Classifying Astronomical Objects

Evan Harrigan
Computer Science
Georgia State University
Atlanta, GA USA
eharrigan1@student.gsu.edu

***Abstract*— In this paper, I discuss classification techniques in the application of classifying astronomical objects (Stars, Galaxies, and Quasars). I will apply Support Vector Machines, Random Forest, and Artificial Neural Network classifiers to data relating to the before mentioned astronomical objects. I will also discuss the process of cleaning and transforming the data into an acceptable state for the classifiers. These classifiers will achieve accuracies ranging from 97-98%, with Random Forest achieving the highest accuracy. I will also compare other metrics of classifier performance such as precision, recall, and f-1 score. Taking all these metrics into consideration, I will conclude by stating and proving that Random Forest classification is the most acceptable model for classifying the data.**

***Keywords*—- Machine learning, Astronomy, Support Vector Machine, Random Forest, Artificial Neural Network**

I. Introduction

Understanding the composition of the universe is key for understanding what it is made of and how it formed. If we can successfully map out a large proportion of the universe, we can discover groups or clusters of specific objects. Having knowledge of the composition of the universe may also help deepen our understanding of physics and allow us to discover new / rethink prior theories. If any type of statistical analysis is wanted to be performed on these classes of astronomical objects, a large sample size is required. Thus, it is helpful to have a way to quickly classify and catalog objects in space for future reference.

The code for this study can be acquired at https://github.com/harriganevan/StellarClassification

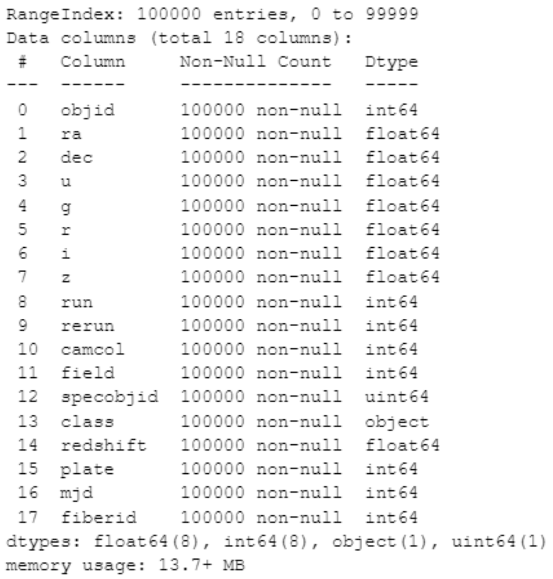
II. Theory

The theory to be tested throughout this study is that there is a way to accurately and efficiently classify astronomical objects using spectral and photometric data.

III. Materials and Methods

Data explanation and characterization

The data being used in this study is acquired from the Sloan Digital Sky Survey [1] Data Release 18. The Sloan Digital Sky Survey captures data from visible astronomical objects across one-third of the sky. The dataset used in this study consists of 100,000 entries described by 18 attributes. Among these attributes are redshift - a measure of how fast an object is moving away relative to Earth, and wavelength filter values ranging from ultraviolet to infrared - a measurement of how much light of a specific wavelength an object emits [2]. The dataset also includes additional photometric attributes.



```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 18 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   objid      100000 non-null  int64
 1   ra         100000 non-null  float64
 2   dec        100000 non-null  float64
 3   u          100000 non-null  float64
 4   g          100000 non-null  float64
 5   r          100000 non-null  float64
 6   i          100000 non-null  float64
 7   z          100000 non-null  float64
 8   run        100000 non-null  int64
 9   rerun      100000 non-null  int64
 10  camcol     100000 non-null  int64
 11  field      100000 non-null  int64
 12  specobjid  100000 non-null  uint64
 13  class      100000 non-null  object
 14  redshift   100000 non-null  float64
 15  plate      100000 non-null  int64
 16  mjd        100000 non-null  int64
 17  fiberid    100000 non-null  int64
dtypes: float64(8), int64(8), object(1), uint64(1)
memory usage: 13.7+ MB
```

Fig. 1 An overview of all attributes used in the dataset.

