Exoplanet Machine Learning Models

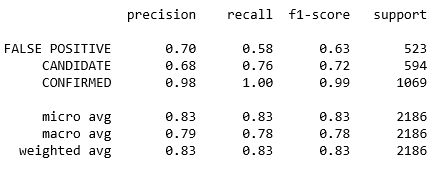
The exoplanet data was split into training and test sets. The training data was used to fit various classifier models and test dataset is used to make predictions. Sklearn library is used for scaling and modeling the data. A classification report was generated using F1 score, precision, and recall for each classifier. Besides, accuracy score was generated using sklearn metrics for each classifier.

F1 score is the measure of a test's accuracy, a weighted average of the precision and recall, with best value at 1 and worst at 0. F1 scores corresponding to every class will tell accuracy of the classifier in classifying the data points in that particular class compared to all other classes. As a rule of thumb, the weighted average of F1 was used to compare classifier models, not global accuracy.

The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The recall is intuitively the ability of the classifier to find all the positive samples. The support is the number of samples of the true response that lie in that class.

Logistics Regression

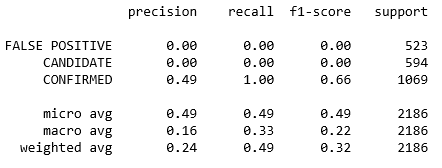
The logistics regression classifier has shown the scores of 0.846447087526685 for training data and 0.833028362305581 for testing data. The classification report was generated for FALSE POSITIVE, CANDIDATE, CONFIRMED classes in the exoplanet data. The report details are below.



The weighted average of f1-score for this classifier is showing 0.83 which is close to the scores observed for training and test data. This score is closer towards 1, which is the best. The sklearn metrics accuracy score is at 83.3028362305581. This score being closer to the f1 score demonstrates the validity of this model.

KNeighborsClassifier

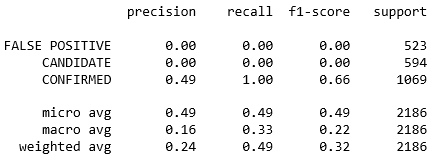
The K-Nearest Neighbors (KNN) classifier model seems to have shown some stability at k=5 but fluctuated at higher k values. The KNN accuracy score was at 0.812. However, the classification report has shown some mixed results as shown below.



The f1 scores were not recorded for FALSE POSITIVE and CANDIDATE classes and the weighted average was only at 0.32

Gaussian Naive Bayes classifier

The Gaussian Naive Bayes classifier (GNB) has shown the metrics accuracy score of 48.9 and the classification report details are as follows.



The GNB f1 score shows a weighted average of 0.32 and could not indicate any scores for FALSE POSITIVE and CANDIDATE classes similar to KNN classifier.

Support Vector Machine

The support vector machine model (SVM) classifier has shown an accuracy score of 45.56 and the classification report details are below.



The f1 score in this model has shown as 0.32 with FALSE POSITIVE and CANDIDATE classes being recognized at scores just greater than zero with CONFIRMED classes scoring higher.

CONCLUSIONS

The logistic regression classifier scored and classified the Exoplanet target classes better followed by the Support Vector Machine (SVM) model.