Sloan Digital Sky Survey Project Report

for

Dr. Alan Dabney
Associate Professor
Texas A&M University
College Station, TX

by

Ian Scarff
STAT 485-500
April 30, 2019

Abstract

For this project, data from the Sloan Digital Sky Survey was used to test various classification methods to classify objects as stars, galaxies, or quasars and identify the best model. The dataset was randomly split into separate training and testing datasets, with the training dataset containing 75% of the original data and the remaining 25% comprising the test dataset. From this dataset, the variables of interest were the class of an object (star, galaxy, quasar), the responses from the five optical filters using the Thuan-Gunn Star Magnitude System (u, g, r, i, z), and the object's redshift. Statistical models used in this analysis are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Basic Trees, and Random Forest. Based on a model's confusion matrix, an accuracy for each class was calculated. Due to an unequal number of observations per class, accuracy was measured by (w * sensitivity) + ((1 – w) * specificity), were w is a tuning parameter. When compared to the other models, Random Forest had the highest accuracies for each class.

Introduction

Astronomy is considered to be one of the most important sciences. Even before the first civilizations, humans studied the stars for insight into the world around them. From astronomy, we learned when the best times to plant crops were, how to navigate the globe, and how it has been incorporated into various early religious practices. For the past few centuries, we have used astronomy to gain insight about the universe as a whole. One of the first steps in studying an astronomical object is to identify what it is. As telescopes have improved, we have gone from solely studying visual light with our eyes to studying the entire electromagnetic spectrum with telescopes interpreting light as data. From this data, we can determine the type of object. Being able to determine what an object is quickly and accurately is key for astronomical research.

For this project, data from the Sloan Digital Sky Survey (SDSS) was used. The SDSS's purpose is to do large scale surveys of the night sky and create detailed three-dimensional maps of the universe. The data used in this analysis came from SDSS's data dump 14 found on the website Kaggle. The dataset contains 10,000 observations, each with 18 variables. A detailed list and description of the variables can be found in Appendix A. The goal of this project is to identify the best model for classifying an observation as a star, galaxy, or quasar. Various classification models were compared against each other to determine the most accurate model. The variables of interest are the class of an object (star, galaxy, quasar), the responses from the five optical filters using the Thuan-Gunn Star Magnitude System (u, g, r, i, z), and the object's redshift, which is equal to the ratio between the observed wavelength and rest wavelength minus 1[$(\frac{\lambda_{observed}}{\lambda_{rest}} - 1)$].

Exploratory Analysis

In this dataset, there are 3 kinds of astronomical objects: stars, galaxies, and quasars. However, there are not an equal number of these objects in the data. As summarized in Table 1, galaxies and stars make up a majority of the data, while quasars make up a significantly smaller amount.

Table 1.

	# of Observation	% of Data
Star	4,152	41.50%
Galaxy	4,998	50%
Quasar	850	8.50%

Figure 1.

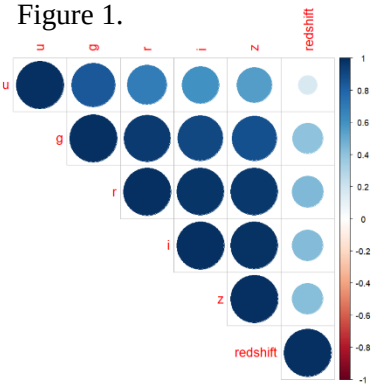


Figure 1 to the left shows the level of correlation across each variable. From this figure, we can see that there is high correlation between the five filter responses, while there is low correlation between redshift and the five filter responses. This is further illustrated by the matrix plot of the variables found in Appendix B. A correlation matrix can also be found in Appendix B. Figure 2 below shows the histograms for each of the five filter responses. We can see that the distribution of the classes in each variable overlap each other and have the same basic shape. It would be hard to classify an object solely based on these variables. However, redshift tells a different story. Figures 3 and 4 below show the histogram for redshift and a magnified histogram for redshift, respectively. We see that there isn't as much overlap in classes and each has distinct ranges. Table 2 displays the summary statistics of these histograms.

