Predicting The Cosmos: A Multiwavelength Approach to Classify Active and Star Forming Galaxies

Saanika Kulkarni[1] and Tony Rodriguez[#]

[1]Dougherty Valley High School, USA
[#]Advisor

ABSTRACT

Distinguishing between active galactic nuclei (AGN) and star forming galaxies (SFG) using spectroscopic analysis has been done since the 1980s using a BPT diagram, using the traditional log(OIII/HB) vs log(NII/HA) lines. Through this paper, I aim to supplement the traditional emission lines used with infrared and ultraviolet photometry to find the best predictors of SFGs vs. AGNs. Successfully distinguishing between AGN and SFGs can inform us about the true demographics of these systems and how they evolve through cosmic time. I use various baseline models to predict SFG vs. AGN. For the traditional emission lines, I achieved an accuracy of 90.0%. However, with multiwavelength data, I was able to achieve an accuracy of 95.15%, indicating that there are better predictors than the ones traditionally used.

Introduction

Galaxies, or large, structured, collections of stars, can be classified in many ways. Depending on their luminosity, structure, and type of emission, they may be classified differently. In this work, I focus on active galactic nuclei (AGN), and are characterized by their compactness and unusual luminosity. There are various theories for AGN luminosity; various studies proposed that AGN are luminous due to the accretion of matter by its supermassive black hole (SMBH) and this theory was backed by later studies. Further X-Ray observations revealed that AGN were major producers of X-Ray emissions, leading to further speculation on AGN energy sources (Salpeter and Zeldovich 1964).

AGN and similar high-energy galaxies can be consolidated into a general "unified" model, which shows the different angles a luminous galaxy can be looked at. Figure 1 shows one such representation of this unified model.

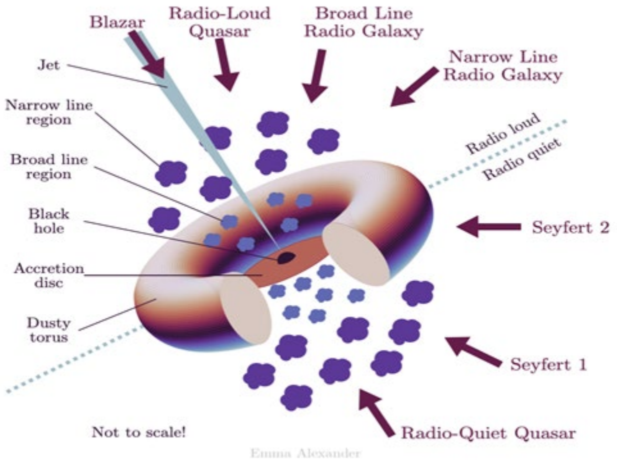

Figure 1. A representation of the unified model of an AGN. Depending on what angle an observer looks at, a supermassive black hole such as the one depicted in the figure may appear as differently classified astronomical objects. For example, looking at the north pole of this figure will give the impression of a radio loud blazar, while looking edge-on will give the appearance of a dusty torus, either radio loud or quiet. (Illustrated by Emma Alexander, developed by Urry & Padovani (1995.)

Here, I investigate the difference between AGN and star forming galaxies (SFG); as evident in their name, SFGs are areas of high star formation, triggered by mechanisms like supernovae, while AGN are not dominated by star formation, but instead by SMBH accretion. The goal of this work is to find the best predictors of AGN vs SFG. In more precise terms, are there certain emission lines that predict AGN vs SFG better than others? I decided to use a supervised classification model to predict these categorical labels given a dataset.

Background

Agostino 2019 is the study most similar to mine; their recreation of the BPT (Baldwin-Phillips-Terlevich) diagram to differentiate between AGN and SFG is something I also aimed to achieve in my study. Agostino 2019 made many unique contributions to this field, such as their novel X-ray AGN selection method, their exclusive use of low-redshift galaxies for their data analysis, and for their insights that LINERS (low-ionization nuclear emission-line regions) can be classified as AGN. Note that while this is significant, LINERS will not be the focus of this study.

