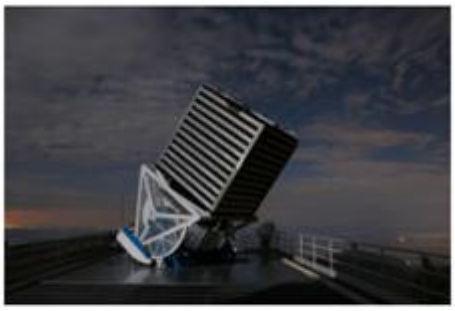
SS4101 Project Report

Vikrant Kumar (17MS089)
Pratap Tanari (17MS129)

The Sloan Digital Sky Survey (SDSS) is one of the most ambitious astronomical surveys that aims to provide detailed optical images of the sky by detecting millions of celestial objects, including galaxies, quasars, and stars. In its current survey phase (SDSS-IV), it has released four sets of data (DR-13 to DR-16) on its observations of the sky. Each consists of 5 or 6 types of data: images, optical spectra, infrared spectra, IFU spectra, stellar library spectra, and catalog data. The catalogue data consists of the parameters measured from images and spectra, such as magnitudes and red-shifts.

Our project is to analyse the catalogue data of DR-14, 15 & 16 to understand the differences in the spectra of the three classes of celestial objects, namely stars, galaxies, and quasars, using Machine learning techniques. To classify the objects, we will use different ML classifiers, and we will try to understand how each classifier of a different structure learns to distinguish between the classes as well as compare their efficiencies.



The SDSS telescope at night

➢ Data Acquisition

Data Release 14 : http://www.sdss.org/dr14/data_access/
Data Release 15 : http://www.sdss.org/dr14/data_access/
Data Release 16 : http://www.sdss.org/dr14/data_access/

Data was acquired through the CasJobs website, an SQL-based interface which lets you query their database for the released data.
Link: http://skyserver.sdss.org/CasJobs/

○ Query used on CasJobs

```
SELECT TOP 10000
p.objid, p.ra, p.dec, p.u, p.g, p.r, p.i, p.z, p.run, p.rerun, p.camcol, p.field, s.specobjid, s.class, s.z as redshift, s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
p.u BETWEEN 0 AND 19.6
AND g BETWEEN 0 AND 20
```

The image table (PhotoObj) contains all image objects and the spectral table (SpecObj) contains the corresponding spectral data. These two tables are joined using the above query.

A table corresponding to each release was downloaded as a .csv file.

➢ Project Overview

We have three datasets, namely dr14, dr15, d16, corresponding to each data release.

We start by training efficient Machine Learning models using the dr14 & dr15 datasets. Note that both dr14 & dr15 contain 10k observations, while dr16 contains 100k observations. This means that we're training on just 17% of the entire dataset to make predictions. The next step is to test these models on the final dr16 dataset to estimate their reliability, accuracy, and resource efficiency.
Many different models were tested. We present three of them in our report - Gaussian Naive Bayes, KNN, and Random forests.

These Models are chosen based on the fact that they can easily be interpreted using our physical understandings of stars, galaxies, and Quasars from theory.

Models that acquire a high accuracy and low resources can help make predictions on unknown data from the Sloan Digitial Sky Survey's upcoming releases.

➢ Data Description

	objid	ra	dec	u	g	r	i	z	run	rerun	camcol	field	specobjid	class	redshift	plate	mjd	fiberid
0	1237648721218633849	152.354824	-0.129744	19.49632	17.59388	16.67482	16.25166	15.95042	756	301	3	244	304055679038023680	GALAXY	0.095048	270	51909	228
1	1237648721218633899	152.289487	-0.063071	19.45274	17.62178	16.68552	16.27155	15.92060	756	301	3	244	304112303886854144	GALAXY	0.095463	270	51909	434
2	1237648721218633973	152.381682	-0.040947	18.45212	16.89779	16.04317	15.65277	15.35365	756	301	3	244	304113953154295808	GALAXY	0.095887	270	51909	440
3	1237648721218633979	152.388011	-0.009529	19.30342	17.38579	16.42225	15.99919	15.64936	756	301	3	244	304109005351970816	GALAXY	0.096564	270	51909	422
4	1237648721218633997	152.408919	-0.159743	19.27068	18.52680	18.12876	17.75474	17.72392	756	301	3	244	304054579526395904	GALAXY	0.137550	270	51909	224

