Project Report(Classification)

Overview:

The given dataset is a multi-class classification dataset containing 719 samples with 18 numerical features. Predict the target class which may be :0 -bus, 1 -Opel Manta, 2 -Saab, 3 -Van.

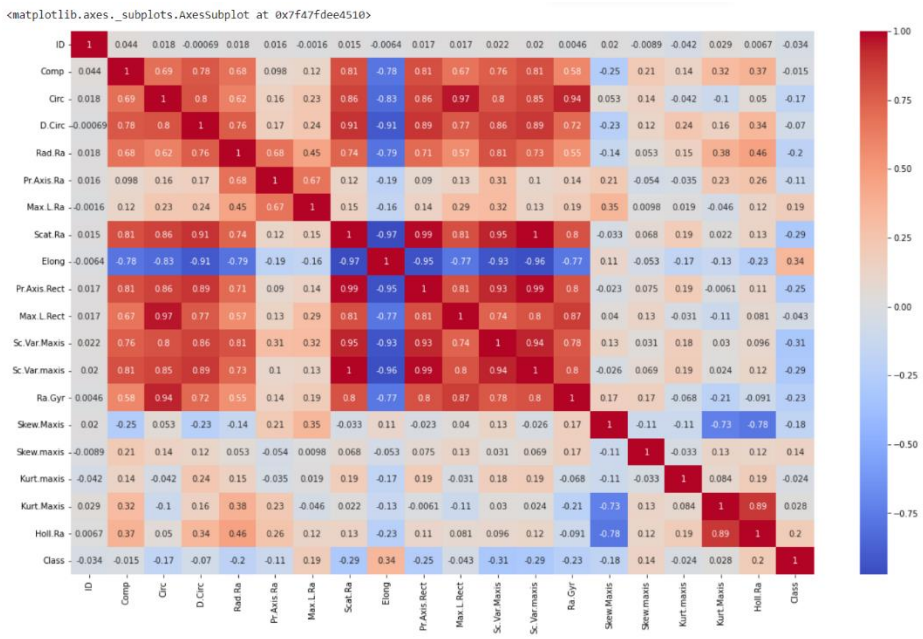
First import required libraries for pre-processing and then load the dataset using Pandas.

Data Pre-processing:

1. There are no missing values found in data using null function.

2. All the features are in numeric form makes better for modelling.

By using Pearson's correlation method features can be selected upon their value of correlation.



From the above correlation matrix we observe that

a. The features Kurt.Maxis and Holl.Ra are highly correlated to each other.

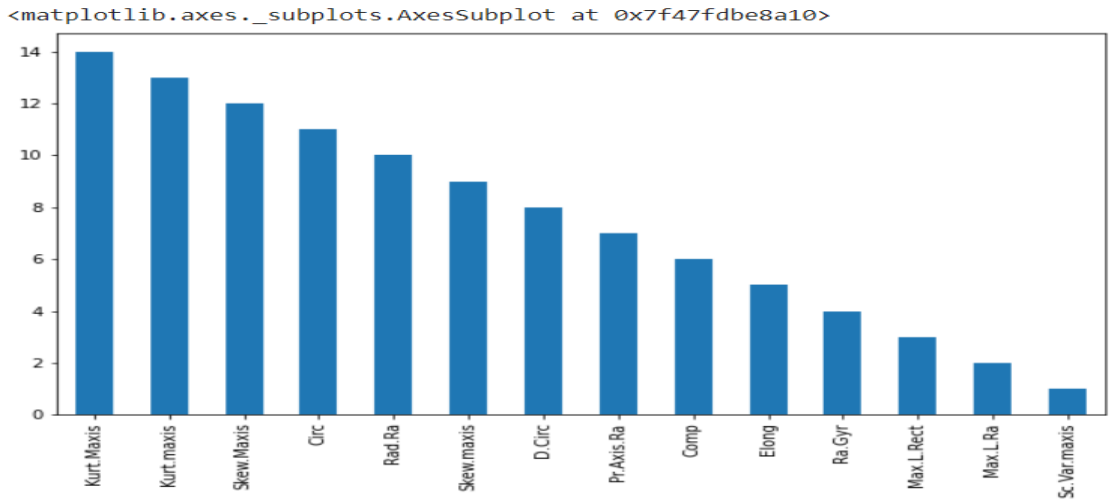
b. Scat.Ra, Pr.Axis.Rect, Sc.Var.maxis, Sc.Var.Maxis are also highly correlated to each other.

When two features have high correlation to each other we can use any one of the feature by dropping the other. From above we use Kurt.Maxis and Sc.Var.maxis.

By using Recursive feature Elimination method we can find the rankings of the feature importance, the lower the rank the higher is the importance.

