HIGH SCHOOL EDITION
Journal of Student Research

Adapting Supervised Machine Learning to Active Galaxy Classification by Application of Multi-Wavelength Data

Elise McHallam

Basis Oro Valley, AZ, USA

ABSTRACT

In this study, we analyze and differentiate Active Galactic Nuclei (AGN), large bright gas regions powered by supermassive black holes, from star forming galaxies, which are galaxies that are actively producing new stars at a significant rate. Despite having physical differences, these AGN and star forming galaxies have very similar traits when it comes to their appearances,making them difficult to distinguish from one another. We formulated various models to differentiate the two, based on the ratios of elements contained within them, as revealed by astronomical spectroscopy. Our primary dataset used spectroscopy from the Sloan Digital Sky Survey, combined with photometry from ultraviolet and infrared space telescopes.The classification models we employed were K-Nearest Neighbors (KNN), Random Forest, and a Linear SVC model to determine the best possible approach for differentiating AGN from star forming galaxies. We show that our best model is as reliable as The BPT diagram, which is currently the state of the art model for differentiating AGN from star forming galaxies. This study shows that machine learning classifiers are able to be efficiently and effectively applied to multiwavelength astronomical datasets.

Introduction

Determining the difference between active galactic nuclei, AGN, and starburst galaxies is crucial for advancing scientific knowledge about galaxy evolution (Padovani et al. 2017). AGN are powered by supermassive black holes, while starburst galaxies are characterized by rapid star formation, and analyzing the defining characteristics of both helps construct a comprehensive narrative of how galaxies evolve over time (Agostino et al. 2019). Furthermore, distinguishing between the two classes of galaxies is effective in categorizing accreting and non-accreting supermassive black holes. This scientific question is extremely valuable due to the formation and evolution of supermassive black holes remains unknown among scientists today (Kewely et al. 2013). In order to address this challenge, the creation of classification models is necessary. The classification models used included K-Nearest Neighbors (KNN), RandomForest, and Linear SVC models.

The BPT diagram, a model used to differentiate AGN and star forming galaxies, was created in the 1980's and was used as the primary model on the issue of AGN and star forming galaxies. The BPT diagram proved useful to many researchers, but the precise line of division separating AGN and star forming galaxies displayed inaccuracy as it constantly shifted (Agostino et al. 2019). In this study, we applied a variety of machine learning models to adapt the understanding of the differentiation between AGN and star forming galaxies. Through the study, we prove the advantage of this technique displaying a modern approach which makes use of large datasets available today.

Dataset

The dataset used in this study is derived from the Sloan Digital Sky Survey dataset of galaxies with spectral line measurements, crossmatched with mid-infrared data from the WISE telescope, and ultraviolet data from the GALEX telescope. AGN emit wavelengths across the electromagnetic spectrum. Their different electromagnetic bands allow for their physical properties to be uniquely analyzed. These bright regions emit most strongly in the X-ray and UV. Major components of the dataset included optical photometry features in the g, r, and i photometric bands, as well as spectroscopically-determined redshift, measurements of far and near UV, measurements of infrared, as well as many types of light. The other half of the dataset focused on spectroscopic values, utilizing flux ratios and spectral lines to distinguish AGN from star forming galaxies. Prior to extensive data cleaning, the entire data set held 169832 rows × 31 columns. The data was initially separated into two major subclasses; AGN and star forming galaxies. When forming the classification models, the dataset split into two subsets, the training set and the test set. The features learned were the spectroscopic line ratios in the dataset. The independent variable was the galaxy class.

Methodology and Models

The main three models used in the study included RandomForest classifiers, K-Nearest-Neighbor (KNN) classifiers, as well as LinearSVC classifiers, as implemented in the scikit-learn package developed for Python. The RandomForest classifier models build multiple decision trees and accumulate their predictions to improve accuracy and reduce overfitting, while KNN models classify a sample based on the majority class among its k nearest neighbors in the space surrounding. Linear SVC models find the optimal line to separate classes by maximizing the margin between support vectors.

