1 Introduction

The stellar classification Data Set is collected from kaggle.
In this data set we study the properties of 100000 where each has been categorized as either as a galaxy or a star or a quasar. These data has been collected by observations of space taken by the SDSS (Sloan Digital Sky Survey) and for each observation 17 features have been recorded. These features are:

- **obj_ID**, Object Identifier, the unique value that identifies the object in the image catalog used by the CAS.
- **alpha**, Right Ascension angle (at J2000 epoch).
- **delta**, Declination angle (at J2000 epoch).
- **u**, Ultraviolet filter in the photometric system.
- **g**, Green filter in the photometric system.
- **r**, Red filter in the photometric system.
- **i**, Near Infrared filter in the photometric system.
- **z**, Infrared filter in the photometric system.
- **run_ID**, Run Number used to identify the specific scan.
- **rereun_ID**, Rerun Number to specify how the image was processed.
- **cam_col**, Camera column to identify the scanline within the run.
- **field_ID**, Field number to identify each field.
- **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).
- **class**, object class (galaxy, star or quasar object).
- **redshift**, redshift value based on the increase in wavelength.
- **plate**, plate ID, identifies each plate in SDSS.
- **MJD**, Modified Julian Date, used to indicate when a given piece of SDSS data was taken.
- **fiber_ID**, fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

I have studied this dataset for interpretation to see which feature is the most impact for classifying an object in the observations either as a galaxy or a star or a quasar.

2 Cleaning and Feature Engineering

One of the features in this dataset was rerun_ID which had only one unique value and which means it doesn't give any useful information, therefore I dropped that column. I also encoded the three possible classes to integers 0, 1, 2.

3 Regression models

First I have checked the most correlated features to classifying the object and they were redshift, i and plate. The first model which has been implemented was k nearest neighbor with n_neighbors=3 which results not so good f1 score and accuracy score for around 0.85.

As we can see the classes are skewed. %60 of the whole observations are classified as galaxies and around %20 each two others, i.e. star and quasar. Therefore I used the stratified test and train split.

Later I used a grid search to find the best random forest classifier. The result was a random forest classifier with maximum depth of 19, maximum features of the square root and with n_estimators: 27. The accuracy score has been improved very much and scored 0.97.

At the end I used the grid search in the gradient boosting classifier. The result was a gradient boosting classifier with maximum features equal to 4 and with n_estimators: 400. The accuracy score has been improved more and scored 0.98.

4 Interpretation

The gradient boosting classifier gives the best model of classifier for this data set. I understand that redshift and the near Infrared filter in the photometric system are the most important features to identify an object in the observations. This model can help use understand the difference of redshift of galaxies and stars and quasars.

Since this data had a large amount of observations (rows) it takes so long for the grid search to find the best model of classification. Therefore I limited my search up to 400 tree. With a more strong computer I believe one can find better model with higher number tree and higher accuracy score and f1 score.