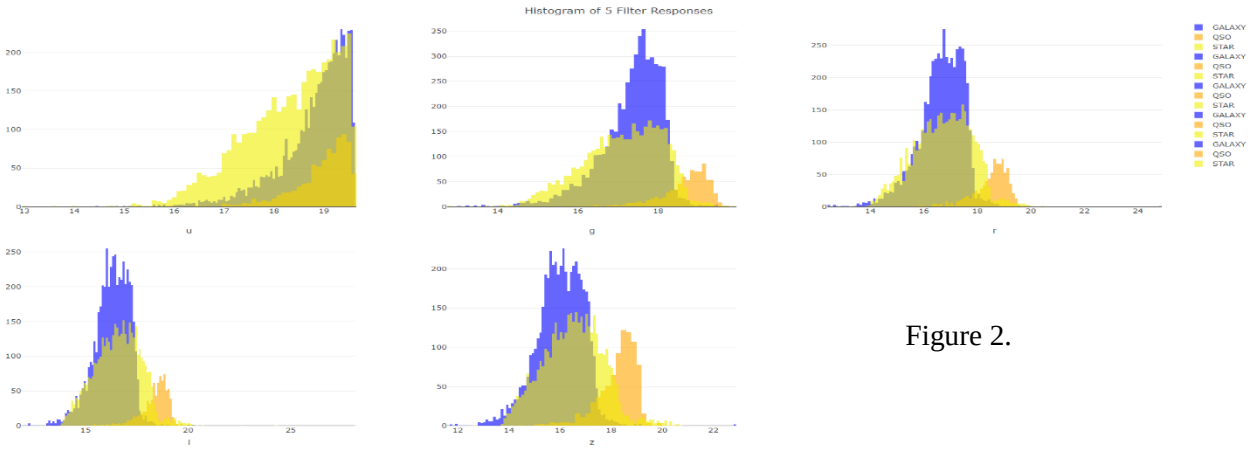


Figure 2.

Figure 3.

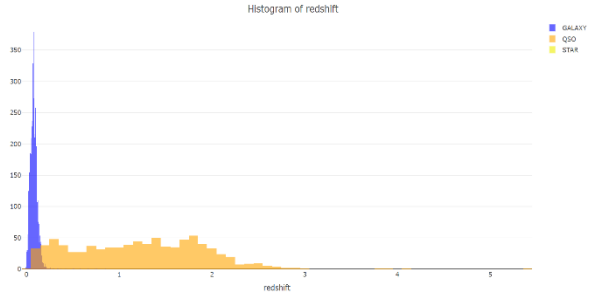


Figure 4.

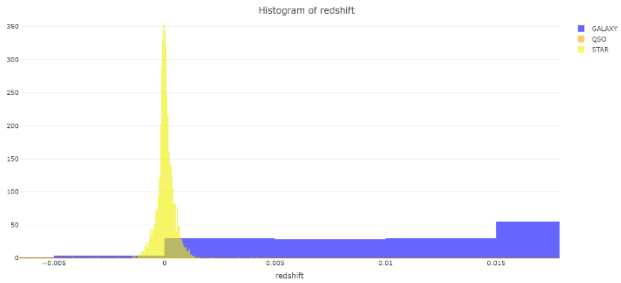


Table 2.

		u	g	r	i	z	redshift
Star	**Min**	12.99	12.8	12.48	13.25	12.61	-4.14E-03
	Median	18.5	17.24	16.79	16.62	16.55	2.45E-05
	Mean	18.33	17.13	16.73	16.59	16.33	4.33E-05
	Max	19.6	19.92	24.8	28.18	20.8	4.56E-03
Galaxy	**Min**	14.46	13.08	12.43	11.95	11.61	-5.13E-08
	Median	18.99	17.5	16.74	16.35	16.07	0.077
	Mean	18.8	17.35	16.65	16.27	16.02	0.08
	Max	19.6	19.68	24.8	24.36	22.82	0.86
Quasar	**Min**	15.93	15.76	15.36	14.82	14.52	4.61E-08
	Median	19.09	18.82	18.66	18.53	18.46	1.23
	Mean	18.94	18.68	18.5	18.36	18.27	1.21
	Max	19.6	19.74	19.87	20.01	20.44	5.35

Finally, Figures 5, 6, and 7 below further show that the majority of observations in each class occupy their own regions. Based on these graphs, we can see where possible dividers could be placed to classify each observation. We can see that the addition of redshift in Figure 7 offers a better look at how each class is spread across a three-dimensional space, with stars and galaxies occupying the bottom space (with galaxies slightly above stars) and quasars spreading up and out.