I first started by testing each of these models with only the spectroscopic values of the dataset as well as default hyperparameters. I then began to add photometric values by first adding in UV magnitudes to my set of features and later incorporating redshift. I increased the accuracy of the KNN models by creating a few grid search plots to see the highest accuracy for the n number of neighbors. After analyzing the accuracy scores of the many models, with and without added features, I recognized that the KNN models showed the highest scores consistently.

Results

Through carefully analyzing the various models created in the study, I concluded KNN is the most accurate model used in the study to classify AGN vs. star forming galaxies. The KNN models with the highest accuracies were the models in which photometric and spectroscopic values were incorporated. These values included those of redshift, FUVmag, as well as NUVmag. The models outputting the lowest accuracies included the RandomForest classifiers. The LinearSVC models performed well, but produced much less consistent accuracy scores when features were included. The LinearSVC models outputted an accuracy score between 92% and 94% for the models that included no photometric or spectroscopic features, redshift, and FUVmag and NUVmag. The LinearSVC model however fell short when W1mag and W2mag features were included, as its accuracy score dropped to 87%. The RandomForest models were the least reliable and least consistent models. The original RandomForest classifier, including no photometric values performed quite well, reaching an accuracy score of roughly 93.5%. The RandomForest classifier for FUVmag and NUVmag however only reached an output score of 80%, and then reached a slightly lower score of roughly 79.9% when including W1mag and W2mag in their model. The lowest output score produced by the LinearSVC model was 78.8%, when including the spectroscopically-determined redshift. The original KNN classifier with no added features obtained an accuracy score of 94%. The highest score obtained by the KNN model was 94.5% when redshift, FUVmag, and NUVmag were included. The lowest accuracy score outputted by the KNN models was roughly 88.5% when W1mag,W2mag, NUVmag, FUVmag, g, and r values were included in the same model.

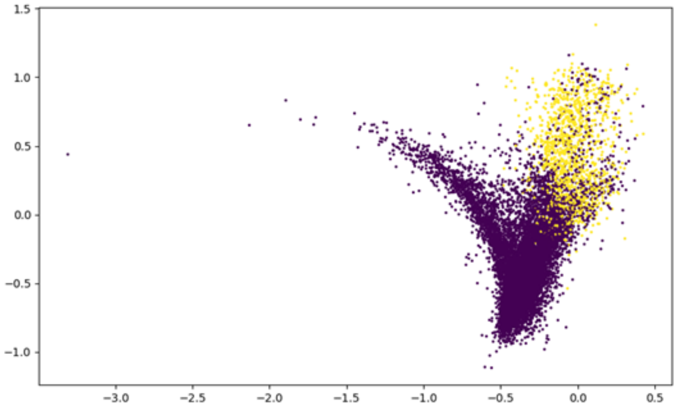

Figure 1. Labeled BPT diagram. AGN are colored in yellow, while star forming galaxies are colored in purple.

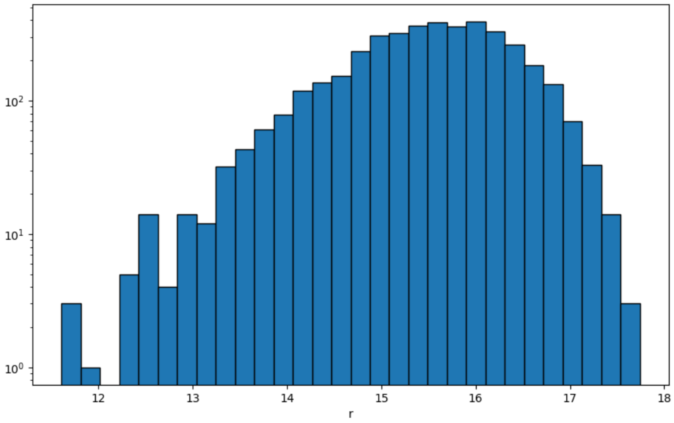

Figure 2. Histogram of g band (visible green color) of the optical photometry features, shown on semi-log plot. The amount of galaxies increases, as you go to more faint magnitudes, and is limited by telescope sensitivity.

