**Summary**

In the era of big data and reliable algorithms, data classification is instrumental in the field of astronomical research. In this project, we explore supervised classification methods such as Random Forest, Decision Tree and Random Forest. We applied these methods to classify the dataset as Star, Galaxy or Quasar along with comparing different classification models to find the superior one.

**Background**

In the branches of astronomy, the classification of celestial objects is essential. The ability to collect data samples of stars, galaxies, and quasars is its main advantage. Even though quasars are crucial to a variety of astronomical studies and research, their sample sizes are still in the relative minority group. Significant increases in the sample sizes of quasars and other astronomical objects can only be produced through improved data classification methods, which will then make it possible to advance study. The use of machine learning models in the task of classification has grown more significant and common as contemporary telescopes capture a growing amount of astronomical data due to their precision and speed.

The classification models of random forest and support vector model (SVM), along with decision tree and Boosting Trees, were used with supervised machine learning models to either exclusively identify and distinguish quasars from other objects or to classify all three of the objects: stars, galaxies, and quasars. Additionally, the fact that numerous supervised classification models were trained using data from kaggle.com shows that the information and variables the dataset supplies are comprehensive, trustworthy, and closely related to the objective of identifying celestial objects.

**Data Description**

The data used in this study are from the Sloan Digital Sky Survey (SDSS), which is a famous astronomical survey that has spent more than 20 years attempting to create incredibly accurate and detailed images and a map of the universe. The data represents 100,000 observations of space collected by the SDSS (Sloan Digital Sky Survey). Every observation is defined by 17 feature columns and a class column that categorizes it as a star, galaxy, or quasar.

1.obj\_ID = Object Identifier, the unique value that identifies the object in the image catalog used by the CAS

2.alpha = Right Ascension angle (at J2000 epoch)

3.delta = Declination angle (at J2000 epoch)

4.u = Ultraviolet filter in the photometric system

5.g = Green filter in the photometric system

6.r = Red filter in the photometric system

7.i = Near Infrared filter in the photometric system

8.z = Infrared filter in the photometric system

9.run\_ID = Run Number used to identify the specific scan

10.rereun\_ID = Rerun Number to specify how the image was processed

11.cam\_col = Camera column to identify the scanline within the run

12.field\_ID = Field number to identify each field

13.spec\_obj\_ID = Unique ID used for optical spectroscopic objects (this means that 2 different observations with the same spec\_obj\_ID must share the output class)

14.class = object class (galaxy, star or quasar object)

15.redshift = redshift value based on the increase in wavelength

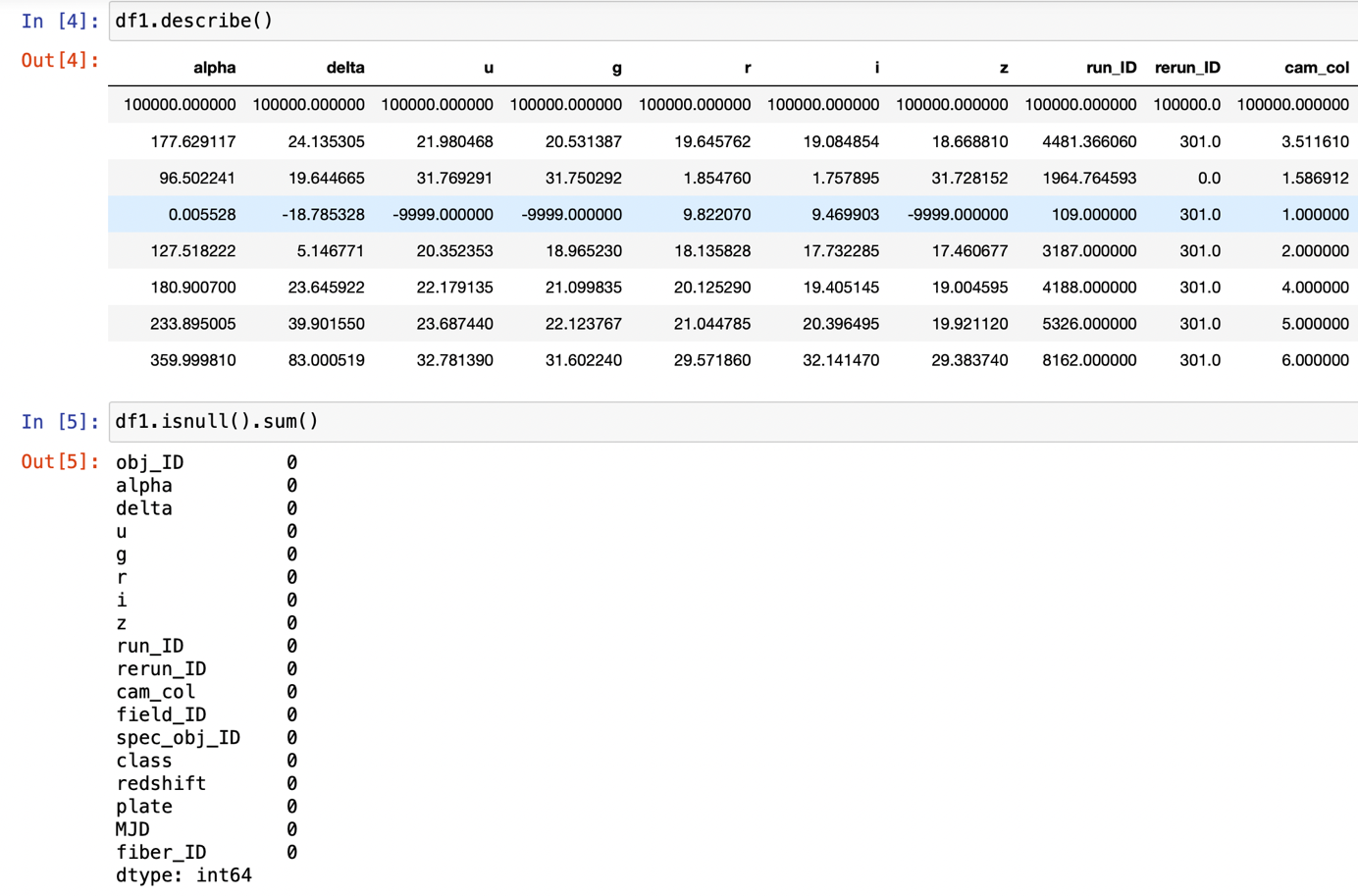
16.plate = plate ID, identifies each plate in SDSS

