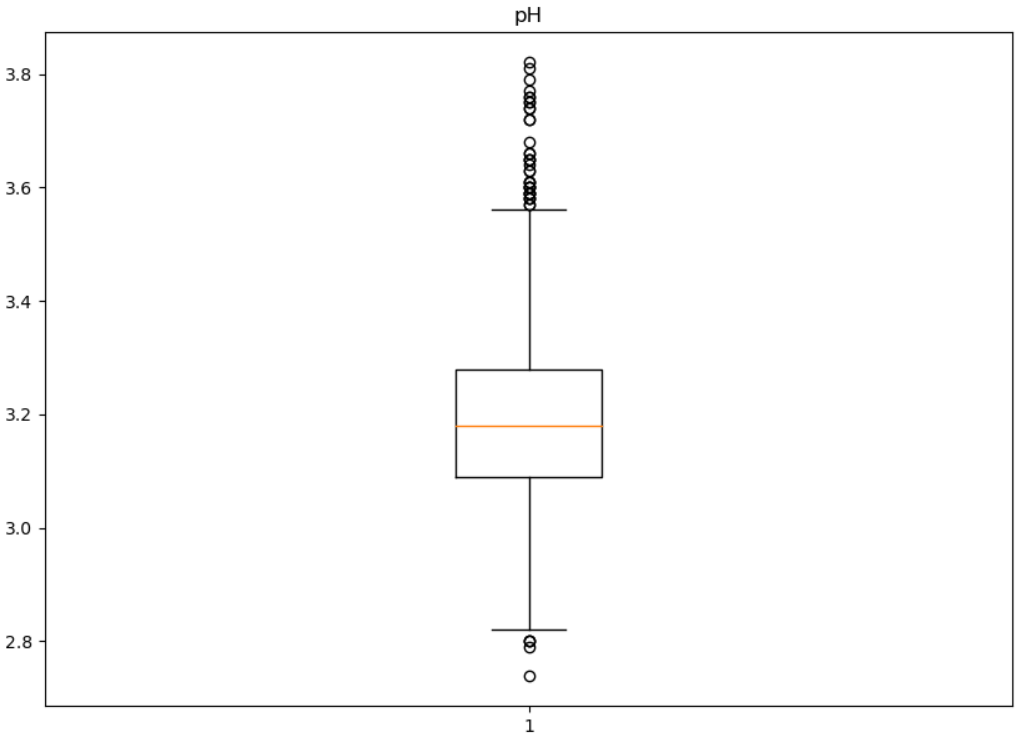

Fig 9. pH

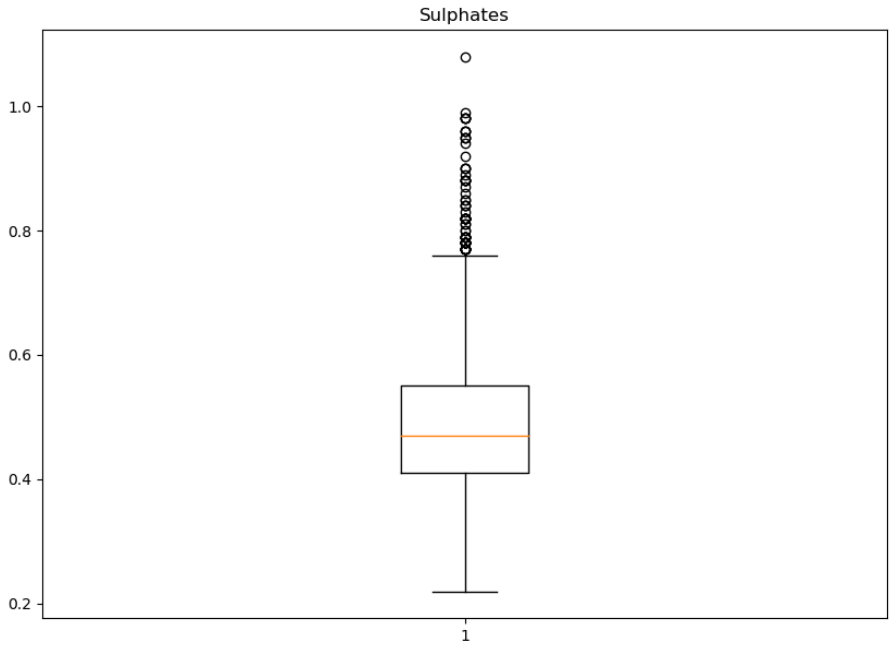

Fig 10. Sulphates

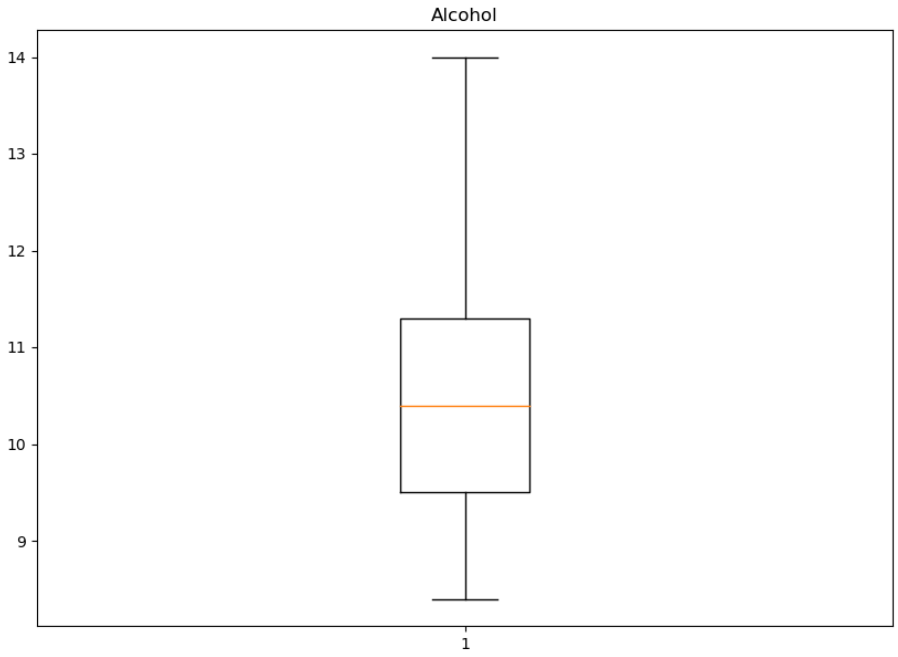

Fig 11. Alcohol

ACCURACY VS PARAMETER PLOTS FOR WINE DATASET

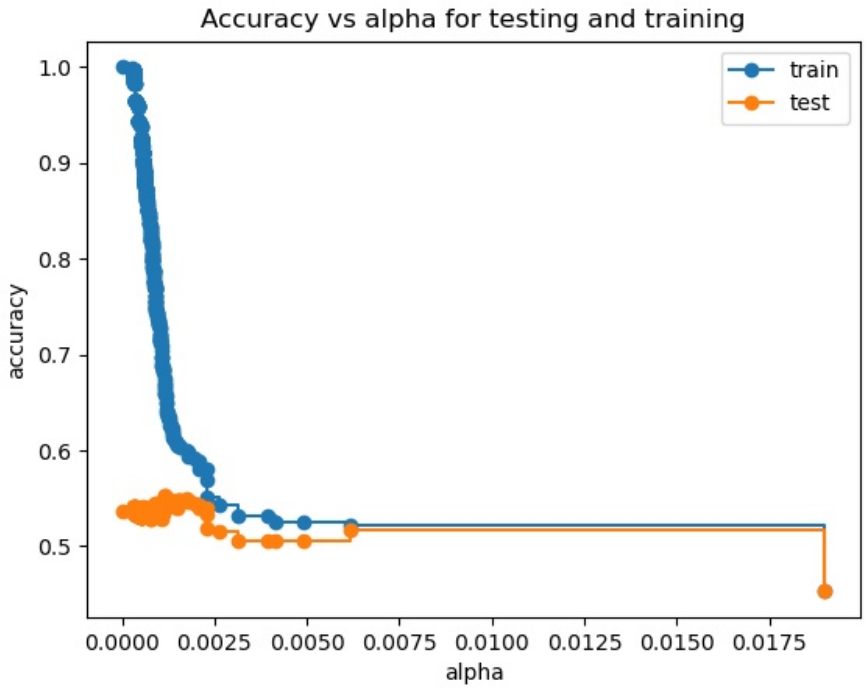

Fig 12. Accuracy vs Alpha for Wine Dataset

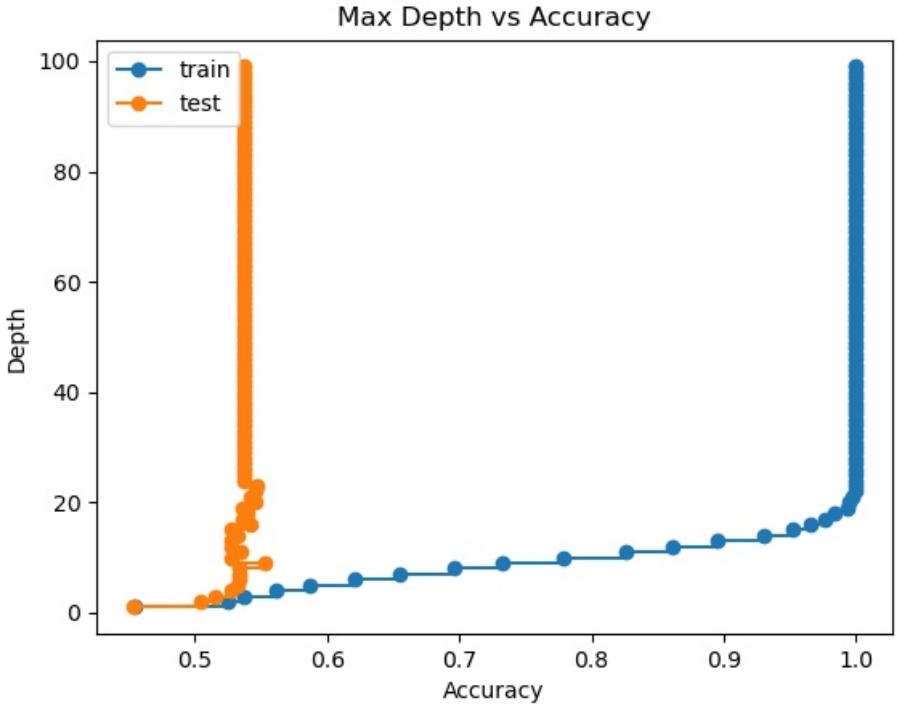

Fig 13. Max Depth vs Accuracy for Wine Dataset

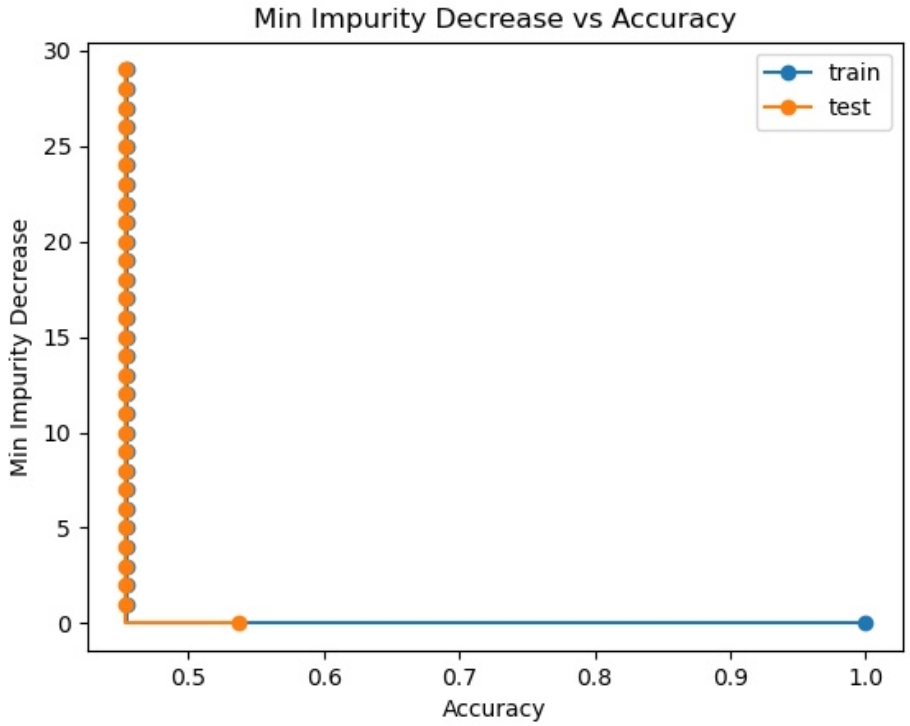

Fig 14. Min Impurity vs Accuracy for Wine Dataset

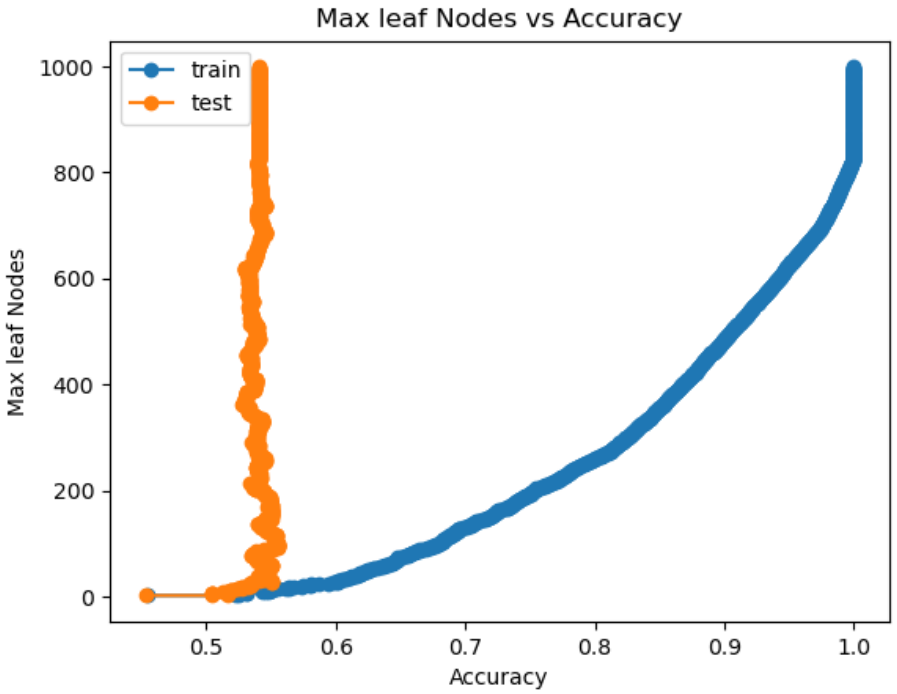

Fig 15. Max Leaf Nodes vs Accuracy for Wine Dataset

