Stellar Classification

Stars, Galaxies and Quasars

Aaron Gupta
School of Mathematical and Computational Sciences
University of Prince Edward Island
Charlottetown, Canada
usgupta@upei.ca

***Abstract*—It is very important in Astrophysics to classify stellar objects. When astrophysicists use telescopes to spot a new object in the sky, they must try to place the object in one of the categories of known object types - such as Stars, Galaxies and Quasars. This paper attempts to conduct such a classification using Machine Learning to see if scientists can hand off this step to an algorithm and continue their research.**

***Index Terms*—astrophysics, classification, stellar objects**

I. Introduction

In this paper, K-Nearest Neighbors (K-NN) Classification and Linear Support Vector Machine (SVM-linear) Classification were used to classify stellar objects based on the red and near-infrared photometric filter values and the redshift values.

The red photometric filter values represent the frequency of light received from the stellar object in the red frequency, which is between 429-462THz. The near-infrared photometric filter values represent the frequency of light received from the stellar object in the near-infrared frequency, which is between 350-429THz. The redshift values represent the amount of shift in frequency of light, which relates to the distance of the object from the observatory. The further away the object, the more the redshift.

First, the data was split into training and testing sets. Then, it was scaled using SciKitLearn's (SKLearn) standard scaler. Next, a K-NN Classifier was trained on the training set, tested on the testing set, tuned, analyzed, and cross-validated. Finally, an SVM-linear Classifier was trained on the training set, tested on the testing set, analyzed, and cross-validated.

The paper begins with a background on why the dataset was chosen and why the specific classification algorithms were chosen. In section III, the distributions and relationships of features in the data are presented and analyzed. Section IV covers the results, parameter tuning, and analysis of K-NN Classification and SVM-linear Classification. Finally, section V has a conclusion and discussion of future research.

II. Background

When a telescope is pointed at a stellar object, the frequencies of bands of light are recorded by the computer in what is known as a spectrograph.[2] Light from different kinds of stars has different distributions of frequencies, each specific to the type, size, temperature, atmospheric composition and distance of the star.[2] Similarly, information about the type, size and distance of galaxies and Quasars can be obtained.[2] This information was present in the dataset chosen, hence the dataset was a good choice for the problem.

The important values to consider in my dataset were the red photometric filter, the near-infrared photometric filter and the redshift of each object. These values provide important information regarding the aforementioned properties of stellar objects, and as will be shown have distinct values over the different classes.

Since the problem in question is a labeled classification problem, K-NN Classification and SVM-linear Classification were the primary choices of algorithms. K-NN is very simple and quick to implement, with only one parameter (the number of neighbors) to be tuned. SVM-linear is a bit more complex and takes longer, with the SVM-poly and SVM-rbf variations taking even longer. SVM-linear has no parameters to be tuned.

III. The Data

The dataset contains n = 100,000 labeled samples, with three possible labels: STAR, GALAXY, and QSO (Quasar Stellar Object). Each sample has m = 3 features: the red photometric filter value, the near-infrared photometric filter value and the redshift value.

A. Feature Distribution

The distribution of features as present were plotted in Fig. 1. The red and the near-infrared filter values show a similar distribution, whereas the redshift values follow their own. From this, not much was inferred.

B. Feature Relationships

The relationships between pairs of features were then plotted in Fig. 2.

- The red and the near-infrared filter values show close to a linear relationship, suggesting that these vary similarly across the dataset.
- The redshift and red filter values have a scattered relationship showing that the red filter values have a high variance when the redshift is low and a lower variance when the redshift is high.
- The redshift and near-infrared filter values show a similar relationship to the previous one, with the exception that the variance drops off quicker.

From these plots, it was inferred that objects closer to the observatory (which are usually stars) show high variance in properties and objects which are further away (which are usually galaxies) show low variance in properties. This coincides with the fact that galaxies are often quite similar to each other with respect to their images obtained on Earth.

C. Correlations between features

The correlation heatmap between features was plotted in Fig. 3. These correspond with the previous scatterplots and show the same relationships.