A view of the data

Contents of the **image table**:

- objid = Object Identifier
- ra = Right Ascension of the object
- dec = Declination of the object

  Right ascension is the angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth. Right ascension and declination together provide the astronomical coordinates of a point on the celestial sphere in the equatorial coordinate system.

- u, g, r, i, z filters

  The five SDSS ugriz filters are a modified Thuan-Gunn astronomic magnitude system. u, g, r, i, z represent the response of the 5 bands of the telescope, which cover an effective wavelength range of ~350 nm to ~920 nm. Central wavelengths of the six filters: u -3551 Å g - 4686 Å r -6166 Å i -7480 Å z - 8932 Å.

  More on imaging techniques of SDSS :
  https://www.sdss.org/dr13/imaging/imaging_basics/

  More on the camera and its filters :
  https://www.sdss.org/instruments/camera/#Filters

- run = Run Number : identifies the specific scan
- rereun = Rerun Number : specifies how the image was processed
- camcol = Camera column : identifying the scanline within the run
- field = Field number

These parameters describe the field that was extracted from an image by the SDSS.

Contents of the **spectral table**:

- specobjid = Object Identifier
- class = object class (galaxy, star or quasar) : our target column
- redshift = Redshift of the object
- plate = plate number
- mjd = (Modified Julian Date) : observation date
- fiberid = fiber ID

The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID.
For more info, refer : https://www.sdss.org/dr16/help/glossary/

➢ Data Exploration

- A descriptive statistics of the **principle features** from our training dataset.

	ra	dec	u	g	r	i	z	redshift
count	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000
mean	171.840789	16.163274	18.628606	17.377191	16.843009	16.582462	16.420090	0.148730
std	48.014533	25.711136	0.822021	0.944800	1.069495	1.143177	1.210852	0.395147
min	8.235100	-8.479532	12.988970	12.565210	11.629170	11.186280	10.874560	-0.004136
25%	151.301233	-0.463345	18.190347	16.825817	16.181415	15.855960	15.617965	0.000121
50%	174.102955	0.624408	18.862705	17.496600	16.853640	16.549470	16.378900	0.046364
75%	199.514506	45.729393	19.261320	18.012950	17.510675	17.253030	17.132850	0.094373
max	348.902530	68.542265	19.599900	19.918970	24.802040	28.179630	22.959940	5.353854

Observations :

- The maximum and minimum value of **ra** (Right Ascension) is **~360** and **~0** respectively, but for **dec** (Declination) it is **-19.5°** and **85°**, therefore we can infer that the observations are mostly of the Northern part of the sky.
- None of the individual observations have any missing feature value.

- Counts of the different classes present in the column **"class"** of each data release.