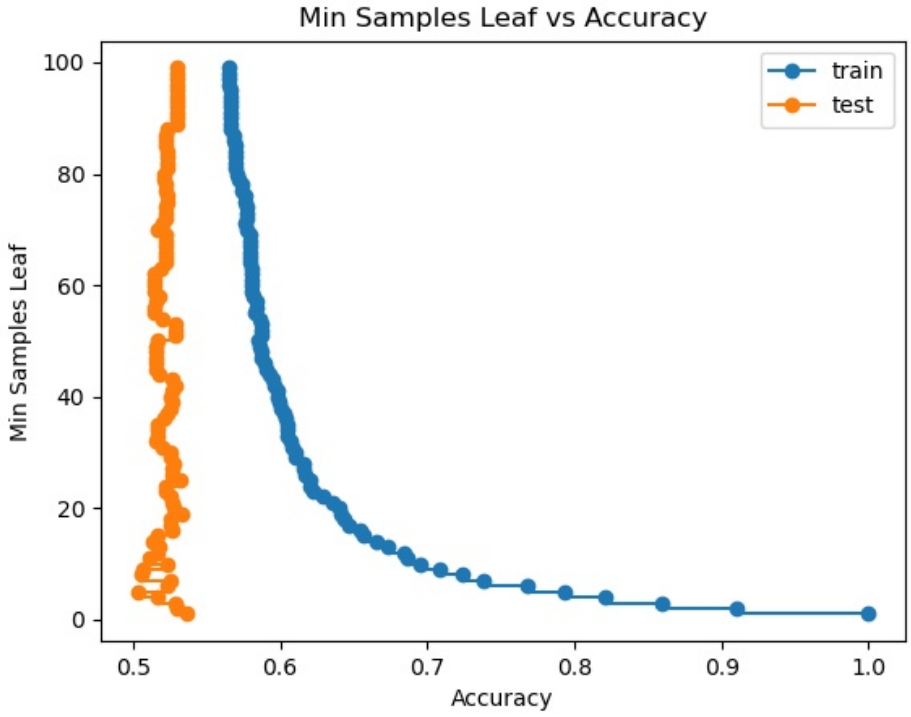

Fig 16. Min Samples Leaf vs Accuracy for Wine Dataset

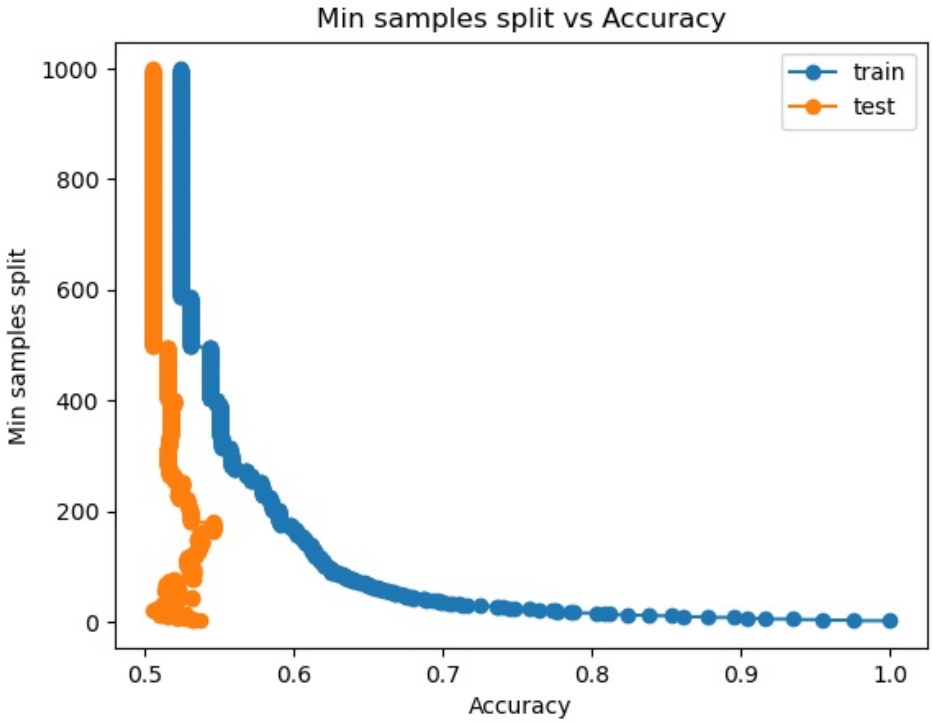

Fig 17. Min Samples Split vs Accuracy for Wine Dataset

Using the parameter values from the plots with alpha 0.0, Max depth 20, Min Impurity Decrease 0, Max Leaf Nodes 50, Min Sample Leaf 90, Min sample Split 200 for Decision tree we got the accuracy of 0.566 for training data and 0.530 for testing data. Which is not any better from the accuracy with default parameter values. The random forest however performs well even with default