MATTER: International Journal of Science and Technology
ISSN 2454-5880

Yulun (Winston) Wu, 2021

Volume 6 Issue 3, pp. 102-122

Date of Publication: 5th January 2021

DOI-https://doi.org/10.20319/mijst.2021.63.102122

This paper can be cited as: Wu, Y. W., (2021). Machine Learning Classification of Stars, Galaxies, and Quasars. MATTER: International Journal of Science and Technology, 6(3), 102-122.

MACHINE LEARNING CLASSIFICATION OF STARS, GALAXIES, AND QUASARS

Yulun (Winston) Wu
11th Grade, Northfield Mount Hermon School, Gill, United States
pwinstonwu@gmail.com

Abstract

The objective of this study was to create a predictive model to classify stars, galaxies, and quasars, along with comparing different classification models to find the superior one. I hypothesized that it was possible to successfully train a machine learning model to classify stars, galaxies, and quasars using astronomical data provided by the Sloan Digital Sky Survey (SDSS). A multinomial logistic regression model has been trained and tested. It had an accuracy of 0.87, a weighted average precision, recall, and an f-1 score of 0.87, and a cross-validation accuracy score of 0.8664. The next model, a decision tree, had an accuracy of 0.99, weighted average precision, recall, and an f-1 score of 0.99, a cross-validation accuracy score of 0.99, and a cross-validation accuracy score of 0.9858. The decision tree model had significantly superior performance compared to the logistic regression model and was a good fit and accurate classifier for stars, galaxies, and quasars, proving my hypothesis to be correct. The model from this study could be used as a reliable classification tool for a wide variety of astronomical purposes to accelerate the expansion of the sample sizes of stars, galaxies, and especially quasars.

Keywords

Logistic Regression, Decision Tree, Stars, Galaxies, Quasars, Classification

1. Introduction

It has always been the pursuit of astronomers to study, analyze, and understand the universe with its countless celestial objects of endless variety. However, due to the celestial objects' differing distances, light-emitting intensities, and positions in space, they often appear as near-identical smudges or dots of light when recorded using photometric data from telescopes, making it difficult to differentiate them from each other (Clarke et al., 2020). Therefore, the classification of celestial objects is very fundamental in astronomy for astronomers to be able to collect meaningful samples to critically study these celestial objects and uncover the complex and mysterious past, present, and future of the universe (Clarke et al., 2020). This study focused on the classification of three celestial objects: stars, galaxies, and quasars.

1.1 Background Information

Stars are burning hot balls of hydrogen and helium. They are formed through processes spanning millions of years where turbulent clouds of gas and dust continuously gather and clump together due to gravity and eventually collapse under their mass and gravity. (Stars and Nebulae). Analyzing stars is important to uncovering many of the complex processes and inner workings of celestial objects as well as better understanding our own Sun.

Galaxies are vast pools of stars, planets, galactic dust, gas, dark matter, and many other celestial objects. Studying galaxies, such as their composition and how they have formed and developed over time, can allow astronomers to learn and understand the functions, processes, and timeline of the universe on a much bigger scale and also better understand our home galaxy, the Milky Way, as well (Galaxies and QSOs).

Quasars, or quasi-stellar-objects, are created by supermassive black holes found in the galactic centers of galaxies (Redd, 2018). This type of galactic nuclei is formed by a disk-like cloud of dust and gas, the accretion disk, swirling around and eventually into the black hole. The friction between dust and gas from the black hole's immense pressure and gravity producing extremely high amounts of heat and light, making quasars one of the brightest celestial objects in the universe (Galaxies and QSOs). Quasars began forming at the early stages of the universe, thus their light and electromagnetic radiations, carrying information from billions of years ago, are a gift to unlock the history and understand the past of the universe. They can also be used to study a very wide range of topics such as galaxy evolutions, intervening intergalactic gas, cosmological evolution, black hole physics, etc. (Carrasco et al., 2015).

1.2 Significance and Literature Review

The classification of celestial objects is of great significance in the fields of astronomy. Its most direct benefit is providing the means to gather data samples of stars, galaxies, and quasars. Particularly for quasars, despite how important they are to a wide range of astronomy studies and research, their sample sizes are still in the relative minority class (Clarke et al., 2020). It is only through improved classification processes can significant increases be made in the sample sizes of quasars and other celestial objects, consequently enabling further progress to be made in research. With modern telescopes recording an increasingly large amount of astronomical data, the usage of machine learning models in the task of classification has become more and more significant and prevalent because of their accuracy and speed (Clarke et al., 2020). While reviews of previous studies with similar objectives of classifications found both supervised and unsupervised machine learning models to be capable of great performance with accuracy and other metrics measured at over 90%, the supervised models had generally higher accuracies in classification, and unsupervised models were shown to be more effective at detecting unknown objects (Viquar et al., 2018; Zhang et al., 2013). With supervised ML models, the classification models of random forest and support vector model (SVM) were used, along with decision tree and logistic regression, to either solely identify and separate quasars from other objects or to classify all three of stars galaxies, and quasars (Carrasco et al., 2015). Furthermore, data from SDSS were used to train many of the supervised classification models, indicating the information and variables that SDSS provides to be detailed, reliable, and highly relevant or connected to the goal of classifying celestial objects.

2. Research Issues

With the classification of celestial objects being so significant to the fields of astronomy in expanding their existing sample sizes for better more comprehensive analysis in the future, fast and reliable classifiers are needed to enhance this progress. To this end, I hypothesized that it was possible to successfully train a machine learning model to classify stars, galaxies, and quasars using astronomical data provided by the Sloan Digital Sky Survey (SDSS). The objective of this study, thus, was to create a predictive model to classify stars, galaxies, and quasars by training and comparing the performance of logistic regression and decision trees machine learning models to identify the superior one.