In this we use any model as an estimator ,here I used decision tree classifier as estimator because it provides better feature importance among models

#We have plotted a bar-plot in order to find out the features to be eliminated as follows:



Above plot provides the ranking of feature importance, here Sc.Var.maxis has least and best rank.

3. Dropping the columns:

From RFE method and from Pearson's correlation method we drop the columns Scat.Ra, Pr.Axis.Rect, Sc.Var.Maxis, Holl.Ra and ID.

4. Finding the outliers using the boxplot, by plot we found there are outliers in the columns Rad.Ra, Pr.Axis.Ra, Max.L.Ra, Sc.Var.maxis.

Using Inter Quantile range method remove ouliers i.e,

Q1=data for quantile 0.25

Q3= data for quantile 0.75

IQR=Q3-Q1

Upperlimit= Q3+(1.5*IQR)

Lowerlimit=Q1-(1.5*IQR)

The data points beyond these limits are outliers which are dropped from columns mentioned before.

5. Spliting the data - Using train_test_split we split 20% of data as test data and remaining for training.

6. Feature Scaling

Features are scaled using standard scaler because when we perform certain algorithms based on Euclidean or Manhattan distance there will be deflection in the model because of the farthest data points. Hence we scaled them by standardisation

New_data=(data-mean)/standard deviation

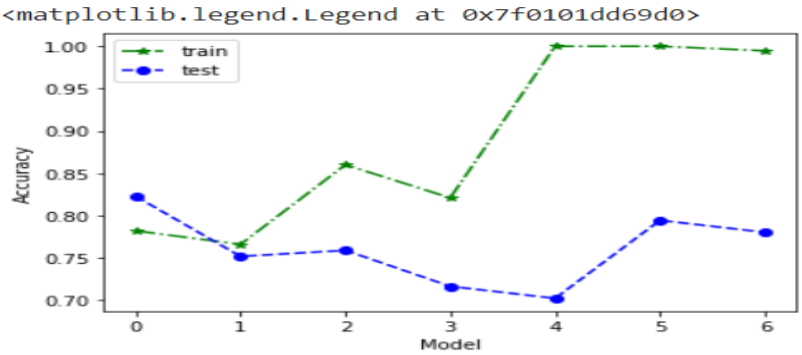
Modeling :

By using different models train the data to get better accuracy. These includes:

LogisticRegression,SGDClassifier,SVC,KNeighborsClassifier,DecisionTreeClassifier,RandomForest Classifier,GradientBoostingClassifier.

Now train the data using all the above models with their default parameters.

The accuracies obtained for train and test data plotted in order to select better model.



Here,

0-Logisticregression,1-SGDclassifier,2-SVC,3-KNN,4-Decision tree,5-Random forest,6-Gradient boosting classifier.

Therefore the better model we obtained is LogisticRegression.

Now tune the parameters of the model to gain better accuracy.

Tuning parameters of model:

```
from sklearn.model_selection import GridSearchCV
parameters={'penalty':['none'],'solver':['newton-cg','lbfgs','sag','saga'],
            'C':[0.1,0.2,0.5,0.7,0.9,1.0],'l1_ratio':[0.1,0.2,0.5,0.8,0.9]}
gscv=GridSearchCV(LogisticRegression(),param_grid=parameters)
gscv.fit(X_train,y_train)
```

Like wise parameters are tuned and got the best parameters as:

{'C': 0.1, 'l1_ratio': 0.1, 'penalty': 'none', 'solver': 'saga'}

Final-Model:

By using above parameters we build a final model for the data .

We got the metrics values as:

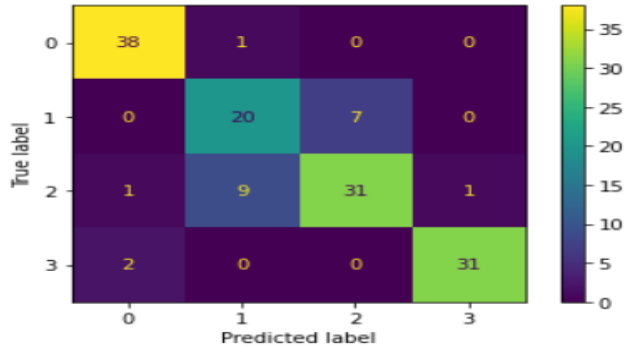
Accuracy_train: 0.79255

Accuracy_test : 0.85106

f1_score : 0.85106

Confusion Matrix as below:

```
/usr/local/lib/python3.7/dist-packages/sklearn/util
  warnings.warn(msg, category=FutureWarning)
<sklearn.metrics._plot.confusion_matrix.ConfusionMa
```



Feature importance:

The importance of different features are plotted in bar plot during RFE.

From the above correlation concepts discussed in RFE, Pearson's correlation method. The important features for the dataset are:

a. Sc.Var.Maxis which has negative correlation(-0.29) for class.

b. Elong has positive correlation of 0.34 with feature class.