**Execution Steps during Mercedes Benz Greener Project.**

Execution steps:

1. Import data
2. Check Outliers for the predicted variable.
3. Check the types of data structure
4. Check for null values.
5. **Check the columns with minimum variance, remove them (Only for Binary columns)**
6. **Variance thresholding**
7. Check multicollinearity (VIF) or correlation method. Check if to remove multicollinear columns. Use the **Correlation method (Kendall)** to remove multicollinearity for Categorical variables. We cannot use VIF for categorical variables.
8. PCA should be used with correlated columns only. as we removed correlated columns, we should not use PCA on the variables. Also, PCA is not recommended to be used for categorical variables as it is based on the variability of the values in the column.
9. Use the ‘get dummies‘ method to convert categorical variables to numeric ones. Limit it to maximum occurring categories. (Only for columns with ‘char’ data type)
10. Split the data: train\_test\_Split.
11. **Using Linear Regression:**
12. Used Linear Regression first, then used Regularization to improve r2 value (Noticed significant change). Used Grid search CV to find the best value of alpha. Also used the Validation curve method.
13. **Using Gradient Boosting Algo:**
14. Hyperparameter tuning by using grid search CV, verifying the scores, and checking the overfitting by cross-validation.
15. Using the Validation curve to check overfitting and parameter tuning.
16. Fit the model with trained parameters on X\_train and Y\_train, Predict the values for X\_test and check the metrics.
17. **Using XGboost:**
18. Using the same procedure for hyperparameter tuning as that used for Gradient boosting. Check the metrics and overfitting.
19. After using XGboost, see the feature importance. If you don’t get the expected metrics, then use only the important features to improve the results.
20. Fit the model with trained parameters on X\_train and Y\_train, Predict the values for X\_test and check the metrics.
21. Check the metrics after selecting important features from feature importance.
22. Feature importance. (To decide important variables that affect the testing time).