3. Method

The machine learning process is explained: data gathering, data analysis, data preprocessing, and training and testing of machine learning models.

3.1 Data

The data used in this study are from the Sloan Digital Sky Survey (SDSS), which is a leading astronomical survey that has been working for more than 20 years to produce extremely precise and detailed imaging and map of the universe. This public dataset, Data Release 14, is the second release of the fourth phase of the survey and had observations through July 2016. It contains 18 variables with 10,000 total entries and no missing values (Abolfathi et al., 2018; Blanton et al., 2017; Doi et al., 2010; Fukugita et al., 1996; Gunn et al., 1998; Gunn et al., 2006). However, 11 of the variables (location of an object on the celestial sphere, the field/area of pixels in the image taken, info and specifications on the spectroscopy, optical fibre, etc.) have no contributions towards the classification of the object and were removed (Blanton et al., 2017). The descriptions of the remaining 6 feature variables and 1 class variable (*Camera*; *Measures Of Flux And Magnitude*; *Redshifts, The Photometric Camera and the CCDs*; *Understanding SDSS Imaging Data*; *Understanding the Imaging Data*), their first 10 entries, and their statistics are shown in Tables 1, 2, and 3 below:

Table 1: *Variables Description*

U	The intensity of light (flux) with a wavelength of 3551Å (Angstrom) emitted by the object
G	The intensity of light (flux) with a wavelength of 4686Å emitted by the object
R	The intensity of light (flux) with a wavelength of 6166Å emitted by the object
I	The intensity of light (flux) with a wavelength of 7480Å emitted by the object
Z	The intensity of light (flux) with a wavelength of 8932Å emitted by the object
Redshift	Measurement of how fast the object is moving away relative to Earth. A result of Doppler's Effect: light emitted from an object moving away increases in wavelength and shifts to the red end of the light spectrum
Class	Classification of the object as star, galaxy, or quasar

Table 2: *First 10 Entries of Dataset*

	U	G	R	I	Z	Redshift	Class
1	19.47406	17.0424	15.94699	15.50342	15.22531	-8.96E-06	STAR
2	18.6628	17.21449	16.67637	16.48922	16.3915	-5.49E-05	STAR
3	19.38298	18.19169	17.47428	17.08732	16.80125	0.1231112	GALAXY
4	17.76536	16.60272	16.16116	15.98233	15.90438	-0.000110616	STAR
5	17.55025	16.26342	16.43869	16.55492	16.61326	0.000590357	STAR
6	19.43133	18.46779	18.16451	18.01475	18.04155	0.000314603	STAR
7	19.38322	17.88995	17.10537	16.66393	16.36955	0.1002423	GALAXY
8	18.97993	17.84496	17.38022	17.20673	17.07071	0.000314848	STAR
9	17.90616	16.97172	16.67541	16.53776	16.47596	8.91E-05	STAR
10	18.67249	17.71375	17.49362	17.28284	17.22644	0.04050813	GALAXY

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is the initial process of data investigation and probing to understand the data through summary statistics and graphical visualizations (Patil, 2018), and, using the file manager Anaconda, the libraries/packages of Matplotlib (Hunter, 2007), Pandas (McKinney, 2010), and Seaborn (Waskom et al., 2017) were installed for this purpose. The number of variables and their types, entries, and missing or null values were examined. Then, the mean, standard deviation, minimum, lower quartile, median, upper quantiles, and a maximum of the feature variables were calculated and organized in Table 3 below. Next, a heat map of variable correlations, a distribution plot, and a count plot of the class variables was created and discussed in the subsections below:

Table 3 : *Dataset Statistics*

	U	G	R	I	Z	Redshift
Mean	18.619355	17.371931	16.840963	16.583579	16.422833	0.143726
Std	0.828656	0.945457	1.067764	1.141805	1.203188	0.388774
Min	12.988970	12.799550	12.431600	11.947210	11.610410	-0.004136
25%	18.178035	16.815100	16.173333	15.853705	15.618285	0.000081
50%	18.853095	17.495135	16.858770	16.554985	16.389945	0.042591
75%	19.259232	18.010145	17.512675	17.258550	17.141447	0.092579
Max	19.599900	19.918970	24.802040	28.179630	22.833060	5.353854

3.2.1 Heat Map of Variable Correlations

Explanation and discussion of the heat map of variable correlations produced form EDA.

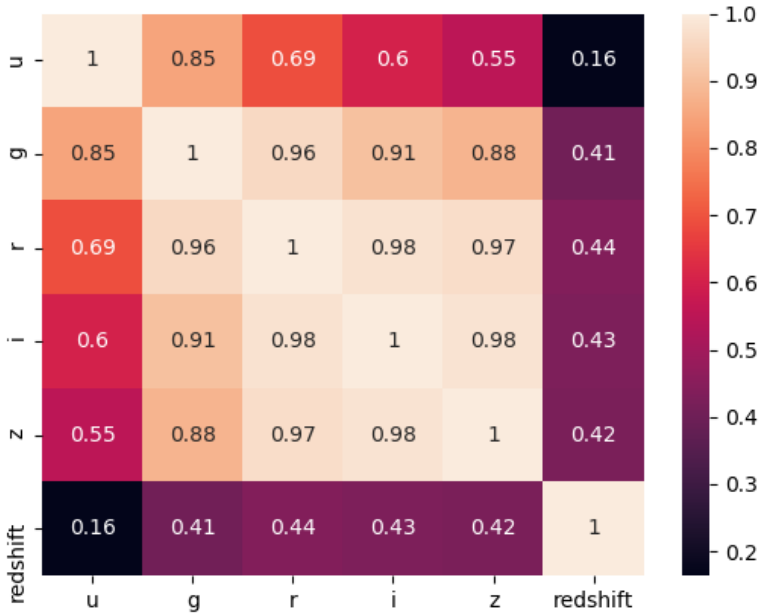

Figure 1*: Heat Map of Feature Variable Correlations*

Figure 1 is a heat map showing the correlation between the 6 features variables. The correlation strength, between -1 (perfect negative correlation) and 1 (perfect positive correlation), of any two features, is indicated by the color. While independent variables with high correlations with each other, creating causal relationships, can damage the accuracy and performance of models

