Using the random forest classifier, we can distinguish possible quasars from stars and galaxies with a 92.31% accuracy.

Using Machine Learning to Improve Efficiency of Quasar Candidates Selection Prior to the Spectroscopic Observations

Qiongwen Mao (Shirley) and Prof. Eilat Glikman

Motivation

Quasars are extremely luminous astronomical objects that are empowered by supermassive black holes. They are crucial for understanding the formation of stars and the evolution of the universe.

To identify quasars from other celestial objects, however, requires spectroscopic observations for each object out of thousands of individual sources, one at a time. Thus, the efficient selection of candidates to be observed is very important.

My project focuses on using a random forest classifier machine learning model to identify quasars before pursuing spectroscopic observation, thus improving the efficiency of the selection of quasar candidates.

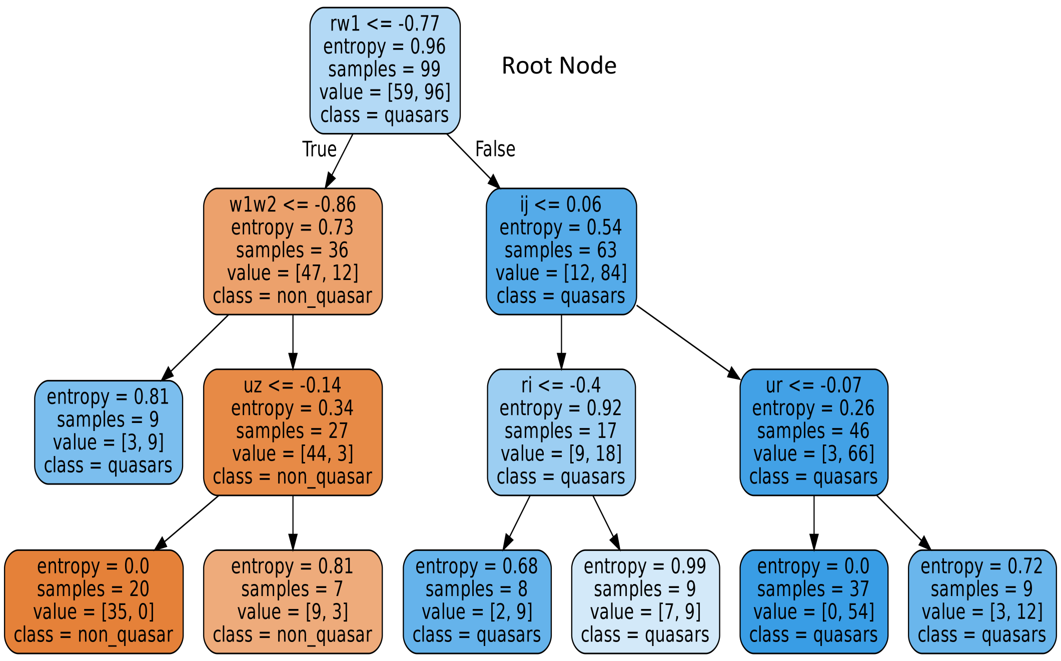
Current Efforts

In this project, we collect the brightness measured at different wavelengths of each celestial object, ranging from ultraviolet to near-infrared, and measure the color, which is the brightness ratio at different wavelengths, for each object.

Color is an important feature to differentiate celestial objects from each other, and, thus, we use it as the main feature to train the random forest classifier.

Random forest is a machine learning algorithm that is widely used in classification problems. It consists of multiple decision tree models that are trained with subsets of the training data to reduce the variance of the output. The random forest will then take the majority vote of the trees to produce the final prediction.

Below is the visualization of one decision tree from my random forest model.



The root node tells us that the tree first classifies the data by setting the decision boundary at the color R-W1 = - 0.77. When R-W1 <= - 0.77, the tree found 59 non-quasars objects and 96 quasars that satisfy this condition. Since more quasars meet the condition, the node is going to classify the objects as a quasar. The decision boundary, as you can see, is not as helpful at distinguishing quasars because the entropy is 0.96, which indicates a high level of disorder. And the same logic applies to the other nodes.

In my model, there is a total of 50 trees with different decision boundaries and trained with different subsets of the data. To predict the class of an unknown object, the random forest classifier counts the decision of each tree and then predicts the class of the object by the majority vote method.

Additionally, **the model has found that the color I-J is the second most important feature to identify quasars.** This is a surprising finding as scientists often use a color that has a larger wavelength difference to select possible quasar candidates. This finding could potentially further improve the quality of the selection.

Middlebury College
Department of Physics

PHYS 0704: Senior Project
Spring 2022