D. Feature-Label relationships

The distributions of each feature across the classes were then plotted in Figs. 4, 5, and 6.

- The red filter values have a lower mean for stars, a medium mean for galaxies and a higher mean for quasars. This means that stars have the lowest intensity of red light, followed by galaxies and quasars respectively.
- The near-infrared filter values follow a similar relationship, with the exception that the variance in means is lower - as noted in previous plots as well. This means that stars have the lowest intensity of near-infrared light, followed by galaxies and quasars respectively with smaller interquartile ranges.
- The redshift values show the highest difference between the classes, with stars having the lowest means (i.e the closest) and quasars having the highest means (i.e the furthest)

IV. K-NN and SVM-linear Classification

The data was initially prepared by only keeping the necessary columns required for analysis. With the random state set to 42, it was split into 80% training and 20% testing sets, which were both scaled using SKLearn's Standard Scaler.

A. K-NN

A K-NN algorithm was set up with k-values from 1 to 42 inclusive. For these values of k, a K-NN classifier was trained on the training set and labels were predicted on the test set. For these tests, mean errors in the true VS predicted labels as well as the lowest mean error and the associated k-value was calculated and stored. The mean error for each k-value is plotted in Fig 7. with the lowest mean error being 0.0356 obtained at k = 8.

Another classifier was trained over the training set with k = 8 and labels were predicted on the test set. For this test, a confusion matrix for true VS predicted labels and an accuracy report was plotted in Figs. 8 and 9 respectively. This classifier was then cross-validated with 5 cross-validation splits and the obtained cross-validation scores were plotted in Fig. 10.

The mean cross-validation score was 95.9%, which is close to perfect.

B. SVM-linear

An SVM-linear classifier with the random state set to 42 was trained over the training set and labels were predicted on the test set. For this, a confusion matrix for true VS predicted labels and an accuracy report was plotted in Figs. 11 and 12 respectively. This classifier was then cross-validated with 5 cross-validation splits and the obtained cross-validation scores were plotted in Fig. 13.

The mean cross-validation score was 95.6%, which is slightly lower than K-NN but still close to perfect.

V. Conclusion and Further Research

The results in this paper show that stellar objects can be very accurately classified into Stars, Galaxies and Quasars using K-NN or SVM-linear Classification based on red and near-infrared photometric filter values and redshift values, with both algorithms returning over 95% mean cross-valdation scores.

Further research should include trying out different classification algorithms such as SVM-poly, SVM-rbf (Gaussian), Decision Tree Classifier, and Linear Discriminant Analysis; all with appropriately tuned parameters and cross-validation.

One could also delve deep into the world of Stellar Physics and transform the current features into compound equations to obtain better classification.

References

[1] fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved January 15th, 2022 from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17.

[2] Geller, Robert M. *Universe*. Eleventh Edition, W. H. Freedman and Company, 2019.