and should be avoided, it is always important for a case-by-case analysis to be done for each high correlation to truly determine if it poses a danger to the performance of the model. As shown by the heat map, high correlations of around 0.9 or above were found between the features of u, g, r, i, and z. These features are extremely crucial information in classifying celestial objects from the photometric data and represent the five key color filters of the SDSS camera CCD system, with wavelengths covered from 3000 to 11000Å (Camera; The Photometric Camera and the CCDs; Understanding SDSS Imaging Data; Understanding the Imaging Data). Consequently, u, g, r, I, and z must be used together in classification to avoid critical information and accuracy loss (Fukugita et al., 1996).

3.2.2 Kernel Density Distribution Plot

Description and discussion of the kernel density distribution plot from EDA.

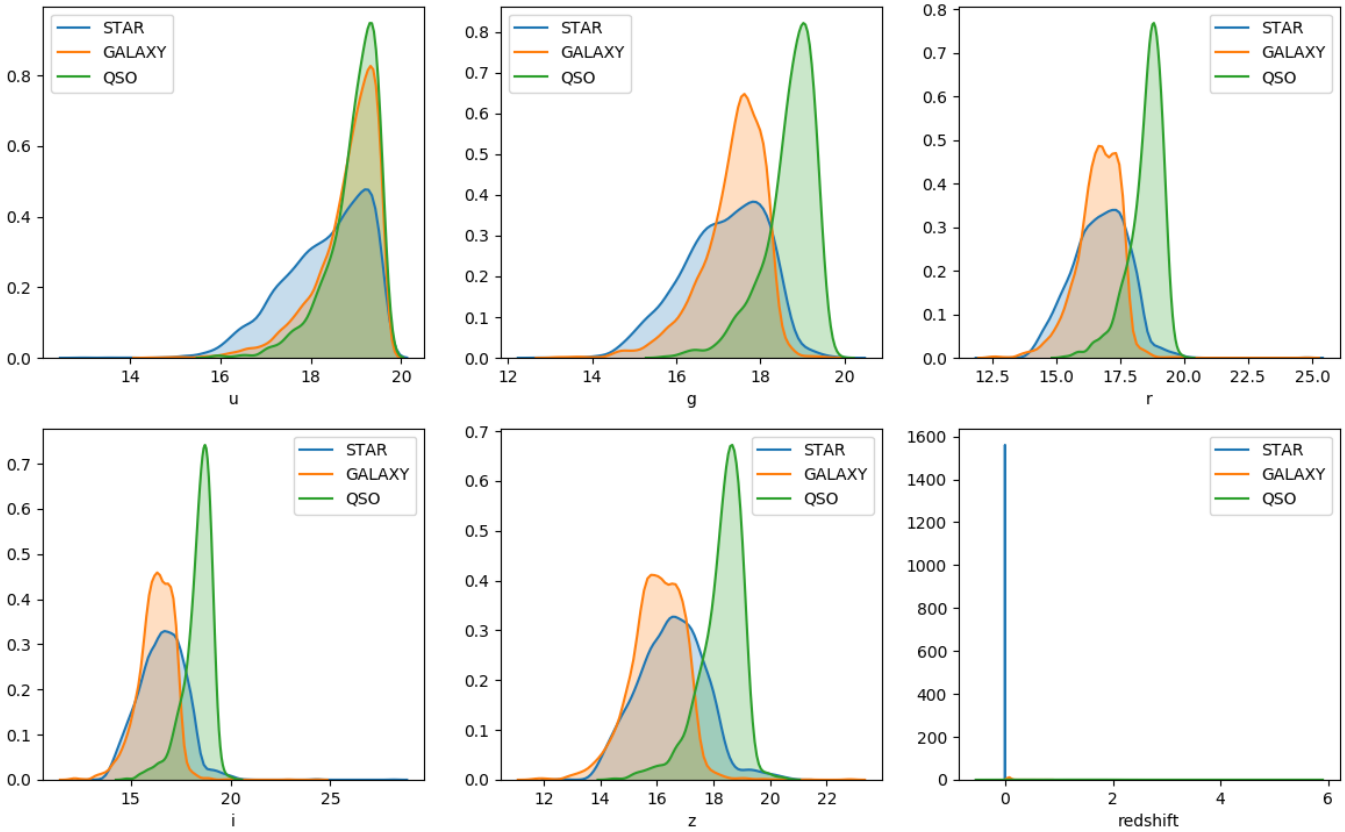

Figure 2: *Distribution Plot of Features, Color by Class*

Figure 2 present the distribution plots of each of the six features, which is further separated by the three classes. Since the distribution plots are kernel density plots, the total area of each density plot is 1, so the values on the y-axises don't represent actual counts but rather magnitudes

relative to the range of values on the x-axises. As shown by the density plots, the dataset also doesn't follow a normal distribution.

3.2.3 Count Plot

Explanation and discussion of the count plot of classes from EDA and their implications.

