Analysis of predicting Potentially Hazardous Asteroids

Using Various Machine learning models

Data Mining

(CCSI 5502)

University of Colorado Boulder

FALL 2023

By

Divya Nallawar
ID: 110988624

Jeet Choksi
ID: 110916558

Tanay Shukla
ID: 110857542

TABLE OF CONTENTS

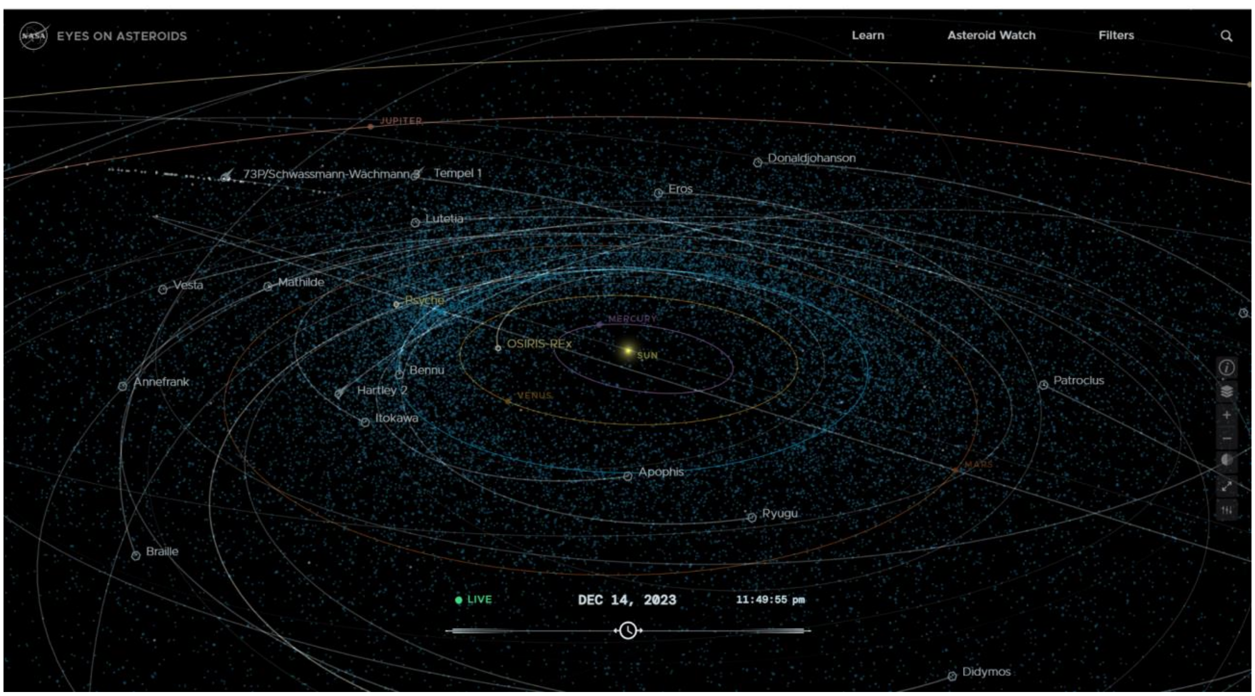
Abstract

Space consists of a lot of matter. Asteroids[1], the small rocky, airless leftovers from the formation of the solar system, are one of the matters that possess potential threat to the earth. This potential threat that might impact earth motivated us to analyze different predictive models and find their accuracy. This project executes different machine learning models to understand their accuracy. The main goal is to determine the precision and effectiveness of various machine learning models in predicting possible effects, which will support proactive planetary defense measures.

Introduction

Asteroids are small, rocky celestial bodies that orbit the Sun, primarily found in the asteroid belt between Mars and Jupiter. They vary in size from small rocks to large bodies several hundred kilometers in diameter. These rocky bodies are not large enough to distinguish as planets. These remnants from the early formation of the solar system come in different compositions, ranging from metallic to rocky or carbonaceous.

The depiction of asteroids in space is shown by the fig.(1). [2]



[1] “Asteroids - NASA Science.”
[2] “Eyes on Asteroids - NASA/JPL.”

Fig. 1 Asteroids in space

Near-Earth Objects (NEOs) are a subset of asteroids or comets whose orbits bring them close to Earth's orbit. They are classified based on their distance from Earth and the potential risk they pose. NEOs come within 1.3 astronomical units (AU) of the Sun and hence within 0.3 astronomical units(AU), or approximately 45 million kilometers, of the Earth's orbit, this close proximity could potentially affect Earth.

