# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 7**

Dt : 23-03-2022

**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.**

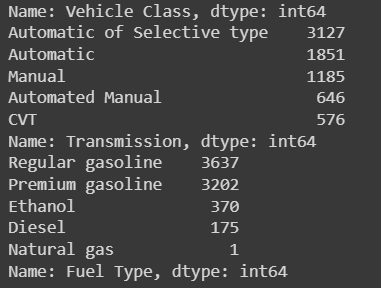
* Feature Engineering and Feature Selection in depth for multicategorical columns.
* Decision Tree Classification.

**Implementation and outcomes.**

CO2 emission rate depends on many factors of the vehicle, and we want to know which feature affects the maximum to CO2 emission rate. Thus, we started with feature analysis and feature engineering for the given data of 5991 unique entries and 12 features in the dataset. Feature engineering will help improve accuracy and reduce overfitting.

From here, Feature Engineering is performed on numerical as well as non numerical multicategorical data columns for extracting features from dataset that might be relevant in CO2 emission prediction.

**Handling Non numeric values :** For columns upto 4-5 categorical non numeric values like Fuel type and Transmission,the single column is converted to multiple binary columns, generating dummy variables in the dataset of fuel types and transmission.



5 categories in the vehicle class and 5 categories for Transmission.

There is only one data of Natural gas and thus can be removed, it will not make any difference.

Added 9 binary columns and removed non numeric columns : Fuel Type and Transmission

If a vehicle uses regular gasoline, value in fuel\_regula gasoline = 1 else 0. For multicategorical values, each non-numerical category was assigned a numeric value using values counted and the .to\_dict() function.

**Outcome**

Positive : No non-numeric columns in the dataset.

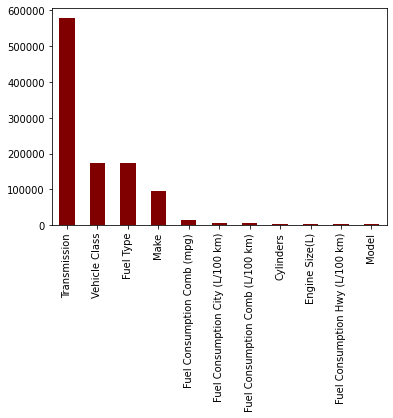
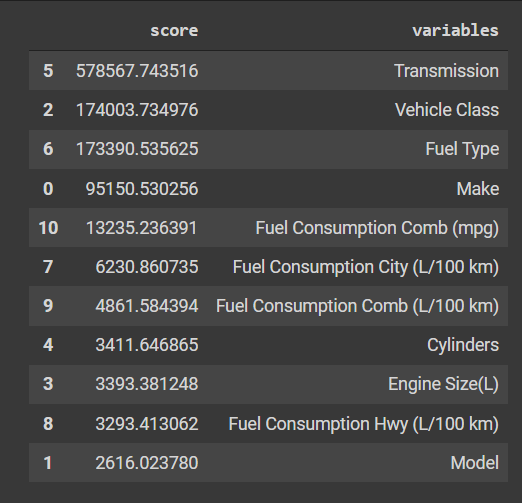
Negative : Increase in binary columns.

Rechecked the dataset changes by not using binary columns and assigning the numeric value by count and to\_dict() function which leads to similar scores in feature selection so the conversion of non numeric values to numeric without appending 9 columns seems a better option to work on further.

**Feature Selection for all the features including non - numeric features**

For future selection we used chi2 score earlier to identify the related features and eliminate the irrelevant features for better accuracy of our model. To select the features that explain the relationship of independent variables with the target variable.

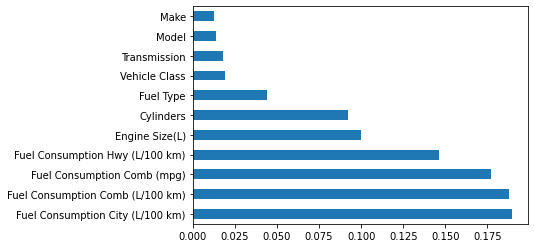
**Outcome**



**Feature Selection Method 2** : Feature Importance Score

Higher the score, more relevant the feature is for the target variable.

Implemented using the Tree Classifier model.



**Feature Selection Method 3** : Correlation among the variables (numeric as well as non numeric)



**Decision Tree Classification**

* Trained the model using training set and using sklearn library

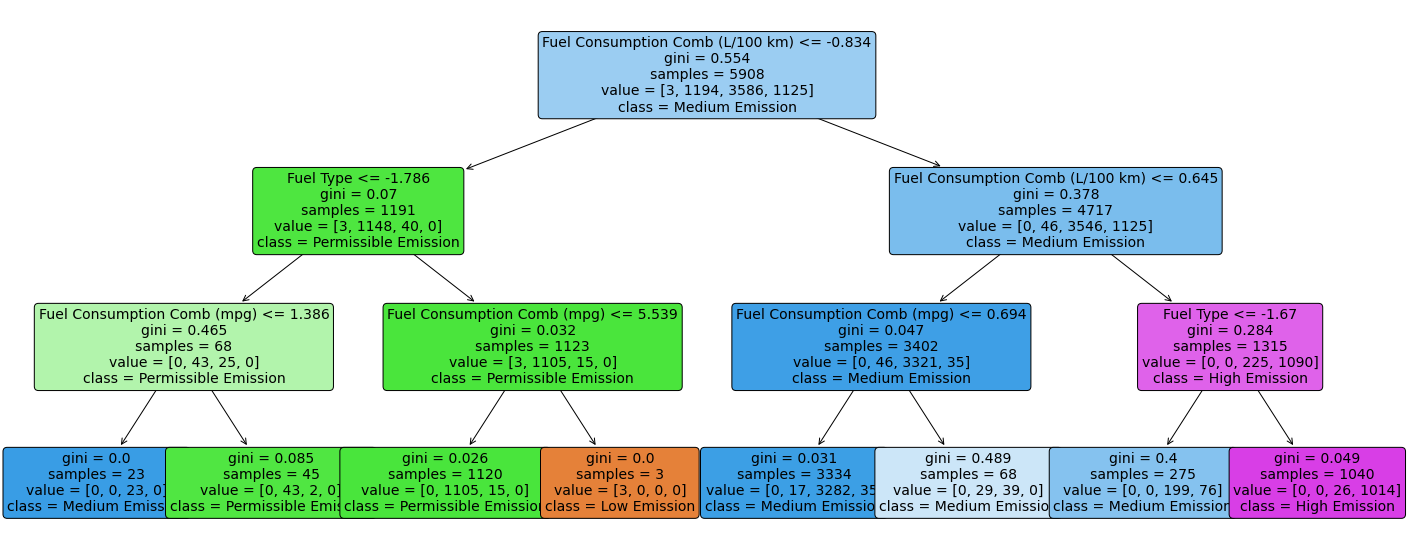
(for all numeric data)

* Plotted Decision Tree graph, which would help in easy interpretations.

**Outcome**

With the model set to predict CO2 Emissions from all the given features, we achieved the highest testing accuracy (0.99 percent) using the DecisionTreeClassifier.

* Decision Tree Classifier Accuracy on dataset (Test Dataset) : 0.99
* Decision Tree Regression R2 score (Test Dataset) : 0.94



Decision Tree classification

**Task to be performed next week.**

* In depth decision tree algorithm with understanding and interpreting decision tree.
* Compare KNN and Decision Tree process of classification for the dataset given.
* Find the most effecting feature on CO2 emissions values.