Assignment [5] Data lab 2022

Random Forest

Mustefa Abrahim
Deprt. Data Science
Indian Institute of Technology Madras (IITM)
ge22m014@smail.iitm.ac.in

***Abstract*—The main goal of this paper is to predict car safety using the Random Forest algorithm. Random Forest is a supervised machine learning algorithm that grows and combines multiple decision trees to form a forest. It can be used in R and Python for classification and regression problems. [3] We predict the safety of cars based on a variety of factors such as population, cost of purchase, number of doors, and maintenance costs, among others. We concluded with 97% accuracy in predicting car safety which was better than the Decision tree algorithm.**

keywords: Modeling, Random Forest, Visualization

I. Introduction

Random Forest is a supervised machine learning algorithm that grows and combines multiple decision trees to form a forest. It can be used in R and Python for classification and regression problems. The random forest algorithm is a bagging method extension that employs both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or the random subspace method, generates a random subset of features, ensuring that decision trees have low correlation. This is a significant distinction between decision trees and random forests. Random forests select only a subset of the possible feature splits, whereas decision trees consider all of them.

In this we have done, first checking of the given data to analyze it's features and Identifies whether it has null values or not, and we got some unrelated

Our main aims of this work are to predict the safety of the car by random forest algorithm and analysis what is the difference between the decision tree and random forest.

Generally, we have done the prediction of car safety by using random forest algorithms, random forest algorithms the combinations of the different decision trees, and the decision of the output was given by the majority vote or average of the overall prediction from the decision tree, so random forest is better than a decision tree.

II. Random Forest

How does the Random Forest algorithm work? Random Forest expands several decision trees, that are combined to make a more accurate prediction. The Random Forest model is based on the idea that multiple uncorrelated models (individual decision trees) perform far better as a group than they do individually. When using Random Forest for classification, each tree provides a category or a vote, and the forest selects the identification with the most votes. When using Random Forest for regression, the forest selects the average of all trees' outputs.