Some of the near earth objects turn out to be potentially hazardous asteroids. To determine if a NEO is a PHA or not, certain conditions are defined. An object is PHA, if its minimum orbit intersection distance (MOID) with respect to Earth is less than 0.05 AU (7,500,000 km; 4,600,000 mi) – approximately 19.5 lunar distances – and its absolute magnitude is brighter than 22, approximately corresponding to a diameter above 140 meters (460 ft). These conditions based on the proximity towards earth and the size makes an object potentially hazardous asteroid.[3]

An article provided by NASA provides a better understanding on the conditions used to determine the PHA. [4]

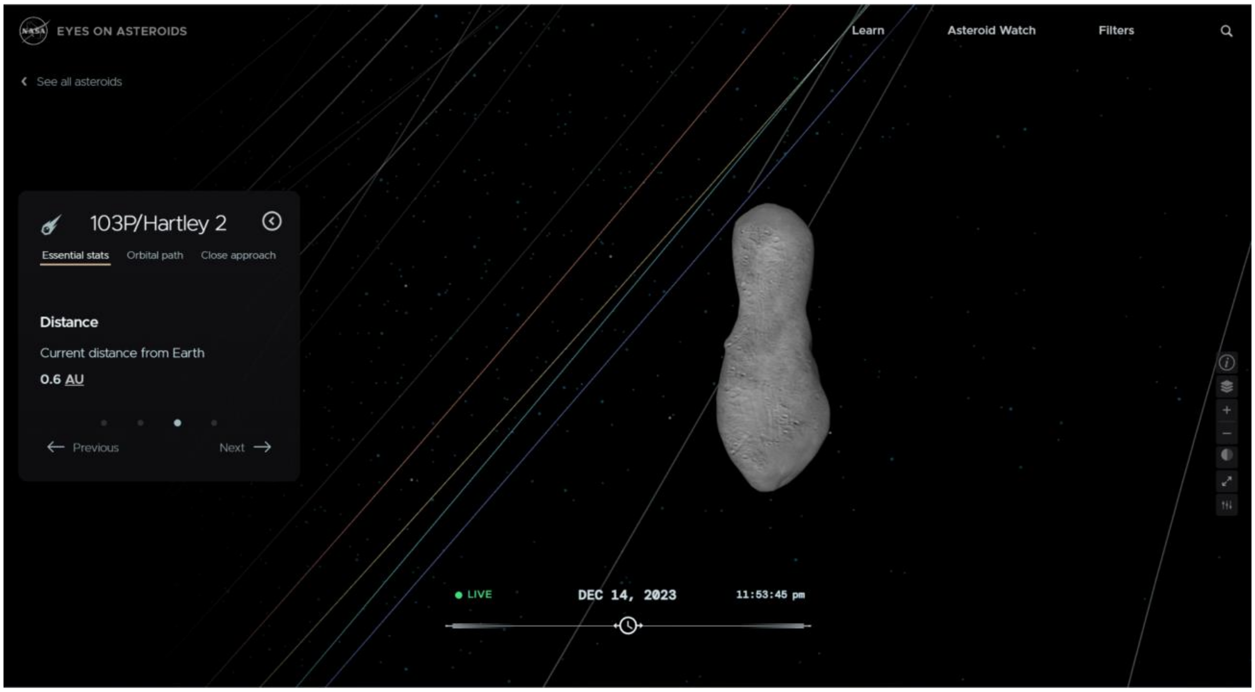


Fig. 2 103P/Hartley 2 Asteroid with current distance from earth 0.6AU

The article[asteroids that hit earth][5], talks about the top 10 asteroids and its specs that hit the earth surface. One of these listed asteroids consists of the Vredefort Crater, which has an estimated radius of 118 miles (190 kilometers), making it the world's largest known impact structure. The crater is believed

[3] Arnold et al., "Analysis of Potentially Hazardous Asteroids."
[4] "NEO Basics."
[5] "Notable Asteroid Impacts in Earth's History."

to have been created over 2 billion years ago. One of the most recent instances involves Sárneczky; in March 2022, a fridge-sized asteroid—about 6 1/2 -feet-long—hit Earth two hours after he initially spotted it.

Looking at one of the recent most mentioned instances, it is crucial to find a way to predict the potentially hazardous asteroids. Asteroids hitting the earth or entering into the earth atmosphere have a huge impact. One of the most recent asteroid impacts is "The Chelyabinsk Event".[6]
In 2013, a 20-meter-diameter asteroid penetrated Earth's atmosphere above Chelyabinsk, Russia. Exploding mid-air, it released energy equivalent to 500 kilotons of TNT. Fortunately, detonation occurred approximately 30 kilometers above the ground, averting direct impact damage. However, a resultant shockwave caused injuries to 1,500 individuals and inflicted damage on 7,200 buildings spanning six cities. The majority of injuries occurred as people, drawn by the bright flash, approached windows to observe. Subsequently, the shockwave, traveling at the slower speed of sound, reached the area and shattered windows, causing harm through flying glass.