Figure 5.

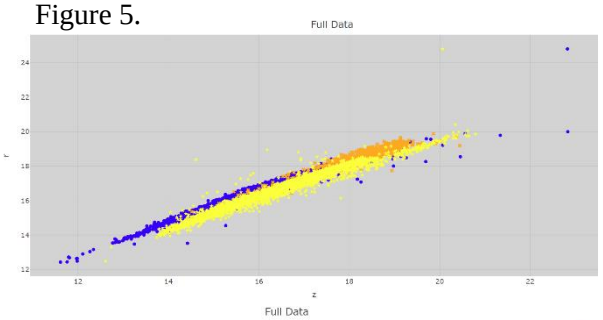


Figure 6.

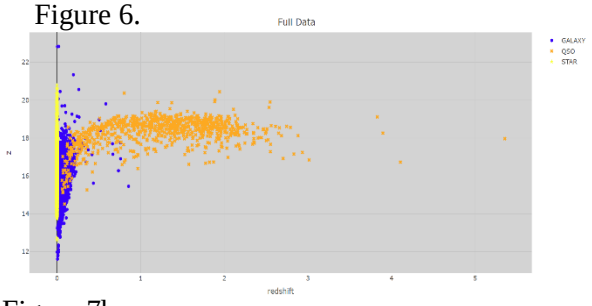


Figure 7a.

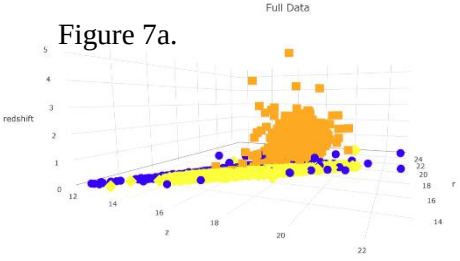
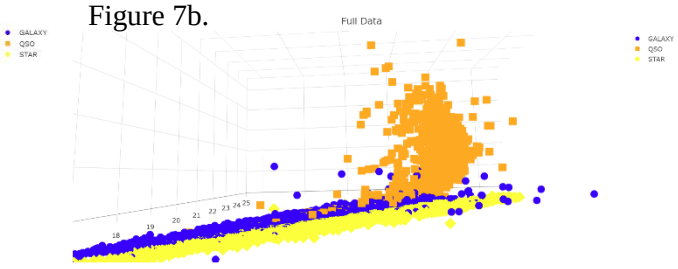


Figure 7b.



Methods & Procedures

To build various classification models, a training data set was constructed by taking a random sample from the full 10,000 observations. This training data comprised 75% of the full data while the remaining 25% was used to create a test dataset. This test data was used to validate the models. Models were built using four different methods: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Basic Trees, and Random Forests. The predictor variables were the five optical filter responses and redshift, with the response variable being class. Using these models, confusion matrices were constructed from predictions made using both the training and test datasets. From these confusion matrices an accuracy for each class was determined. Accuracy was calculated by (w * sensitivity) + ((1 – w) * specificity), where w is a tuning parameter with a range from 0 to 1, with 0.01 increments.

Results

From comparing each model's confusion matrix and accuracy measurements, the Random Forest model was found to be the most accurate model. Tables 3 and 4 to the left show the confusion matrix for the Random Forest predictions using the test data and the resulting accuracy measurements, respectively. When compared to the confusion matrices and accuracy measurements of the other models, which can be found in Appendix B, we see that each class had the highest accuracy when using the Random Forest model. Figure 8 below shows the varying accuracies per weight. When compared to similar graphs of the other models, which can also be found in Appendix B, we can see that the optimal accuracy lies around a weight of 0.15. In addition, when

Table 3.

