Neshat Jalali Heravi December 2021
Metis – Classification Project Summary

# Binary Classification of Asteroids:

# Is the Asteroid Coming Towards Earth Hazardous?


## Abstract

As of November 2021, there are 2,223 known Potentially Hazardous Asteroids (PHAs) (about 8% of the total near-Earth population), of which 160 are estimated to be larger than one kilometer in diameter! [1] If there's a need for intervention, an enormous budget and planning is needed to stop an asteroid from hitting Earth. The resources needed depend on the characteristics of the asteroid. The objective of this project is to explore the data collected from the NASA open API and build a classification model to be able to identify if an asteroid is hazardous or not! NASA and government agencies can use this information as part of their planetary-defense strategies.

After feature engineering, handling class imbalance, and tuning hyperparameters, two models were selected based on their performance metrics:

- Decision Tree Classifier model with 99.47 % accuracy, 97.96 % Recall, and 0.99 AUC score.
- XGBoost Classifier model with 99.47% accuracy, 98.64% Recall, and 0.99 AUC score.

The top important features according to the predictive models in this study are:

- "Est Dia Miles(min)" - minimum estimated diameter of the asteroid in miles.
- "Minimum Orbit Intersection" - MOID is defined as the distance between the closest points of the osculating orbits of two bodies.


## Design

This analysis seeks to establish a predictive classification model for identifying "hazardous" asteroids using features such as its mass, velocity, and some orbital characterizations. The goal is to select a meaningful classification model with acceptable accuracy, recall, precision, and AUC score to provide insights to NASA and other federal agencies in charge and help identifying the hazardous asteroids.

### Data

The data about asteroids for this project has been collected from the NASA open API and available on Kaggle. The target is whether an asteroid is hazardous or not (True or False). The feature columns include asteroid's speed, some dimensions, and other orbital information. After cleaning and preliminary EDA on data and handling categorical variables, the set used for modeling contained 4,687 rows and 87 features. Each row represented an asteroid.

Neshat Jalali Heravi December 2021
Metis – Classification Project Summary

## Algorithms

- Feature Engineering: A preliminary EDA was performed to look for missing and null values. Target's binary categorical values were label encoded and other categorical features were converted to dummy variables. Feature Correlation heatmaps and Variance Inflation Factor (VIF) were used to select the best numerical features.
- Modelling: Logistic Regression, SVM, Decision Tree, Random Forest and XGBoost models were tested before Decision Tree and XGBoost were selected as the best performing models based on the designated performance metrics. Class imbalance was handled using SMOTE (synthetic minority oversampling technique), and GridSearchCV was used for hyperparameter tuning.
- Model Evaluation: The entire dataset was split to train and test sets (80/20). The models were evaluated based on their performance metrics including accuracy, recall, precision, F1 score and ROC-AUC curve and AUC score.

## Tools

- Numpy and Pandas Python libraries for data manipulation
- Scikit-learn Python package for statistics and modelling
- Matplotlib and seaborn libraries to create visualization plots

## Communication

You can find the presentation slides and details on the analysis, and the notebook on my GitHub.

## References

**[1]** Reference 1