The importance of predicting PHAs lies in their potential impact on Earth. By identifying and tracking these celestial objects, we gain crucial time to devise strategies for potential deflection or mitigation. Early detection allows for adequate preparation and, if necessary, implementation of measures to avert potential impacts, safeguarding both human lives and critical infrastructure.

In this project, our objective is to analyze an asteroid dataset and develop a predictive model for Potentially Hazardous Asteroids (PHAs). We aim to explore multiple models to identify the most effective one for predicting PHAs, while also assessing the critical parameters influencing these predictions.

Data Description

The dataset has been sourced from the Datastro.eu website. The dataset consists of nearly 33,360 rows of data, with around 34 columns providing insights of different asteroids.
The dataset contains many important columns which gives out a lot of insight regarding an asteroid. Some of the columns are Absolute Magnitude(H), Slope parameter(G), Argument of perihelion, Semilatus rectum distance (AU), Mean daily motion(n), Aphelion distance (AU), Mean anomaly(M), Orbital eccentricity(e), NEO flag, One km NEO flag.
The NEO flag, One km NEO flag contains binary values.
NEO flag- the value is 1 if the object seems to be a near earth object, otherwise it has bee kept empty.
One km NEO flag- the value is 1 if the object seems to be just one km away from earth surface, otherwise it has been kept empty.

[6] "Asteroid Impacts."

