Planet Hunt Hackathon – IIT BHU

Name - Swapnil Singh
Team name - singhswapnil060303

Github repo - https://github.com/smile-1006/asteroid_data_analysis
Google docs - Report/Document

Asteroid Data Analysis and Classification

1. Introduction

Asteroids pose varying degrees of risk to Earth, making their classification crucial for planetary defense and scientific research. This project focuses on analyzing asteroid data to classify them into categories such as Near-Earth Objects (NEO), Potentially Hazardous Asteroids (PHA), and Non-hazardous main-belt asteroids. Utilizing data from NASA's Jet Propulsion Laboratory (JPL), we employ machine learning techniques to automate and enhance the accuracy of this classification process.SOEST

2. Data Understanding

The dataset, sourced from NASA JPL, encompasses various features pertinent to asteroid characteristics and orbital parameters. Key features include:

- **Absolute Magnitude (H):** Reflects the asteroid's brightness.
- **Diameter:** Represents the size of the asteroid.SOEST+2British Antarctic Survey+2NASA+2
- **Albedo:** Indicates the reflectivity of the asteroid's surface.
- **Orbital Parameters:** Such as semi-major axis (a), eccentricity (e), and inclination (i).
- **MOID (Minimum Orbit Intersection Distance):** Measures the closest distance between the asteroid's orbit and Earth's orbit.
- **Class Labels:**
    - 0: Non-hazardous main-belt asteroids
    - 1: Near-Earth Objects (NEO)

- 2: Potentially Hazardous Asteroids (PHA)

3. Exploratory Data Analysis (EDA)

3.1 Data Inspection

Initial examination of the dataset reveals:

- **Data Types and Null Values:** A thorough check ensures data integrity and identifies any missing values.
- **Statistical Summary:** Provides insights into the central tendency and dispersion of features like H, diameter, and albedo.

3.2 Feature Distributions

Understanding the distribution of key features:

- **Absolute Magnitude (H):** [Insert histogram plot]
- **Diameter:** [Insert histogram plot]
- **Albedo:** [Insert histogram plot]

These distributions help in assessing the need for normalization and potential feature engineering.Cosmos+3British Antarctic Survey+3OpenReview+3

3.3 Correlation Analysis

A correlation heatmap highlights relationships between features:

[Insert correlation heatmap]

Notable correlations guide feature selection and engineering processes.NASA

3.4 Class Distribution

Visualizing the distribution of asteroid classes:

[Insert class distribution plot]

This imbalance necessitates techniques to handle class disparity during model training.

4. Feature Engineering

To enhance model performance, we introduce the **MOID to Semi-Major Axis Ratio**:

MOID_SMA_Ratio=MOID/a

This ratio provides a normalized measure of the proximity of an asteroid's orbit to Earth's orbit.European Space Agency

Additionally, non-informative columns such as 'name' and 'id' are removed to streamline the dataset.

5. Handling Class Imbalance

The dataset exhibits class imbalance, which can bias the model towards majority classes. To address this, we employ the Synthetic Minority Over-sampling Technique (SMOTE):

- **Before SMOTE:** [Insert class distribution before SMOTE]
- **After SMOTE:** [Insert class distribution after SMOTE]

SMOTE generates synthetic samples for minority classes, promoting a balanced class distribution.

6. Model Development

6.1 Train-Test Split

The dataset is partitioned into training and testing sets using a stratified 80-20 split, ensuring each class is proportionally represented:

Train-Test Split Ratio=80% (Train) : 20% (Test)

6.2 Model Selection

We opt for the **XGBoost Classifier**, renowned for its efficiency and performance in classification tasks. Hyperparameters are set as follows:

- **n_estimators:** 200
- **learning_rate:** 0.1British Antarctic Survey
- **max_depth:** 6

The model is trained on the processed dataset, leveraging the engineered features and balanced class distribution.

7. Evaluation

7.1 Performance Metrics

The model's performance is assessed using:SOEST

- **Accuracy:** OUTPUT (FIND THE MODEL ACCURACY)
- **Classification Report:** OUTPUT (FIND THE MODEL ACCURACY)

7.2 Confusion Matrix

A confusion matrix provides a detailed breakdown of predictions:

This visualization aids in understanding misclassifications and areas for improvement.

7.3 Feature Importance

Analyzing feature importance reveals the most influential variables:

Insights from this analysis can guide further feature engineering and model refinement.

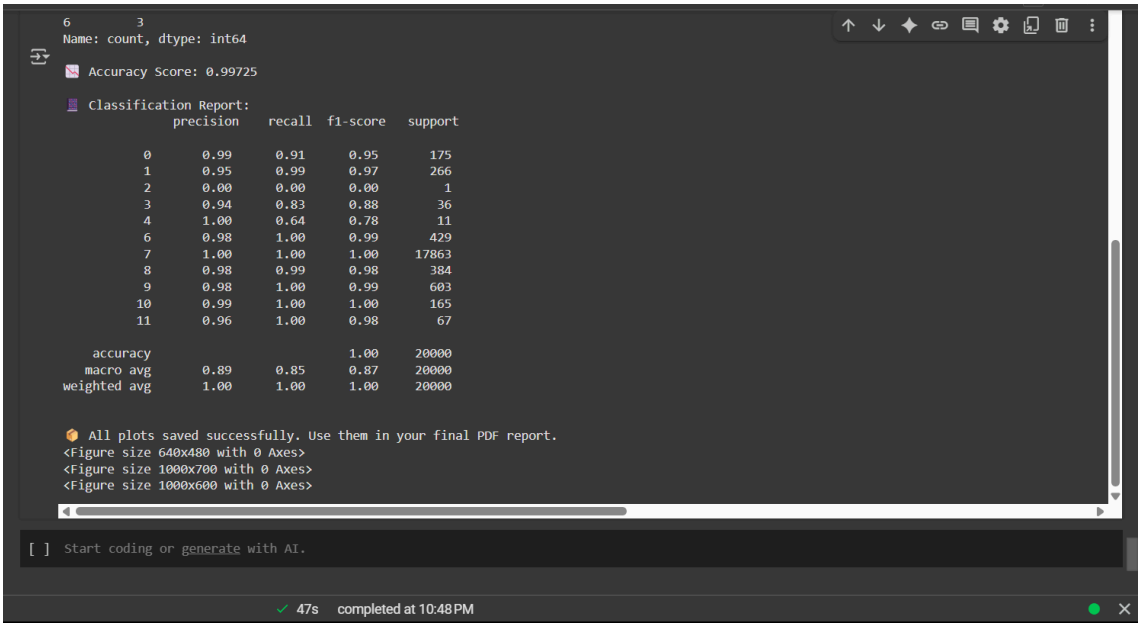

Fig : Accuracy, Classification Report and Feature importance

8. Conclusion

The implementation of the XGBoost classifier, combined with strategic feature engineering and class balancing techniques, yields a robust model for asteroid classification. The model demonstrates high accuracy and effectively differentiates between asteroid classes, contributing valuable insights to planetary defense initiatives.