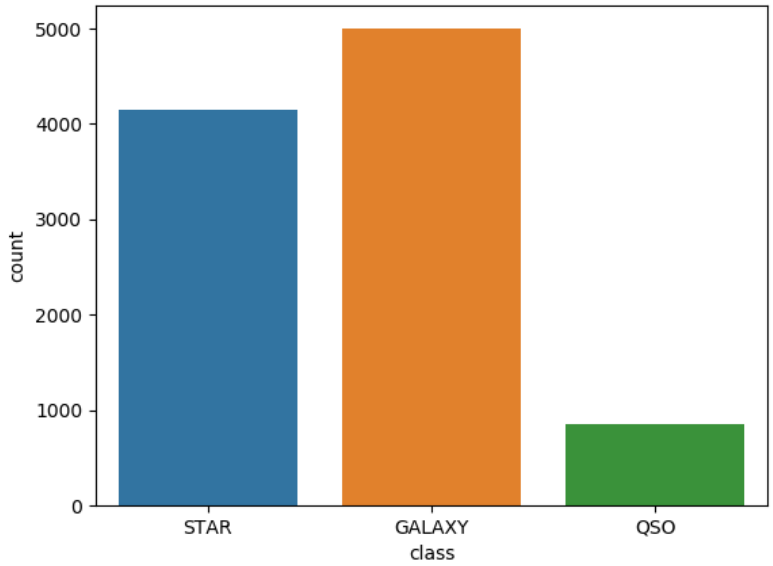

Figure 3: *Count Plot of Each Class*

Figure 3 shows the count plot of each class. The purpose of the count plot was to investigate whether the data entries for the three classes were relatively balanced. An imbalanced dataset, such as a three-class dataset with a ratio of 8:1:1, could cause a classification model, like logistic regression, to over-favor predictions of the class or classes with the overwhelming majority of the data, crippling the model's ability to classify others with the small minority (Chang et al., 2018). However, as shown by the figure, the ratio of data entries for stars: galaxies: quasars in the SDSS dataset is about 4.5: 5.5: 1, which is not considered to be a significant imbalance. Nevertheless, this possible problem of imbalance is further addressed in two ways. First, decision tree models perform very well with imbalanced training data, as they work by learning an impurity hierarchy of if/else questions formed from the features (Kat, 2019). This forces all three classes of stars, galaxies, and quasars to be equally favored in classification with similar performances (Logan & Fotopoulou, 2020). Second, the variety of performance metrics, including the precision, recall, and f-1 scores in the classifications reports; the precision-recall curves; and the confusions matrix, could comprehensively gauge the performance of the logistic regression and decision tree models, detecting any inaccuracies resulting from imbalanced data (Kohli, 2019). Therefore, the

high performances of both models, as discussed later, signalled that the possible data imbalance is not a problem.

3.3 Preprocessing Data

To preprocess and train models, the package of Scikit-learn (Pedregosa et al., 2017) was installed. During preprocessing, the dataset was standardized to ensure the performance of the models (Dorpe, 2018). Since the feature variables have different magnitude scales because of their different units of measurement, if they were not standardized, variables on a much larger scale would dominate over others, making the model unable to learn from those with smaller scales and consequently inaccurate. Since the variables are not normally distributed, the MinMax scaler was used to center the variables and bring them all up to the same scale using equation 1 (Dorpe, 2018):

$$\frac{x-min(x)}{max(x)-min(x)} \quad (1)$$

This process was repeated for every value of every feature variable. After standardization, the data were separated into two groups: the training sample with 80% of the data and the testing sample with 20%. The training sample would be used to fit or train the model, and the testing sample would be used to test the accuracy and performance of the model.

3.4 Machine Learning

The first machine learning model used is the logistic regression model. Since the dependent class variable is categorical, with three values or classes, the multinomial logistic regression model was used, which used the softmax function and cross-entropy loss function to predict multiple classifications (Chauhan, 2018). Then, a classification report, including precision, recall, f1-score, their macro and weighted averages, and accuracy, along with the cross-validation score were calculated. A confusion matrix was also created and visualized using a corresponding heat map.

The next machine learning model used is the decision tree model. A decision tree model is a series of nodes, conditions, that are connected by branches, which are the outcomes of the nodes' conditions. Starting at the root node, each node has two or more branches that lead more nodes or leaf nodes (the resulting predictions/classification). The conditions of the nodes in the decision tree were formed from the feature variables and their values, and their placements were determined by the Gini index, which measured the heterogeneity/impurity of the data resulting from the separation by the condition, the smallest of which was chosen for each node (Sanjeevi, 2018). The classification report and cross-validation score for the decision tree were also calculated. The tree visualization and the heat map of its confusion matrix were also created.

4. Results and Discussions

The performance metrics of the two machine learning models are discussed and compared.

4.1 Multinomial Logistic Regression

The classification report, cross-validation score, and confusion matrix heat map of the multinomial logistic regression model are described and explained.

4.1.1 Classification Report

The classification report of the multinomial logistic regression is displayed, and its significance is explained.

Table 4: *Classification Report of Logistic Regression Model*

	Pecision	**Recall**	**F1-Score**	**Support**
STAR	0.84	0.87	0.85	804
GALAXY	0.87	0.87	0.87	1018
QSO	0.98	0.85	0.91	178
Accuracy			0.87	2000
Macro Avg	0.90	0.86	0.88	2000
Weighted Avg	0.87	0.87	0.87	2000

ConvergenceWarning: lbfgs failed to converge (status=1):

STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Table 4 shows the classification report of the multinomial logistic regression. The classification report shows the main performance metrics of the machine learning model in predicting each class, including the precision, recall, f1-score, their macro and weighted averages, their counts, and the overall accuracy (Kohli, 2019):

True positives (TP): Predicted positive and actually positive.

False positives (FP): Predicted positive but actually negative.

True negatives (TN): Predicted negative and actually negative.

False negatives (FN): Predicted negative but actually positive.

- Accuracy - Percentage of true predictions:

$$\frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

- Precision - Percentage of predicted true positive predictions

$$\frac{TP}{TP+FP} \quad (3)$$

- Recall - Percentage of the positive cases that were predicted true:

$$\frac{TP}{TP+FN} \quad (4)$$

- F1-score - Weighted average of precision and recall; the best score of 1 and worst score of 0:

$$\frac{2}{\frac{1}{precision}+\frac{1}{recall}} = \frac{2*precision*recall}{precision+recall} \quad (5)$$

- 'macro avg' - Unweighted average of each metric. This did not account for class imbalance.
- 'weighted avg' - Weighted average of each metric. This accounted for class imbalance.

