# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 8**

Dt : 30-03-2022

**Group Name : Discover Decipher**

**Group Members**

Nimisha Patel - AU1940146 - Btech Computer Science and Engineering.

Sakshi Shah - AU1940213 - Btech Computer Science and Engineering.

Astha Patel - AU1940312 - Btech Computer Science and Engineering.

Kareena Matwani - AU1940314 - Btech Computer Science and Engineering.

**Task performed this week.**

* **Interpret the decision tree classifier**

The accuracy of the decision tree classifier on the test set obtained is 0.99. The decision tree represents a test on an attribute at each node. A decision tree checks if an attribute or set of attributes satisfies a condition, and according to the results, the tree splits the data. Gini is the measure of difference between two variables (that is incorrect classification). Like if the graph of two variables are straight lines the value of gini will be zero. Gini value used to split nodes. DecisionTreeClassifier considers gini impurity of smaller partitions of the same attribute, to maximize the accuracy of classification. The probability of misclassification times the probability of element chosen over all 4 classes defined here gives gini impurity.

Algorithm here tries to find all the possible boundaries among the dataset and will choose the decision boundary with the lowest gini impurity. On the root node, all the data points are mixed. 3 : low emission, 1194 permissible emission, 3586: Medium emission, 1125 : high emission. The medium emission range is more frequent in the data and tree choses as the root. The algorithm continues to create homogenous groups by splitting data until the gini value is 0, where we cannot have more homogenous groups.

The fuel consumption and fuel type values are mostly used attributes for decision tree classifiers and we will know using further algorithms or by feature engineering if these two features are the major factors impacting CO2 emissions.

* **Decision Tree Regression Based Model**

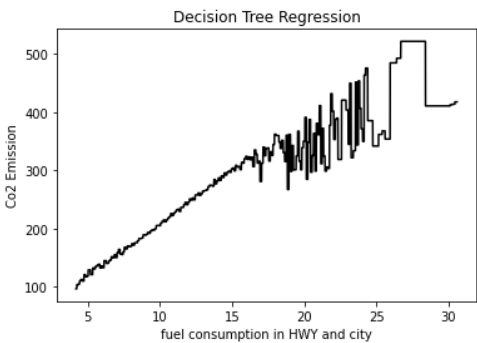
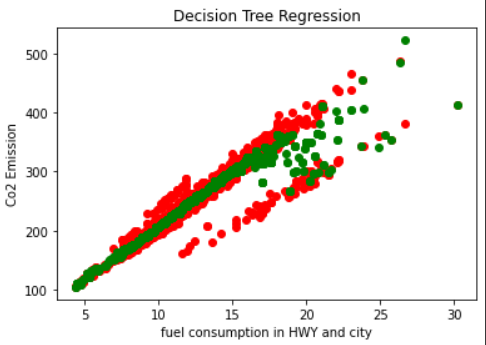
This model represents the correlation between Fuel Consumption and CO2 emission of the dataset based on Fuel Consumption in Highway and Fuel Consumption in city.

Here variable X is a 2 dimension array with Fuel Consumption in highway and city. And y is a co2 emission.

**Outcome:**

Visualizing the result using the graph.

The real value of co2 emission is represented in red color and predicted value of co2 emission is represented in green color. Decision Tree Regression R2 score (Test Dataset) : 0.94



The plot given by the decision tree regression model is drawn below in black color.

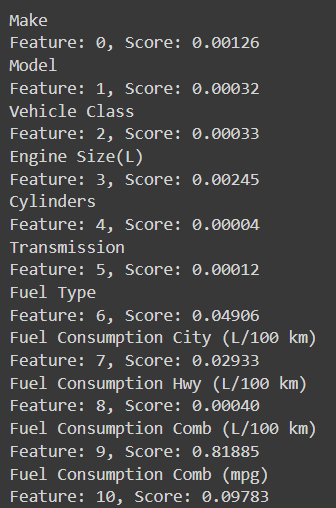
**Disadvantage:**

Sometimes this model can cause overfitting so it should have minimum leaf nodes. So to solve this problem we will do the Random Forest Model.

**Feature Importance scores using decision tree regression**

The features we use influence the result of prediction as well as classification and thus feature engineering can supplement the algorithms towards better accuracy and knowing the influence of any particular feature in the CO2 emission, which can be controlled by manufactures at the time of manufacturing and this can be an important step in green manufacturing.

Feature Importance scores can help figuring out the most relevant feature to the target. Importance scores are supported by few predictive models like decision tree, logistic regression, linear regression and more. So along with decision tree regression, we implemented the model for importance score to know importance of features in prediction.



Feature Importance scores

**Task to be performed next week.**

* Similarities and differences in the feature selection methods (implemented and to implement some) (implement : ANOVA score and Feature importance score using different models) and interpretations.
* Explore dimensionality reduction and its impact on the results.
* Random Forest method in depth along with testing variation in data.