	Last ob	r.m.s. r	Argum	Other designations	Tp	Orbit ty	NEO fla	One km	Epoch (	Referer	Node	Slope p	Name	Perturb	Absolut	Mean a	Numbe	Perturb	Orbital	Uncert	Number of observations	Arc yea	Semi
2	2023-11-0	0.54	178.9078	1956 PC	2460446	Amor	1	1	2460201	E2023-V4	304.286	0.15	Eros	3Ek	10.42	222.7629	57	M-v	1.760714	0	14163	1893-202	1.45
3	2022-08-0	0.57	156.2302	2000 JW8	2459956	Amor	1	1	2460201	E2023-V0	183.8539	0.15	Albert	3Ek	15.59	56.29428	21	M-v	4.280747	0	2024	1911-202	2.6
4	2023-10-1	0.83	26.64832		2459867	Amor	1	1	2460201	E2023-TI3	171.3194	0.15	Amor	3Ek	17.38	123.5335	21	M-v	2.658518	0	585	1932-202	1.9
5	2023-07-1	0.6	31.4379		2460009	Apollo	1	1	2460201	E2023-P11	87.95271	0.15	Icarus	3Ek	16.59	168.6494	40	M-v	1.119375	0	1696	1949-202	1.07
6	2023-11-0	0.7	159.5887		2460736	Amor	1	1	2460201	E2023-V4	62.23069	0.15	Betulia	3Ek	14.69	198.1138	30	M-v	3.257587	0	1466	1950-202	2.19
7	2023-10-2	0.47	276.9536	1983 CY3	2460192	Apollo	1	1	2460201	E2023-UJ4	337.1719	0.15	Geograph	3Ek	15.27	5.84271	41	M-v	1.390314	0	8133	1951-202	1.24
8	2023-10-3	0.65	286.0512		2460143	Apollo	1	1	2460201	E2023-UT	35.55766	0.09	Apollo	3Ek	16.09	31.77158	35	M-v	1.783685	0	2848	1930-202	1.47
9	2021-10-2	0.65	325.2963		2460227	Apollo	1	1	2460201	E2023-D5	212.8889	0.15	Cerberus	3Ek	16.79	337.0999	26	M-v	1.122251	0	2117	1971-2021	1.07
10	2004-10-0	0.71	347.7923		2460912	Amor	1	1	2460201	MPO6915	162.9368	0.1	Quetzalco	3Ek	18.05	187.2434	6	M-v	4.057111	1	48	1953-200	2.5
11	2023-06-2	0.63	194.5473		2460489	Amor	1	1	2460201	E2023-V2	188.2843	0.15	Cuyo	3Ek	14.4	269.8173	30	M-v	3.151492	0	3207	1954-202	2.14
12	2023-11-0	0.57	338.4106		2459924	Amor	1	1	2460201	E2023-V2	246.2977	0.15	Anteros	3Ek	15.71	159.5255	21	M-v	1.711179	0	5038	1973-202	1.4
13	2023-05-1	0.39	115.4929		2460521	Amor	1	1	2460201	MPO7453	246.5465	0.15	Tezcatlipo	3Ek	13.87	218.713	27	M-v	2.235432	0	6665	1950-202	1.70
14	2023-07-1	0.63	55.34924		2460325	Apollo	1	1	2460201	E2023-R01	33.0593	0.15	Bacchus	3Ek	17.23	250.3202	21	M-v	1.119495	0	940	1977-202	1.07
15	2021-12-0	0.53	40.08952		2460275	Aten	1		2460201	MPO7453	211.3045	0.15	Hathor	3Ek	20.43	264.7781	12	M-v	0.775065	0	1283	1976-2021	0.84
16	2022-09-1	0.57	38.02418		2459486	Amor	1	1	2460201	MPO7190	167.0916	0.15	Seneca	3El	17.48	175.1913	10	M-v	4.019259	1	83	1978-202	2.52
17	2023-01-1	0.65	154.7892		2459827	Amor	1	1	2460201	E2023-UR	172.0482	0.15	Krok	3Ek	16.17	116.7585	22	M-v	3.153803	0	2270	1981-202	2.15
18	2023-01-2	0.76	63.56324		2460405	Apollo	1	1	2460201	MPO7453	242.466	0.15	Syrinx	3Ek	16.04	307.8761	17	M-v	3.874696	0	543	1981-202	2.46
19	2023-02-0	0.63	316.5836		2461463	Amor	1	1	2460201	E2023-C1	349.9382	0.15	Don Quixo	3Ek	13.07	218.6988	23	M-v	8.809306	0	1337	1983-202	4.26
20	2022-08-1	0.68	296.698		2460312	Apollo	1		2460201	E2023-V0	157.8987	0.15	Vishnu	3Ek	18.53	258.9445	19	M-v	1.09071	0	694	1986-202	1.05
21	2023-09-1	0.59	237.0425	1986 VT7;1987 MB	2460103	Apollo	1	1	2460201	E2023-UJ4	294.3499	0.15	Cuno	3Ek	14.17	34.47466	32	M-v	2.788454	0	3386	1959-202	1.98
22	2022-12-2	0.71	291.8413	1937 CA	2460263	Apollo	1	1	2460201	E2023-D5	311.7702	0.15	Pan	3Ek	17.28	324.6296	25	M-v	1.731779	0	1461	1987-202	1.44
23	2023-08-2	1.54	121.4282		2460240	Apollo	1	1	2460201	E2023-QF	325.5266	0.15	Castalia	3Ek	17.51	324.4221	17	M-v	1.096106	0	449	1989-202	1.06
24	2023-09-0	0.57	52.57874		2460635	Amor	1	1	2460201	E2023-R9	358.4188	0.15	Eric	3Ek	12.56	208.6971	26	M-v	2.831608	0	4041	1975-202	2.00
25	2023-11-0	0.59	228.1365	1962 PG	2460486	Apollo	1	1	2460201	E2023-V4	309.1856	0.15	Heracles	3Ek	14.08	246.9389	32	M-v	2.48628	0	4029	1953-202	1.8
26	2023-11-0	0.78	321.1481		2460068	Amor	1	1	2460201	E2023-V4	352.2444	0.15	Lyapunov	3Ek	15.36	25.34206	14	M-v	5.157252	0	333	1987-202	2.98
27	2023-10-2	0.54	86.81017	1990 UV12	2460208	Amor	1	1	2460201	E2023-UR	189.9194	0.15		3Ek	13.9	357.8967	29	M-v	3.712868	0	4259	1953-202	2.39
28	2022-12-0	0.7	153.7825		2459661	Amor	1	1	2460201	MPO7275	128.6464	0.15	Jasonwhe	3El	17.12	167.7384	18	M-v	3.171392	0	512	1955-202	2.15
29	2023-05-2	0.6	231.8252	1970 RA	2459940	Amor	1	1	2460201	E2023-UR	173.1038	0.15	Melissabr	3Ek	14.35	78.8766	26	M-v	3.251169	0	4095	1970-202	2.19
30	2023-09-2	0.61	48.25893		2460337	Apollo	1	1	2460201	E2023-SK	45.71211	0.15		3Ek	17.24	274.4648	18	M-v	1.57646	0	1268	1974-202	1.3
31	2023-03-1	0.67	335.7678		2460744	Amor	1	1	2460201	E2023-F72	14.10607	0.15		3Ek	15.51	189.3792	23	M-v	3.138581	0	2283	1990-202	2.14
32	2023-11-0	0.59	122.6158		2460444	Amor	1	1	2460201	E2023-V4	9.89558	0.15	Camarillo	3Ek	16.23	259.9377	22	M-v	2.40333	0	3527	1974-202	1.7
33	2023-07-1	0.57	218.4477		2460025	Apollo	1	1	2460201	E2023-UJ4	280.5858	0.15	Zeus	3Ek	15.47	50.70462	26	M-v	3.411772	0	1333	1988-202	2.26
34	2023-10-2	0.67	8.3832		2460284	Apollo	1	1	2460201	E2023-UL	161.2911	0.15	Talos	3Ek	17.19	287.0055	26	M-v	1.124688	0	1165	1991-202	1.08