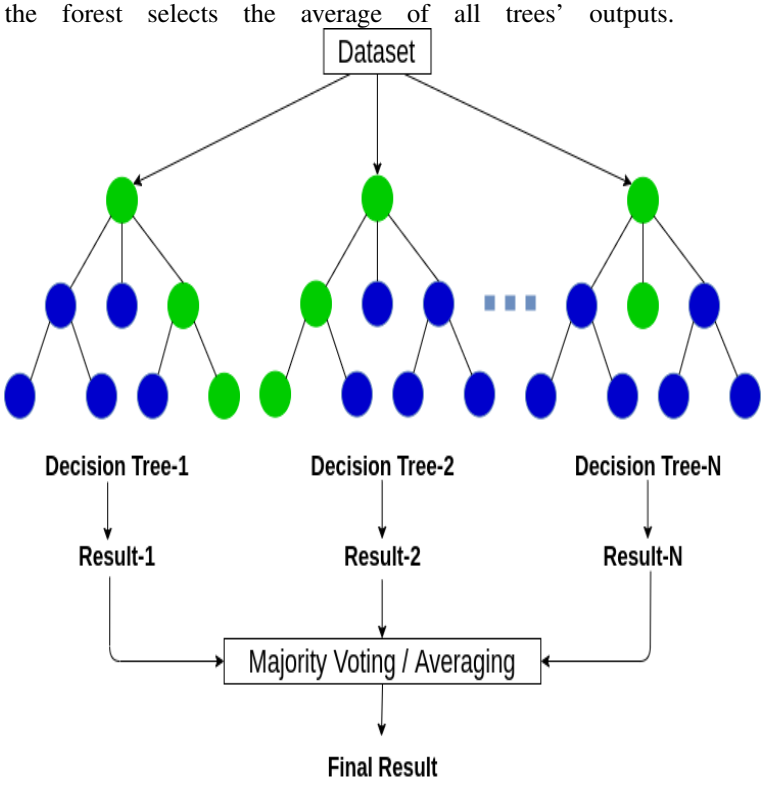

Figure 1: Random Forest model

III. What are the advantages of Random Forest?

Random Forest is extremely popular, and with good reason! It has a number of advantages, ranging from accuracy and efficiency to relative ease of use. Scikit-learn provides a simple and efficient random forest classifier library for data scientists who want to use Random Forests in Python. The most convenient advantage of using random forest is its ability to automatically correct for decision trees' tendency to overfit to their training set. When executing this algorithm, using the bagging method and random feature selection almost completely eliminates the problem of overfitting, which is great because overfitting leads to inaccurate results. Furthermore, even when some data is missing, Random Forest usually retains its accuracy. When analyzing a large database, a random forest is

far more efficient than a single decision tree. Random Forest, on the other hand, is less efficient than a neural network. A neural network, also known as a neural net, is a collection of algorithms that reveal the underlying relationship in a dataset by mimicking how the human brain thinks.

Neural nets are more complex than random forests, but they produce the best results possible by adapting to changing inputs. Unlike neural nets, Random Forest is designed to allow for rapid development with few hyper-parameters (high-level architectural guidelines), resulting in less setup time. So, the key benefits of using Random Forest are:

- Ease of use
- Efficiency
- Accuracy
- Versatility – can be used for classification or regression More beginner-friendly than similarly accurate algorithms like neural nets

IV. Data

The Car Evaluation Database was derived from a simple hierarchical decision model. The prediction task is to classify a car based on its safety. [3] The car dataset is divided into two parts. The first is Car Acceptability, and the second is Technical Characteristics. The overall price of buying and the cost of maintenance are two aspects of car acceptability. The number of doors (doors), the capacity in terms of people carried, the size of the luggage boot, and the car's estimated safety. [2]

- Number of occurrences: 1727
- The number of attributes: 7.
- Null Value: no.

  The table below shows some attributes of our dataset in the Car evaluation we used to predict car safety. As we can see in this table, we have some integer values and some character values with different data types, so to do the decision tree we have to make the same type otherwise it will be difficult to give the decision on the different data types, so we encoded to the number and made the decision on it. [2]

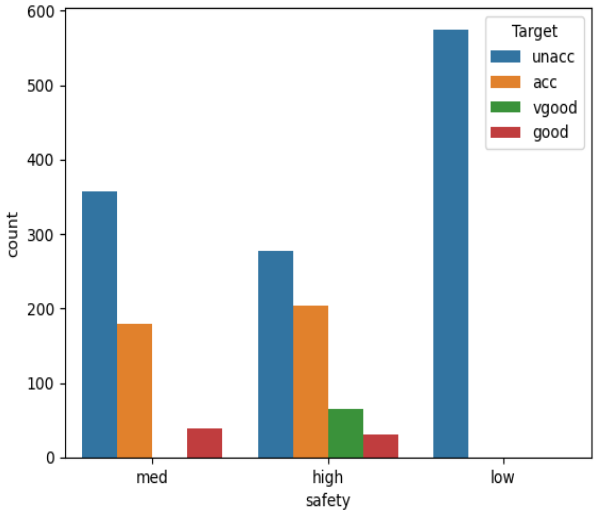

Figure 2: Visualization of original Data

Attribute	Attribute values
Buying	high,high,med,low
maint	high,high,med,low
doors	2,3,5more
persons	2,4,,more
lug_boot	2,3,4,more
safety	small,med,med,big
Target	uacc,acc,good,vgood

Table 1: Attribute and its values in Car Evaluation

V. Our Problem

We solved the prediction of car safety by the random forest algorithms, which used the 1000 estimator and 50 max_depth and go the results as the following