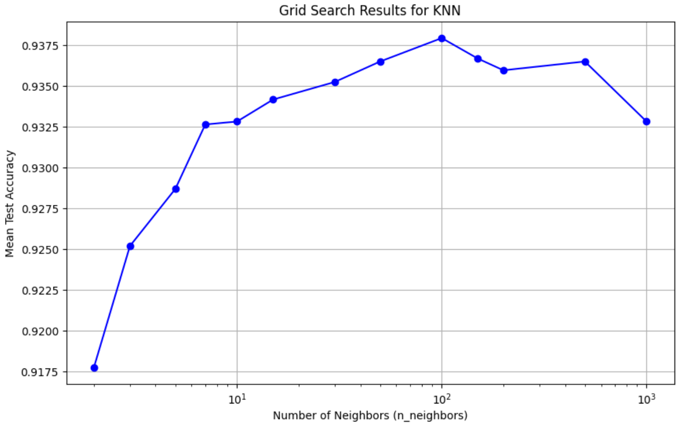

Figure 3. Grid search plot for KNN model with no photometry features. The optimal model is the model with 100 neighbors.

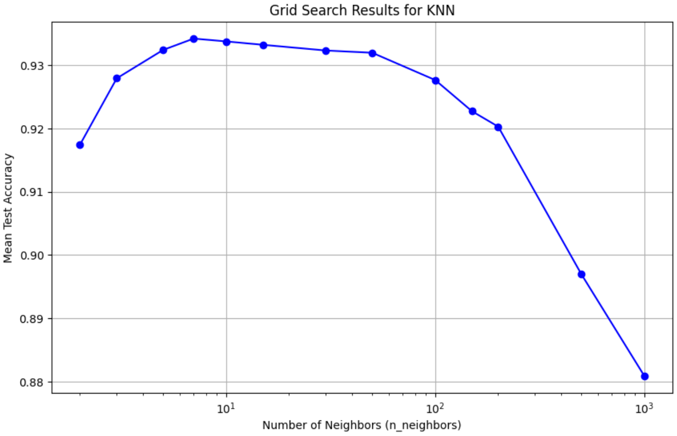

Figure 4. Grid search plot for KNN model with added features, FUVmag and NUVmag.

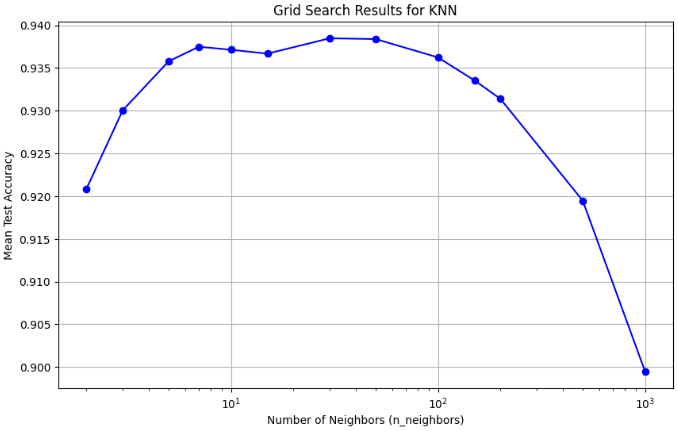

Figure 5. Grid search plot for KNN model including W1mag and W2mag features.

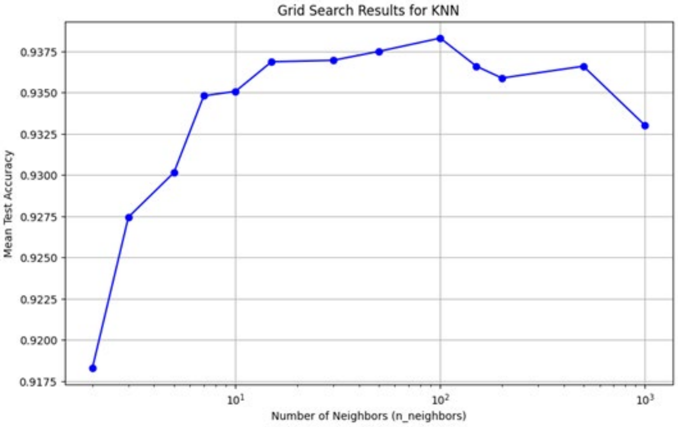

Figure 6. Grid search plot for KNN model with redshift feature.

