Analysis of Near-Earth Asteroids

Data Analytics (UE19CS312) Final Project

Presented by: Vishisht Rao, Sahil Gupta & Sai Mihir J

PROBLEM STATEMENT

- Performing the analysis of NASA's JPL Asteroid dataset for the detection of Near-Earth Asteroids and how certain parameters they possess classify them as Potentially-Hazardous Asteroids.

ABOUT THE DATASET

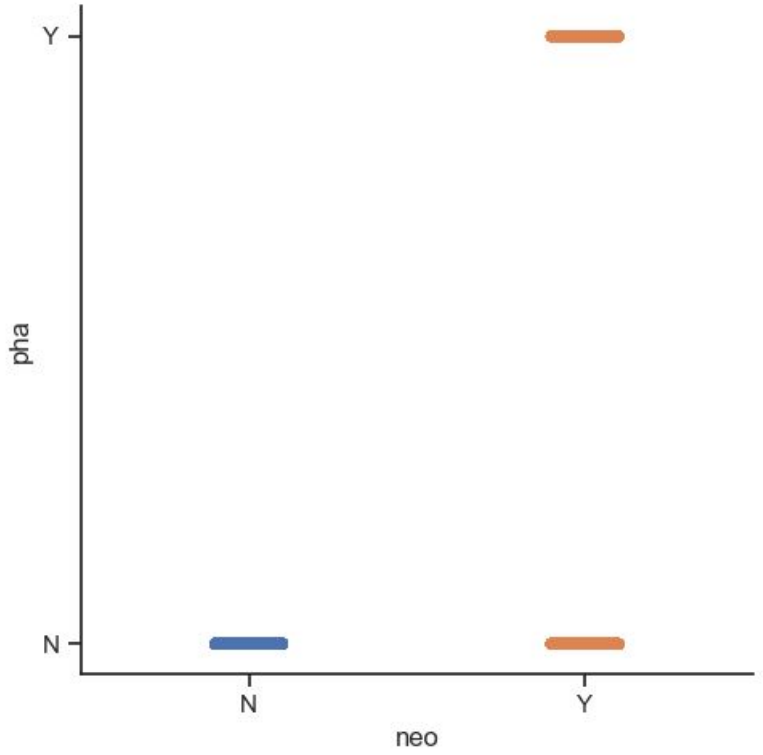
- **JPL is a federally funded research and development center managed by Caltech for NASA**. It is responsible for exploration of space, other planets, the sun, etc. and aims to understand our place in this universe and to search for the possibility of life beyond earth. The dataset that has been selected is **NASA's Jet Propulsion Laboratory of California Institute of Technology**.
- The entire dataset has been **extracted directly from the JPL website and all JPL-authored documents are sponsored by NASA under Contract NAS7-030010**.
- It contains **958,524 unique values** which correspond to characteristics of as many asteroids. It has **45 columns** of which some important ones are **NEO flag, PHA flag, absolute magnitude (H), diameter, geometric albedo, eccentricity, perihelion distance, inclination angle with respect to x-y ecliptic plane, minimum orbit earth intersection distance**, etc.

IMPORTANCE OF OUR PROJECT

- Some noteworthy asteroid impacts on Earth involve the collision 65 million years ago which caused a global firestorm, a global warming that eventually led to the extinction of dinosaurs. In 1490, 10,000 people were killed in the Chinese city of Chiling-Yang when an asteroid broke overhead. In 1908, an asteroid exploded 50 meters above the surface of the Earth above Tunguska, Siberia resulting in the loss of 1000 reindeers and trees over the span of 2000 square kilometers.
- Considering that such severe outcomes are possible in the case that an asteroid collides with the Earth, it is **important to explore the possible preventions of such an event**. In order to prepare oneself best in the case of an asteroid impact, it is **crucial to accurately detect Potentially Hazardous Asteroids(PHAs) as early as possible**. Hence, the aim of this project is to **clearly determine which classification model works best** in terms of accuracy and performance to **predict whether a given asteroid is potentially hazardous or not**.

HOW DID WE APPROACH THE PROJECT?

