Beyond The Horizon: Part_1 Report

Project Overview:

Flux data of various star present in universe is provided from which to predict whether that star have a planet revolving around it or not.

Installation:

- NumPy
- Pandas
- Scikit
- Imblearn

Methodology:

1. First with help of pandas extracted the “labelled_data.csv” data into jupyter notebook after that separating “LABEL” and rest of the data. Forming Train and test dataset from give data.
2. After that going through given data, I have observed that in 5087 instances provided in which 5050 instance are label 1 and 37 instance are of label 2 which show that the give dataset is imbalanced dataset which need to be handle before putting it into the Machine Learning model.
3. To handle the imbalanced dataset, use the SMOTE (Synthetic Minority Over-sampling Technique) which is oversampling technique that is applied to training set to create synthetic samples for minority class in these it was the Label 2 and this thing was import from imblearn.
4. Since the huge amount of data, various model was getting overfitted to apply constraint Linear Regression model was used train the linear regression model.
5. After training the model threshold value “1.4” is used to distinguish the value of “2” and “1”.