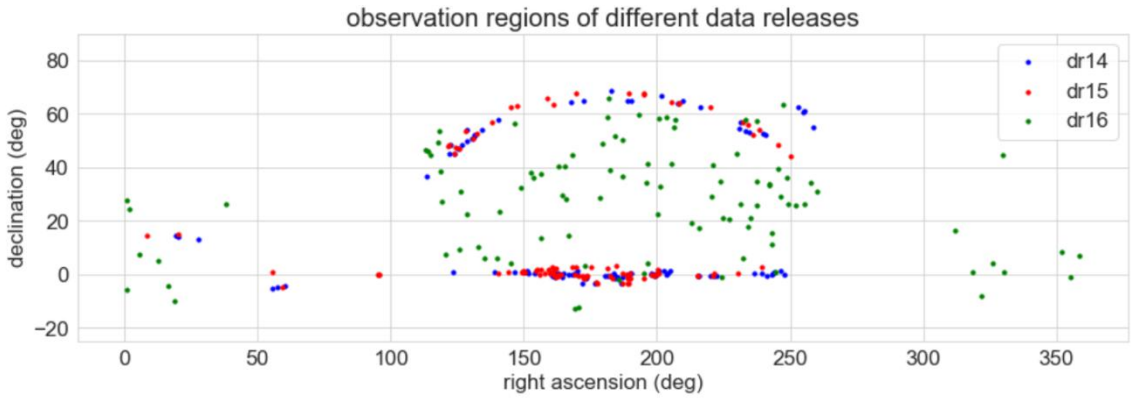
	data release 14	data release 15	data release 16
GALAXY	4998	5386	51323
STAR	4152	3713	38096
QSO	850	901	10581

Name: class, dtype: int64

Observation:

- Most of the objects are galaxies(>50%), stars constitute ~40% of the data and only ~10% are Quasars.

- A depiction of the regions of the sky that were detected in every release.



Observation:

- The dr14 and dr15 scans were mostly on the equatorial axis and along a certain arc while dr16 was spread out, mostly between 100 - 275 degrees.

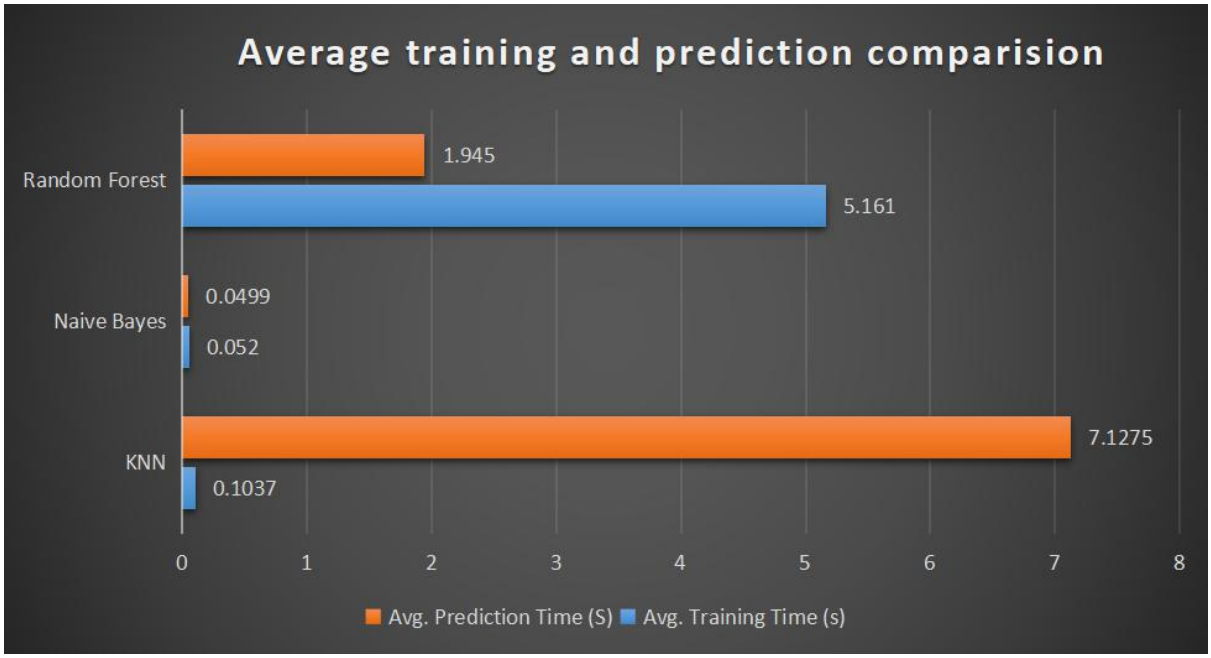
➢ Application of ML Models

link to our code : https://drive.google.com/drive/folders/1oQfYpG-VfQw1vIkEOxZpLzlmXCeegdjJ?usp=sharing

Note : It is assumed that the reader has knowledge of the ML models used below.

