CSS 677- Data Science with Python Final Project

Understanding Exoplanets with Data Science

Prakhar Varshney - [prakhar@bu.edu](mailto:prakhar@bu.edu)

**Introduction**

Understanding Exoplanets with Data Science is about testing efficiency of Machine learning Classifiers on highly imbalanced data such as Kepler-planet-catalog and Nasa’s Exoplanet Archive.

The Project consists of two parts

1. **Predicting the presence of Exoplanets**
2. **Predicting the Habitability of confirmed Exoplanets**

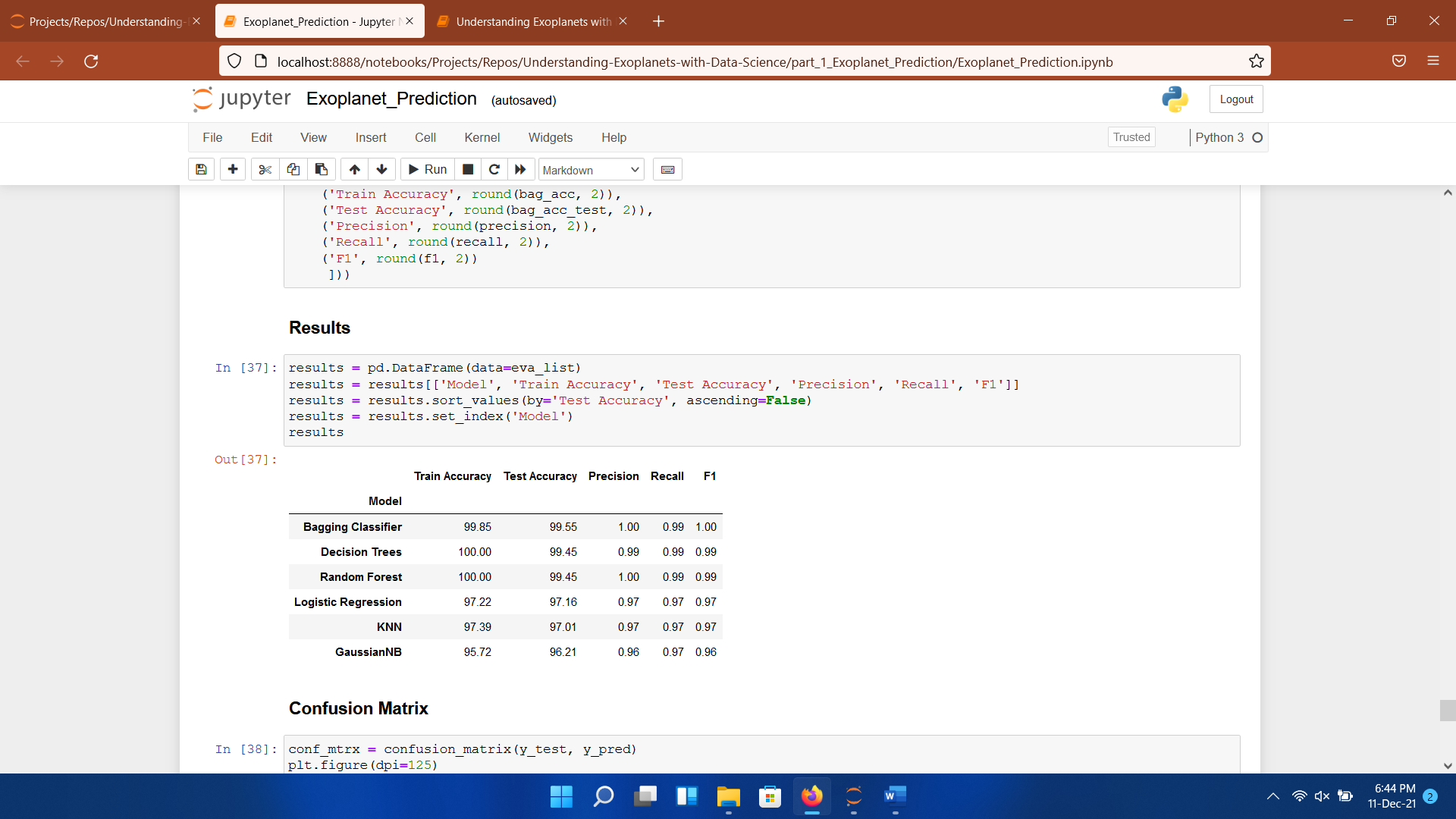
**Part 1: - Predicting the presence of Exoplanets**

For this part, I have used Kepler’s Cumulative Planet archive -2021 which contains all the potential and confirmed planets with their parameters.

