CLASSIFICATION OF EXOPLANETS

Decision Tree Model

Problem Statement

The data in this model comes from the Planetary Habitability Laboratory's expanded version of NASA's catalogue of confirmed exoplanets. The problem statement included classifying the exoplanets into their respective classes of habitability (hypopsychroplanet, psychroplanet, mesoplanet and thermoplanet). The dataset included characteristics of each planet such as temperature, planet mass, planet radius, and the gravitational pull on the surface.

ML Concepts Employed

Data Pre-processing: Data pre-processing refers to the techniques of transforming noisy data i.e. data with inconsistent data values, into clean data that could be used efficiently in a classification model. Our implementation of data pre-processing included removal of attributes with 75% missing values, imputing of missing values with mean, median and mode values and encoding categorical attributes in the dataset.

Decision Tree: A Decision tree is a classification tool that is used to efficiently classify values in the datasets based on the determined splits in the attributes. We employed a decision tree model to help classify the exoplanets. The decision tree model takes more time than most models to complete its classification but also outputs classifications with exceptional accuracy, which is why we could obtain an accuracy of 99%

Confusion Matrix: Confusion matrices are used as accuracy measures to efficiently enumerate the correct and incorrectly predicted values from the model used. The values that are predicted correctly are labelled as True Positives and True negatives and they contribute towards the accuracy of the model.

Results

After running our Decision Tree model on the cleaned dataset, we could obtain a predicted set of values from the test split which showed a 99% accuracy at the output.

```
Predicted values:
['non-habitable' 'non-habitable' 'non-habitable' ... 'non-habitable'
 'non-habitable' 'non-habitable']
Confusion Matrix:  [[   0    0    0    1]
 [   0   11    0    0]
 [   0    0 1144    0]
 [   0    1    0    6]]
Accuracy :  99.8280309544282
Report :                    precision    recall  f1-score   support

hypopsychroplanet       0.00      0.00      0.00         1
       mesoplanet       0.92      1.00      0.96        11
    non-habitable       1.00      1.00      1.00      1144
    psychroplanet       0.86      0.86      0.86         7

        micro avg       1.00      1.00      1.00      1163
        macro avg       0.69      0.71      0.70      1163
     weighted avg       1.00      1.00      1.00      1163
```

Figure 1: Confusion Matrix

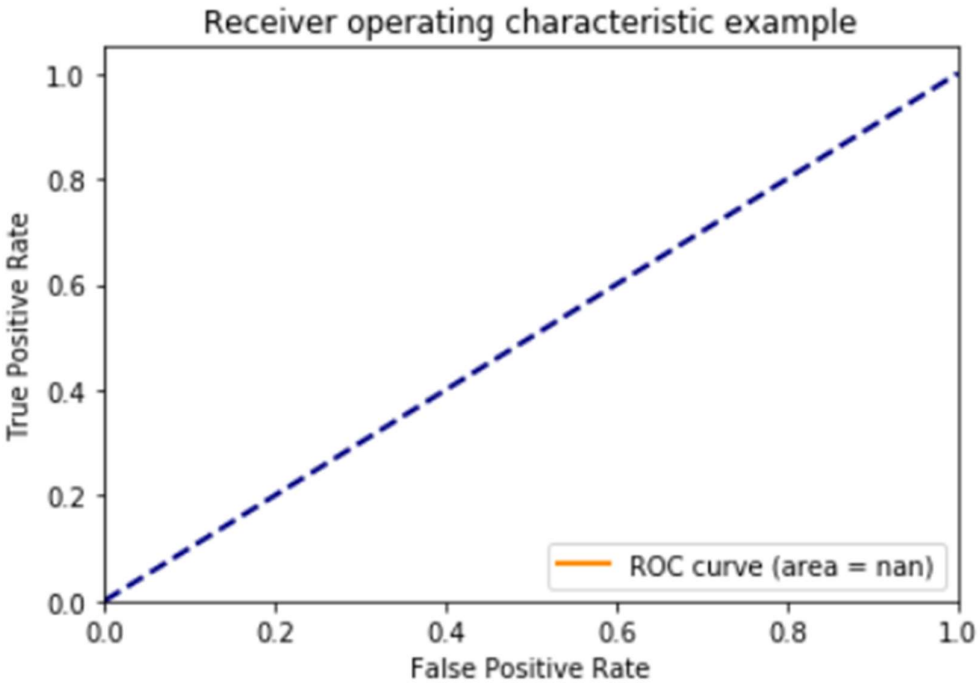

Figure 1: Area-under-Curve Graph

Analysis

After careful analysis of the predicted values outputted from the model, we were in a position to make a few deductions regarding the dataset.