**CLASSIFICATION OF SKY OBJECTS**

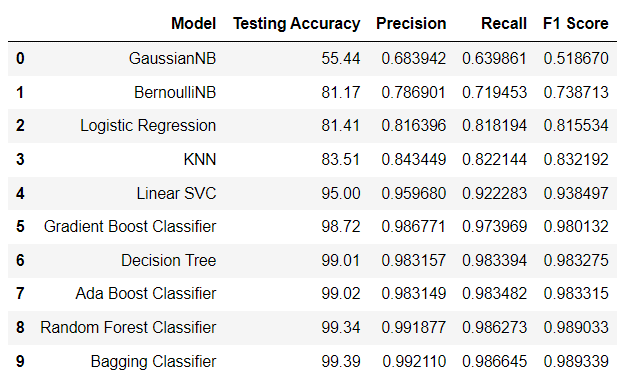
“All models are wrong, but some are useful.” You just have to find the most useful model for your problem & data.

**Project Description:**

This project attempted to use multiple classification algorithms to identify sky objects such as stars, galaxies, and quasars based on their spectroscopic (redshift or spectral type) and photometric (flux or size, brightness) properties, and to select the best performing among them. The dataset is obtained from Sloan Digital Sky Survey. The data set consists of 22 columns and 500000 rows with non-missing values. It is having three unique values in the target column: Galaxy, Star, and Quasar. The balance between the three classes of the target variable was 54% Galaxy, 35.4% Star, and 10.6% Quasar. To construct a well-working classification model, 11 features were identified as critical which are redshift, flux density and other features related to electromagnetic spectrum colour indexes.

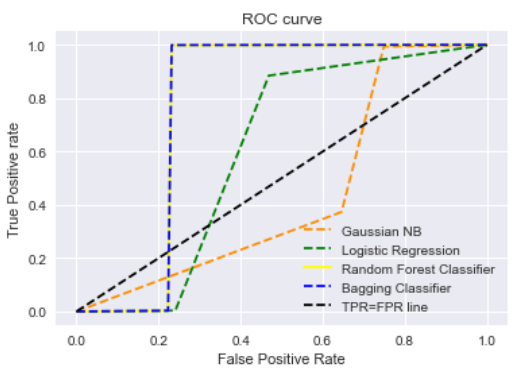
**Algorithms used:**

1. Logistic Regression
2. Gaussian Naïve Bayes
3. Bernoulli Naïve Bayes
4. Linear SVC
5. Decision Tree
6. KNN Classifier
7. Bagging Classifier
8. Random Forest Classifier
9. AdaBoost Classifier
10. Gradient Boosting Classifier
11. Which algorithm provided the best results for multi-class classification?

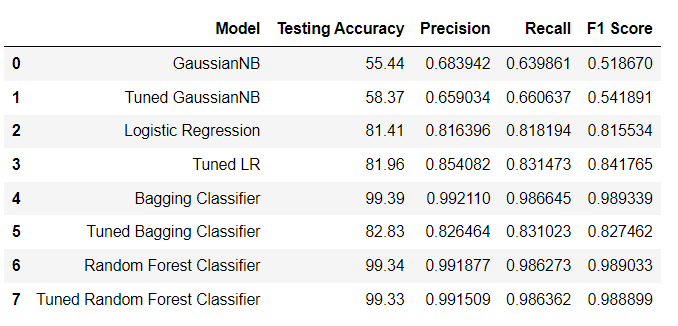


The accuracy of categorization, rather than minimizing False Negatives, is the most relevant information in this dataset. As a result, we can observe that Random Forest Classifier and Bagging Classifier perform the best based on the table obtained after all algorithms ran their predictions.

From the ROC curve we can conclude that Gaussian NB, Logistic Regression models are not good models because the threshold score of True Positive rate is not closer to 1 and threshold score of Negative rate is high, Bagging Classifier and Random Forest Classifier models are good because the threshold score of True Positive rate is closer to 1 and threshold score of Negative rate is low,



1. Was hyper parameter tuning successful in improving the metrics?



On the two algorithms that performed the worst and best, we tried tuning the hyperparameters. As can be seen, the accuracy of algorithms that performed poorly in their original form has improved. The Random Forest Classifier algorithm is tuned using Grid Search and the observed result is that tuned Random Forest Classifier algorithm has slight increase in accuracy when compared to its base model. Bagging algorithm is tuned by changing base estimator to linearSVC from Decision Tree and the result is it performed poorly compared to its base version.

1. What is the accuracy of the best model on Holdout data?

Accuracy on holdout data using bagging algorithm is 99.44

**Final Conclusions**:

The best results that we got came from ensemble models, which reduce variance and improve performance over their constituent learning models. Gaussian NB and Logistic Regression models had relatively smaller accuracy values, Bagging Classifier with Decision Trees and Random Forest Classifier models had relatively larger accuracy values in this project.