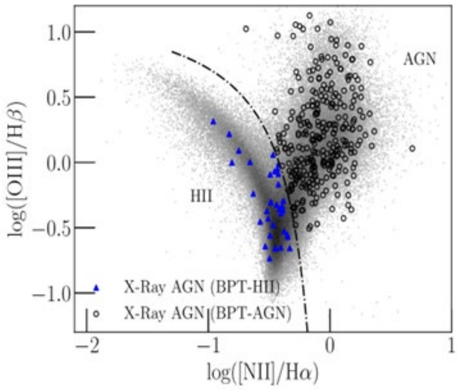

Figure 2. The BPT diagram from Agostino 2019. This shows the usage of the two empirical ratios from Kaufmann et al. 2003 and the differentiating line. HII are SFGs, while AGN are symbolized by circles. Blue triangles indicate verified AGN that are located in the SFG region of the BPT diagram and would be misclassified based on the BPT diagram alone.

While exploring optimal emission lines for classifying Active Galactic Nuclei (AGN) and Star-Forming Galaxies (SFGs), I encountered research from the NASA JWST team on utilizing infrared observations to penetrate dust surrounding these objects. This sparked the idea that infrared bands might offer improved differentiation between AGN and SFGs. However, the unified model emphasizes the limitations of solely relying on infrared data for classification. Therefore, a multi-wavelength approach, incorporating data from various parts of the electromagnetic spectrum and advanced analysis techniques, could potentially yield more accurate and nuanced classification results.The angle at which an observer is looking at an object could be ambiguous, and since both AGN and SFGs have dust at certain angles, this could lead to misleading conclusions about the nature of the object.

Dataset

For my research, I used the GALEX Sloan Digital Sky Survey Wise Legacy Catalog (GSWLC), developed by Salim et al. 2016. It contains 700,000 low-redshift galaxies (Salim et al. 2016), and is mainly useful for (infrared) IR and (ultraviolet) UV analysis. I worked with numerical data on the flux of multiple bands such as r (red), g (green), and i (infrared). There were also additional bands I used for my data preprocessing; I wanted to keep all observations that didn't have a spectrotype of STAR, ensuring that my sample only consisted of AGN and SFG type objects. Another feature, SUBCLASS, indicates the object's nature; for example, SFG is a subclass of galaxy, and so is starburst and AGN. For this reason, I removed observations that had a null subclass. I also kept the SN_MEDIAN and quality flag features to make sure that my data was reliable. After both of these actions, I had a total of 8815 observations to split into training and test data.

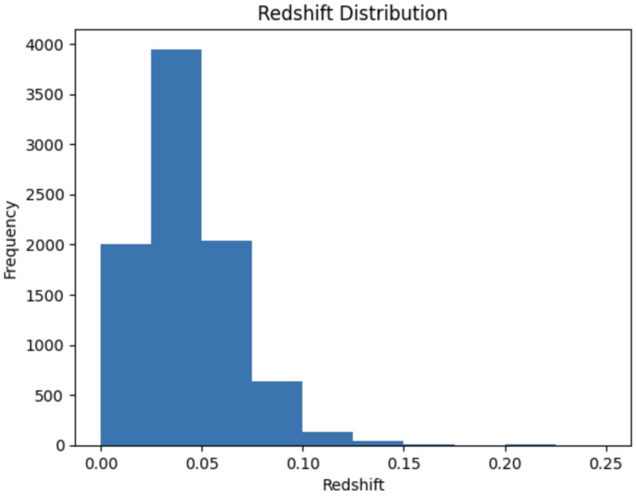

Figure 3. A histogram showing the redshift (how much light has been stretched away from an observer) of the ~8000 systems in my sample. About half of the sample has a redshift of around 0.03, while a few outlier redshifts are greater than 0.2.

First, I decided to recreate the BPT diagram (Agostino 2019) using my data. I had line flux values for NII, Hα, OIII, and Hβ, and took the log of ratios to compute values in the BPT Diagram. This is visible in Figure 4.

