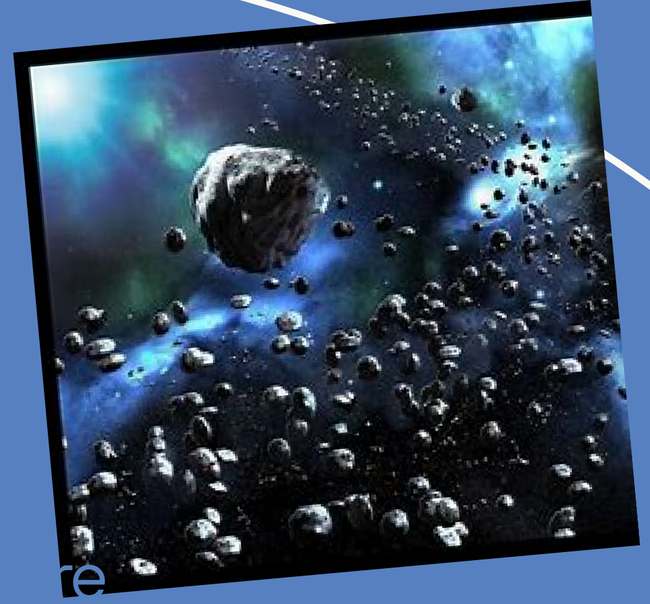
TROUBLESOME

PALLAVI
AABHA
SAEE



AIM :

TO DEVELOP A MACHINE LEARNING MODEL WHICH CAN TAKE THE GIVEN FEATURES OF THE DATASET TO PREDICT IF AN ASTEROID IS HAZARDOUS OR NOT

PROCESS

1.Data Preprocessing and Handling Missing Values:

i. First we dropped columns with more than two third NaN values as can cause noise.

ii. Visualized correlations by correlation matrix between features and drooped more columns like : 'pha', 'name', 'data_arc', 'q'

iii. Filled remaining NaN values:

a. Numerical: Filled with mean.

b. Categorical: Filled with mode.

iv. Categorical Encoding:

Encoded categorical variables using label encoding.

2. Train - Test Split:

Split whole given data set into training and test sets in ratio of 7:3 ,ensuring representation of all categories to maintain categorical features amomg both

3. Feature Scaling:

I applied standard scalar to standardize numerical features

4. Model training and Performance Evaluation:

We tried following models and obtained following results :-

Model name	Training Accuracy	Test Accuracy	ROC-AUC Score	Cross -Validation score	Run Time(s)
Naive bayes	98.20%	98.21%	1	99.96%	57
SVM	99.82%	98.82%	1	99.96%	900
Random Forest	100%	99.99%	1	99.96%	120
XGboost	100%	99.97%	1	99.96%	65

5. Model Comparision:

We first eliminated naive bayes and svm due to their less accuracy.

Accuracy of random forest and xgboost was almost same but since run time of random forest is twice than that of xgboost so we finally decided to go with xgboost.

6. Model Application

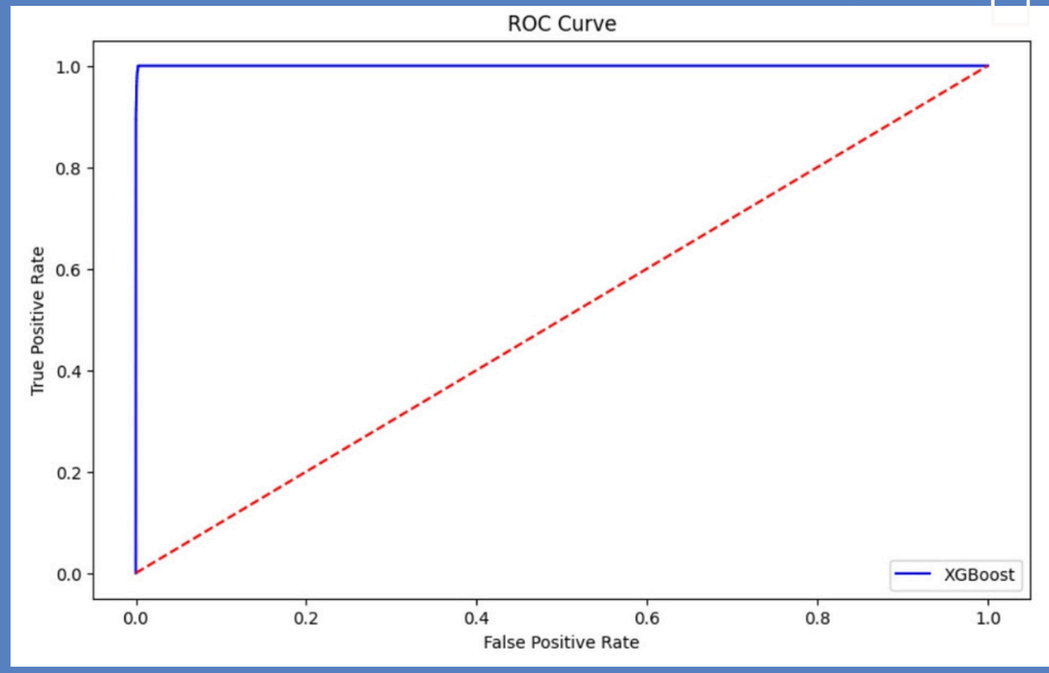
We saved the model for train.csv using joblib and then used the same for predicting pha values for test.csv file and put the predictions in test_predictions.csv file.

RESULT (XGBOOST):

1. Printed confusion matrix: for detailed breakdown of classification outcome.

```
Confusion Matrix:
[[221734      5]
 [    62    483]]
```

2. Plotted ROC curve area under which is equal to 1.



3.	Training Accuracy	100%
	Test Accuracy	99.97
	ROC-AUC Score	1
	Cross -Validation score	99.96%

THANK *YOU*