mikaal10802@gmail.com

[Company name]  [Company address]

report assignment 2 fa21-bse-143

**Question 1:**

The dataset in this question was about datasauras. In this basically, if we plot the points representing a dataset, for example all the points representing the dataset dino, so a dinosaur will be made. With help of those points.

For the prediction of the data, and training, we were asked to use the Decision Tree classifier and random forest classifier, classifier because, the output value, the **dataset** is not a numeric value.

First using the Decision Tree classifier, without any argument it was identified, that the accuracy was just 41.12 %. To improve this accuracy some variations were done in constructing the decision tree, in first implementation using the Entropy calculation, the accuracy was less than that of 41.12% it came out to be 40.5 percent whereas the gini method provide us the accuracy of 40.8%

Then I utilized the grid methodology which helped the accuracy to improve to 43%.

Best Parameters: {'max\_depth': 150, 'min\_samples\_split': 2}

Best Accuracy: 0.4302244617498855

This was the highest accuracy I could avail using the decision Tree classifier.

The Random Forest classifier on the other hand by default parameters by python was more accurate than that of the variations I made in the decision Tree classifier. The accuracy I got was **44.32%** then by using the n\_estimators, and max\_depth to 14 this accuracy was improved to 45.13%.

Later on keeping the max\_depth to 14 and then increasing the n\_estimators to 120, the accuracy increased upto **48%**

<https://colab.research.google.com/drive/1E6__48ALJPc_vSN42CWVGAPvXAqpoEKz?usp=sharing>

**Question 2:**

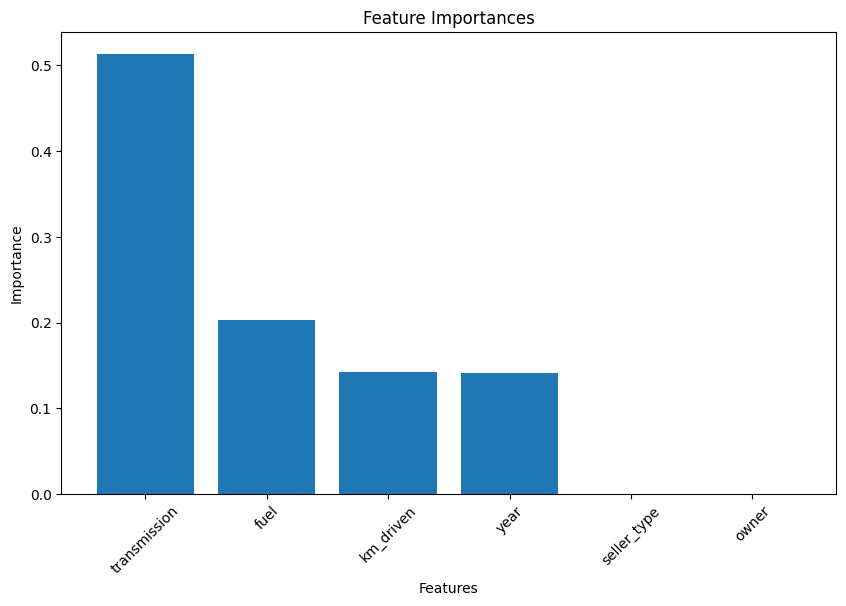
**The data of this question was about the price prediction of car depending upon the model, seller type etc. Based upon which the price of the cars was given.**

<https://colab.research.google.com/drive/1o5ic2d6xLhF9QlNvRbkjcBg5H20S7_2Z?usp=sharing>

The price was in numbers and they ranged from 20000 to 8 million, and thus because it is price, it was a case of regression not classification

Using the Decision tree regressor, the mean squared error was found, that was Mean Squared Error: 109006644437.49913

Later on using this was identified that using the decision tree regressor, the owner, and seller types had negligible importance.



Using the variations by the grid method

Best Parameters: {'max\_depth': 5, 'max\_features': 'auto', 'min\_samples\_leaf': 4, 'min\_samples\_split': 10}

Best MSE: 85886556818.41817

Mean Squared Error: 85886556818.41817

Thus the Mean squared error improved a lot.

Meanwhile using the Random Forest regressor, it was found that the Mean Squared error was much more lesser than the best parameters of Decision tree regressor

Mean Squared Error: 73075221663.76144

Using the random forest the following importance was found:

Feature Importance

0 year 0.236479

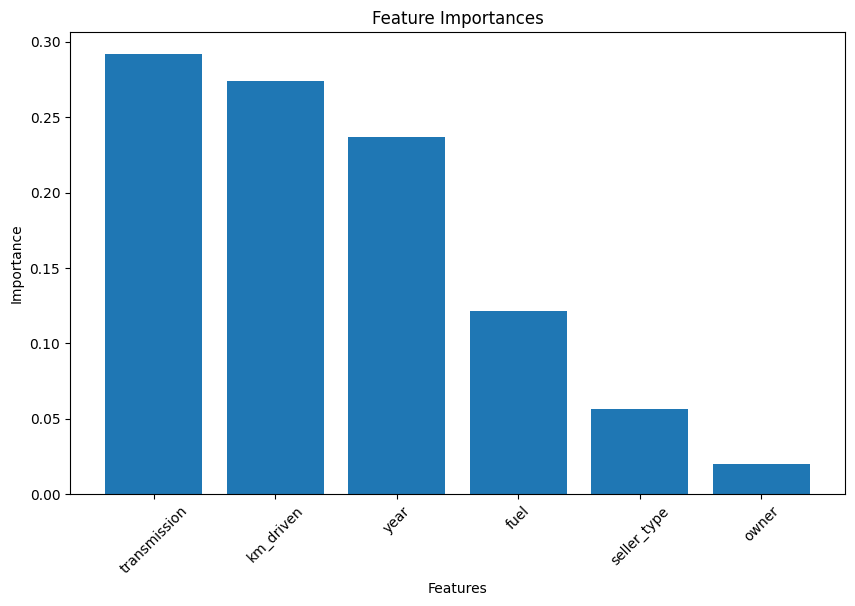
1 km\_driven 0.273748

2 fuel 0.121639

3 seller\_type 0.056713

4 transmission 0.291663

5 owner 0.019759



After wards I removed the owner feature from the data and then the following mean square error was obtained

Mean Squared Error: 203464527554.2972