INTRODUCTION:

The NASA Kepler space telescope has been out on a planet-hunting mission to discover hidden planets outside of our solar system. This data collected for over nine years of period is being used to process and predict scores.

PROJECT DETAILS:

Initially, the raw dataset was preprocessed and removed unnecessary columns and dropped null value rows. Then the data was scaled using MinMaxScaler followed by splitting the data into test and train data to create a machine learning model.

Models and their scores:

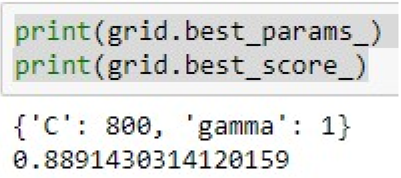
- Support Vector Machine Model (SVM):

We used SVM to fit the model and found the scores as below:

```
Training Data Score: 0.8502592253735896
Testing Data Score: 0.838975297346752
```

- ➢ GridSearchCV to tune the model

Later, we used GridSearch Model to tune the parameters (C and Gamma) to see improvements in prediction scores. See the scores below for better scores at gamma:1 and C: 800:

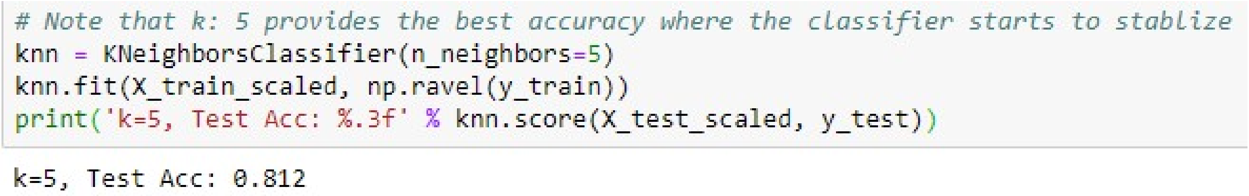


```
print(grid.best_params_)
print(grid.best_score_)
```

```
{'C': 800, 'gamma': 1}
0.8891430314120159
```

- ➢ To evaluate further, we used Random Forest classifier and K-Nearest Neighbor Models and was able to see slightest improvement in scores of Random Forest.
    - Random Forest Model showed a result of 89.56 % of score. But, when a feature “koi_time0bk_err2” is dropped, it showed a slightly better score of 89.70%
    -

    (can be considered negligible).
    - KNN did not prove to be a better model as it resulted in less scores than SVM.



```
# Note that k: 5 provides the best accuracy where the classifier starts to stablize
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train_scaled, np.ravel(y_train))
print('k=5, Test Acc: %.3f' % knn.score(X_test_scaled, y_test))
```

```
k=5, Test Acc: 0.812
```

OBSERVATIONS:

- When tried to compare other models with SVM, though Gridsearch has shown improvement with tuned parameters, Random Forest Model proved to be best model than SVM (with more than 5% increase in scores and mostly similar to hyper tuned parameters) and KNN model with more than 8% increase in scores) and GridSearchCV Model.
- Also, we noticed that excluding the weakly correlated features have shown some increase in scores, which suggests that lesser the features (that are weakly or not correlated), more is the accuracy.