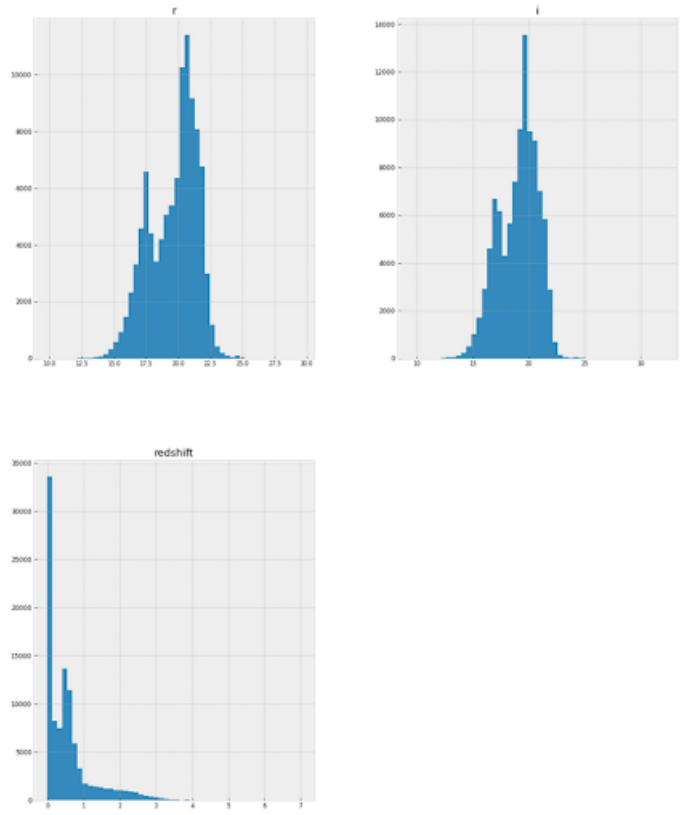


Fig. 1. Distribution of features.

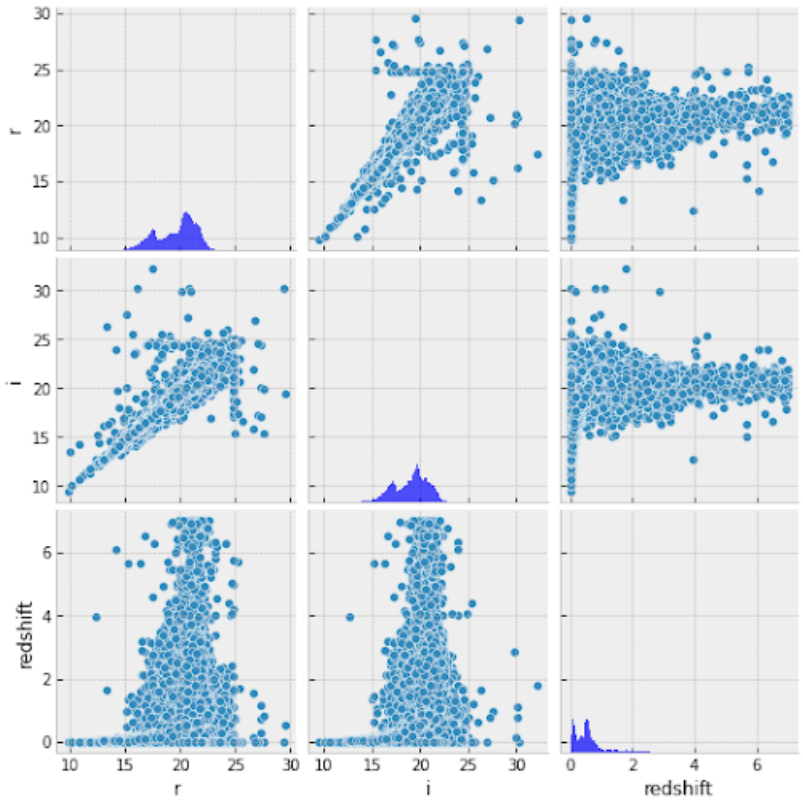


Fig. 2. Relationships between features.

	r	i	redshift
r	1.000000	0.962868	0.433241
i	0.962868	1.000000	0.492383
redshift	0.433241	0.492383	1.000000

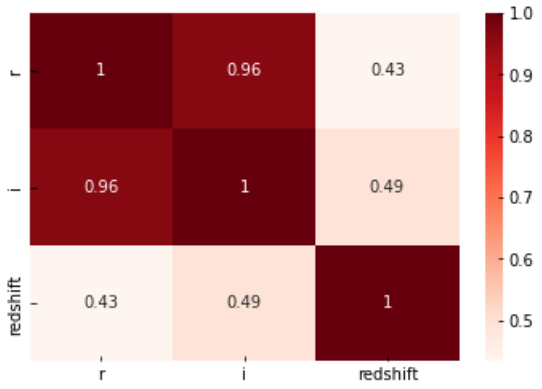


Fig. 3. Correlations between features.