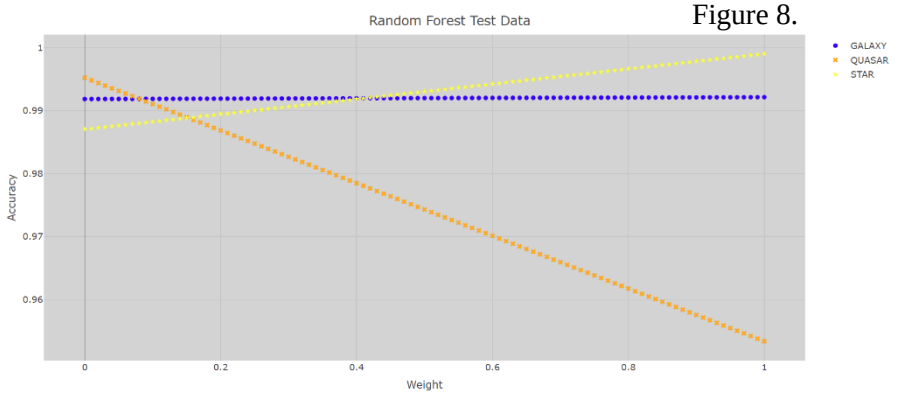
RF Confu. Matrix (Test Data)			
	Galaxy	**Quasar**	**Star**
Galaxy.pre	1263	9	1
Quasar.pre	7	184	0
Star.pre	3	0	1033

Table 4.

	RF Accuracy Measurements (Test Data)			
	Sensitivity	Specificity	$Weight_{MaxAccu.}$	Max Accuracy
Galaxy	0.992	0.991	1	0.992
Quasar	0.953	0.995	0	0.995
Star	0.999	0.987	1	0.999

compared to other models, the difference between the optimal accuracy's is minimal with the Random Forest model.

Figure 8.



In addition to the Random Forest model, the Basic Tree model also presented something of interest. Figure 9 to the right shows the plot of the Basic Tree. It chose redshift as the divider for classes. Any object with a redshift below ≈0.0042 is a star, any object with a redshift between ≈0.0042 and ≈0.2177 is a galaxy, and anything with a redshift greater than ≈0.2177 is a quasar. Figure 10 to the right shows these dividing lines (zoomed in to show stars clearly). We can see that practically all stars lie in their specified region. While there is a little overlap between galaxies and quasars, the majority of them lie in their own specified regions as well.

Figure 9.

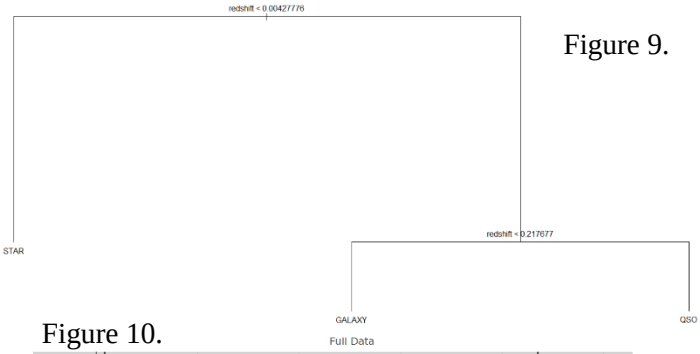
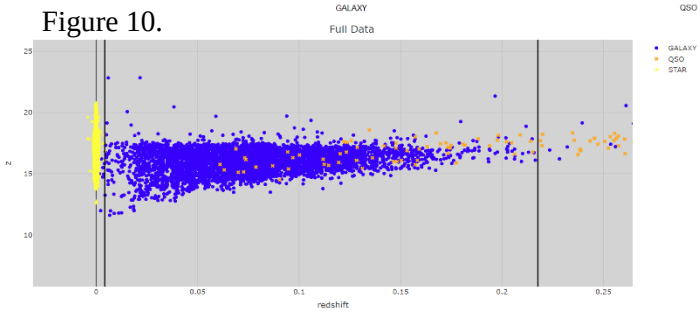
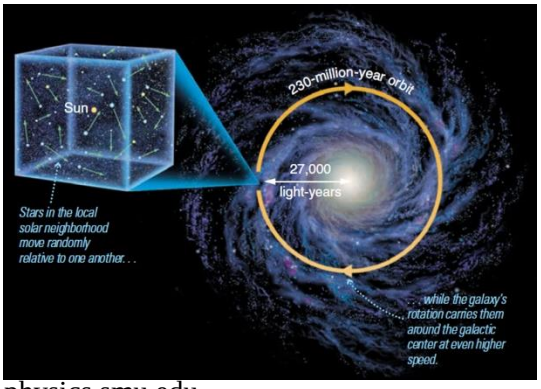