Data Preprocessing

Data preprocessing is essential to ensuring accurate results from a model. The first step in my preprocessing process was to check for null values and duplicates. Simple calls on a Pandas DataFrame object can be used to return the number of null values and duplicates present. Such calls were made and it was determined that there were no null values nor duplicates. All attributes are either float or integer data types. This means that there was no need to do any encoding. The next step in the preprocessing process was removing unnecessary attributes. The 'rerun' attribute has the same value for all entries (301). Therefore, it provides no information and can be removed from the dataset. Certain IDs (objid and specobjid) are unique identifiers used to label each object. These IDs provide no information and produce noise which could lead to overfitting. For this reason they are removed from the dataset. The remaining 15 attributes will remain in the dataset and used by the classification algorithms.

The next step is outlier detection / removal. Since the data was of low dimensionality, I opted to view boxplot distributions of the attributes and detect outliers using an intuitive approach, opposed to removing or replacing all data points that fell outside of a predetermined threshold (1.5 IQR). Fig. 2 shows an example of an obvious outlier within the *g* attribute. This entry had the same outlier value (-9999) in *u, g,* and *z*. Since this entry had obvious outlier values in 3 of the attributes, I decided to drop this entry. This will have negligible effect since the sample size is large. Fig. 3 shows the *g* attribute after removing the outlier.

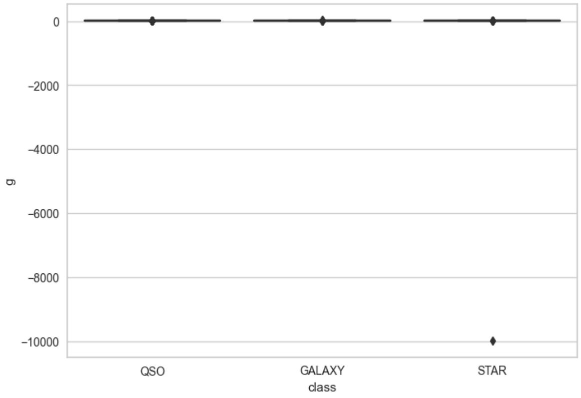


Fig. 2 A boxplot of *g* (a wavelength filter value corresponding to ultraviolet). Shows a clear outlier in a 'STAR' class object.

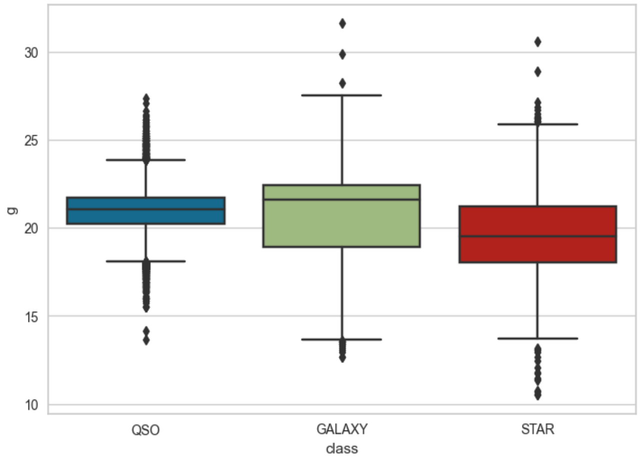


Fig. 3 A boxplot of *g* after removing the outlier in Fig. 2.

After removing the outlier I ran the same boxplots and concluded that there were no necessary outliers to remove anymore.

The original data used is imbalanced. Meaning there are not equal amounts of entries for each class. This imbalance is not of a large factor, with the largest imbalance being 3:1 between Galaxies and Quasars respectively. This likely won't have a large effect on the models performance but in order to ensure that model performance metrics (such as accuracy) are correct, it is important to balance the data. This was accomplished using SMOTE, Synthetic Minority Oversampling Technique. This oversampling generates synthetic samples for the minority classes (Stars and Quasars).

Principal Component Analysis is a dimensionality reduction technique used with the purpose of reducing the dimensions of a dataset while retaining as much variance as possible.