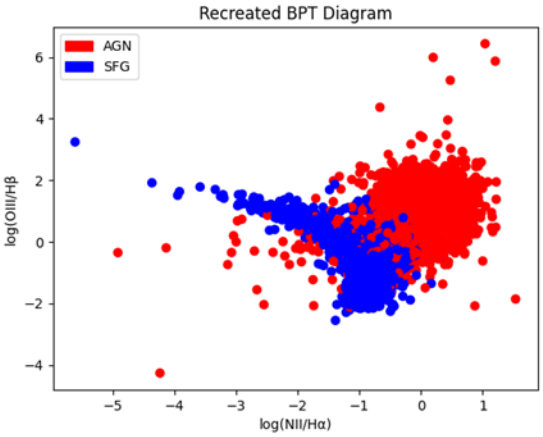

Figure 4. My recreated BPT diagram. As apparent in the legend, blue dots are SFGs, while red dots are AGN. The differentiating line (as mentioned in Agostino 2019) is clearly visible in this diagram. However, it's also apparent that there are quite a few AGN on the general SFG side. Because of this, I sought to find better predictors of SFG vs AGN to achieve more true positives while classifying. To calculate the log values for both axes, I utilized the line flux values of NII, Hα, OIII, and Hβ.

The features of this dataset that were important to me were (apart from g, r, and i) ultraviolet photometric data (FUV and NUV) and infrared photometric data (W1mag and W2mag). Additionally, I kept features like SN_MEDIAN which helped me determine the quality of observations. These features are important to my study because they provide a wide range of multiwavelength data I can use to classify with.

Methods

For my study, I chose to use various machine learning models to predict AGN vs. SFGs. I split the data of ~8.8 thousand observations 80/20, meaning that 80% of data was used for training the model, while 20% was used for testing.

One model used was a Logistic Regression model, which works well when classifying categorical data, such as mine. To carry this model out, I inputted the data into the sklearn module for Logistic Regression, and then calculated various scores based on accuracy, such as precision, recall, and F1 (described and explained in the following section).

Another model I used was KNN Neighbors Classifier. This model is unique because unlike regression, it tries to group similar things instead of making a curve of best fit. I experimented with K, or the number of neighbors the model seeks to find around data points.

The model that I thought was the best for my study turned out to be a Random Forest classifier, which uses multiple decision trees to classify categorical data. Although this is a supervised (data is already labeled as AGN vs SFG) model, the results from this model turned out to be the most accurate for my study. The Random Forest classifier uses a hyperparameter called Max Depth which represents the longest length of the root node (beginning of the decision tree) to the leaf node (end of the tree). It uses multiple decision trees (specified in random state) to make multiple classifications and then averages it out for a final classification.

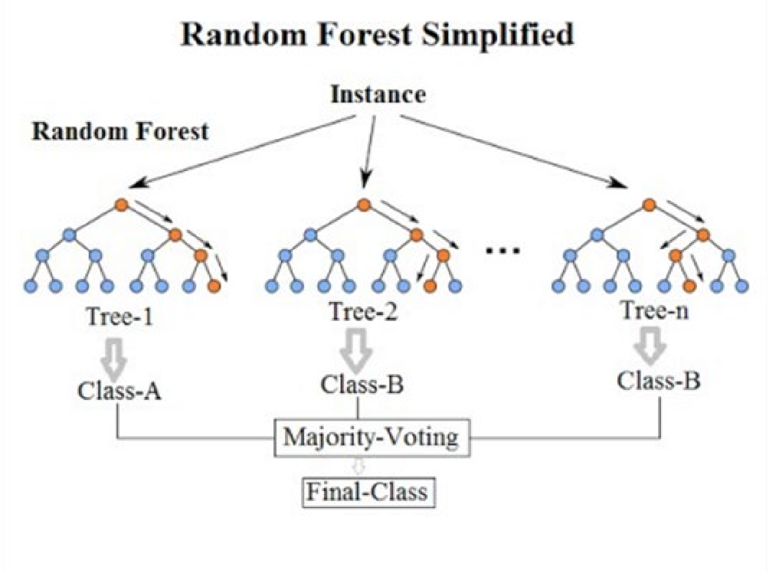

Figure 5. A simplified cartoon of the mechanics of the Random Forest classifier. Depending on the number of trees indicated, the classifier makes *n* classifications and takes the majority to decide the final output (Will Koehrson).