Figure 10.

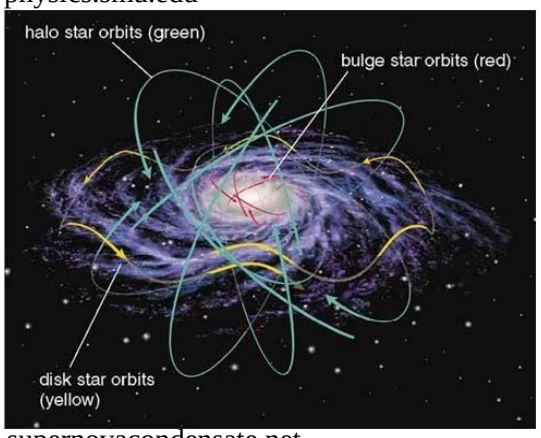


Furthermore, based on the Random Forest model, redshift was determined to be the most important variable, according to the Gini Index. Astronomically, the importance of redshift and where the dividers are placed this makes sense. We can only see individual stars that lie in our own galaxy. All the stars in the disk of Milky Way travel in the same general direction around

the center. However, each star moves in a random direction. This means some stars move towards us and some move away. Stars in the galactic halo and galactic bulge have major inclinations (an angled orbit) and will have more significant redshift. This is illustrated in the pictures to the left. So, we would expect to see stars lie in a very small, tight region in a plot of redshift near zero, which we do. For galaxies outside the Local Group, due to the accelerating expansion of the universe, we would expect the redshift to be greater than stars. This means that the higher the redshift, the farther the galaxy. Most quasars are very old objects, so they are farther away, and therefore have extreme redshifts.



physics.smu.edu



supernovacondensate.net

Discussion

Future work for this project would include the testing of more classification methods, such as Boosting and Support Vector Machines. Deeper knowledge of the classification methods used would further aid in model building and determining which is most accurate. In addition, perhaps there is a better measurement for accuracy. These models should also be tested using future data dumps from SDSS. This project is the first stepping stone in aiding astronomical and astrostatistics research.

Acknowledgments

I would like to thank Dr. Alan Dabney for all his support and helpful comments throughout the duration of this project.

Works Cited

Bessell, Michael S. "Magnitude Scales and Photometric Systems." *ENCYCLOPEDIA OFASTRONOMY ANDASTROPHYSICS*. Basingstoke: Nature Publishing Group, 2001. 2,6-7. Document. <http://hea.iki.rssi.ru/AZT22/RUS/eaa_bessel.pdf>.

—. "Standard Photometric Systems." *Annual Review of Astronomy and Astrophysics* (2005): 13-14.

Breiman, Leo, et al. "Package ‘randomForest’." *Breiman and Cutler's Random Forests for Classification and*. 25 March 2018. <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>.

Masci, Frank. *MAGNITUDE AND COLOR SYSTEMS*. 3 November 2014. Document. <http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/magsystems.pdf>.

Ripley, Brian. "Package ‘tree’." *Classification and Regression Trees*. 17 March 2018. <https://cran.r-project.org/web/packages/tree/tree.pdf>.

Ripley, Brian, et al. "Package ‘MASS’." *Support Functions and Datasets for Venables and Ripley's MASS*. 31 March 2019. <https://cran.r-project.org/web/packages/MASS/MASS.pdf>.