Fig. 3 Dataset for Near-Earth asteroids

__Description of columns in the dataset__[7]

Attribute	**Type**	**Description**
Name	string	Name, if the asteroid has received one
Number	string	Number, if the asteroid has received one; this is the asteroid's permanent designation
Principal_desig	string	Principal provisional designation (if it exists)
Other_desigs	string	Other provisional designations (if they exist)
H	float	Absolute magnitude, *H*
G	float	Slope parameter, *G*
Epoch	float	Epoch of the orbit (Julian Date)
a	float	Semimajor axis, *a* (AU)
e	float	Orbital eccentricity, *e*
i	float	Inclination to the ecliptic, J2000.0 (degrees)

[7] "Glossary."

Node	float	Longitude of the ascending node, ☊, J2000.0 (degrees)
a_p	float	Argument of perihelion, ω, J2000.0 (degrees)
M	float	Mean anomaly, M, at the epoch (degrees)
n	float	Mean daily motion, n (degrees/day)
U	string	Uncertainty parameter, U (integer with values 0–9; but refer to entry in Table 1 for other possible values)
Ref	string	Reference
Num_obs	integer	Number of observations
Num_opps	integer	Number of oppositions
Arc_years	string	Only present for multi-opposition orbits (year of first observation – year of last observation)
rms	float	r.m.s. residual (")
Perturbers	string	Coarse indicator of perturbers used in orbit computation
Perturbers_2	string	Precise indicator of perturbers used in orbit computation

Attribute	**Type**	**Description**
Last_obs	string	Date of last observation included in orbit solution (YYYY-MM-DD format)
Hex_flags	string	4-hexdigit flags (refer to entry in Table 1 for explanation; in JSON format this information has been decoded and is supplied in individual keywords)
Computer	string	Name of orbit computer (be it a person or machine)
orbit_type	string	Possible values: • Atira • Aten • Apollo • Amor • Distant Object

NEO_flag	integer	Value = 1 if flag raised, otherwise keyword is absent
One_km_NEO_flag	integer	Value = 1 if flag raised, otherwise keyword is absent
Perihelion_dist	float	Perihelion distance (AU)
Aphelion_dist	float	Aphelion distance (AU)
Semilatus_rectum	float	Semilatus rectum distance (AU)
Orbital_period	float	Orbital period (years)
Synodic_period	float	Synodic period (years)

Table(1) Data description

Methodology

Exploratory data analysis

When simply looking at the data, we found that around 33,330 rows shows that the objects detected are nearly earth objects. Furthermore, the dataset has 1378 rows of data which indicates near-earth objects just 1Km away.

```
sns.catplot(x = 'PHA', y = 'One km NEO flag', data = neo, hue = 'PHA')
plt.show()
```