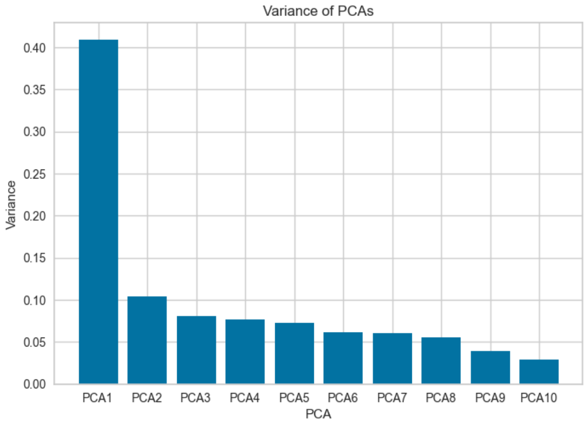


Fig. 4 A graph of the amount of variance each PCA retains (up to 10 PCA's)

Fig. 4 shows how much variance each PCA will retain. Using 5 PCAs will retain 75% of the original variance. In order to retain 98%, 10 PCAs must be used. Tests were done to see the effect of using PCAs on the accuracy of the models. Using 5 PCAs as input reduced the average accuracy of the models by ~25% compared to using the original data dimensions. This was an unacceptable amount so a test using 10 PCAs was done. 10 PCAs reduced the average accuracy by ~5%. Although this is significantly better than 5 PCAs, I determined that this too was unacceptable, especially considering that using 10 PCAs only reduces the dimensionality of the data by 5, hardly speeding up the model training time. Because of the unacceptable PCA results, I opted to not do PCA and instead stick to using the original data.

Data analysis

Three classification algorithms were tested and compared: support vector machine, random forest, and artificial neural network. Each algorithm used a 70/30 training-testing split. The support vector machine uses the rbf kernel. The random forest uses the entropy function to measure the quality of the split and considers log2 features when looking for the best split. The artificial neural network uses the logistic activation function and has one hidden layer of 9 nodes. Fig 5. describes the shape of the neural network used. [3]

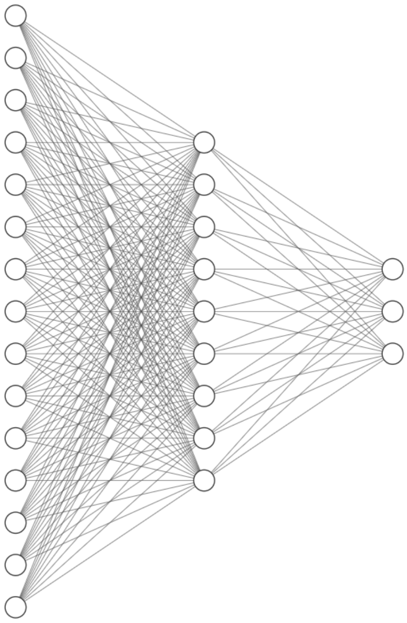


Fig. 5 Shape of neural network.

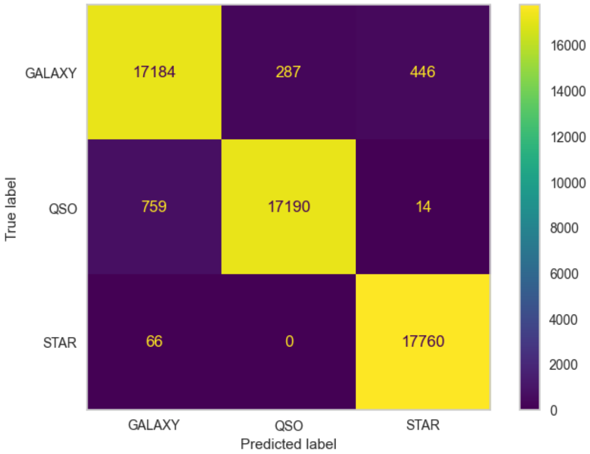
Evaluation and interpretation

The models were compared using three different metrics: Confusion matrices, ROC curves, and classification reports (precision, f-1, recall).

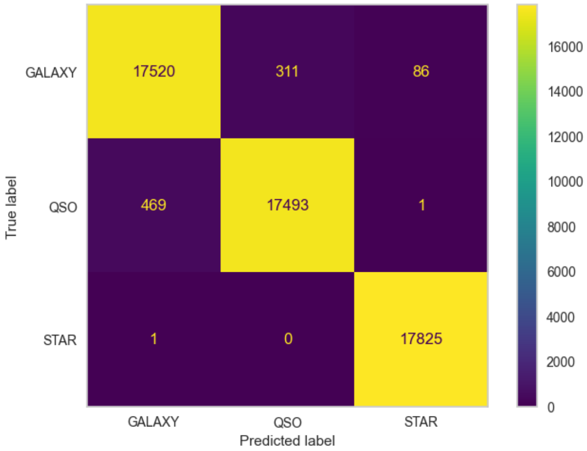
IV. Results

Confusion matrices:

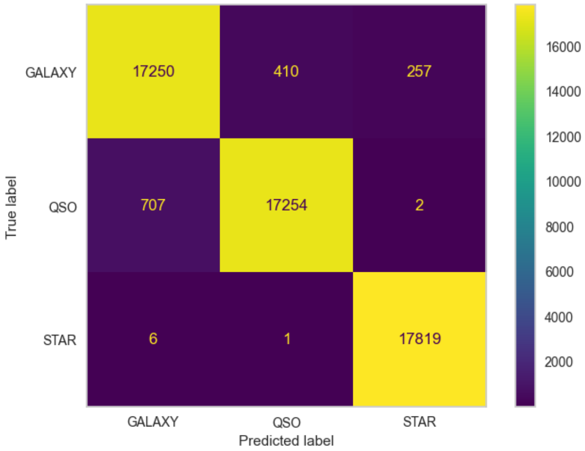
Support vector machine:



Random forest:

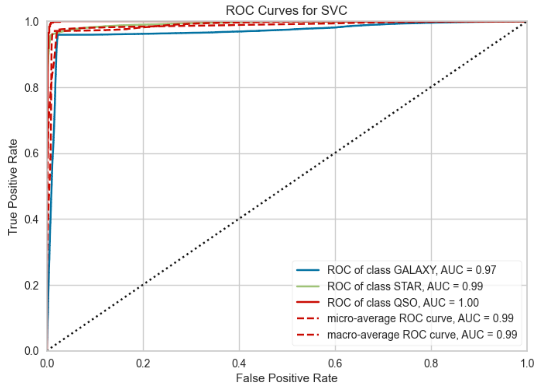


Artificial neural network:

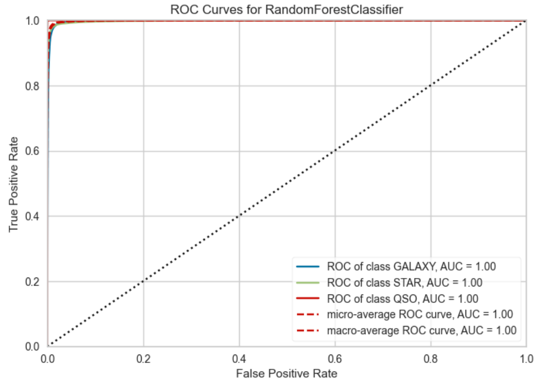


ROC curves:

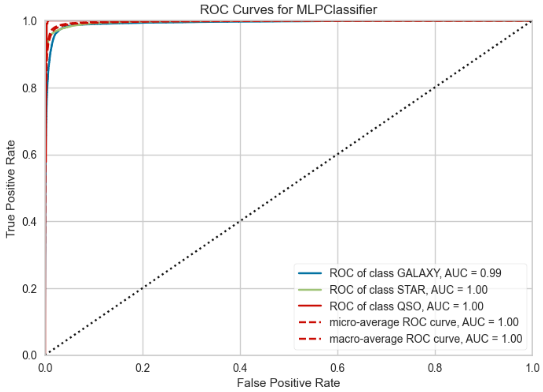
Support vector machine:



Random forest:

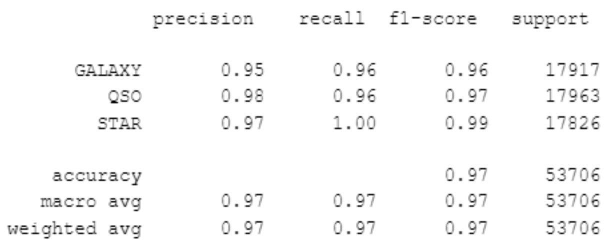


Artificial neural network:



Classification reports:

Support vector machine:

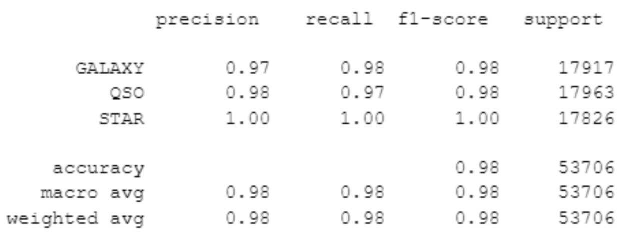


```
              precision    recall  f1-score   support

      GALAXY       0.95      0.96      0.96     17917
         QSO       0.98      0.96      0.97     17963
        STAR       0.97      1.00      0.99     17826

    accuracy                           0.97     53706
   macro avg       0.97      0.97      0.97     53706
weighted avg       0.97      0.97      0.97     53706
```

Random forest:

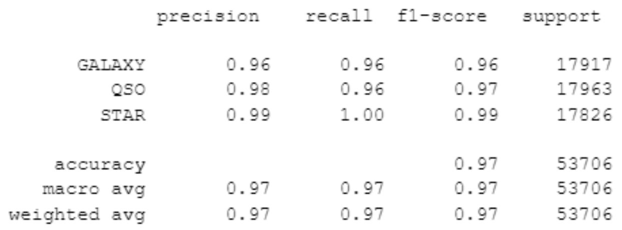


```
              precision    recall  f1-score   support

      GALAXY       0.97      0.98      0.98     17917
         QSO       0.98      0.97      0.98     17963
        STAR       1.00      1.00      1.00     17826

    accuracy                           0.98     53706
   macro avg       0.98      0.98      0.98     53706
weighted avg       0.98      0.98      0.98     53706
```

Artificial neural network:



```
              precision    recall  f1-score   support

      GALAXY       0.96      0.96      0.96     17917
         QSO       0.98      0.96      0.97     17963
        STAR       0.99      1.00      0.99     17826

    accuracy                           0.97     53706
   macro avg       0.97      0.97      0.97     53706
weighted avg       0.97      0.97      0.97     53706
```

V. Discussion and Conclusion

Analyzing the confusion matrices we can see the predicted and actual class labels for each classifier when using the testing split of the dataset. We can see that each model was very good at classifying stars correctly but struggled minorly differentiating between galaxies and quasars. Just by looking at the confusion matrices, we can see that random forest provided the best fit for the data. We will now look at other classifier metrics to see if that claim holds true. An ROC curve is the plot of the true positive rate against the false positive rate at a variety of thresholds. Taking the area under the curve (AUC) will provide an idea of how well the model is capable of distinguishing between classes. Analyzing these curves we can see that the random forest ROC curve has the highest AUC, again showing that it is the best model for the data. Lastly we will examine some classification reports which provide precision, recall, and f-1 scores for each classifier. (ex. Precision is a measurement about how precise a model is out of all positive predictions. Looking at the random forest confusion matrix, we see that galaxies have a true positive of 17520. It has a false positive of 469 + 1. Precision is calculated as:

$$\frac{True\ positive}{True\ positive + False\ positive} = \frac{17520}{17990} = .974)$$

The other metrics are calculated in different ways but they all provide some idea of how well the model performs. Looking at the macro and weighted averages of each report, we see that random forest has the highest values at .98, while the other classifiers have values of .97. Comparing all three of these metrics (confusion matrices, ROC curves, and classification reports), we can see that random forest provides the most ideal model for classifying our data.

Each model has the opportunity to increase its ability to fit the data. There are many parameters for each model that will affect its performance. For example, the shape of the neural network can be refined to optimize the accuracy, support vector machines can use different kernels, and random forest can set splitting criterion and more. There are a lot of parameters that need to be taken into consideration when tuning a model. It is possible that there exists a set of parameters for one or more of these models that when used will produce better results than what is shown in this study.

References

[1] "SkyServer," *Home - Skyserver SDSS*. [Online]. Available: https://skyserver.sdss.org/dr18. [Accessed: 27-Apr-2023].

[2] "A comparison of the passband (AKA transmission functions, AKA response curves) of the SDSS ugriz filters of the Sloan Digital Sky Survey (SDSS) and CFHTLS ugriz filters of the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS).," *UNLV Physics & Astronomy*. [Online]. Available: https://www.physics.unlv.edu/~jeffery/astro/photometry/photometry_sdss.html. [Accessed: 27-Apr-2023].

[3] "NN-svg," *NN SVG*. [Online]. Available: https://alexlenail.me/NN-SVG/. [Accessed: 27-Apr-2023].