For all three models, something I wanted to avoid was overfitting, which is when the model makes predictions too close to the training data, but cannot make those same predictions for the test data (Mathworks). As a result, the model loses accuracy and other metrics based on it become skewed as well. For this reason, tuning hyperparameters for each model carefully is crucial. One way I could tell the hyperparameters needed tuning is if the accuracy, precision, recall, and F1 scores were significantly off from each other. If, for example, accuracy is very high but the precision isn't, this can indicate that the model is overfitting and giving too many false positive classifications. Likewise, other metrics differing from the accuracy by a significant amount could lead to other classification problems.

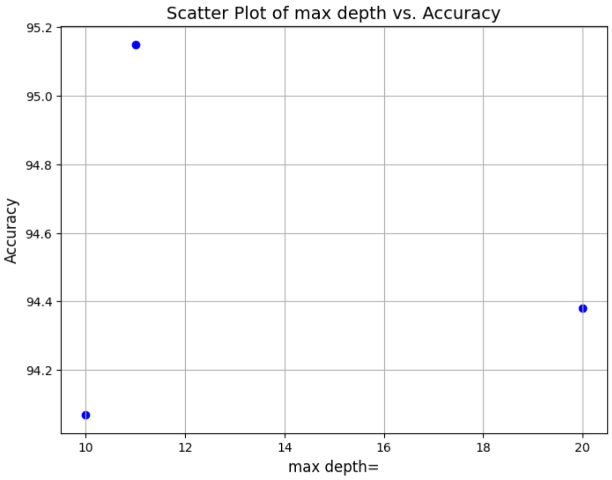

Figure 6. My hyperparameter grid search plot. This shows the accuracy of the Random Forest Classifier as a function of max depth. The classifier peaked at max depth = 11, which gave an accuracy of just below 95.2%.

The model can also be underfitted. Underfitting can result from insufficient data, which can lead to poor accuracy. While reducing the data, I initially wanted to keep rows that had a SN_MEDIAN (signal to noise ratio) of greater than 50, which would make sure that the data was reliable. However, after reducing that, I was left with only around 6 thousand data rows, which would make the test/train split groups pretty small. Therefore, I sacrificed data reliability by lowering the SN_MEDIAN criteria to greater than 30.0 so as to not risk underfitting the model.

Results and Discussion

With just the nominal features present in the BPT diagram, I achieved an accuracy of 90.00% with the Random Forest Model. I also calculated three other metrics, as mentioned in the previous section: precision, recall, and F1. Precision is the fraction of true positives as compared to all positive classifications; recall is how well the model did while testing, or the fraction of how many true positives the model predicted out of all actual positives; F1 is the mean of precision and recall (Performance Metrics in Machine Learning (Bajaj 2022). For the Random Forest, the precision, recall, and F1 was 91.58%, 88.98%, and 90.26%, respectively.

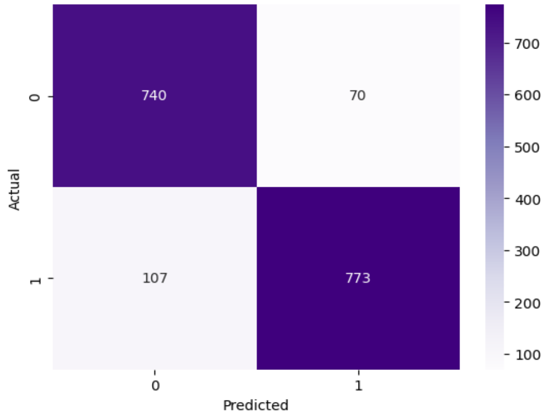

Figure 7. The confusion matrix for the nominal Random Forest Model. 0 corresponds to SFGs, while 1 is AGN. Here, the model classified 70 galaxies as AGN when they were actually SFGs, and 107 galaxies as SFGs when they were actually AGN.

Similar results were achieved for the other two models I tried: the Logistic Regression and the KNN Neighbors Classifier. The Logistic Regression model had an accuracy of 90.24%, and its corresponding metrics were within 0.7% of the accuracy. For the KNN Neighbors Classifier, an accuracy of 89.7% with just the nominal features on the BPT diagram with similar metrics was calculated.

