Decision Tree V.S Random Forest V.S Extra Trees classifiers

Kaleb Nim  
Diploma in Applied Artificial Intelligence and Analytics   
School of Computing   
Singapore Polytechnic

***Abstract: This technical paper aims to compare the predictive power of the different Tree based classifiers. Notably Descision Tree classifier ,Random Forest classifier and Extra Tree classifiers. The performance of these algorithms will be compared based on: 1.scoring metric, 2.tendency of overfitting and 3.processing time. This will be conducted on 3 varied synthetic datasets generated.***

***Keywords— comparison, machine learning, python, Decision Tree, Random Forest, ExtraTrees classifier, supervised learning, classification***

# Introduction

Descision Tree and Random forest are very popular maching learning algorithm used for both classification and regression problems. This is due to the nature of these algortirhms being simple yet accurate in predicting. Descision Trees algorithms have been around for a long time. Compared to random forest only being introduced by Breiman [1] in 2001.

# Related works

In the past there have been similar works of comparing between these ensemble methods and descision tree

An earlier investigation of this problem by Prajwala T R. [2]. He used compared only randomforest and decision tree topredict whether it is going to rain the next day of not. He compared the two models interms of Redistribution error rate (i.e., error rate computed on the training sample) and time taken to construct the model.

Another investigation was conducted by Jehad A, Rehanullah K, Nasir A, Imran M [3]

# Decision Trees

## How it works

Abit of background behind descision trees. A decision tree generally speaking is a binary tree that recursively splits the dataset until we are left with pure leaf nodes ( aka only one type of class) .There are two nodes in a decision tree, a decision node and a leaf node. A decision node conatins a condition that splits the dataset. The leaf node helps us decide the class of the new data point. The logic to determin the class will be explained further in the section III part B.

* 1. Entropy and Information Gain

So how does the decision node decide on the optimal split? The answer to that is the model will split the data in which the information gain is maximized. To calculate information gain, we need to figure out the information contained in a state. The way to quantify this is to use Entropy. If Entropy is high then we are very uncertain about the randomly picked split.

pi = probability of class i

A state that has the minimum entropy of 0. is called a pure node.

Now to find the information gain corresponding to a spilt. We will subtract combined entropy from the child node from the entropy of the parent node.

E = Entropy

The model compares every possible split and takes the one that maximized information gain. In other words, the model traverses through

condition that splits the dataset. The leaf node helps us decide the class of the new data point. The logic to determin the class will be explained further in the section III part B.

# Random Forest

# How it works

The some stuff write here

## How it worksPt2

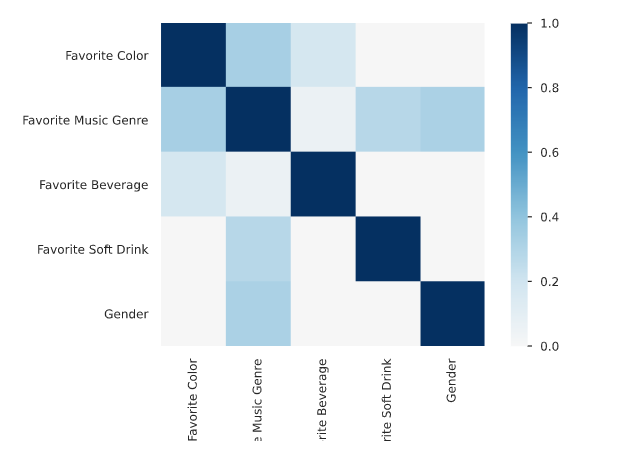
One uses bagging the other doesn’t

1. Dataset

We will compare our models on a very interesting dataset: Gender Classification. To classify gender based on personal preferences. Note that this is a very small dataset



Figure1: Peek into dataset



Gender most correlated to Favourite Music Genre.

Since the categorical features are nominal. I’ll perform one-hot-encoding for all features. then Label Encode Gender 0: Male 1: Female.

1. Results and Discussion

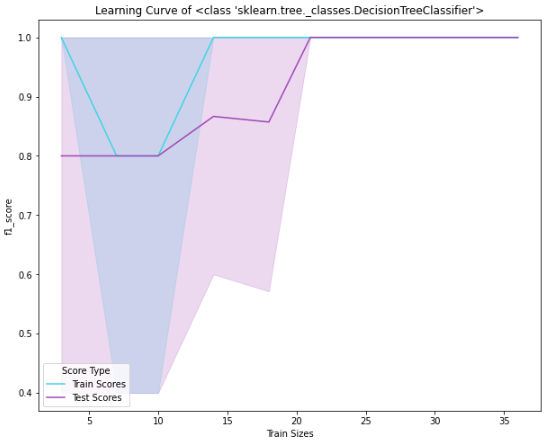
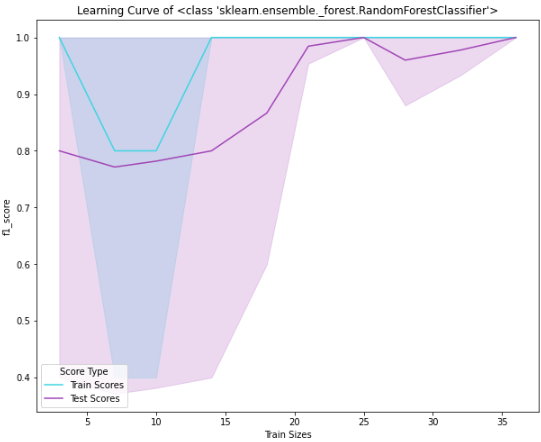
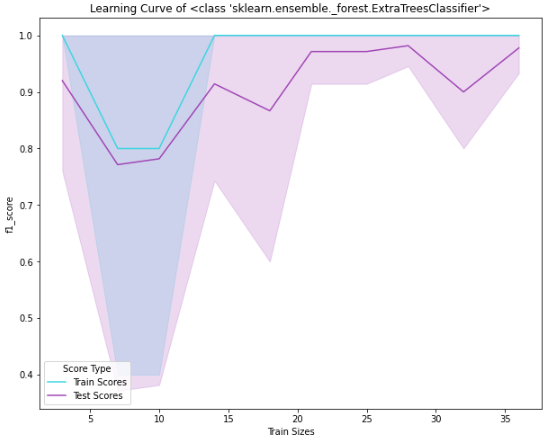


Figure3 : Learning Curve of Descision Tree Classifier

Figure 3: Learning Curve of Random forest Classifier



*Figure 4 Learning Curve of ExtraTreesClassifier*

**X Conclusion**

Through this tiny dataset with only 66 instances, and 5 features, From the Learning Curves, we can see that all ensemble have very high bias but were able to generalise to dataset. All 3 models were able to get a perfect F measure of 1.

##### Acknowledgement

I thank Dr Wilson and Dr Peter, who implemented this technical paper section into the assignment.

##### References

1. Prajwala T R., 2015. A Comparative Study on Decision Tree and Random Forest Using R Tool [online] Department of Information Science and Engineering, CMRIT, Bangalore. Available at: <https://www.researchgate.net/publication/272425234\_A\_Comparative\_Study\_on\_Decision\_Tree\_and\_Random\_Forest\_Using\_R\_Tool> [Accessed 25 May 2022].
2. Leo Breiman. 2001, RANDOM FORESTS [Online].Statistics Department University of California Berkeley Available at: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf> [Accessed 25 May 2022].

# 

## 