17.MJD = Modified Julian Date, used to indicate when a given piece of SDSS data was taken

18.fiber\_ID = fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

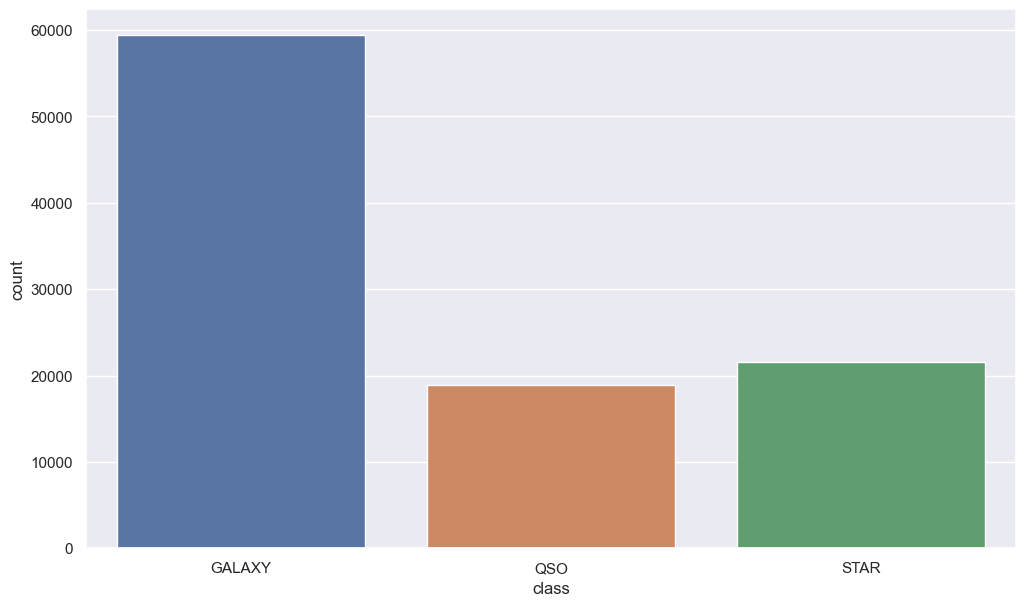
**Exploratory Data Analysis**

The dataset consists of 180000 records and a few null values with 18 attributes. The data set statistics have been obtained by using a summary function in R. Furthermore, it has been identified that the dataset has three classes- Galaxy, QSO and Star having 59445, 18961 and 21594 records.