As shown by the classification report, all of the performance metrics, with 1 as the maximum and 0 as the minimum, are well above 0.84 and mostly around 0.9, which is a good performance from a machine learning model by any standards and indicates that the logistic regression model fitted the data well. It is also worth considering the difference between accuracy and f1-score, as shown by equations 2 and 5: while accuracy focuses on the true positives and true negative, the f1-score focuses on the false positive and false negative and is a better measure of the falsely classified cases than accuracy. Also, accuracy is generally fit for when the dataset is balanced, whereas the f1-score is a better metric for when the dataset is imbalanced. Since imbalance is a possible issue for the SDSS dataset, the f-1 score is a better metric than accuracy in this study (Huilgol, 2019).

Directly below the classification report is a corresponding convergence warning that appeared during the data analysis. This warning could indicate several different issues: the dataset was not properly scaled, the hyperparameter (like C) needed to be changed, or the default max iteration was too low. Since the dataset was already scaled by the MinMaxScaler previously, not properly scaling could not be the problem. The hyperparameters or the max iteration could be further adjusted, but the fact that max iteration was too low, meaning the logistic regression model needed more iterations to converge, indicates that the multinomial logistic regression model is not optimal for the dataset.

4.1.2 Cross-Validation

Cross-validation is a method of model validation that splits the data in creative ways to obtain better estimates of the "real world" model performance and to minimize validation error. It tests whether the model is overfitted to the specific training and testing samples, which could result

in bad performance with any other dataset that the model has not seen yet (Shaikh, 2018). The k-fold cross-validation method, as shown below in Figure 4, was specifically used:

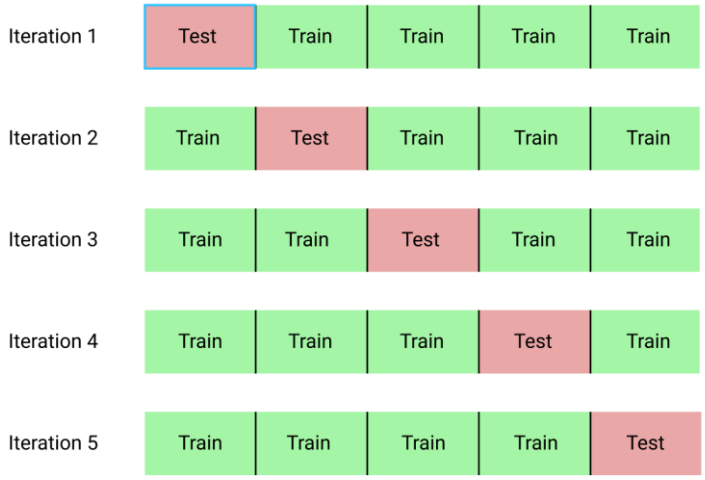

Figure 4: *k-Fold Cross-Validation Method, k=5*

The k parameter in this method determined the number of groups that a data sample was to be split into. In this case, k=5, so the dataset would be correspondingly split into 5 equal parts and the below process would run 5 times, each time with a different test data set (Shaikh, 2018):

1. Choose one of the groups as the test data set
2. Use the remaining groups as the training data set
3. Train a model on the training set and evaluate its accuracy on the test set
4. Retain the accuracy score and discard the model

The score of the k-fold cross-validation for the multinomial logistic regression model was 0.8664, representing the accuracy of the model as the average of the accuracy of each fold or iteration, which indicates that the model would perform/generalize well with other datasets with a low validation error.

4.1.3 Confusion Matrix Heat Map

The confusion matrix heat map of the logistic regression is displayed and explained.

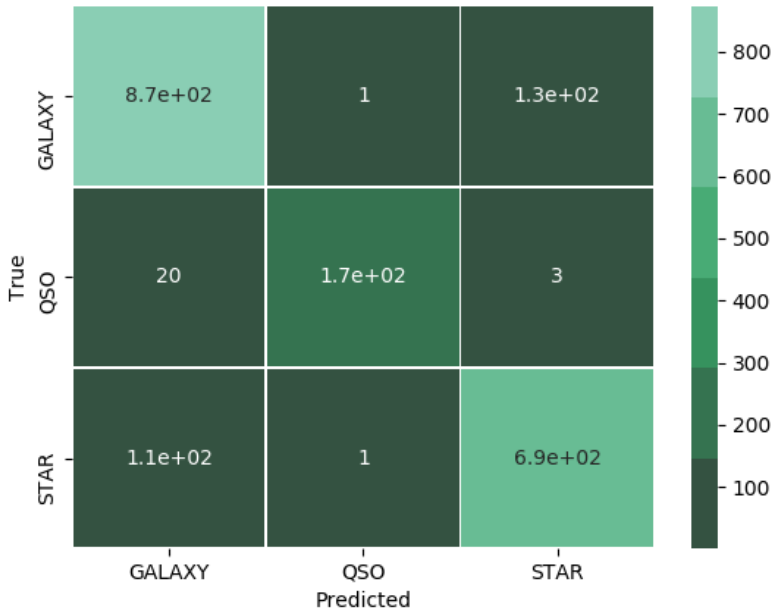

Figure 5: *Heat Map of Confusion Matrix of Logistic Regression Model*

Figure 5 shows the heat map of the multinomial logistic regression model's confusion matrix. A confusion matrix presents the count of the true positives, false positives, true negatives, and false negatives of each class. The three squares at the top left, the center, and the bottom right show the counts, as indicated by the color intensity, of the true positives of the classes, with the rest of the squares showing the counts of false classifications. Overall, the majority, around eighty-five to ninety, of the cases in each class were correctly classified, further showing the model's performance is fairly well. However, upon further inspection, it can be seen that each of those ten to fifteen percent of wrong classifications for each class was centered around one particular predicted class. This centering of wrong classifications around one wrong label could indicate intrinsic problems or internal unfitness of the multinomial logistic regression model.

