Car Class Prediction Document

- Car Class Prediction dataset consist of 719 rows and 20 columns. All the columns in this dataset are numerical columns. This dataset does not contain any null values.
- It is a balanced data. That is all the car class have more are less same the amount of data.
- From the pair plot we can see that most of the features are highly corelated to each other some are positively correlated scat.Ra and Comp. Some are negatively correlated elong and comp
- From the box plot we can see that there are outliers in some of the columns such as Rad.Ra, Pr.Axis.Ra, Max.L.Ra, Sc.Var.Maxis, Sc.Var.maxis, Skew.Maxis, Skew.maxis, kurt.maxis
- For handling the outliers, I used IQR (Inter Quantile Range) method
  - ➢ For each column I find the IQR
  - ➢ Then replaced the lower and upper values with its median value of that column

- As the dataset is small with 716 data, I don't want to lose the data, so I replace the lower and upper values with its median for each column
- With the help of heat map, I have seen that most of the features are highly correlated to each other
- To reduce the correlation and dimensionality of the dataset I decided to do PCA (Principal Component Analysis)
- For that I split the data into X and Y
- Then using Sklearn libraries' StandardScaler, I standardized the data
- After that I did feature extraction using PCA method
  - ➢ Using the PCA I find eigen values
  - ➢ With that eigen values I find that using the first 8 components I can explain the 99% of variance
  - ➢ I used those 8 components for further model building
- Using the model selection module form sklearn split the data into X_train, X_test, y_train and y_test
- Basic model that used are
  - ➢ Logistic Regression
  - ➢ Decision Tree Classifier
  - ➢ KNN Classifier
  - ➢ Random Forest Classifier
  - ➢ Support Vector Classifier

- From the above model I selected the three best model with their accuracy and F1_score, they are
    - Logistic Regression
    - KNN Classifier
    - Support Vector Classifier
- Using the GridSearchCv from sklearn l tunned the hyper parameters of the above best 3 models for improving the accuracy
- From the scores obtain, SVC has the best score, before parameter tuning it has score of 81%
- Final Model for the Prediction of Car Class is Support Vector Classifier with
    - Accuracy of 84%
    - F1_score of 84%