Sievert, Carson, et al. "Package ‘plotly’." *Create Interactive Web Graphics via 'plotly.js'*. 10 April 2019. <https://cran.r-project.org/web/packages/plotly/plotly.pdf>.

Sloan Digital Sky Survey. *The Sloan Digital Sky Survey: Mapping the Universe*. 2018. <https://www.sdss.org/>.

Southern Methodist University. *Physics*. 2019. <http://www.physics.smu.edu/jcotton/ph1311/ch14b.htm>.

Wei, Taiyun, et al. "Package ‘corrplot’." *Visualization of a Correlation Matrix*. 17 October 2017. <https://cran.r-project.org/web/packages/corrplot/corrplot.pdf>.

Xan, Invader. *Galactic orbits*. 22 July 2012. <https://supernovacondensate.net/2012/07/22/galactic-orbits/>.

Appendix A

Sloan Digital Sky Survey Data Variables

- Objid
  - Object Identifier.
- Ra
  - J2000 Right Ascension (r-band).
  - Right ascension (abbreviated RA) 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 in question.
- Dec
  - Declination (abbreviated dec).
  - The angular distance of a point north or south of the celestial equator.
- u, g, r, i, z
  - Variables from the Thuan-Gunn Star Magnitude System.
  - Thuan-Gunn System is a type of star magnitude system using various set of broad-band filters. The variables u,g,r,i,z are the responses from the five optical filters. It is an extension of the UBVRI system, with filters optimized for faint galaxies by rejecting night sky lines.
- Run
  - Run Number.
- Rerun
  - Rerun Number.
- Camcol
  - Camera Column.
- Field
  - Field Number.
  - Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by: - run number, which identifies the specific scan, - the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and - the field number. The field number typically starts at 11 (after an initial ramp up time), and can be as large as 800 for particularly long runs. An additional number, rerun, specifies how the image was processed.
- Specobjid
  - Object Identifier.
- Class
  - Object Class
  - Galaxy, Star, or Quasar.
- Redshift
  - Final Redshift.
  - Redshift occurs when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.
- Plate
  - Plate Number.

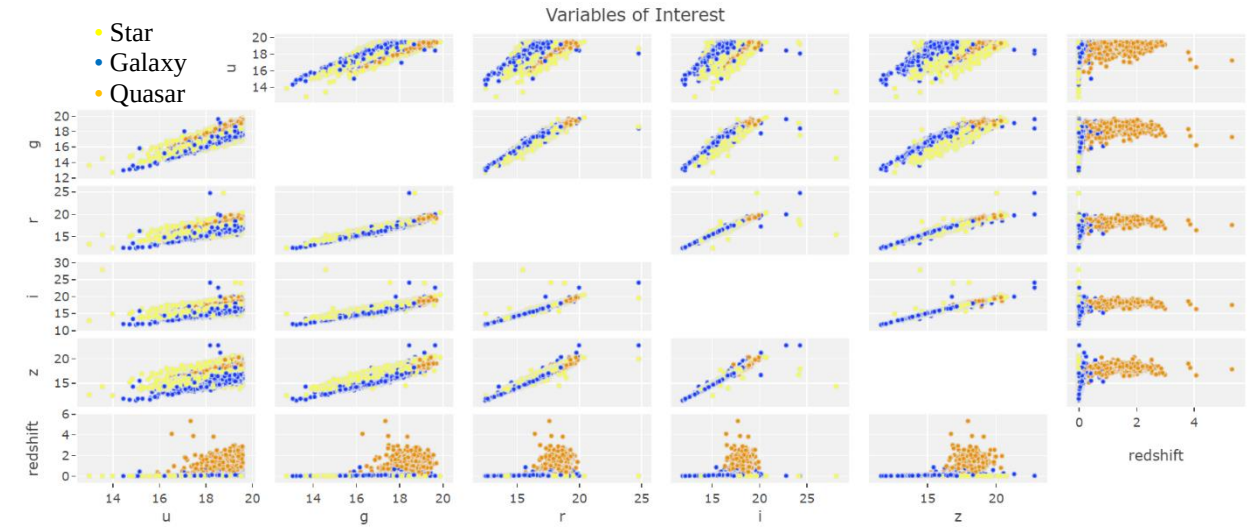
- Mjd
  - Modified Julian Date
  - Used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.
- FiberId
  - Fiber ID
  - The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID.