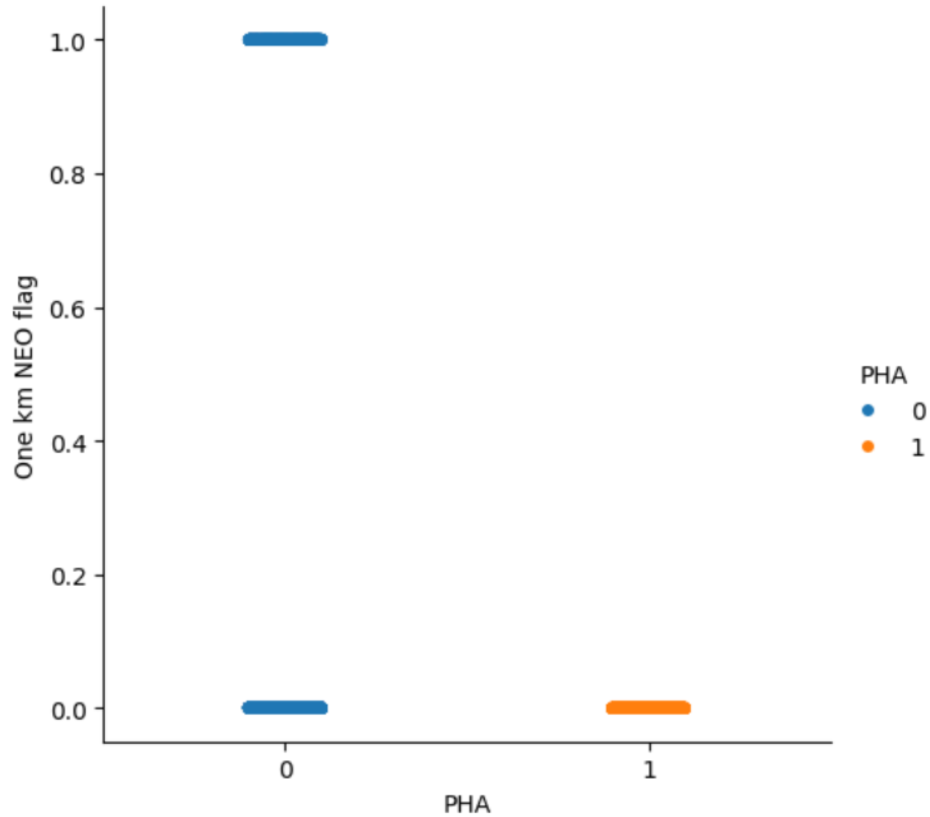


Fig. 4 Correlation of One Km NEO flag and PHA flag

For further understanding of the data, we dropped a few columns which do not have any impact on the data, like name, arc year, computer, last observation, other designation, etc.

To dig deep into the dataset and to understand it, we tried to correlate each parameter with one another.

Correlation of the dataset helped us to understand and verify that the absolute magnitude has a huge impact on determining an asteroid as PHA. Similarly, there are few parameters that affect the decision on assigning an asteroid as a PHA. The parameters are orbital eccentricity, Semilatus rectum distance (AU).

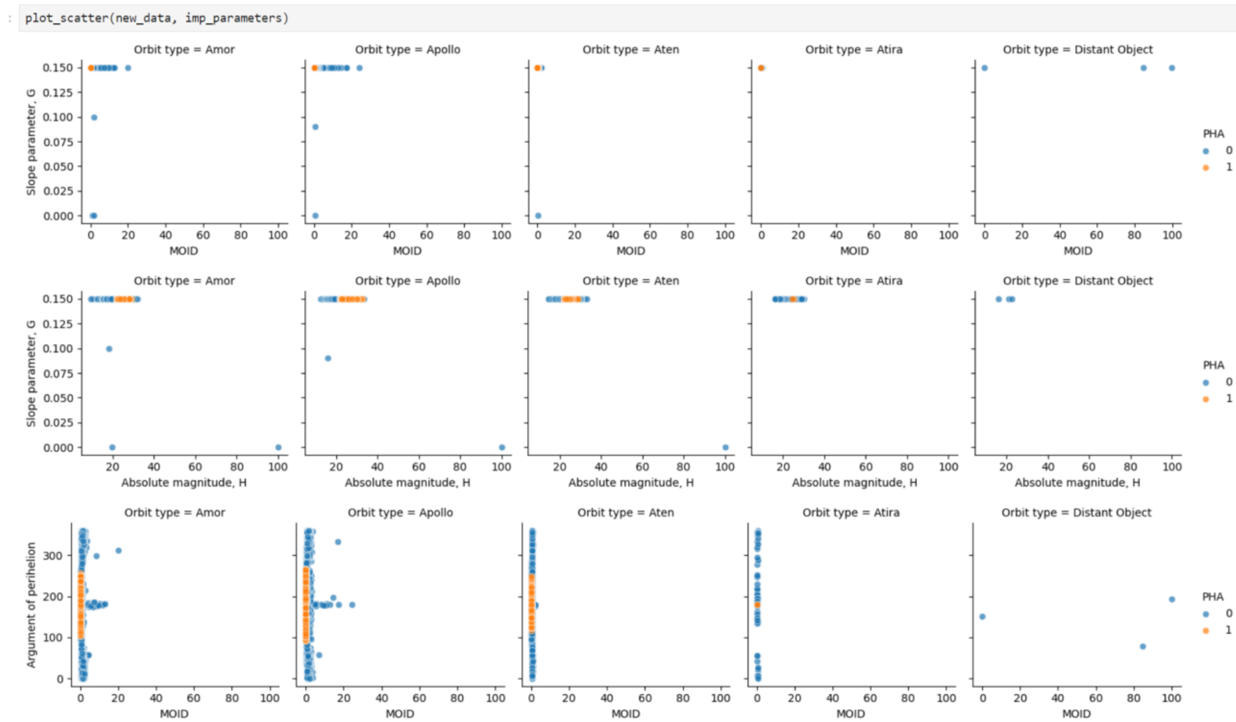
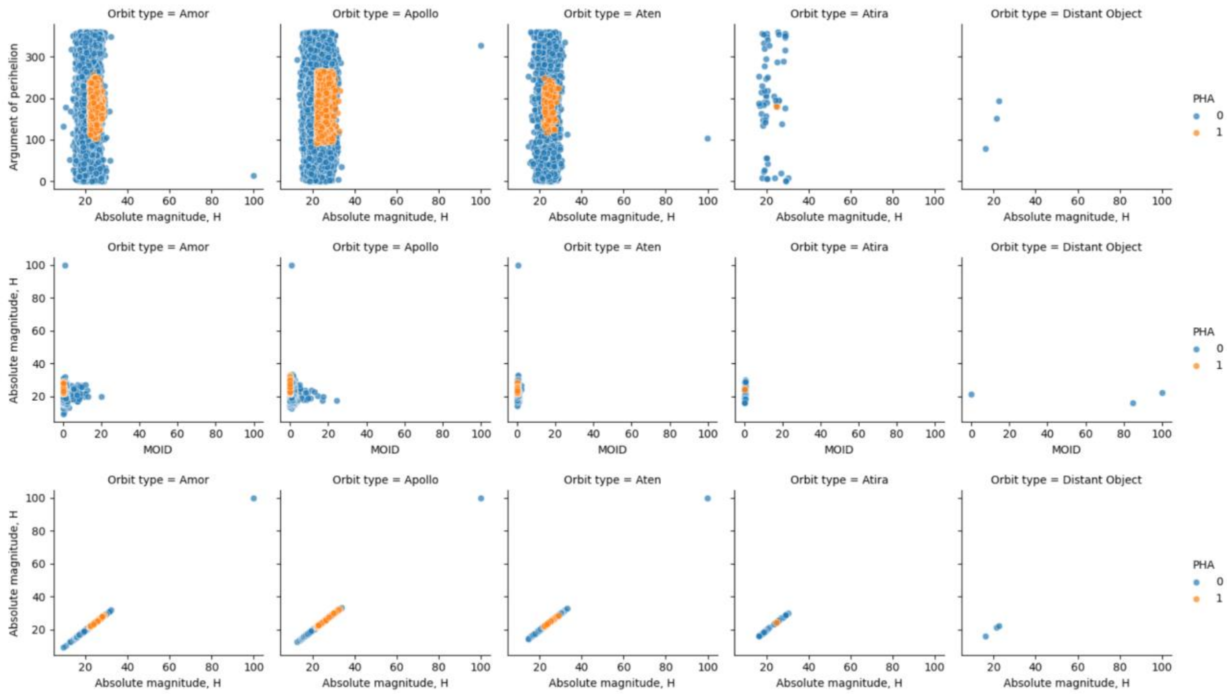