All the columns excluding *uqriz* filters and redshift were removed from the training data set since in theory they should have no correlation with the target predictions. Removing these columns help in optimizing the resource & time efficiency of the algorithms. These also help in increasing the accuracy of our dataset by reducing the overfitting of the relatively smaller training dataset. For example, removing the right ascension and declination columns from our principle features helped in increasing the accuracy of KNN algorithm by 20% !

Time complexity of the Algorithms

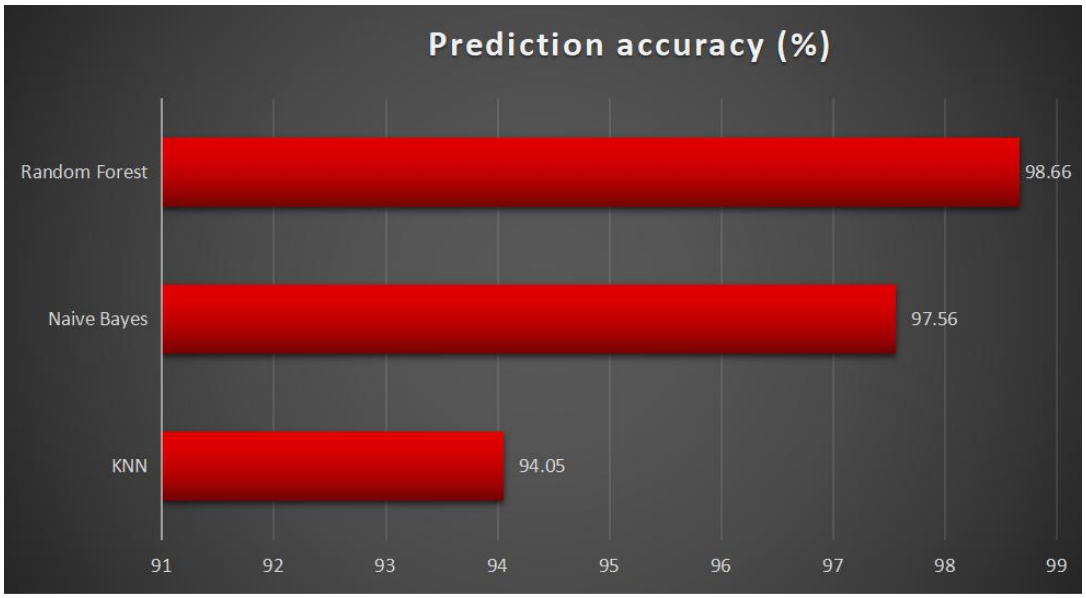


Observations:

i. The average prediction time of KNNs are much higher than the other two models.
Explaination: This is because the KNN simply memorizes the entire dataset during training. During the prediction phase, it has to find the N nearest neighbors among all the training samples in the K dimensional space of features for every query. (N is the n_neighbours parameter of the algorithm and K is the number of features)

ii. Naive bayes requires least resources among all the Models.
Explaination: This is because the Naïve bayes classifier just needs to compute the statistical mean and standard deviation of each of the features of every class and then make predictions from the Gaussian curves thus obtained.

Accuracy of the Algorithms



Observations :

- Random Forests and Naive Bayes achieved 98.66% & 97.56% accuracy respectively.
- KNN achieves a lower accuracy of 94.05%.