Appendix B

Correlation Matrix

	u	g	r	i	z	redshift
u	1.000	0.849	0.692	0.603	0.551	0.164
g	0.849	1.000	0.958	0.907	0.880	0.408
r	0.692	0.958	1.000	0.978	0.969	0.441
i	0.603	0.907	0.978	1.000	0.982	0.431
z	0.551	0.880	0.969	0.982	1.000	0.424
redshift	0.164	0.408	0.441	0.431	0.424	1.000

Matrix Plot



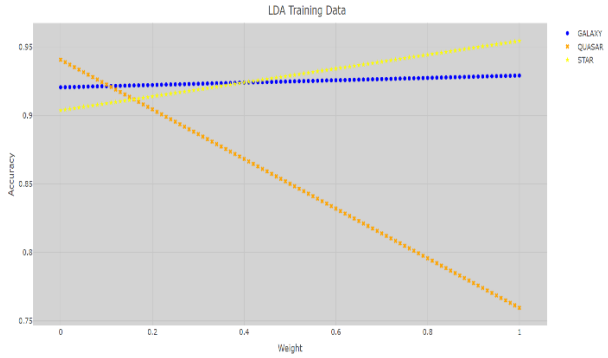
LDA

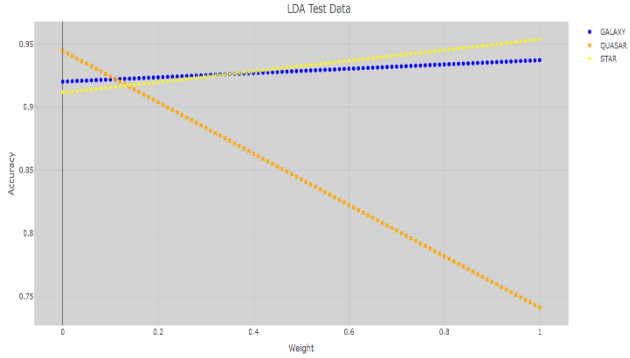
	LDA Confu. Matrix (Training Data)		
	Galaxy	**Quasar**	**Star**
Galaxy.pre	3461	146	139
Quasar.pre	4	499	3
Star.pre	260	12	2976

	LDA Confu. Matrix (Test Data)		
	Galaxy	**Quasar**	**Star**
Galaxy.pre	1193	48	48
Quasar.pre	1	143	0
Star.pre	79	2	986

	LDA Accuracy Measurements (Training Data)			
	Sensitivity	**Specificity**	**Weight**	**Max Accuracy**
Galaxy	0.929	0.921	1	0.929
Quasar	0.76	0.941	0	0.941
Star	0.955	0.904	1	0.955

	LDA Accuracy Measurements (Test Data)			
	Sensitivity	**Specificity**	**Weight**	**Max Accuracy**
Galaxy	0.937	0.92	1	0.937
Quasar	0.741	0.945	0	0.945
Star	0.954	0.911	1	0.954