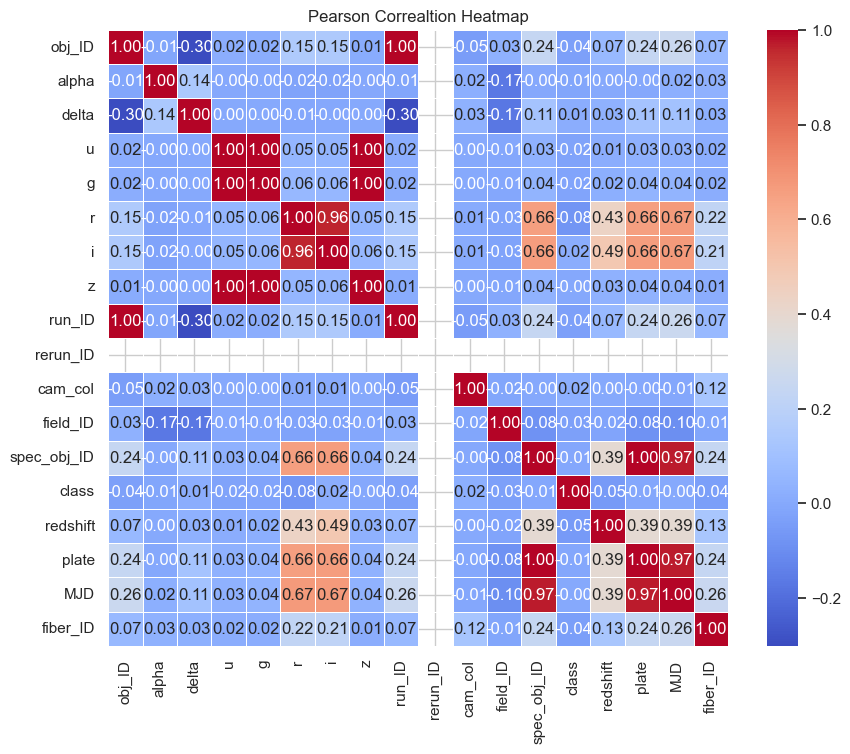


The count plot was used to determine whether there was a reasonable amount of balance between the data entries for the three classes. This chart shows the relationship between different kinds of numbers (numeric variables) and what they categorize (categoric variables). Each of the categories in the categorical variable is represented by a bar. The bar on the chart has a number next to it. This number represents the numeric value of the bar. This bar plot represents distribution of the three different classes- Galaxy, QSO and Star. Additionally, it was also concluded that data imbalance is not a problem.



**Check for correlation:-**

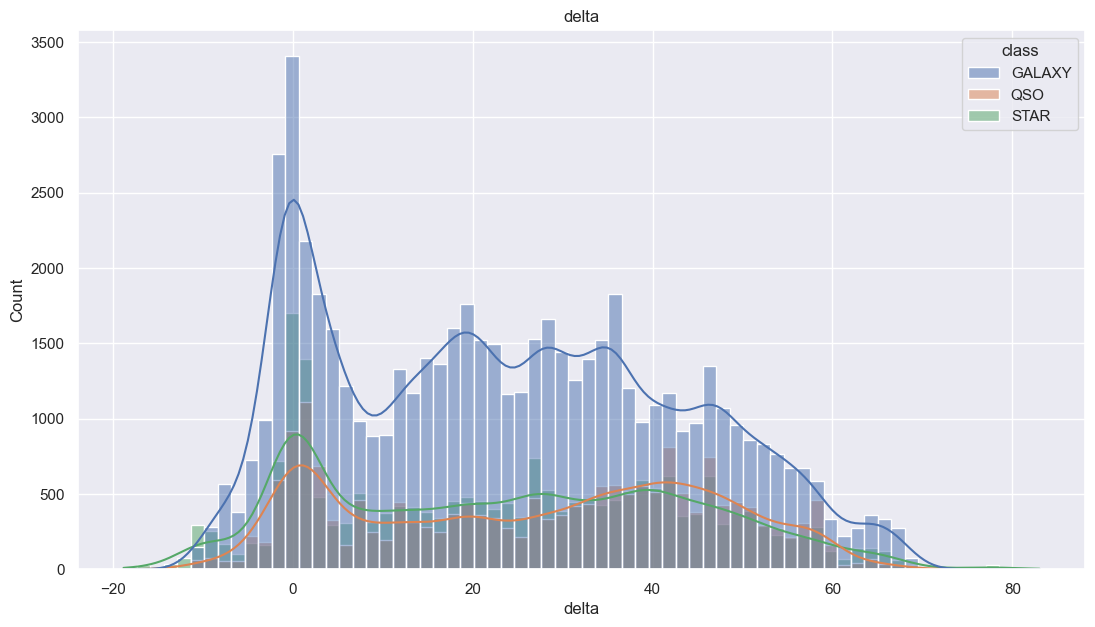
Initially we need to check for collinearity between the variables in the dataset which could lead to a 'Multicollinearity' problem. Multicollinearity tells a linear dependence between the independent variables. The existence of multicollinearity generates a high variance of the estimated coefficients and hence, the coefficient estimates corresponding to those interrelated independent variables will not be accurate in the classification model. They can become very sensitive to small changes in the model. Collinearity tells us how two variables are correlated with each other. In other words, if one variable is highly correlated with another, then most of the variance of one variable is explained by another (shares similar information). So, we may not include the highly correlated variable in the model. One way to extract correlation between variables is using Pearson's correlation coefficient. This correlation coefficient does not prove the 'Causality' (cause and effect). Causality tells the actual relation between variables unlike correlation which is just a possibility.