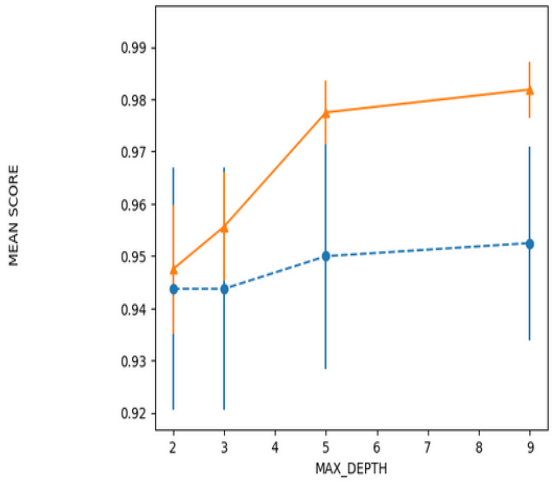

Figure 3:1st Mean vs Max_depth

This visualization is when we took the first mean vs max depth as we see from the graph are closest to each other at the initial point and far from each other at the endpoint.

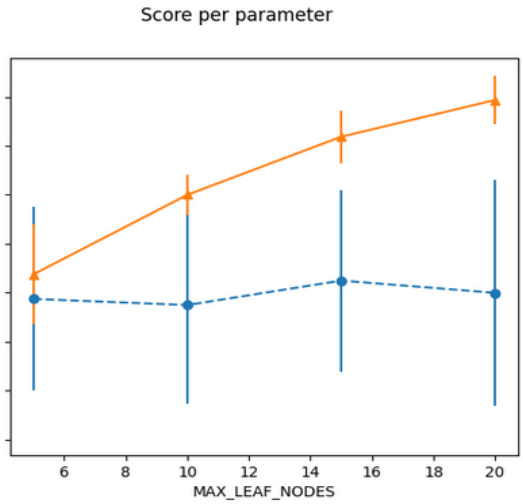

Figure 4:2nd Mean vs Max_depth
This visualization is when we took the first mean vs max depth as we see from the graph are closest to each other at the initial point and far from each other at the endpoint, the same as the first one.

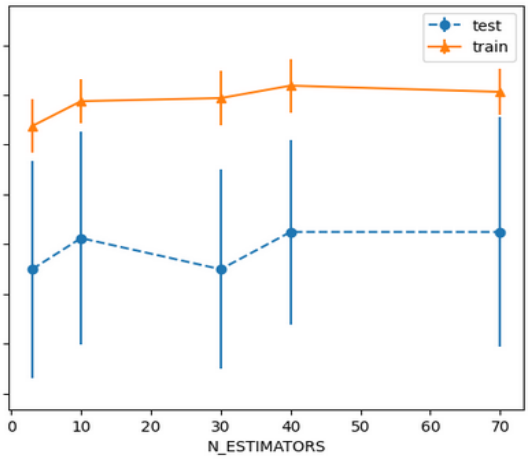

Figure 5:Nth Mean vs Max_depth (average)
This visualization is the average of the nth decision, it does not look like the 1st and 2nd closest to each other at the initial point and far from each other at the endpoint.
From this, we observe the Random forest is better than the decision tree algorithm as it uses the average of the many decision trees.

VI. Conclusions

In this paper, we did the prediction of Car safety based on the average output of different decision trees by combining different Decision tree algorithms, and finally, we got 97% accuracy for the Cars safety, we used the train and test classification by the ratio of 80 to 20, the accuracy we got in the random forest was better than the Decision tree.

References

[1] Hastie T.J. Springer; New York: 2008. The elements of statistical learning: data mining, inference, and prediction. [Google Scholar]
[2] Amit, Y., Blanchard, G., Wilder, K. (1999). Multiple randomized classifiers: MRCL Technical Report, Department of Statistics, University of Chicago.
[3] Boccaletti S. Modeling and forecasting of epidemic spreading: the case of COVID-19 and beyond. Chaos Solitons Fractals. 2020 [PMC free article] [PubMed] [Google Scholar]