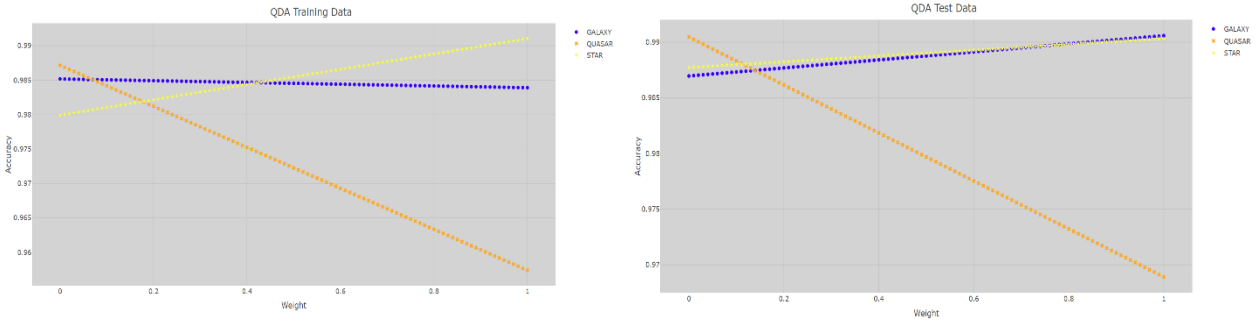
QDA

	QDA Confu. Matrix (Training Data)		
	Galaxy	Quasar	Star
Galaxy.pre	3665	27	12
Quasar.pre	39	629	16
Star.pre	21	1	3090

	QDA Confu. Matrix (Test Data)		
	Galaxy	Quasar	Star
Galaxy.pre	1261	6	1
Quasar.pre	10	187	9
Star.pre	2	0	1024

	QDA Accuracy Measurements (Training Data)			
	Sensitivity	Specificity	Weight	Max Accuracy
Galaxy	0.984	0.985	0	0.985
Quasar	0.957	0.987	0	0.987
Star	0.991	0.98	1	0.991

	QDA Accuracy Measurements (Test Data)			
	Sensitivity	Specificity	Weight	Max Accuracy
Galaxy	0.991	0.987	1	0.991
Quasar	0.969	0.991	0	0.991
Star	0.99	0.988	1	0.99



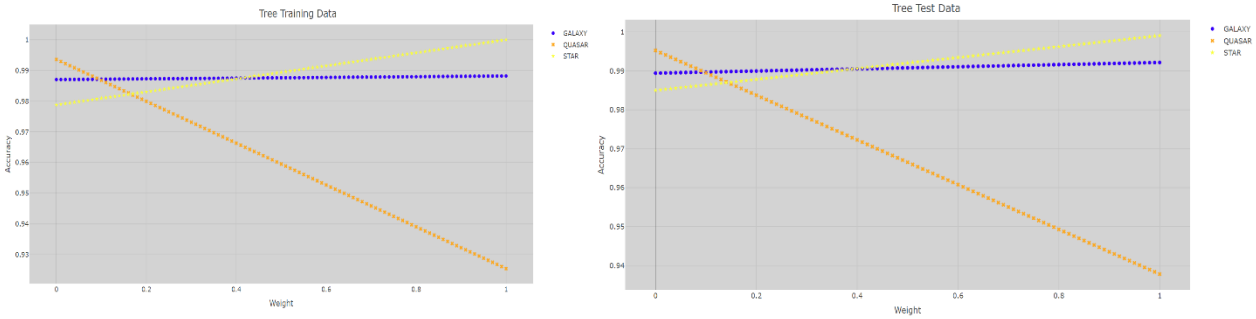
Basic Tree

	Tree Confu. Matrix (Training Data)		
	Galaxy	Quasar	Star
Galaxy.pre	3681	48	0
Quasar.pre	19	608	0
Star.pre	25	1	3118

	Tree Confu. Matrix (Test Data)		
	Galaxy	Quasar	Star
Galaxy.pre	1263	12	1
Quasar.pre	6	181	0
Star.pre	4	0	1033

	Tree Accuracy Measurements (Training Data)			
	Sensitivity	Specificity	Weight	Max Accuracy
Galaxy	0.998	0.987	1	0.988
Quasar	0.925	0.994	0	0.994
Star	1	0.979	1	1

	Tree Accuracy Measurements (Test Data)			
	Sensitivity	Specificity	Weight	Max Accuracy
Galaxy	0.992	0.989	1	0.992
Quasar	0.938	0.995	0	0.995
Star	0.999	0.985	1	0.999



Random Forests

	RF Confu. Matrix (Training Data)		
	Galaxy	Quasar	Star
Galaxy.pre	3725	0	0
Quasar.pre	0	657	0
Star.pre	0	0	3118