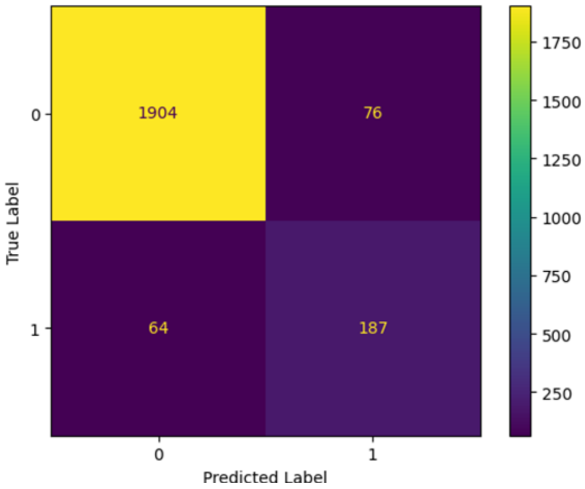

Figure 7. Confusion Matrix for KNN with no added photometry or spectroscopy features. The number of false positives and false negatives are low, indicating the model is reliable and accurate. (Zero corresponds to star forming galaxies and one corresponds to AGN)

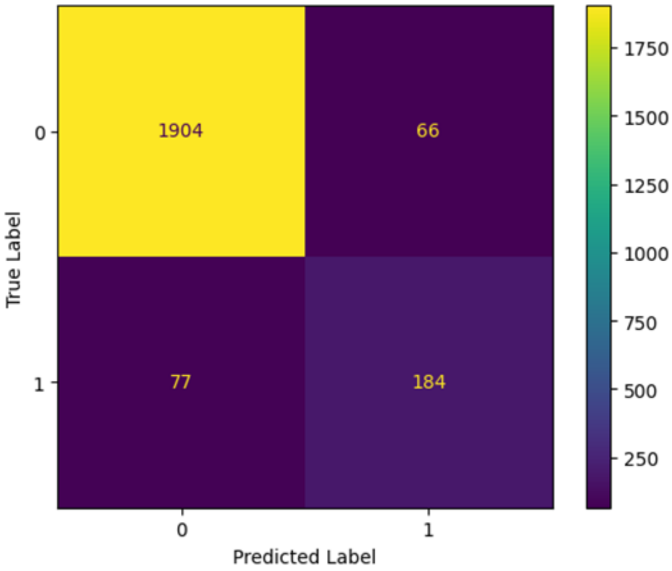

Figure 8. Confusion Matrix for RandomForest with no added photometry or spectroscopy features. (Zero corresponds to star forming galaxies and one corresponds to AGN)

Conclusion

Through the study we have analyzed the different ways in which it is possible to challenge the BPT model, classifying AGN from star forming galaxies. Many models, including various features, were tested against one another, in order to determine which form of classification would be most accurate for the dataset. We primarily focused on three classification models, including KNN, LinearSVC, and RandomForest models. We have been able to adapt photometric and spectroscopic values to KNN classifier models in order to obtain an accuracy score as high as 94.5%. This machine learning model was proven reliable in this study, and can be accurately applied to larger data sets from various telescopes. Overall, the study has shown the capabilities of machine learning algorithms to reshape and rewrite many classifying models from past decades that we still use today.

Acknowledgments

I would like to extend my gratitude to the Inspirit AI program and Tony Rodriguez, my AI + X mentor.

References

Kewley, Lisa et al. 2013. The Astrophysical Journal. doi:10.1088/2041-8205/774/1/L10

Agostino, Christopher J. and Salim, Samir et al. 2019. The Astrophysical Journal. doi: 10.3847/1538-4357/ab1094

Padovani, Paolo et al. 2017. Frontiers in Astronomy and Space Sciences. doi: 10.3389/fspas.2017.000