After filtering data for missing values and preprocessing we test our dataset for prediction in several Classifiers such as

* Bagging Classifier
* Decision Trees
* Random Forest
* Logistic Regression
* KNN
* Naïve Bayesian

We get the Following Results: -

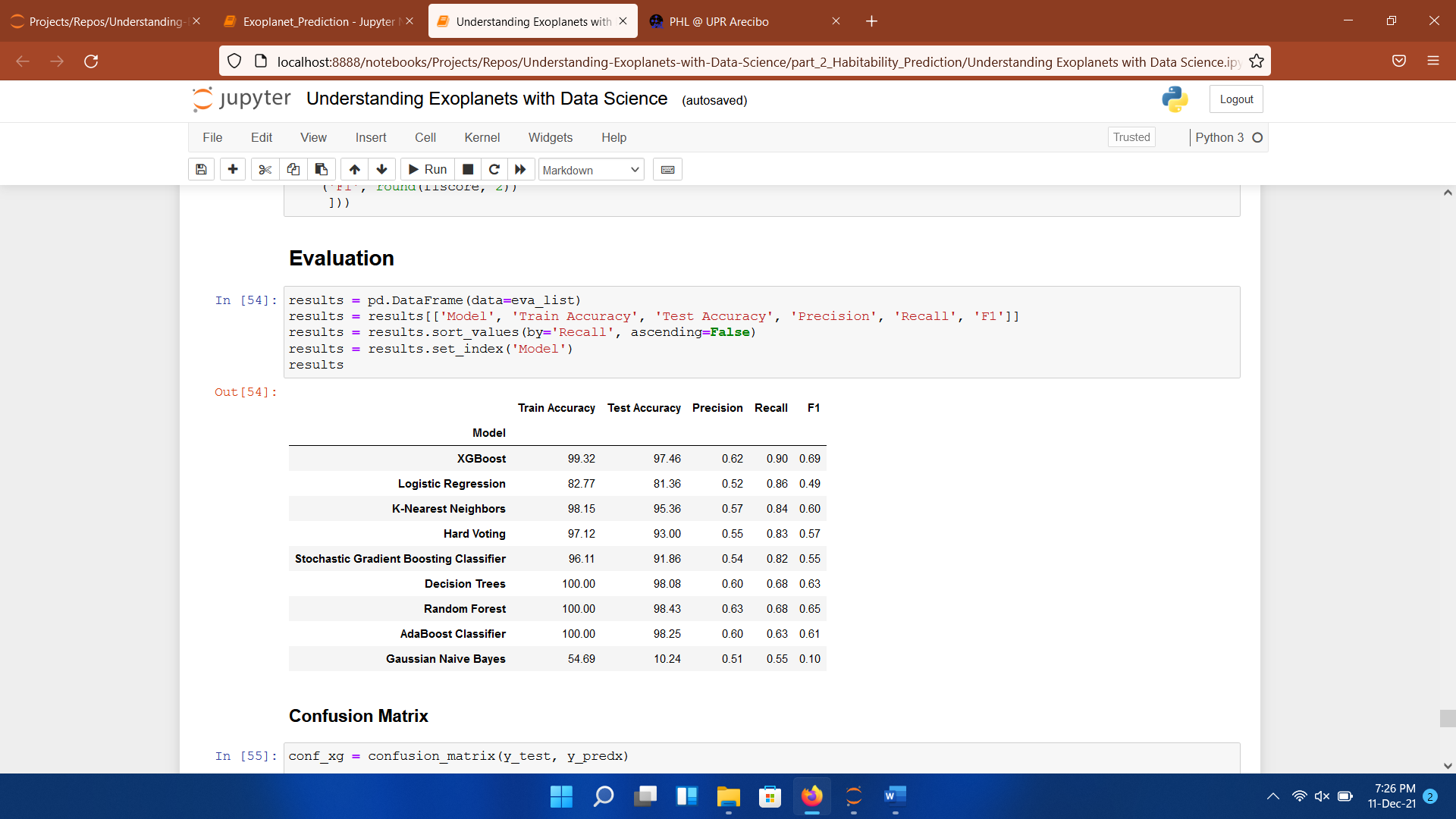


**Part 2: - Predicting the Habitability of confirmed Exoplanets**

To Predict the habitability of exoplanets we have merged two datasets which are 1. Nasa’s Exoplanet Archive 2. PHL (Planet Habitability Laboratory) Catalogue.

To fix missing data I have used Imperative Imputation and after that I have tested the data against the various Classifiers.

Here is the accuracy result



**Results: -**

**For Planet Detection Confirmation Bagging Classifier/ Decision Trees works the best**

**For Habitability Prediction XGBoost is the best Classifier.**

**How to Run: - The project only requires some packages which are used for analysis and visualization**

* **In the project Folder**
* **run: pip install -r requirements.txt**
* **It will install all the dependencies required for this project**
* **All the data files are separately stored in “data” folder for each part.**