4.2 Decision Tree

The visualization, classification report, cross-validation score, and confusion matrix heat map of the decision tree model are described and explained.

4.2.1 Tree Model Visualization

The visual representation of the decision tree model is shown and its features are described.

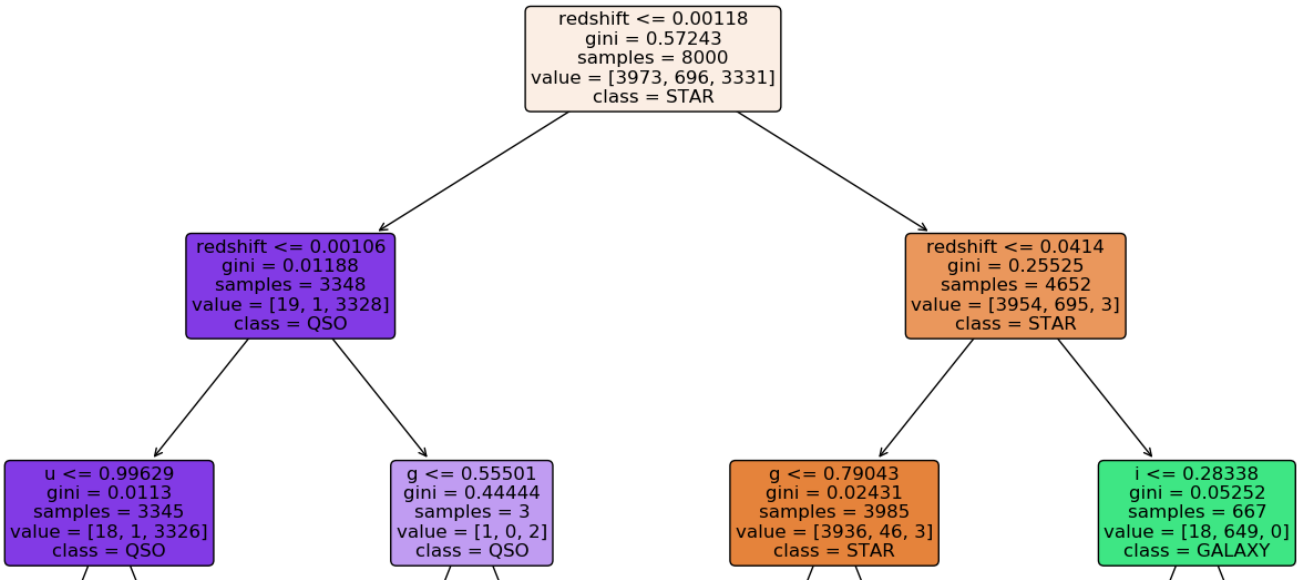

Figure 6: *First Two Levels of Decision Tree Model*

Figure 6 is a visualization of the first two levels of the decision tree model. In the visualization, there are five pieces of information for every node. The first piece of information is a condition or question. The structure of the condition consists of selecting a feature variable and choosing a threshold value using the Gini index method. The incoming objects will be separated by the comparison of their corresponding feature variable value to the threshold value. The next piece of information is the Gini impurity index. The Gini index represents the weighted impurity, heterogeneity, of the training data after it is split by the current node's condition (Ramadhan et al., 2017), with 0 denoting that all objects in each split are one class and 1 denoting that the objects are randomly distributed. This is the method used to determine which feature variable and threshold are used for each node position by order of smallest Gini index (more desirable) to largest. While the overall Gini index of the entire decision tree decreases with each successive node split, the Gini index for each node can both increase and decrease as compared to the previous node depending on its sample size and value distributions (Lee, 2016). The third piece of information shows the sample size of the data that passes through the node during the training process. The fourth piece of information shows the counts of the classes in the training data sample that passes through the node. The fifth piece of information is the supposed classification of the node, based on the class that has the majority count in the fourth piece. Each classification has a corresponding color: orange for stars, green for galaxies, and purple for quasars, with the color intensity matching the majority percentage. Leaf nodes, which are on the bottom ends of the tree,

have no conditions, a definite Gini index of 0, zeroes for two of the three classes, and the fifth piece with a definite classification and highest color intensity.

4.2.2 Classification Report

The classification report of the decision tree model is displayed. Its scores and metrics are described and compared to those in the classification report of the logic regression.

Table 5 : *Classification Report of Decision Tree Model*

	Precision	**Recall**	**F1-Score**	**Support**
STAR	1.00	0.99	1.00	856
GALAXY	0.99	0.99	0.99	987
QSO	0.95	0.93	0.94	157
Accuracy			0.99	2000
Macro Avg	0.98	0.97	0.98	2000
Weighted Avg	0.99	0.99	0.99	2000

Table 5 is the classification report for the decision tree model. All of the metrics in the report are the same as those for the logistic regression model to more effectively compare the two models and evince the superior one. When comparing the two classification reports, the decision tree model's performance can be seen to be significantly better than that of the multinomial logistic regression model. The precision, recall, f-1 score, their macro, and weighted averages, and accuracy of the decision tree model are mostly within proximity to the highest value of 1, signalling that the decision tree model is far superior to the logistic regression in its performance of classifying stars, galaxies, and quasars.

4.2.3 Cross-Validation

Often with performance metrics that are extremely close to or are equal to 1, a model may be suspected of being over-fitted to the data, which is why the cross-validation accuracy score is extremely critical in this case to mimic the model's possible ʻreal-world ʼperformance on unseen data and to ensure that it has actual significance (Shaikh, 2018). The cross-validation score of the decision tree model results is 0.9858, which is even significantly higher than the logistic regression model's cross-validation score of 0.8664. With such a high cross-validation score, the decision tree model can be confidently concluded to be extremely well fitted to the data and that its high

performance with the testing data will translate smoothly to unseen data.