Since the goal of my study is to find better predictors of AGN vs SFG, I used additional photometric features to gain better accuracy and related metrics. Combining optical, UV, and infrared data gave an accuracy of 95.15%, which is a significant difference than just using the nominal features. In the context of the study, a 5.15% difference would entail classifying nearly 500 more galaxies correctly, which can help us understand more about the chemical composition and cosmic evolution of AGN vs SFGs. Additionally, upcoming large surveys such as those from the Sloan Digital Sky Survey and others will produce spectra for millions of galaxies, meaning that this 5% increase would correspond to 50,000 improved classifications.

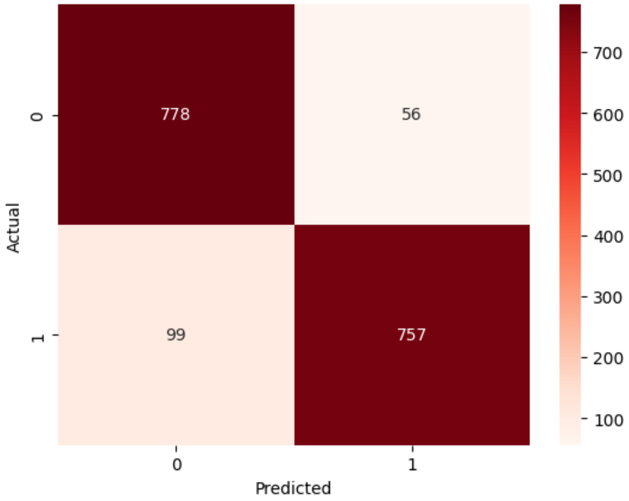

Figure 8. The confusion matrix for the updated Random Forest model with all of the wavelength data. Again, 0 represents SFGs and 1 represents AGN. This model is clearly better, and there are significantly less false positives/false negatives.

Conclusion

Throughout this study, multiple models have indicated that there are better predictors than the nominal emission lines used in the BPT diagram (Agostino 2019). The best accuracy was achieved using additional optical, infrared, and ultraviolet wavelength data, 95.15%, with the Random Forest model (optimized for the best hyperparameters), as compared to the 90% accuracy rate with the nominal BPT diagram features. By no means is this study fully conclusive. A future step could be exploring beyond low-redshift galaxies, and also using other line measurements such as those of silicon, sulfur, magnesium, carbon, and nitrogen seen in a subset of these galaxies. Currently, the results of this study can only be generalized to low-redshift galaxies in the SDSS GALEX-WISE dataset, but whether this can be generalized to higher redshift galaxies is still unknown.

Acknowledgments

I would like to thank Inspirit AI and Tony Rodriguez for their immense support with my research.

References

Agostino, C. J., & Salim, S. (2019). Crossing the line: Active galactic nuclei in the star-forming region of the BPT Diagram. The Astrophysical Journal, 876(1), 12. https://doi.org/10.3847/1538-4357/ab1094

Bajaj, A. (2022, July 21). Performance metrics in machine learning [complete guide]. Neptune.Ai. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

Emma L. Alexander. (n.d.). Retrieved August 17, 2024, from https://emmaalexander.github.io/resources.html

Kauffmann, G., Heckman, T. M., Simon White, D. M., Charlot, S., Tremonti, C., Brinchmann, J., Bruzual, G., Peng, E. W., Seibert, M., Bernardi, M., Blanton, M., Brinkmann, J., Castander, F., Csábai, I., Fukugita, M., Ivezic, Z., Munn, J. A., Nichol, R. C., Padmanabhan, N., … York, D. (2003). Stellar masses and star formation histories for 105galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 341(1), 33–53. https://doi.org/10.1046/j.1365-8711.2003.06291.x

Koehrsen, W. (2020, August 18). Random forest simple explanation - Will Koehrsen. Medium. https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

Mathworks. (n.d.). Overfitting. MATLAB & Simulink. Retrieved August 17, 2024, from https://www.mathworks.com/discovery/overfitting.html

Salpeter, E. E. (1964). Accretion of interstellar matter by massive objects. The Astrophysical Journal, 140, 796. https://doi.org/10.1086/147973

Urry, Megan, C., Padovani, & Paolo. (1995). Unified schemes for radio-loud active galactic nuclei. Publications of the Astronomical Society of the Pacific, 107. https://doi.org/10.1086/