parameters as compared to the decision tree. This is because Random Forest is quite robust to noise and we do not really have to prune the random forest [2].

So, further implementing random forest with parameters random state 10, criterion gini, max depth 22, n_estimators 1000 and ccp alpha 0.0 we get the accuracy of 0.664. Higher the depth of the tree more are the chances for overfitting. So, taking the depth to be 22 and pruning the tree we get good accuracy. These are the parameters used for the random forest with the Kaggle dataset which gave the accuracy of 64.80 on Kaggle dataset.

GENE DATASET:

The gene dataset focuses on recognizing exon/intron boundaries (EI sites), and intron/exon boundaries (IE sites) or neither (N). The dataset is taken from UCI machine learning repository [3].
There is total 61 attributes for this dataset and one index column.
The 60 attributes are sequential DNA nucleotide positions and the 61st column is the class (IE, EI or N)
The letters representing the nucleotide have been mapped to numbers for this dataset.
The Accuracy for decision tree with default parameter is 1.00 for training data and 0.910 for testing data.
Based on the plots below if we change the values for the parameters as follows:
random_state=5, ccp_alpha=0.0014, max_depth=9, min_impurity_decrease=0, max_leaf_nodes=20, min_samples_leaf=1, min_samples_split=50 the accuracy for training data came to be 0.931 and for testing data it came to be 0.918 which is better than the values we got with the default parameters. Also, the values for testing and training with modified parameters have less difference in between which shows reduced overfitting. Below are the accuracy plots with various parameters for gene dataset:

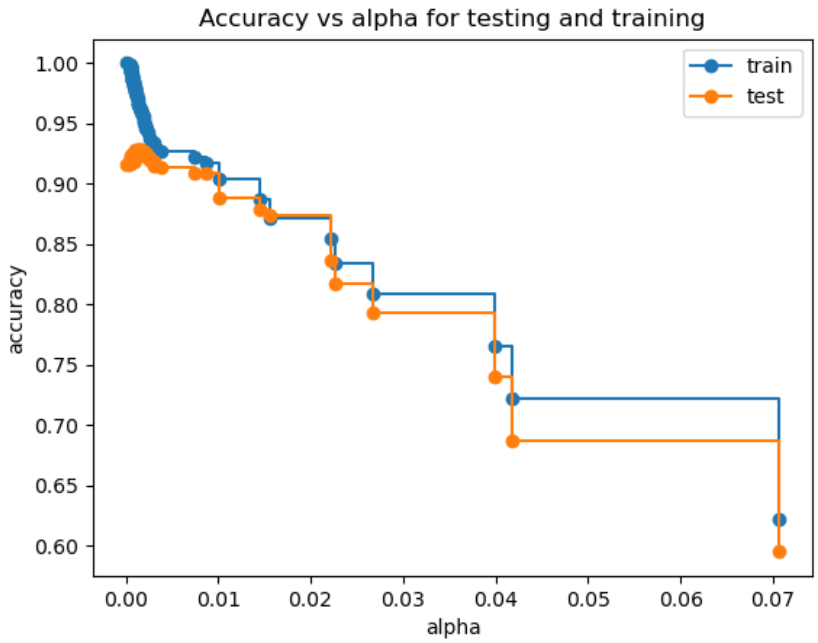

Fig 18. Accuracy vs Alpha

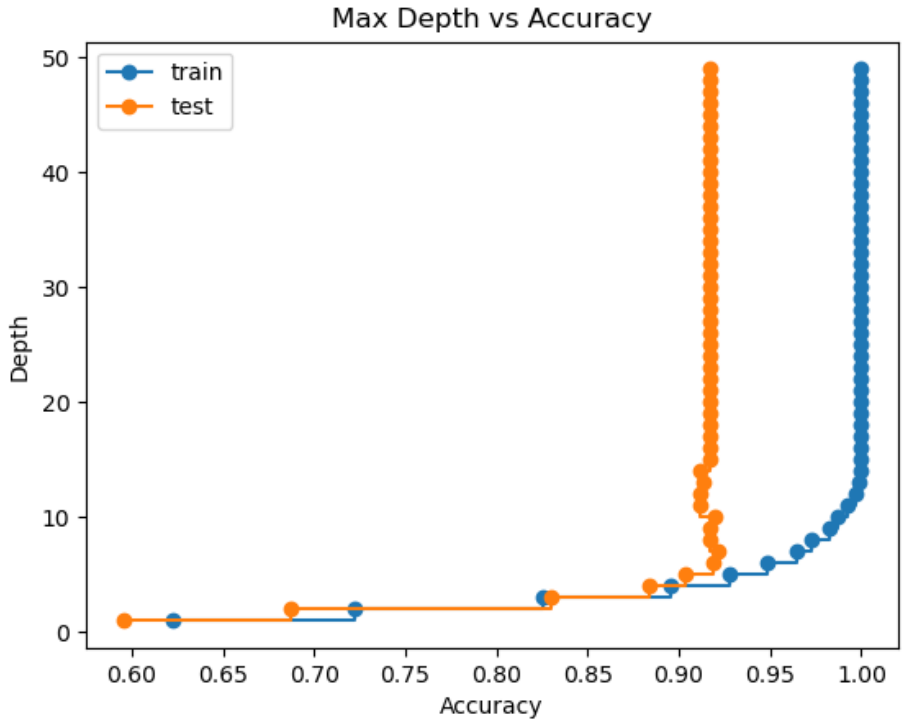

Fig 19. Max Depth vs Accuracy

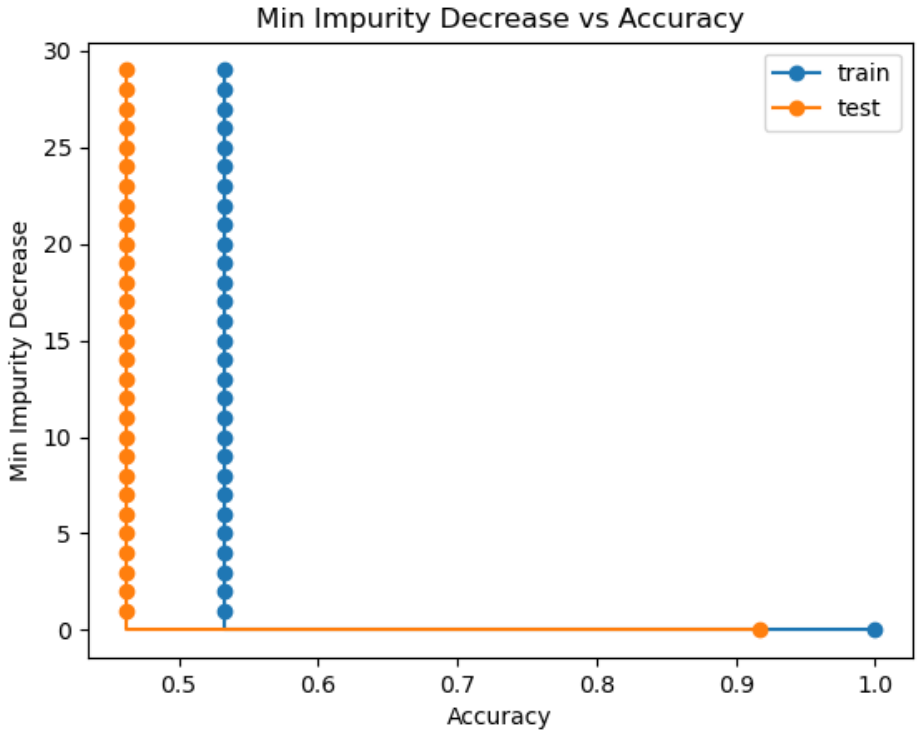

Fig 20. Min Impurity vs Accuracy

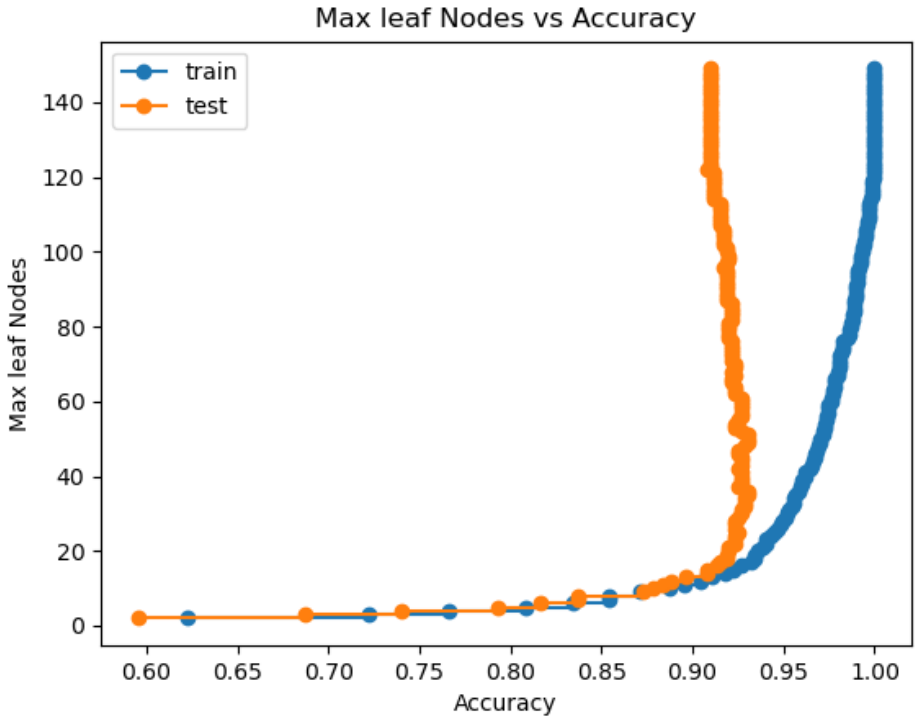

Fig 21. Max Leaf Nodes vs Accuracy

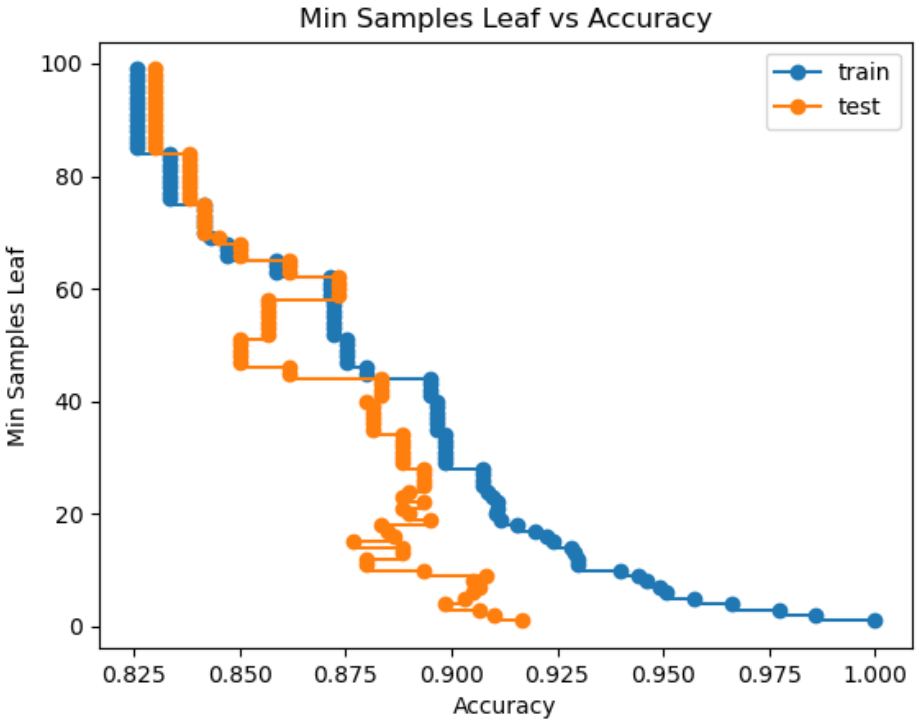

Fig 22. Min Samples Leaf vs Accuracy

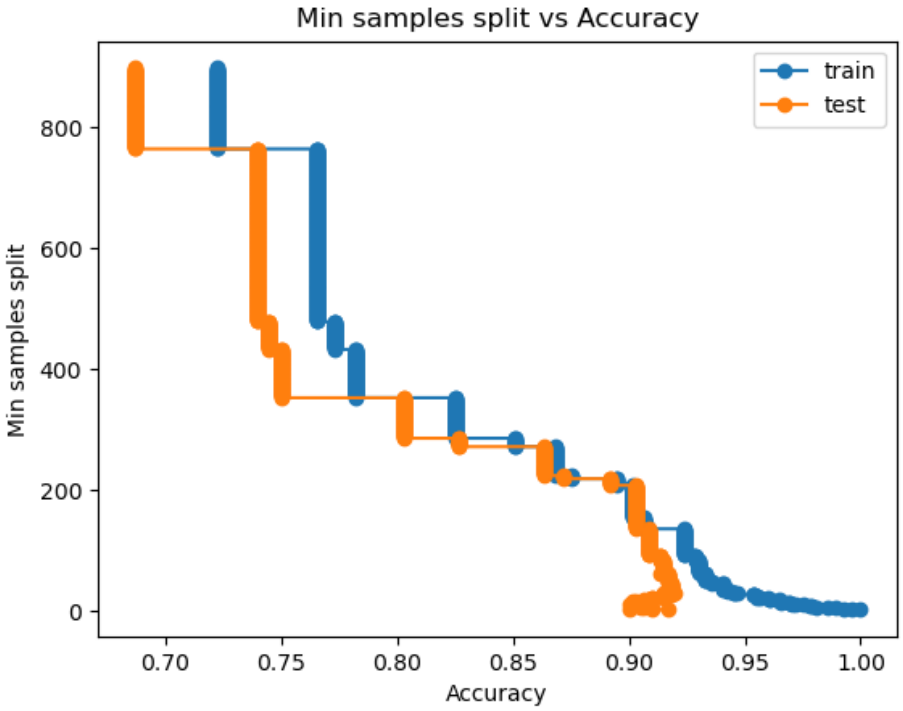

Fig 23. Min Samples Leaf vs Accuracy

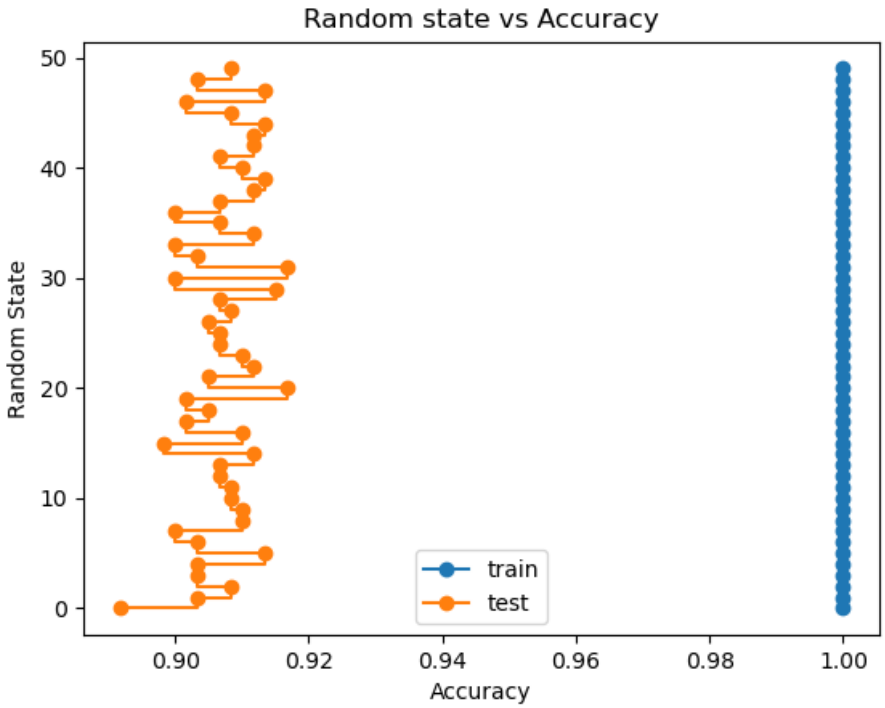

Fig 24. Random State vs Accuracy

The implementation of Random forest gave good results even with default parameters. The accuracy for training data set came to be 1.00 and for testing data set it came to be 0.948 which shows that random forests are much better than decision tree.

When the parameters for random forest modified as criterion='gini', max_features='log2', random_state=10, max_depth=15, n_estimators=1000 we got the training accuracy to be 1.00 and the testing accuracy to be 0.95 which is not a huge change from what the values were with the default parameters but is still an improvement. The above stated parameters were used with the Random Forest on the Kaggle dataset which gave the accuracy of 96.33

From the plots above it can be seen that the alpha value for both the datasets does not make much difference on the accuracy. The only point where the accuracy is seen to be high is when the alpha value is low or 0.0

Same is with minimum impurity decrease, accuracy for this parameter is high when the value for minimum impurity decrease is zero. Max depth however is a crucial parameter to tune. Because it is directly related to tree pruning so it impacts the accuracy. Same is with the maximum leaf nodes and minimum samples leaf. Minimum values for these parameters are seen to be beneficial.

References

[1] D. a. G. C. Dua, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[2] V. M. Sebastian Raschka, Python Machine Learning, Packt Publishing, 2017.

[3] D. a. G. C. Dua, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29.