SUMMARY

Data set details

1. Dataset file: 'cars_class.csv'
2. This is a multi-class classification data set.
3. The data set has 719 samples.
4. There are 20 numerical features including the target feature.
5. The target variable is the class of the car which may be one of : 0 –bus, 1 – Opel Manta, 2 Saab, 3 – Van.

OVERVIEW

Comp: Compactness
Circ: Circularity
D.Circ: Distance Circularity
Rad.Ra: Radius ratio
Pr.Axis.Ra: pr.axis aspect ratio
Max.L.Ra: max.length aspect ratio
Scat.Ra: scatter ratio
Elong: elongatedness
Pr.Axis.Rect: pr.axis rectangularity
Max.L.Rect: max.length rectangularity
Sc.Var.Maxis: scaled variance along major axis
Sc.Var.maxis: scaled variance along minor axis
Ra.Gyr: scaled radius of gyration
Skew.Maxis: skewness about major axis
Skew.maxis: skewness about minor axis
Kurt.maxis: kurtosis about minor axis
Kurt.Maxis: kurtosis about major axis
Holl.Ra: hollows ratio

PERFORMING THE FOLLOWING TASKS IS INVOLVED

1. Initially I Imported all the Libraries whatever I want
2. Reading the Dataset used by pandas
3. Then perform all required steps to analyze the data frame like shape, describe, info, etc.,
4. In this dataset all data points are having numerical values only, so I may not perform any encode the categorical values.

5. Some of the pre-processing techniques and visualization were done on the dataset.
6. List of columns was dropped due to those having too many unique values. So I thought it's better to remove it for my model performance.
7. I used Heatmap to visualize the strength of correlation among the variables.
8. Declaring feature column and target in X and y respectively.
9. Split data into separate training and test set.
10. I will carry out feature engineering on different types of variables for transforming raw data into useful features that help us to understand our model better and increase its predictive power.
11. This is a multiclass classification data frame, so I build relevant models like Logistic Regression, Support Vector Classifier**,** KNeighbors Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Random Forest Classifier, GaussianNB.
12. Among those classification models Random Forest Classifier and Decision Forest Classifier gives the best score. Hence, I took the Random Forest Classifier as the final model.
13. Some visualization is applied between the Train score and the Test score.
14. The following metrics were used in 'Final_model' on the test data (Accuracy score, F1 score, Displaying the confusion matrix).
15. Random Forest Classifier was tunned by the hyperparameter libraries Gridsearch CV AND Randomized Search CV.
16. Feature importance was reported and visualized.

Conclusion

In this project, I used 7 classification Machine learning techniques out of which Random Forest Classifier gives the best Train and Test scores 1.0 and 0.7361111111111112 respectively, so I build a model for Random Forest Classifier to predict the Class of the car.