**Feature Distribution Analysis:**

As alpha, delta, u, g, r, i, z, redshift, plate, MJD are astronomical quantities, therefore, we'll keep them as our primary features. Rest of the features are irrelevant or have no significant difference in distribution by class.





A screenshot of a computer

Description automatically generated

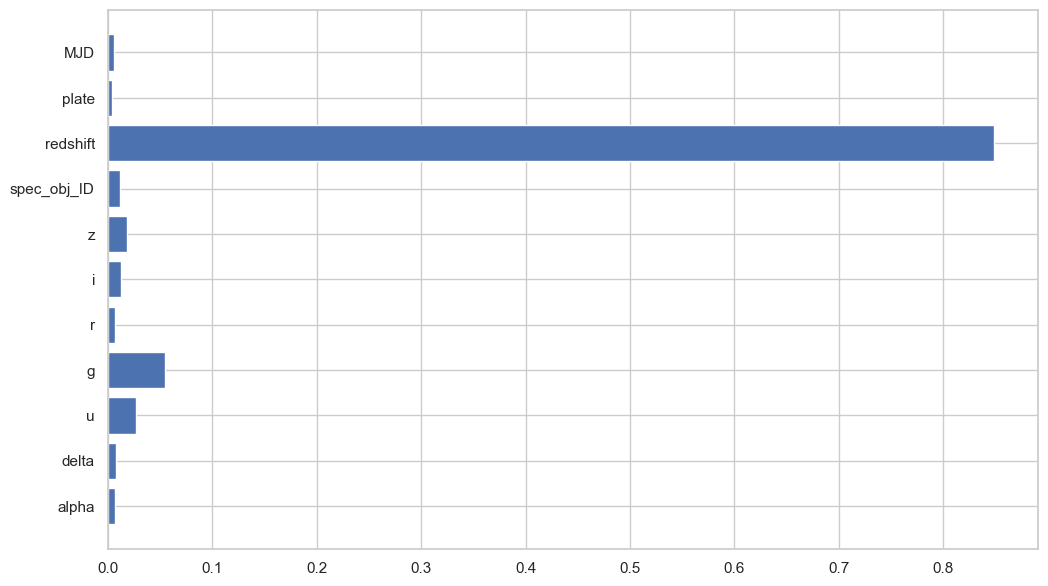
**Models and Performance Evaluation**

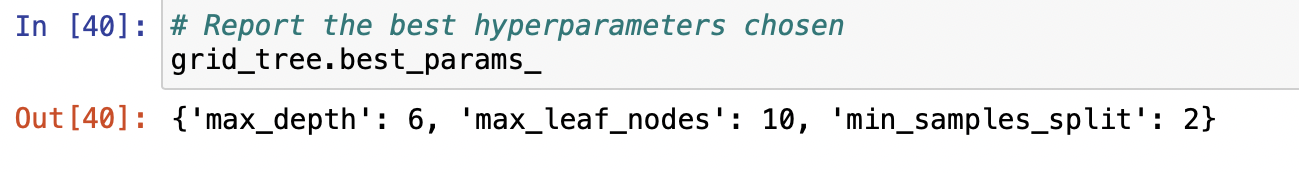
The first model used is Decision Tree. A decision tree model consists of a number of nodes, or conditions, which are joined by branches, or the results of the conditions of the nodes. Each node has two or more branches leading to additional nodes or leaf nodes that begin at the root node (the resulting predictions/classification). The feature variables and their values were used to create the conditions for the nodes in the decision tree, and the Gini index was used to determine where each node should be located. The Gini index measures the heterogeneity/impurity of the data as a result of the separation of the data by the condition, and the smallest value was chosen for each node. Additionally produced were the tree visualization and its confusion matrix.