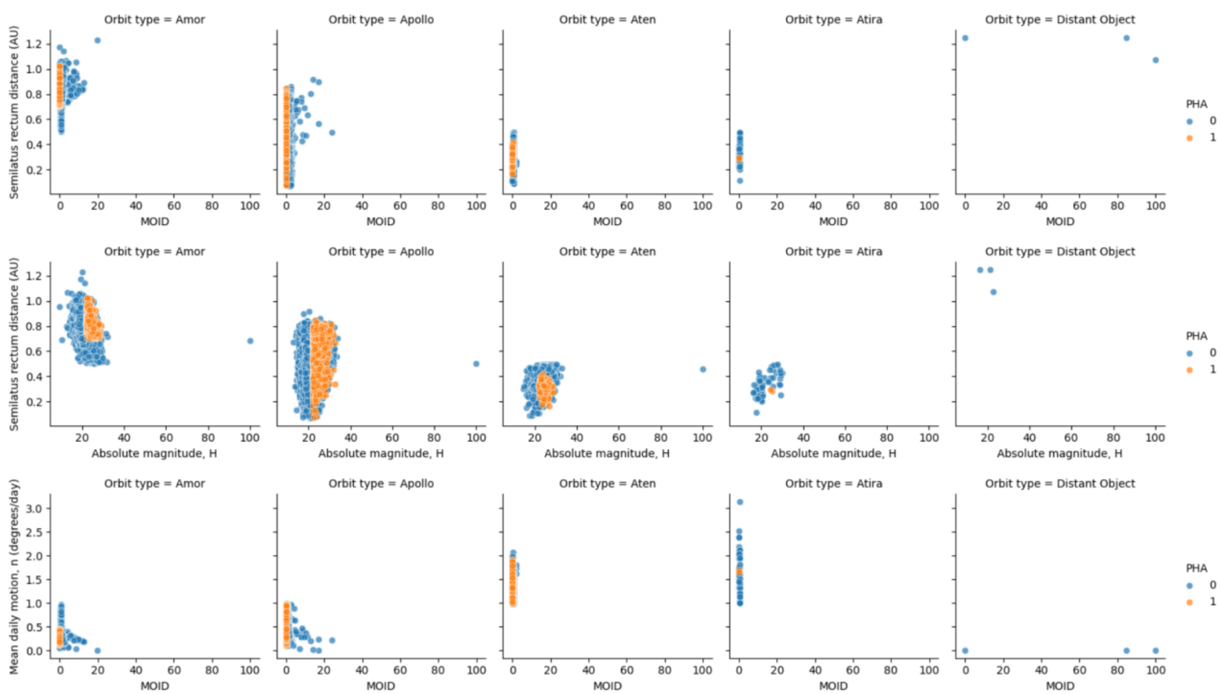
Fig. 5 Correlation of different parameters