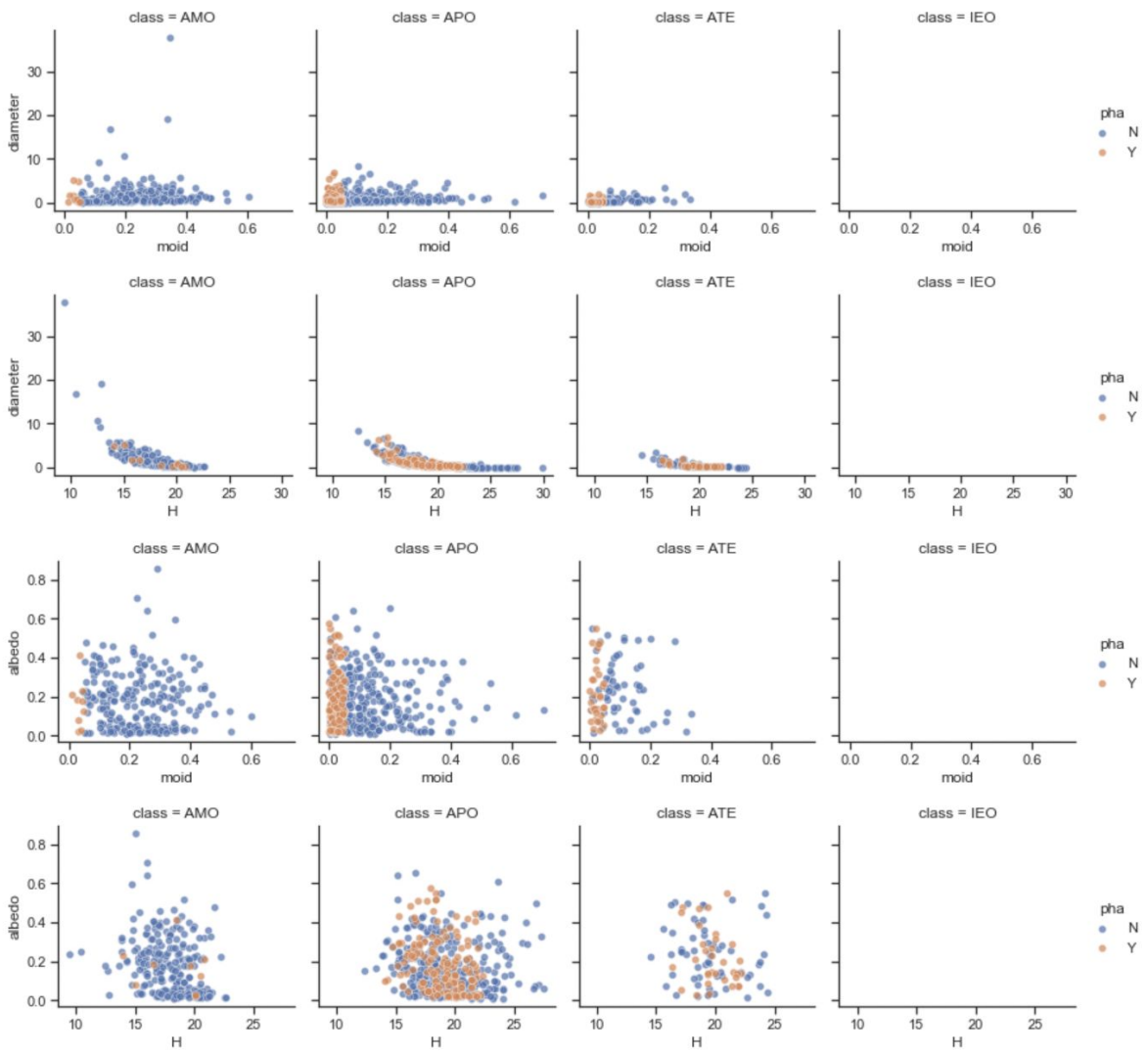
- We first **analyzed the proportion of Near-Earth Asteroids being Potentially Hazardous**, and we noticed that every Near-Earth Asteroid accounted as Potentially Hazardous.
- **A near-earth asteroid is considered potentially hazardous** based on 2 parameters:
  - **Minimum orbit earth intersection distance (moid) < 0.05**, and
  - **Absolute magnitude (H) < 22.0**

