A Novel Method in Improving the Accuracy of Exoplanet Classification Using Machine Learning

Zhang, Brian (School: William G. Enloe High School)

Since the Kepler Space Telescope entered service in 2009, the data on exoplanets has increased significantly. Astronomers must manually analyze this data to identify which objects of interest are exoplanets, which is very time-consuming. Creating a method that precisely and accurately classifies exoplanet candidates is essential to advancing exoplanet research. In this study, data from the Kepler Objects of Interest table was analyzed using six common machine learning models: Naive Bayes (NB), Logistic Regression (LR), K Nearest Neighbors (KNN), Random Forest (RF), Boosting Tree (BT), and Voiting Classification (VC). BT, RF, LR, and VC were all able to achieve F! scores above 90%. This study assessed various data preprocessing methods. Z-score normalization was used in tandem with imputation to deal with missing values and variations in the magnitude of the data. Normalization improved the metrics of the data, while imputation decreased the metrics. Feature importance was calculated using Gini impurity, and the most important features were transit light curve data and exoplanet orbit characteristics. Recursive Feature Elimination with Cross Validation (RFECV) was also used to help improve the metrics of the LR, RF, and BT models. Usage of RFECV showed improvements across all metrics and found the optimal number of metrics for LR, RF, and BT was 39, 28, and 26 features, respectively. This project demonstrates a novel method for improving the accuracy of exoplanet classification and proves that a comprehensive machine learning pipeline is essential to increasing the accuracy of exoplanet classification.