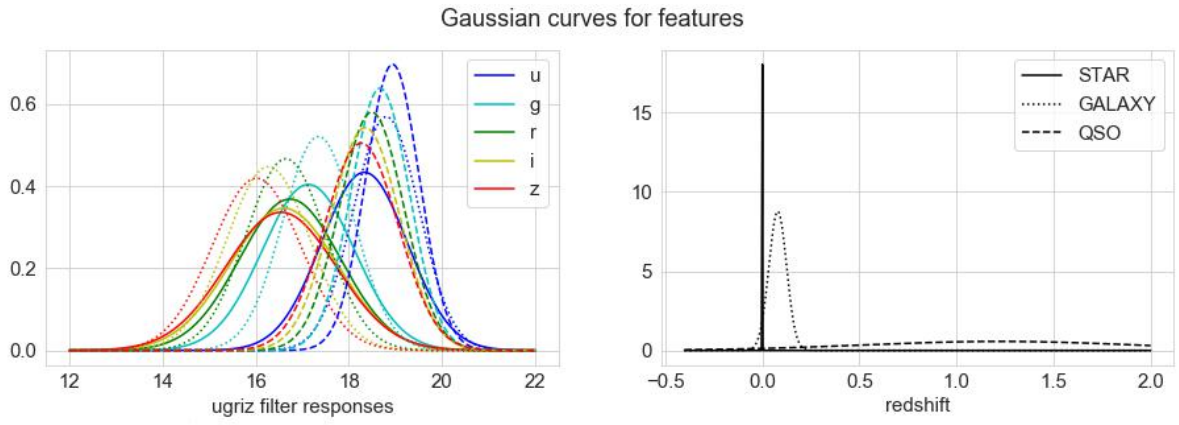
  Possible explaination:
  The low efficiency of KNN can be due to the fact that they are not robust against outliers.

➢ Analysis of features & algorithms :

- **Gaussian Naïve Bayes**

The NB classifier is a probabilistic classifier based on the application of Bayes' theorem while assuming that the features occur independently of each other. The Gaussian NB assumes that the continuous values are distributed normally with a mean and standard deviation.

To understand how the gNB is able to achieve high accuracies we take a look at the data by plotting a Gaussian function using statistical analysis for every feature of every class from our training dataset. We present two plots below :



Observations:

i. The mean of *ugriz* filter responses of Galaxies and stars are similar but lower than that of Quasar.
ii. The redshift distribution functions :
   a. Stars = Sharpest , lowest mean
   b. Galaxies = intermediate standard deviation, slightly higher mean than stars
   c. Quasars = High standard deviation, higher mean than both stars and galaxies

Physical Explaination : Quasars are extremely luminous, therefore it is expected that they have high responses for every filter. Also because of this reason we expect that Quasars can be detected for wider and much higher ranges of distances by our Telescopes. If the distances are higher, then by the Hubble distance-redshift relationship, we should expect that objects with greater ranges of redshifts can be observable and will mostly be Quasars, followed by galaxies and then by stars.