4.2.4 Confusion Matrix Heat Map

The confusion matrix heat map of the decision tree is displayed and compared to that of the multinomial logistic regressions.

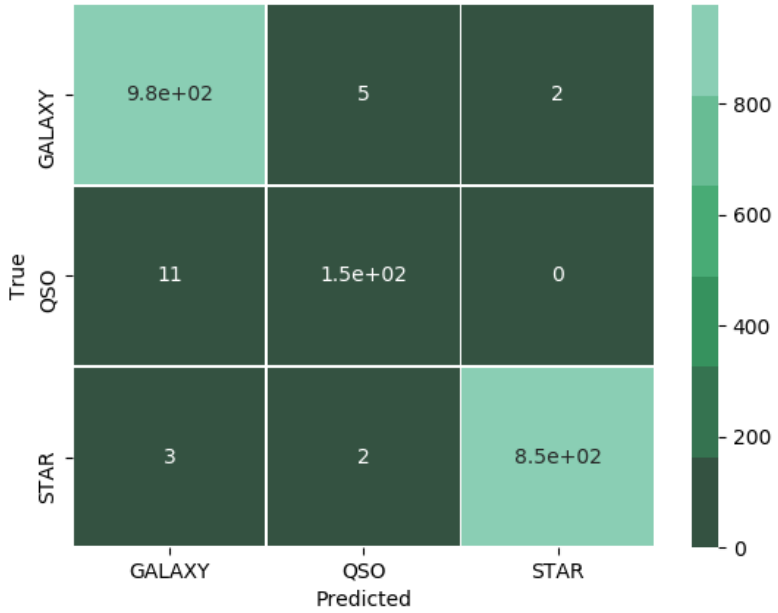

Figure 7: *Heat Map of Confusion Matrix of Decision Tree Model*

Figure 7 shows the heat map of the decision tree model's confusion matrix. When compared to that of the logistic regression model, the confusion matrix for the decision tree has both significantly less false positive and false negatives for all three classes, which has already been evinced by the higher scores on the classification reports. Furthermore, the decision tree also solved a noticeable problem with logistic regression. As previously discussed, the false classifications of the logistic regression were all primarily concentrated around a predicted class. The decision tree, on the other hand, mostly solved this issue by having fairly balanced false classifications and small false classification sizes so that their distribution becomes insignificant even if the wrong classifications were concentrated like those of the quasar class.

4.3 Research Limitations and Future Improvements

One of this study's limitations is that while there are far more than three types of light-emitting celestial objects in the universe, the final decision tree classification model was trained on data from only stars, galaxies, and quasar and thus can only classify stars, galaxies, and quasars. This means that if a real-world dataset contains data from foreign objects, like comets, large gas clouds, etc, the model would always wrongly classify them as either stars, galaxies, or quasars.

A second limitation, and a future improvement at the same time, lies in training more classification models. Aside from multinomial logistic regression and decision trees, there is still a wide variety of classification models and variations, such as the random forest, support vector machine, or neural network, that are not used in this study due to time constraints. Although the decision tree model's performance in classifying stars, galaxies, and quasars was high and seemed to be the most fitting model, there is still room to improve. Each model has its strengths and weaknesses and utilizing more models can enable a much fuller understanding of their respective strengths and weaknesses, thus lead to a model that fits the data even better.

Another improvement is to solve the logistic regression model's convergence warning. Although the warning indicates that logistic regression is possibly just not fit for the dataset, there is still a possibility that, by adjusting the hyperparameter and the max iterations so that the warning is resolved, the model's performance can be significantly improved, leading to a much more meaningful and interesting comparison with the decision tree model.

5. Conclusion

In this study, machine learning models of multinomial logistic regression and decision trees were built and trained using SDSS data to classify stars, galaxies, and quasars. While other studies similarly used data from SDSS to classify stars, galaxies, and quasars, this study compared the performance between two different machine learning models while other studies often only use one classifier of mostly random forest (Yoong 2018; Carrasco et al., 2015). The data used are from the Sloan Digital Sky Survey's Data Release 14. The dataset consists of 17 feature variables and 1 class variable. Of the 17 feature variables, only 6 are relevant to the classification of celestial objects and the rest were removed. After EDA, the dataset was subsequently scaled using the Min-Max scaler and split into training and testing samples. The logistic regression model and the decision tree model were subsequently fitted to the training data and tested by a variety of performance metrics using the testing sample. By analyzing and comparing the variety of performance metric results for the multinomial logistic regression and decision tree models, it can be concluded that the decision tree model is significantly superior to the logistic regression model and is an extremely well fit for classifying stars, galaxies, and quasars. Therefore, my hypothesis of being able to successfully train a machine learning model for the classification of celestial objects using SDSS data is proven to be correct. The decision tree model created in this study,

which has accuracies all within proximity to one hundred percent, can be used as a competent and reliable classification tool in all astronomical purposes with the need for classifying stars, galaxies, and quasars, potentially being able to allow astronomers to increase the sample sizes of stars, galaxies and particularly quasars with great speed and accuracy.

Acknowledgement

I wish to express my deepest appreciation and gratitude to Dr. Qin Taylor. This research study could not have been possible without her guidance and mentoring. She responded to every single one of my questions with an answer of depth and clarity and maintained a strong and reciprocal process of research, feedback, and suggestions. I absolutely cannot thank her enough for how open she was through the entire process to my input and thoughts and working with me.