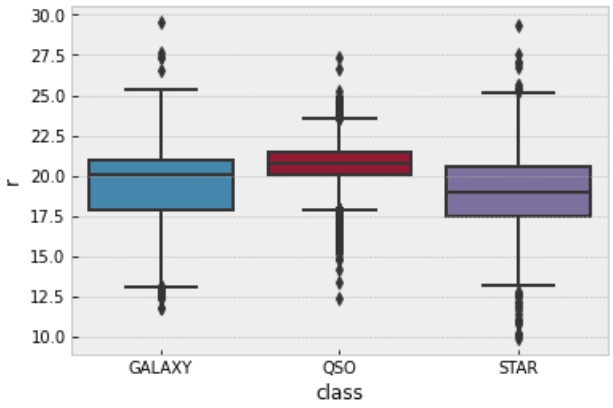


Fig. 4. Red Photometric Filter Values Across Classes.

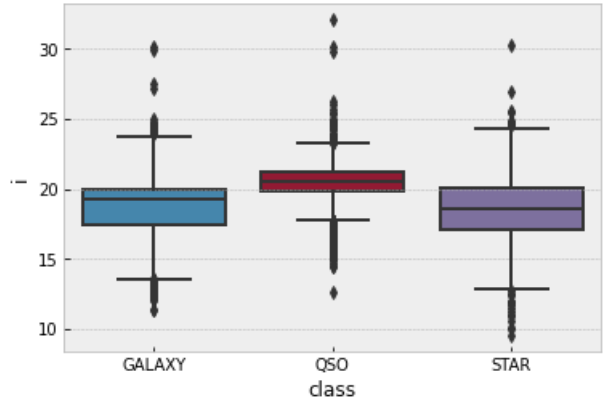


Fig. 5. Near-infrared Photometric Filter Values Across Classes.

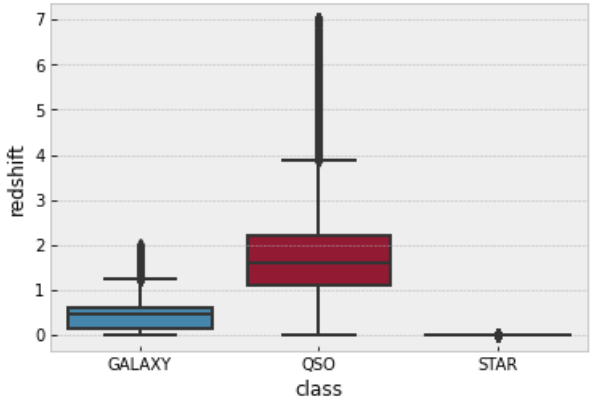
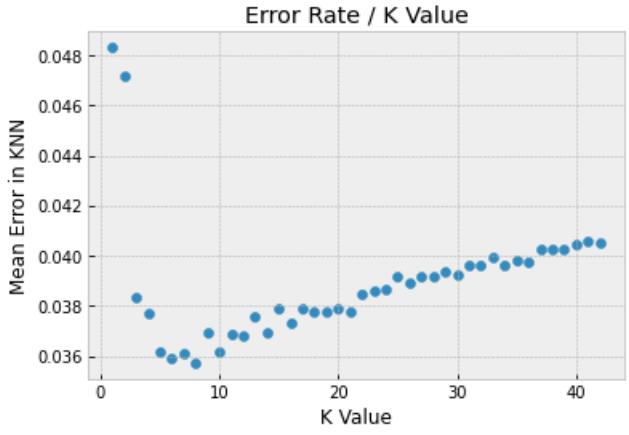


Fig. 6. Redshift Values Across Classes.




```
Minimum Mean Error: 0.036
Associated K-Value: 8
```

Fig. 7. K-NN Mean Errors for Tuning K-Value.