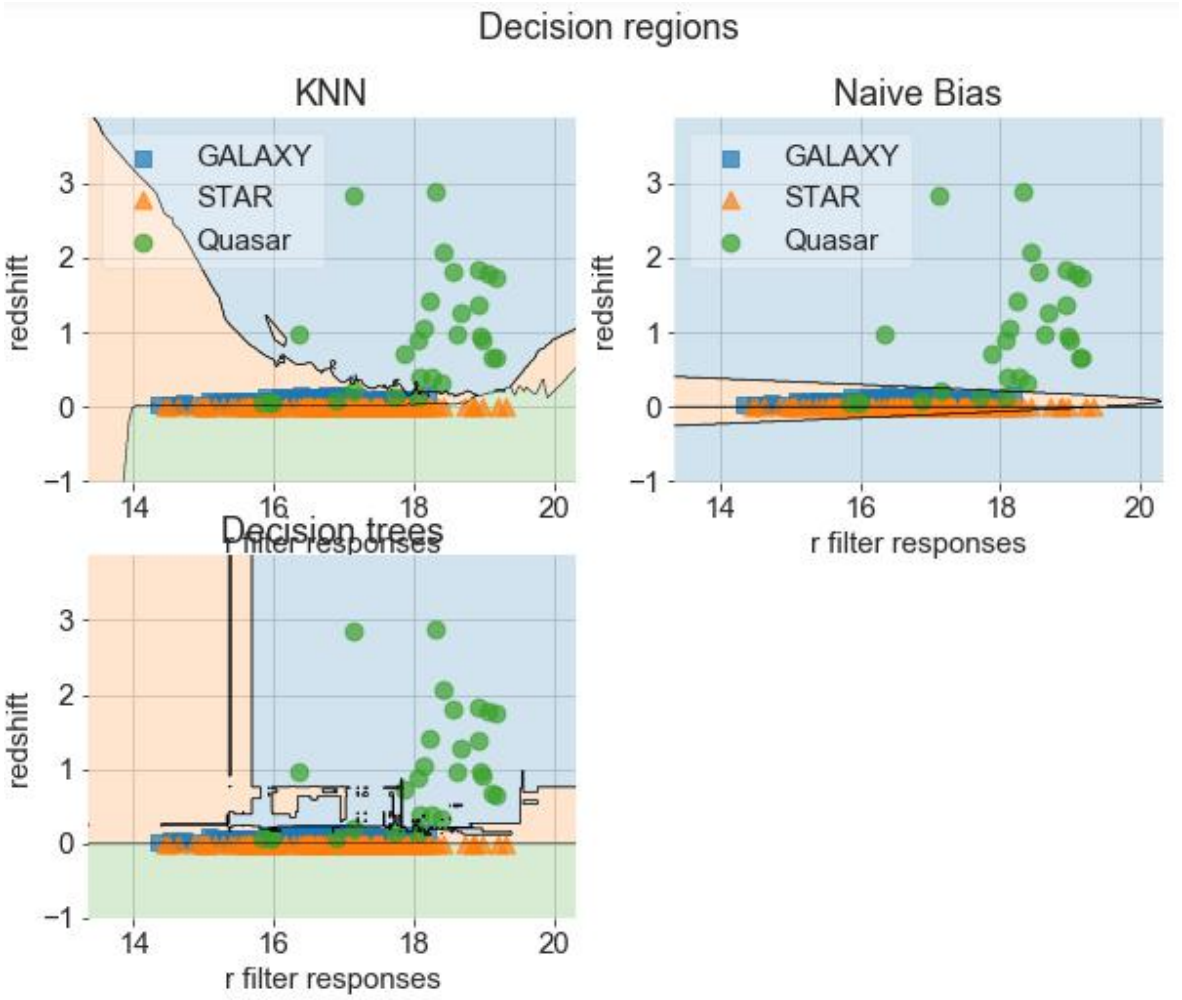
We observe that because of these remarkable differences between stars, galaxies and Quasars, the normal distributions of the features of these classes are different for different classes and thus immensely help in categorizing the datasets in our NB model.

- **Comparison between the Decision boundaries of the Models**

Decision boundaries can help in partitioning the underlying vector space of all the features for each class to be distinguished from each other.

The KNN Models working can be understood easily by showing the decision boundaries of the features in the whole space. Since the space will have as many dimensions as the number of features, it can be difficult the depict decision boundaries for more than 2 features.

To solve this we used two of the six features of the dataset.We observe from previous section that most of the *ugriz* filters had good correlation for each other for a particular class. So we took the middle filter 'r' and assumed that the other filters would have similar trends to some approximation. The second feature that was taken is the 'redshift'.

