## Abstract

## 1. Introduction

*- what is quasars? And why is it important to study them?*

*- motivation of the project*

*- who else made measurements like this?*

*- include a description of the theory*

Quasars are

## 2. Data

* *Photometric data from SDSS, UKIDSS, and WISE(?)*
* *Petrosian data*
* *Data Processing: Normalize the data*
* *Feature processing: color*
* *Motivation of using random forest classifier 🡪 non-linearity and clutter in the data*

## 3. Methodology

* *Random Forest Classifier*
  + Bagging for imbalanced data: class\_weight for training trees
* *Hyper-parameter Optimization* (table of comparison before and after the optimization)

3.1 Data Processing (Standard Scaler 🡪 Normalization)

3.1 Random Forest Classifier

In this project, we used a random forest model to classify quasars and non-quasars objects. The model is an ensemble of multiple decision tree models and makes the prediction by taking the majority consensus of classification across all trees’ predictions. Each tree is trained with a random subset of the training data, which is a mechanism called bootstrapping that is used for reducing model’s sensitivity to the original training data, and with a random selection of features to decrease the correlation between the trees. With the training, the tree will grow in a top-down manner, shown in Figure X (visualization of a tree), that starts with an internal node that has a conditional boundary of a certain feature, splits into branches based on the outcome of the boundary. When the end of the branch does not split anymore, each leaf node represents a decision or class label. To determine the best feature on which the samples must be split is decided by the entropy, which is a measurement of disorder (See Appendix).

We randomly divided the data into training and testing datasets. The former accounted for 80% of the whole and the latter for 20%. The training set is used to train the random forest classifier, and the testing set for assessing the performance of the model by giving the model unseen data points.

3.3 Performance Evaluation & Metrics

## Results & Discussion

* *Precision, recall, f1*
* *Feature importance & Gini impurity explained*

## Conclusion

## Appendix

* *Visualization of growth of a tree estimator*
* *Entropy explained*