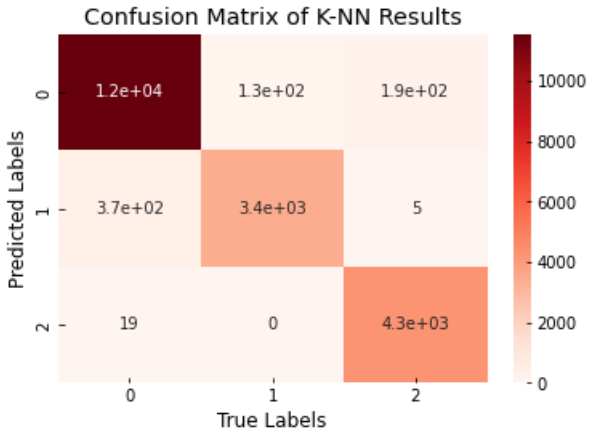


Fig. 8. K-NN True VS Predicted Labels Confusion Matrix

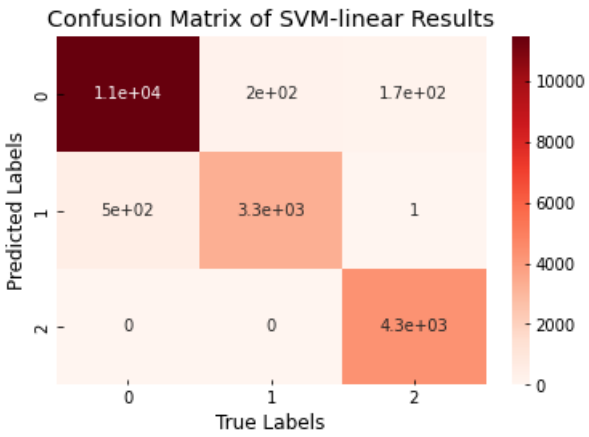


Fig. 11. SVM-linear True VS Predicted Labels Confusion Matrix.

	precision	recall	f1-score	support
GALAXY	0.97	0.97	0.97	11860
QSO	0.96	0.90	0.93	3797
STAR	0.96	1.00	0.98	4343
accuracy			0.96	20000
macro avg	0.96	0.96	0.96	20000
weighted avg	0.96	0.96	0.96	20000

Fig. 9. K-NN Accuracy Report.

	precision	recall	f1-score	support
GALAXY	0.96	0.97	0.96	11860
QSO	0.94	0.87	0.90	3797
STAR	0.96	1.00	0.98	4343
accuracy			0.96	20000
macro avg	0.95	0.95	0.95	20000
weighted avg	0.96	0.96	0.96	20000

Fig. 12. SVM-linear Accuracy Report.

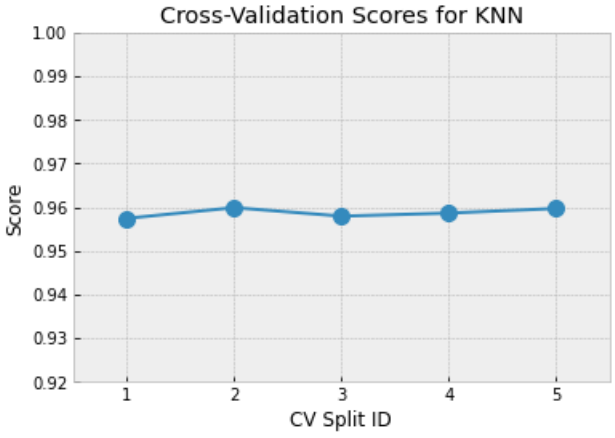


Fig. 10. K-NN Cross-Validation Scores.

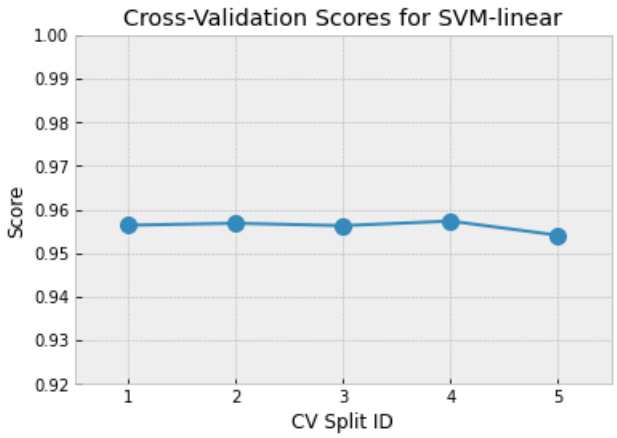


Fig. 13. SVM-linear Cross-Validation Scores.