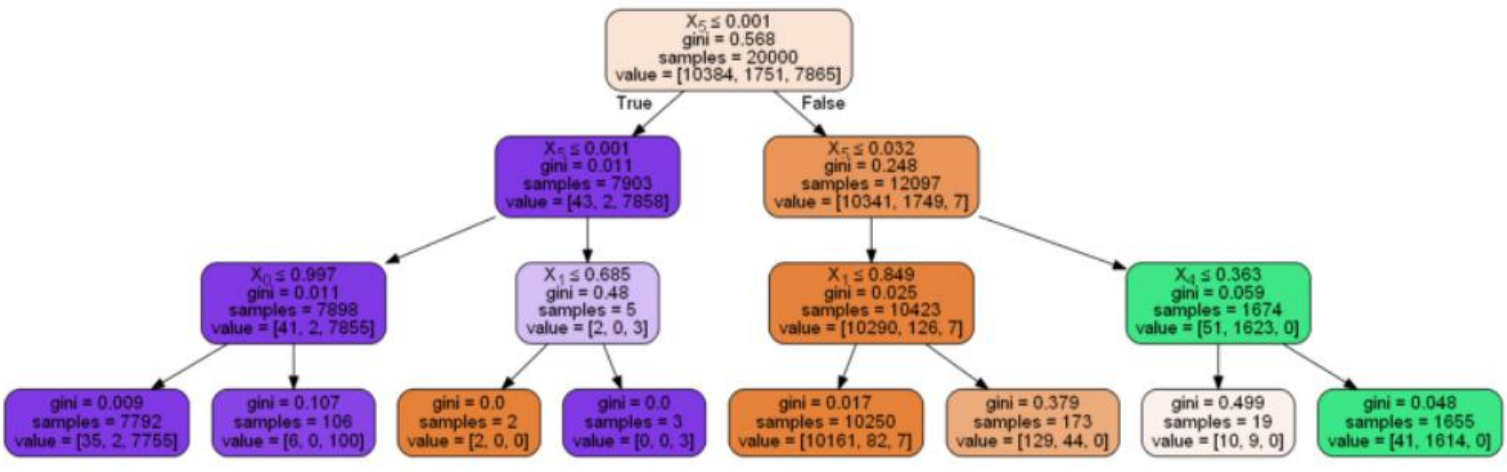


Observations:

i. All three algorithms have allotted a large diffused space for predicting quasars, even though they are only 10% of the dataset.
ii. The boundaries of Decision trees and KNN seems to look similar, and different to Naïve Bayes since Naïve bayes has allotted very small area to prediction of Stars.
iii. By observing the small cuts inside the Quasar regions of Decision trees, we can say that the Random forests are probably over-fitting to the training dataset.

- **Visualization of Random Forests**:

Random forests mix multiple decision trees together to predict the classes. Therefore to understand how the Random forests work, we made a simplistic Decision tree model with a maximum depth of just 3 nodes. This decision tree was able to achieve an accuracy of 98.34%. We provide a visual depiction of the decision tree :



Here $X_{0\text{-}4}$ denotes the *ugriz* filters and X_5 denotes the redshift. We see that the Decision tree distinguishes between the classes in first two steps by only using the redshift values, which indicates that the red shift (and in turn the distance of the observable object) plays a dominant role in the decisions. It is able to separate most of the Quasars from the stars and galaxies solely from this method which leads us to same physical conclusion as before that most stars, galaxies observed are closer than Quasars in the training dataset.

➢ Conclusion

Random Forest and Naive Bayes model's prediction accuracy are the best as they managed to predict 98.66% & 97.56% accuracy respectively, of the data correctly of the successive datasets. Here, the KNN models lose to the other two as it could only predict 94 % of the data. Though the training time of KNN is less than the random forest, it took much time in predicting the data, which accounts for its inefficiency.

To compare Naive Bayes and Random Forest, we observe that the prediction accuracy is a little higher for random forests. Naive Bayes provides a much better time complexity, as seen from the training and prediction timings, compensating its inefficiency by small margins. Therefore, Bayes model suits the best for future SDSS dataset modeling and predictions.