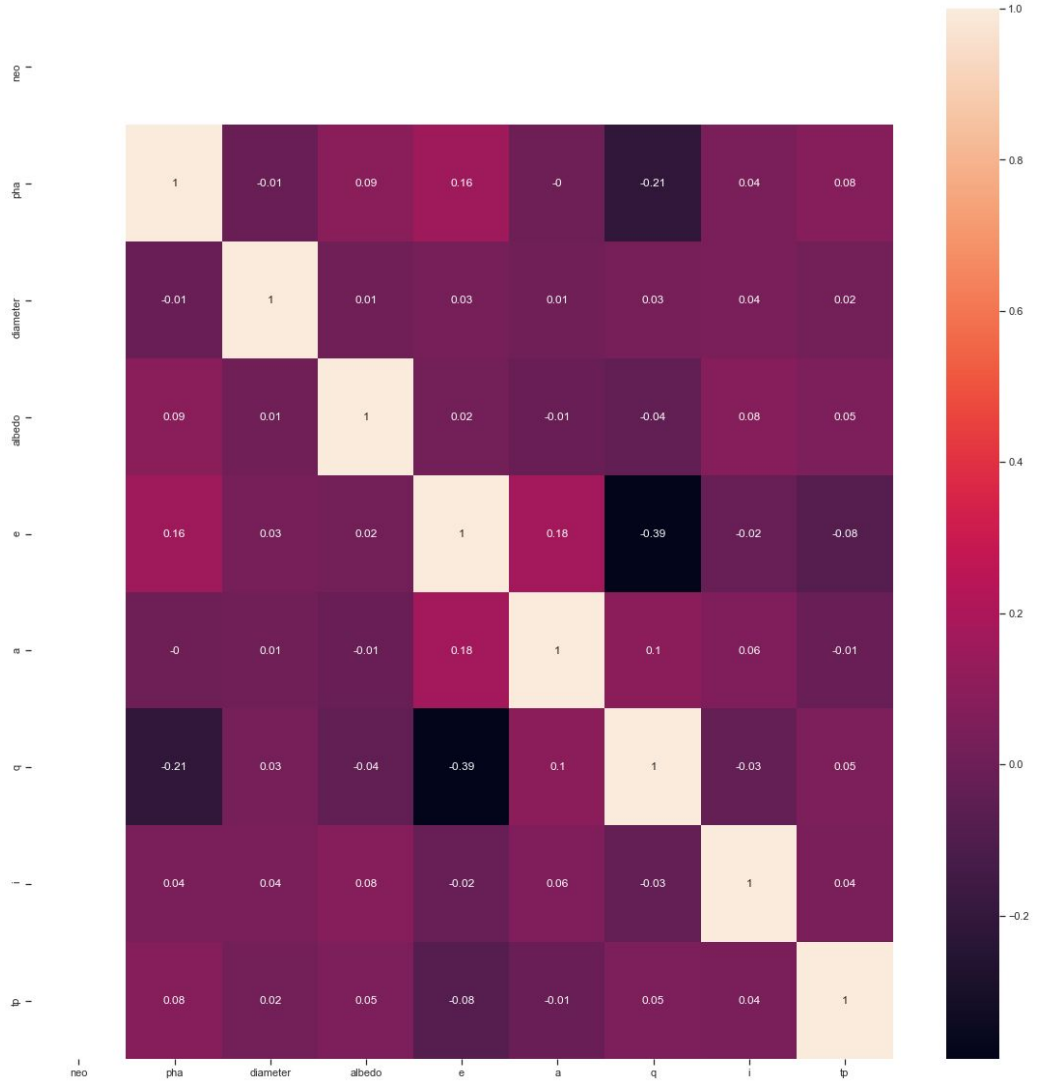


HOW DID WE APPROACH THE PROJECT? (Contd.)