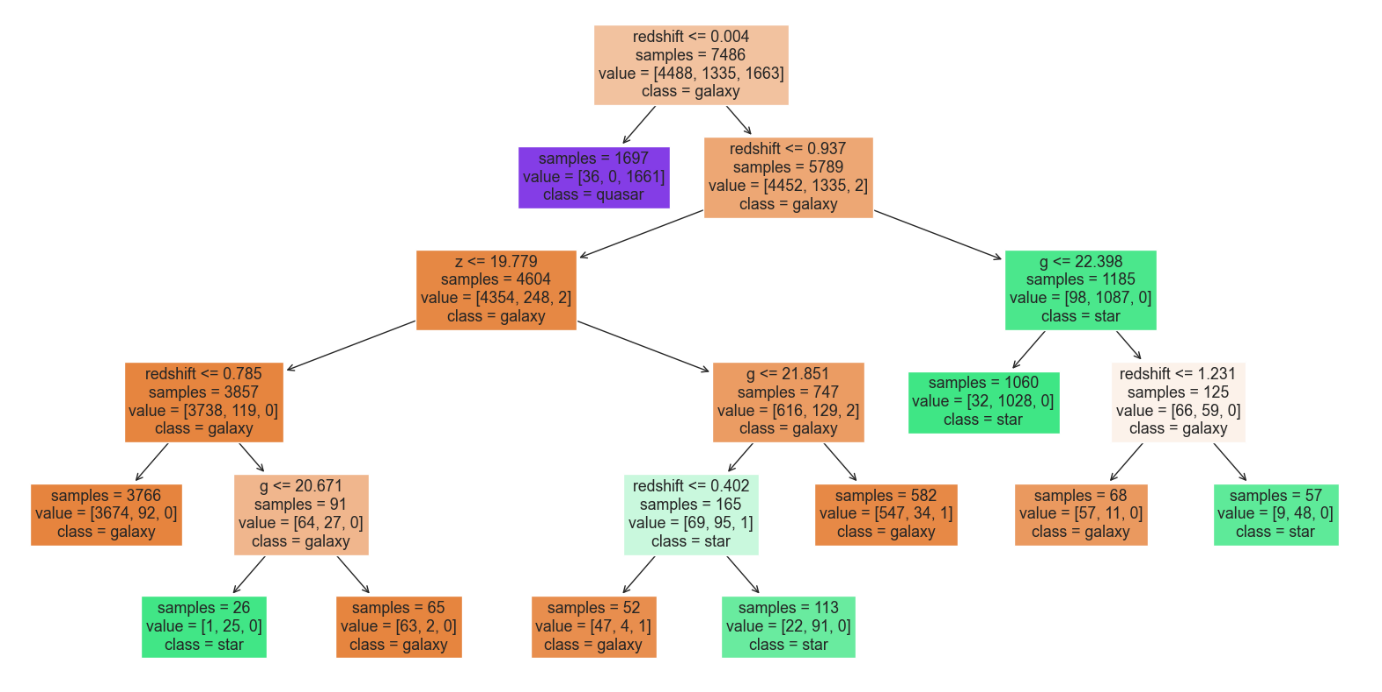
The second Model used is Random Forest. This is an algorithm for collective tree- based learning and is the random forest classifier. A collection of decision trees was drawn from a randomly chosen portion of the training set make up the random forest classifier. In order to determine the final class of the test object, it combined the votes from many decision trees. Additionally, the algorithm was conceived as a confusion matrix to help in classification.

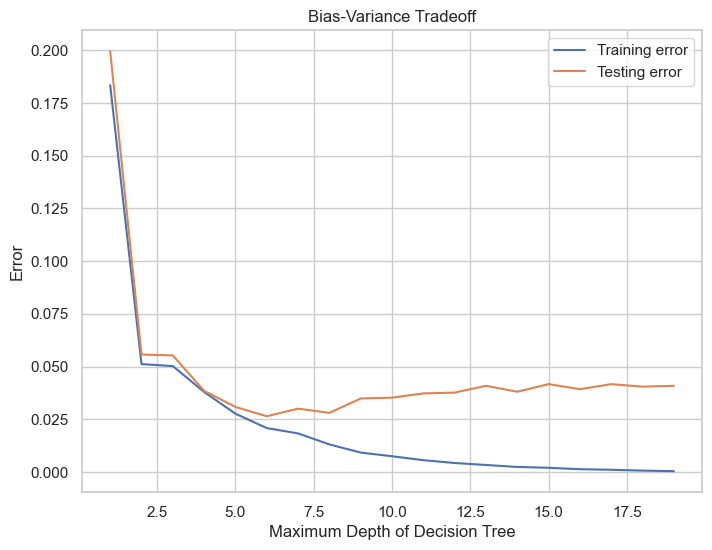
The last model used is Support Vector Machine. SVM is particularly effective in high-dimensional spaces and is used to find a hyperplane that best separates the data into different classes in feature space