I would also like to take this chance to thank Mrs. Betty. She helped me to get in touch with Dr. Taylor has also helped me countless times since I started this journey of big data analysis more than a year ago.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Abolfathi, B., Aguado, D. S., Aguilar, G., Allende Prieto, C., Almeida, A., Ananna, T. T., … Zou, H. (2018, April). *The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/2018ApJS..235...42A/abstract

Blanton, M. R., Bershady, M. A., Abolfathi, B., Albareti, F. D., Allende Prieto, C., Almeida, A., … Zou, H. (2017, July). *Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/2017AJ....154...28B/abstract

Carrasco, D., Barrientos, L. F., Pichara, K., Anguita, T., Murphy, D. N. A., Gilbank, D. G., … López, S. (2015, August 24). Photometric classification of quasars from RCS-2 using Random Forest⋆ . Astronomy & Astrophysics. https://doi.org/10.1051/00046361/201525752

Chang, M., Dalpatadu, R. J., Phanord, D., & Singh, A. K. (2018). A Bootstrap Approach For Improving Logistic Regression Performance In Imbalanced Data Sets. *MATTER: International Journal of Science and Technology*, *4*(3), 11–24. https://doi.org/10.20319/mijst.2018.43.1124

Chauhan, G. (2018, October 10). *All about Logistic regression*. Medium. https://towardsdatascience.com/logistic-regression-b0af09cdb8ad

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. (2020, May 21). *Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra*. arXiv.org. https://doi.org/10.1051/0004-6361/201936770

Doi, M., Tanaka, M., Fukugita, M., Gunn, J. E., Yasuda, N., Ivezić, Ž., … French Leger, R. (2010, April). *Photometric Response Functions of the Sloan Digital Sky Survey Imager*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/2010AJ....139.1628D/abstract

Dorpe, S. V. (2018, December 13). *Preprocessing with sklearn: a complete and comprehensive guide*. Medium. https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. (1996, April). *The Sloan Digital Sky Survey Photometric System*. NASA/ADS.

https://ui.adsabs.harvard.edu/abs/1996AJ....111.1748F/abstract
https://doi.org/10.1086/117915

Gunn, J. E., Carr, M., Rockosi, C., Sekiguchi, M., Berry, K., Elms, B., … Brinkman, J. (1998, December). *The Sloan Digital Sky Survey Photometric Camera*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/1998AJ....116.3040G/abstract

Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., … Wang, S.-i. (2006, April). *The 2.5 m Telescope of the Sloan Digital Sky Survey*. NASA/ADS. https://ui.adsabs.harvard.edu/abs/2006AJ....131.2332G/abstract

Huilgol, P. (2019, August 24). *Accuracy vs. F1-Score*. Medium. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kat, S. (2019, August 10). *Logistic Regression vs. Decision Tree - DZone Big Data*. dzone.com. https://dzone.com/articles/logistic-regression-vs-decision-tree

Kohli, S. (2019, November 18). *Understanding a Classification Report For Your Machine Learning Model*. Medium. https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Lee, C. (2016, October 3). *Logistic Regression versus Decision Trees*. The Official Blog of BigML.com.https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/

Logan, C. H. A., & Fotopoulou, S. (2020, January 23). *Unsupervised star, galaxy, QSO classification - Application of HDBSCAN*. Astronomy & Astrophysics. https://www.aanda.org/articles/aa/full_html/2020/01/aa36648-19/aa36648-19.html

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). K https://doi.org/10.25080/Majora-92bf1922-00a

Patil, P. (2018, May 23). *What is Exploratory Data Analysis?* Medium. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Ramadhan, M. M., Sitanggang, I. S., & Anzani, L. P. (2017). Classification Model For Hotspot Sequences As Indicator For Peatland Fires Using Data Mining Approach. *MATTER:*

International Journal of Science and Technology, *3*(2), 588–597. https://doi.org/10.20319/mijst.2017.32.588597

Redd, N. T. (2018, February 24). *Quasars: Brightest Objects in the Universe*. Space.com. https://www.space.com/17262-quasar-definition.html

SDSS SkyServer . *Galaxies and QSOs*. SDSS SkyServer DR14. http://skyserver.sdss.org/dr14/en/astro/galaxies/galaxies.aspx

SDSS SkyServer. *Redshifts*. SDSS SkyServer DR12. https://skyserver.sdss.org/dr12/en/proj/advanced/hubble/redshifts.aspx

SDSS SkyServer. *Stars and Nebulae*. SDSS SkyServer DR14. http://skyserver.sdss.org/dr14/en/astro/stars/stars.aspx

SDSS. *Camera*. SDSS. https://www.sdss.org/instruments/camera/

SDSS. *Measures of Flux And Magnitude*. SDSS. http://www.sdss3.org/dr8/algorithms/magnitudes.php

SDSS. *Understanding SDSS Imaging Data*. SDSS. http://www.sdss3.org/dr9/imaging/imaging_basics.php

Shaikh, R. (2018, November 26). *Cross Validation Explained: Evaluating estimator performance.* Medium.https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

Viquar, M., Basak, S., Dasgupta, A., Agrawal, S., & Saha, S. (2018, April). *Machine Learning in Astronomy: A Case Study in Quasar-Star Classification*. ResearchGate. https://www.researchgate.net/publication/324536351_Machine_Learning_in_Astronomy_A_Case_Study_in_Quasar-Star_Classification. https://doi.org/10.1007/978-981-13-1501-5_72

Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

Yoong, T. (2018, December 15). *Predicting Stars, Galaxies & Quasars with Random Forest Classifiers in Python*. Medium. https://towardsdatascience.com/predicting-stars-galaxies-quasars-with-random-forest-classifiers-in-python-edb127878e43

Zhang, Y., Zhao, Y., Zheng, H., & Wu, X. (2013, January). *Classification of Quasars and Stars by Supervised and Unsupervised Methods*. 2013IAUS..288..333Z Page 333. https://doi.org/10.1017/S1743921312017176