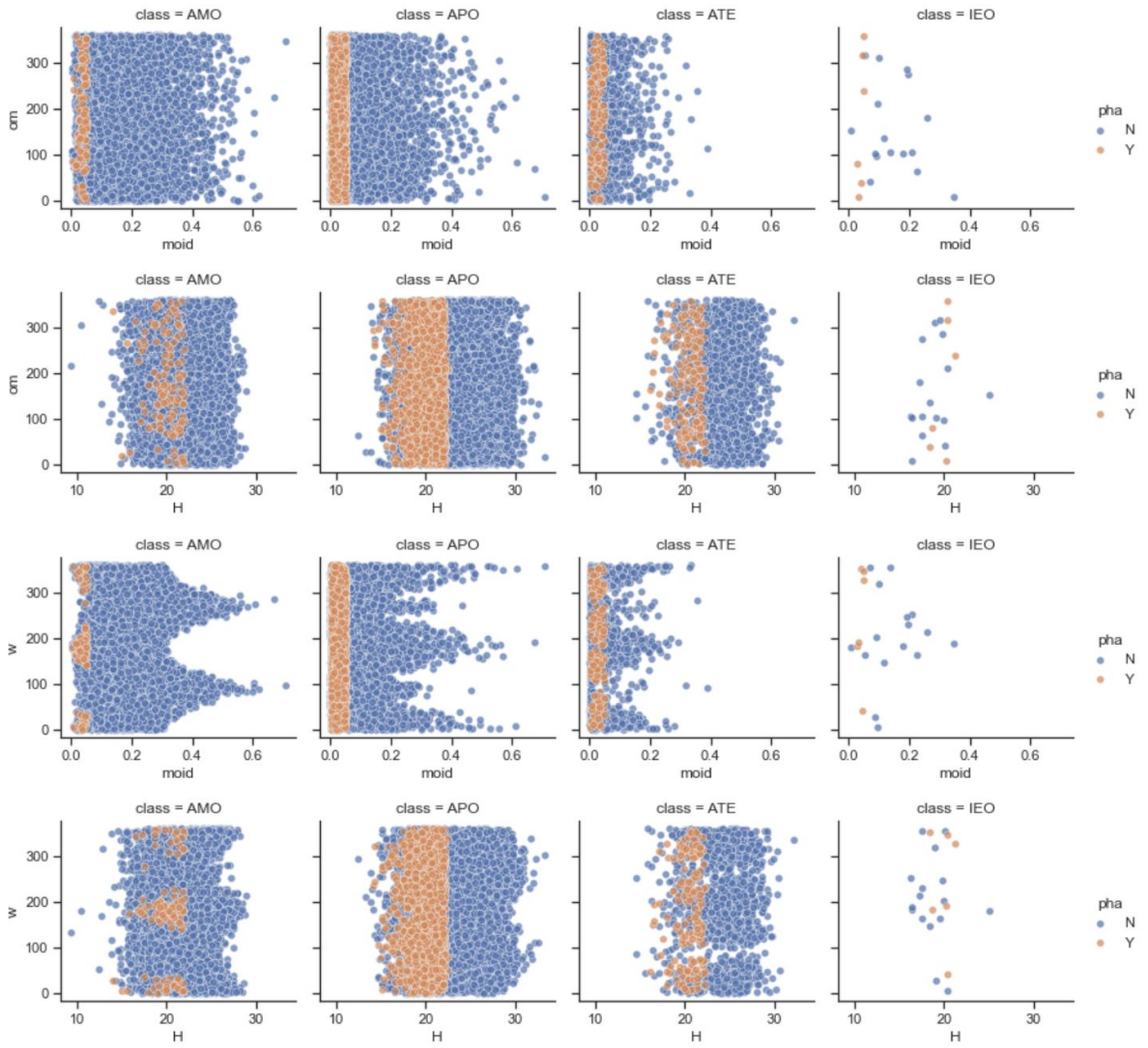
1. We then split the parameters into 2 parts: one being what we felt were important and the others which we felt didn't make much of the impact. We compared both the parameter sets with the moid value and H value to see how they affected these values for each classes of asteroids.
2. Once we split the important and non-important parameters into different datasets, we ran it under different prediction models and chose the top 4 models that gave us the better precision score. We also used a neural network to see how it would perform against the top 4 models we ran both the datasets in.
3. Lastly, using feature importance, we extracted the importance of each feature in both the datasets to see how much they affected the behaviour of the near-earth asteroid.