1. **Decision Tree**

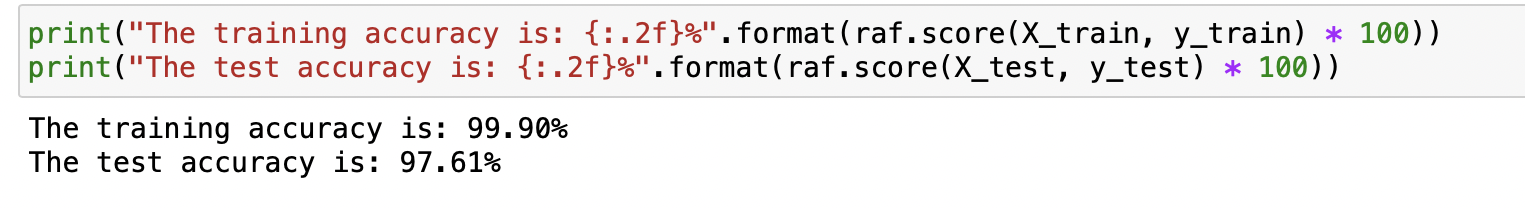


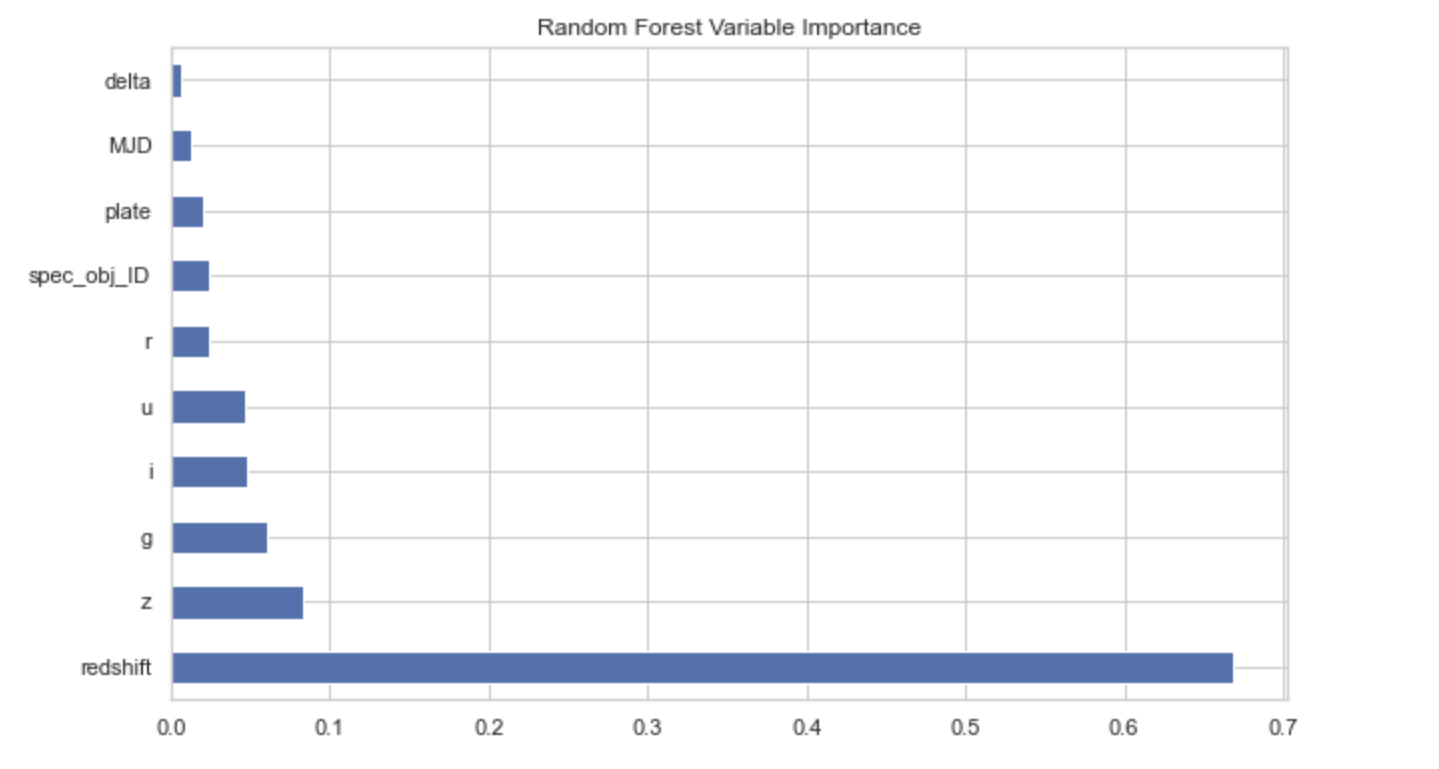


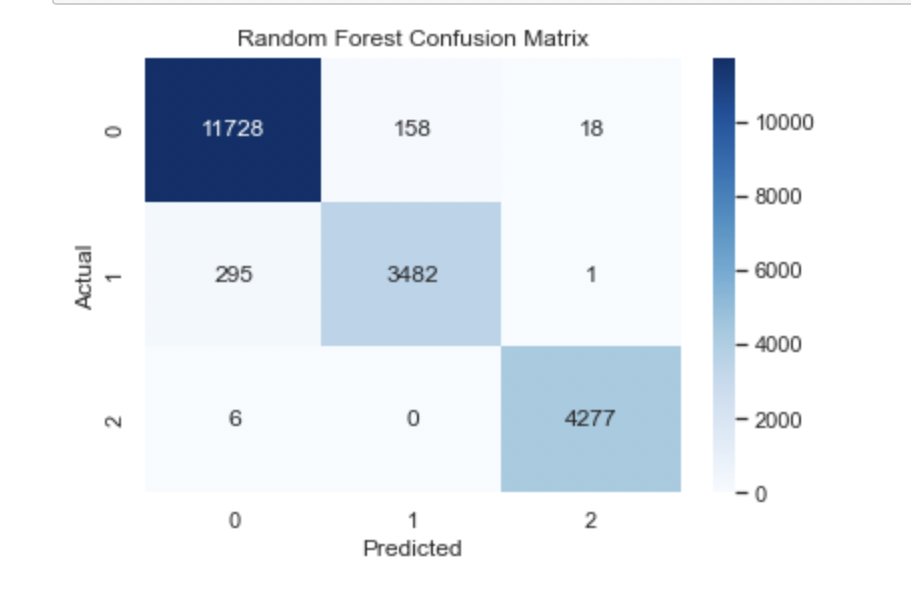


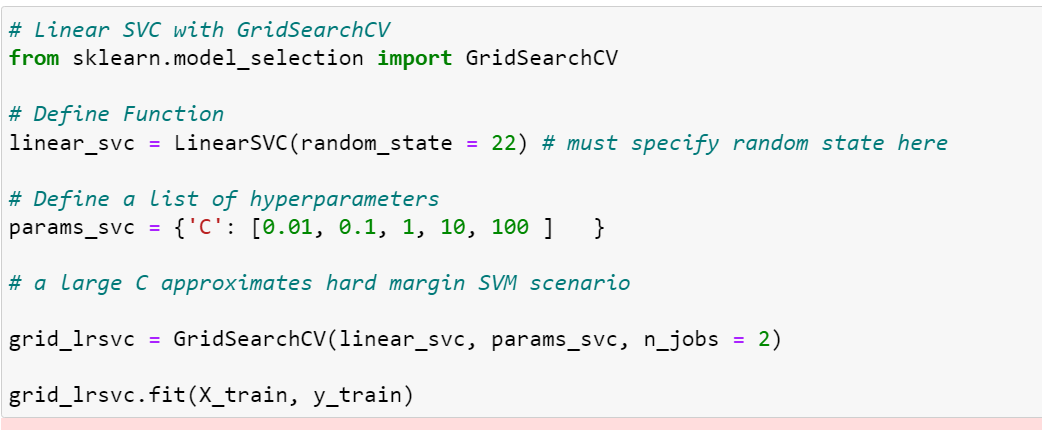


1. **Random Forest**

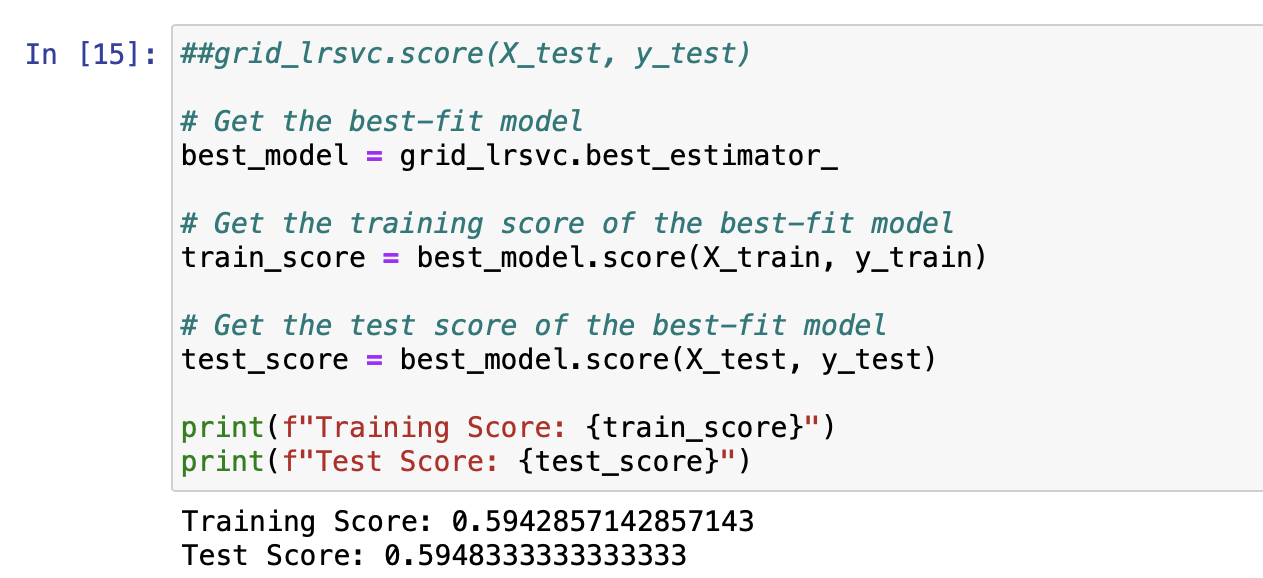


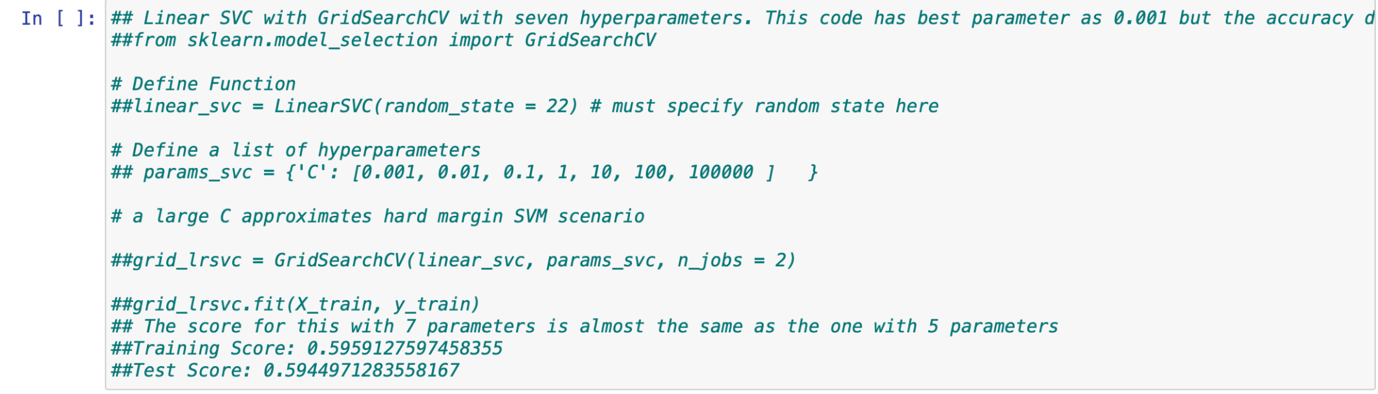


