SIEMEN

SUMMER OF INNOVATION

Space Data Science

Objective

The purpose of this project was to teach the machine what an exoplanet data set looks like so that it can tell if a planet is an exoplanet when given enough information.

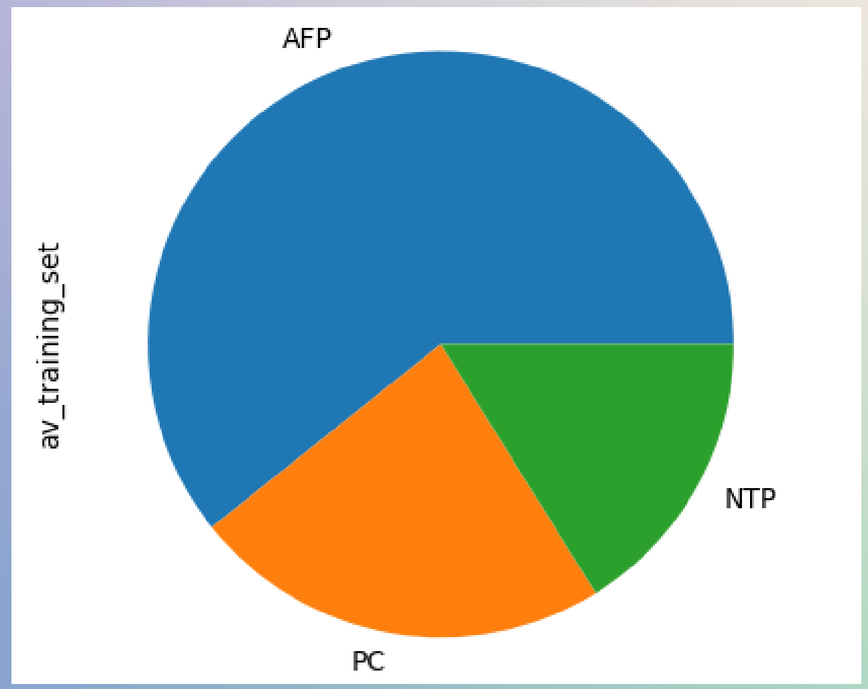
Description

Our algorithm accepts a CSV file from the user and determines whether the data is from an exoplanet. It also creates graphs to show the ratio of different exoplanet kinds.

Library used

Numpy
Matplotlib
Pandas
Sklearn
XGboost

To read and clean the given exoplanet data, we utilized **Pandas.**



The dataframe was cleaned up by removing empty columns.

We eliminated rows in the av_training_set column with values of "UNK" since we don't want "UNK" as a result.

To classify the data into three discrete groups, we replaced "AFP," "NTP," and "PC" with 0, 1, and 2, respectively.
We removed columns like "kepid" "tce plnt num" , which aren't useful in determining whether a planet is an exoplanet.

We created set of independent features - X and target variable - y consisting of av_training_set.
Using Sklearn, we generated a train and test set. The train set received 70% of the dataset, whereas the test set received 30%.

We attempted to use multiclass Logistic Regression at first, but were unable to achieve a satisfactory accuracy score.

We tried another method, Random Forest Classifier, and obtained better results.

However, we discovered that the XGBoost algorithm provides superior accuracy but takes a little longer to run.

We used matplotlib to create a pie chart based on the number of "AFP," "NTP," and "PC" values in result of the input data.

Result

For this model, we looked at a few metrics such as F1 Score, Precision, and Confusion Matrix, but our major focus is Accuracy.

Logistic Regression (default hyperparameter) gives about 75% accuracy.

RandomForestClassifier(n_estimators=200, criterion='gini',max_depth=13) gives accuracy of about 81.85%.

XGBClassifier(learning_rate=0.02, n_estimators=600, objective='multi:softprob', silent=True, nthread=8) gives a accuracy of 83.1024%.