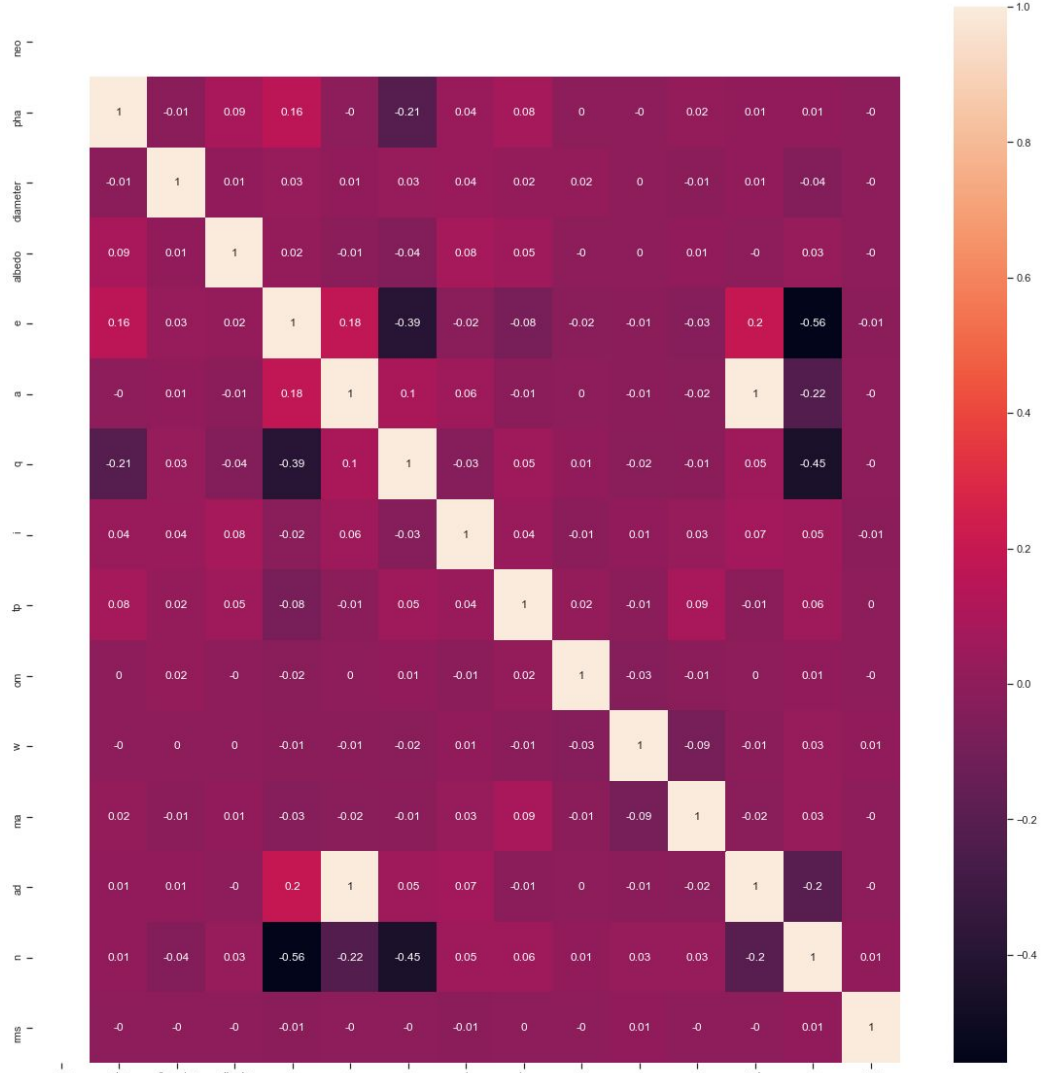
```
param_imp = ['diameter', 'albedo', 'e', 'a', 'q', 'i', 'tp']
```





```
param_oth = ['om', 'w', 'ma', 'ad', 'n', 'rms']
```





MODEL SELECTION

- For dataset with **important parameters**

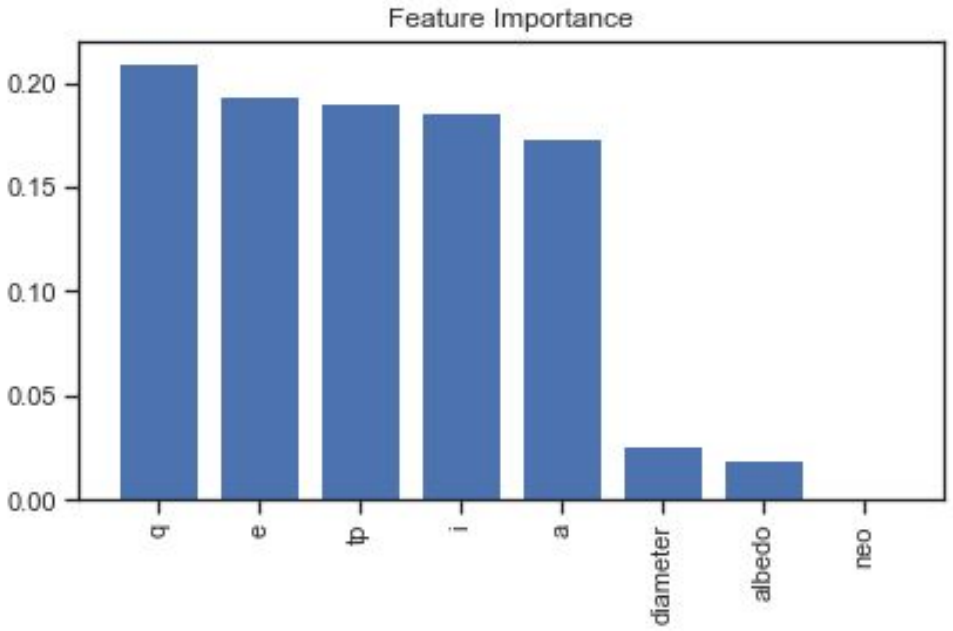
	Score
RandomForestClassifier	0.914173
SVC	0.913955
GradientBoostingClassifier	0.913737
LogisticRegression	0.911990
AdaBoostClassifier	0.910898
KNeighborsClassifier	0.902817
DecisionTreeClassifier	0.865036