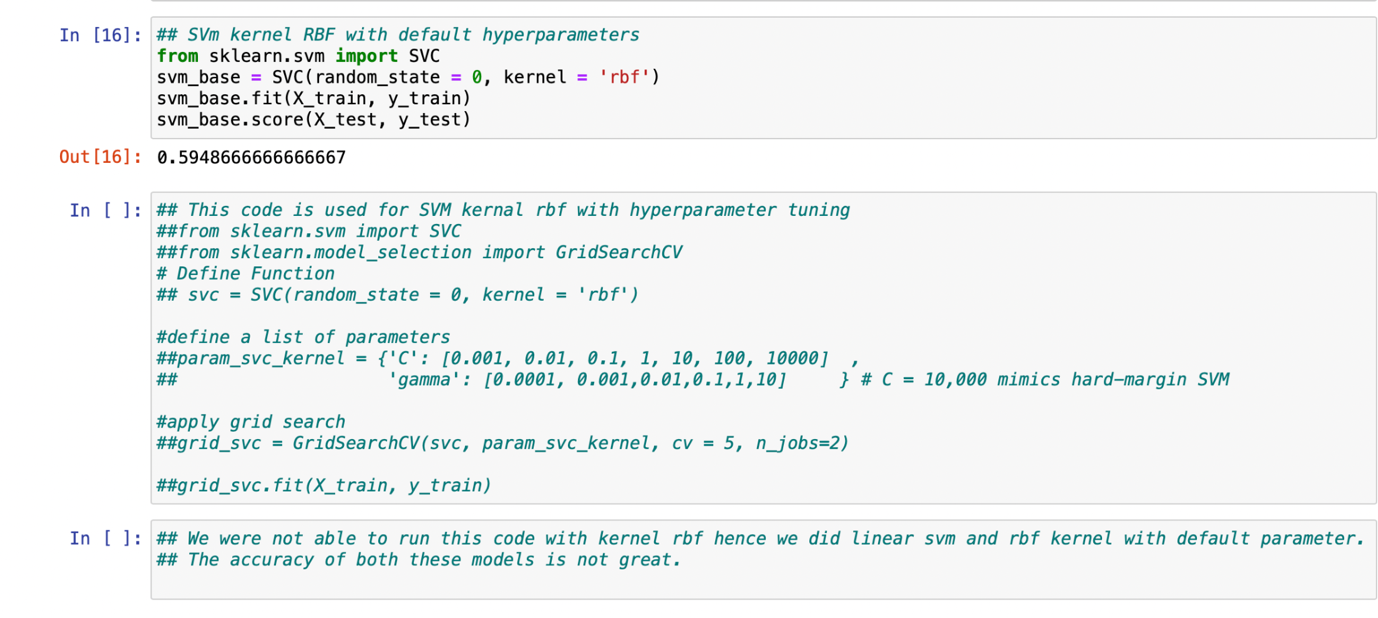




1. **Support Vector Machine**







A screenshot of a computer program

Description automatically generated

**Conclusion**

The Three models were fitted to the training data and evaluated using testing sample for a range of performance criteria. When it comes to classifying stars, galaxies and quasars the random forest model developed in this project has a best accuracy, can be used as a competent and reliable classification tool.

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Decision Tree | Random Forest | SVM |
| Training | 99.11 | 99.44 | 59.48 |
| Testing | 96 | 97.61 | 59.42 |