# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 10**

Dt : 13-04-2022

**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.**

**Hyperparameter Tuning on Random Forest Regression model**

Using available inbuilt algorithms in python, we found the hyperparameter for random tree forest. Hyperparameters further were tuned to increase the accuracy of the model. Here we are fining the features like:

* n\_estimators - number of decision trees in the random forest
* max\_features - number of features to consider at every split
* max\_depth - maximum number of levels in a tree
* min\_samples\_split - minimum number of samples required to split a node
* min\_samples\_leaf - minimum number of samples required at each leaf node
* From our Co2 emission dataset,
* **Outcome :**

**best number of estimators: 1600**

**best max\_features: sqrt**

**Max\_depth: 20**

**Min\_samples\_split : 2**

**Min\_samples\_leaf : 1**

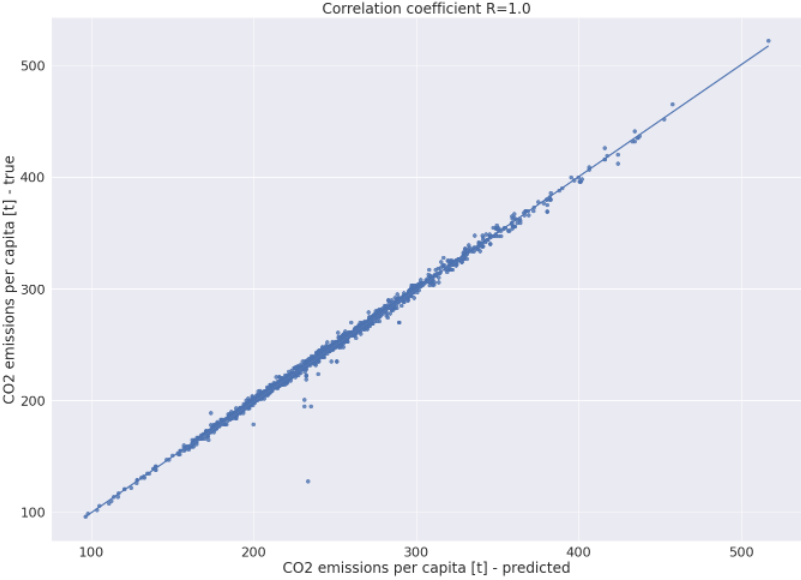
* Now we tune these estimators (in model) for our unseen(test) data set to get higher prediction accuracy for our model.
* Here we had to use k-fold to measure the accuracy of the test data set in folds. We divided 10 folds.
* Here each of the 10 folds exhibits R2 scores between 0.992 and 0.998 with a very small standard deviation of 0.002. This shows that the model has achieved good predictions for all the folds without any particular one that stands out with a worse result.

**R2 = 0.9947**

**Mean Squared Error: MSE = 18.0808**

**Root Mean Squared Error: RMSE = 4.2522**

* We got the graph of unseen dataset with r square value 0.99



**Random Forest Classification**

Performed **random forest classification** on the dataset including non categorical feature values,The accuracy of the model turned out to be as follows :

**Accuracy of Random Forest classifier on training set: 1.00**

**Accuracy of Random classifier on test set: 0.98**

**Confusion matrix :**

**[[ 2 0 0 0]**

**[ 0 290 4 0]**

**[ 0 9 875 5]**

**[ 0 0 8 284]]**

The class1 true predictions are less due to the smaller frequency of the particular class label.

The **cross validation for feature selection** using random forest classifier is not able to recursively eliminate any feature except transmission from the given 11 features and thus the number of feature input cannot be reduced with help of feature selection for classification.

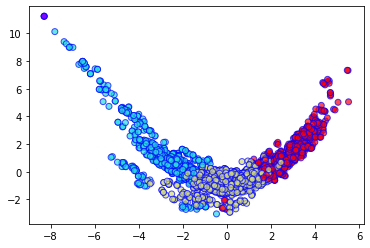
**Feature ranking after RFECV:**

**[1 1 1 1 1 2 1 1 1 1 1]**

**Chosen important features:**

**['Make', 'Model', 'Vehicle Class', 'Engine Size(L)', 'Cylinders', 'Fuel Type', 'Fuel Consumption City (L/100 km)', 'Fuel Consumption Hwy (L/100 km)', 'Fuel Consumption Comb (L/100 km)', 'Fuel Consumption Comb (mpg)']**

Thus, we performed  **Linear Discriminant analysis,** dimensionality reduction technique (commonly used for supervised classification problems).Here, we have 4 classes we need to separate them efficiently and using only one feature may lead to overlap in the class labels.



4 class - random forest classifier after Linear Discriminant analysis

Accuracy of Random Forest classifier on training set: 0.91

Accuracy of Random classifier on test set: 0.90

[[ 0 2 0 0]

[ 0 266 22 6]

[ 0 20 814 55]

[ 0 0 41 251]]

Due to LDA dimensionality reduction the accuracy of random forest classifiers decreased.

**Task to be performed next week.**

Try PCA for dimensionality reduction.

Analyze the effects of dimensionality reduction by LDA and PCA.