- For dataset with **all the parameters**

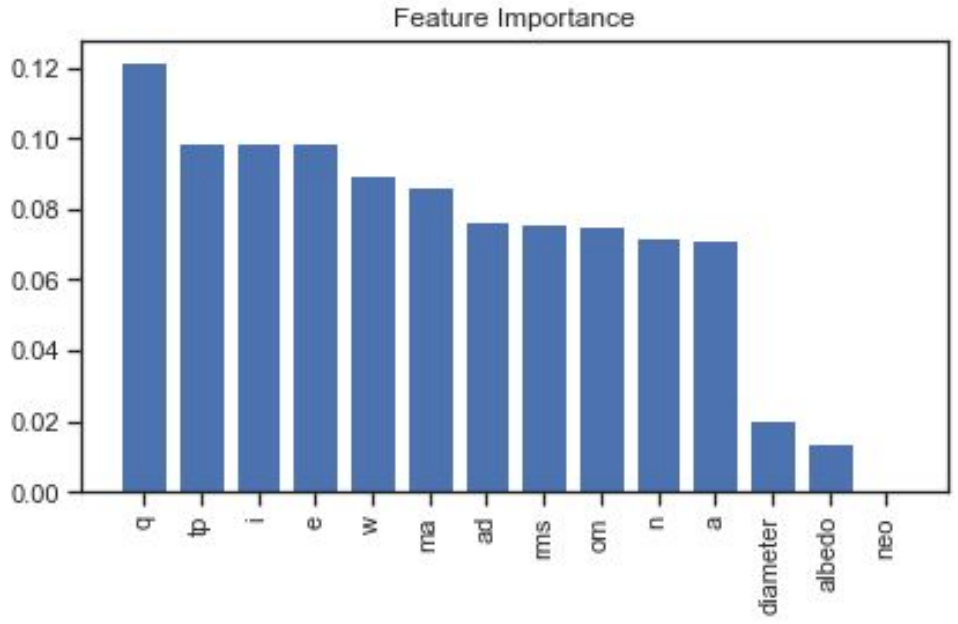
	Score
RandomForestClassifier	0.915921
GradientBoostingClassifier	0.913737
SVC	0.913300
LogisticRegression	0.911334
AdaBoostClassifier	0.910898
KNeighborsClassifier	0.904128
DecisionTreeClassifier	0.868749

FEATURE IMPORTANCE EXTRACTION

- For dataset with **important parameters**



- For dataset with **all the parameters**



CONCLUSIONS

1. **For the Model Selection:** We concluded that among the top 4 models chosen, **Random Forest Classifier (RFC)** gave us the highest precision score for both the datasets.
2. **Feature Importance using RFC:** We concluded that our **assumptions of choosing certain parameters as important turned out to be right** as it had a equal influence for both the datasets and overall impact on the influence of near-earth asteroids.
3. **Future Works:** For the dataset that we have chosen, we could perform a better analysis by **testing it with some more non-linear models or maybe neural networks and tweak the code a little bit**. We could also try **applying this to real time data to observe the efficiency of our model** in real world applications.

PROJECT MEMBERS & CONTRIBUTIONS

- **Vishisht Rao (PES1UG19CS572):** Project Idea, Initial Analysis of the dataset, Building the ANN and Model analysis.
- **Sahil Gupta (PES1UG19CS415):** Building the SVC Model and Feature Importance of the parameters.
- **Sai Mihir J (PES1UG19CS418):** Cleaning the dataset, EDA and Visualization of the dataset.

We would like to thank our project supervisor and professor, Dr. Gowri Srinivasa for her constant advice and mentorship throughout the entire course of this project.