In this figure, we can say that the absolute magnitude or MOID are ot at all correlated with the slope parameter of an asteroid. On the other hand, we can see that the asteroids which are potentially hazardous has an argument of perihelion ranging from 200-250 degrees.



Orbit type = Amor
Orbit type = Apollo
Orbit type = Aten
Orbit type = Atira
Orbit type = Distant Object
Argument of perihelion
Absolute magnitude, H
MOID
PHA
0
1

Fig. 6



Orbit type = Amor
Orbit type = Apollo
Orbit type = Aten
Orbit type = Atira
Orbit type = Distant Object
Semilatus rectum distance (AU)
Mean daily motion, n (degrees/day)
Absolute magnitude, H
MOID
PHA
0
1

Fig. 7

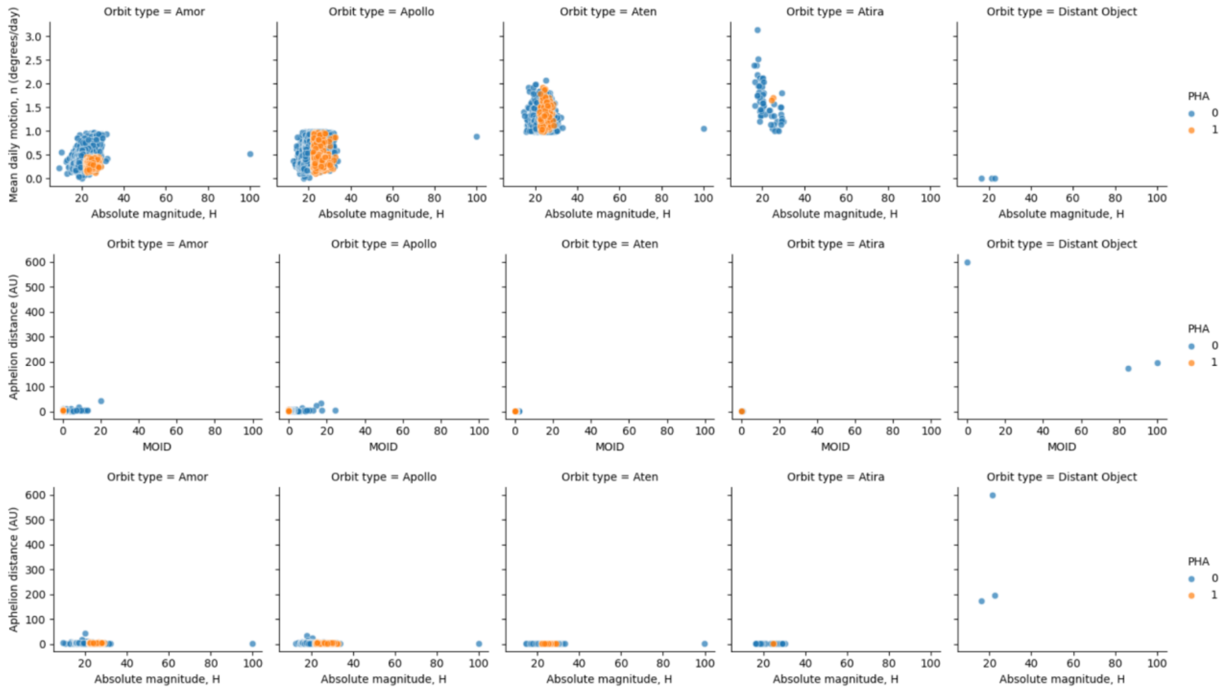


Fig. 8

From the above figures we can say that its difficult to classify the dataset based on the aphelion distance, mean daily motion and slope parameter.

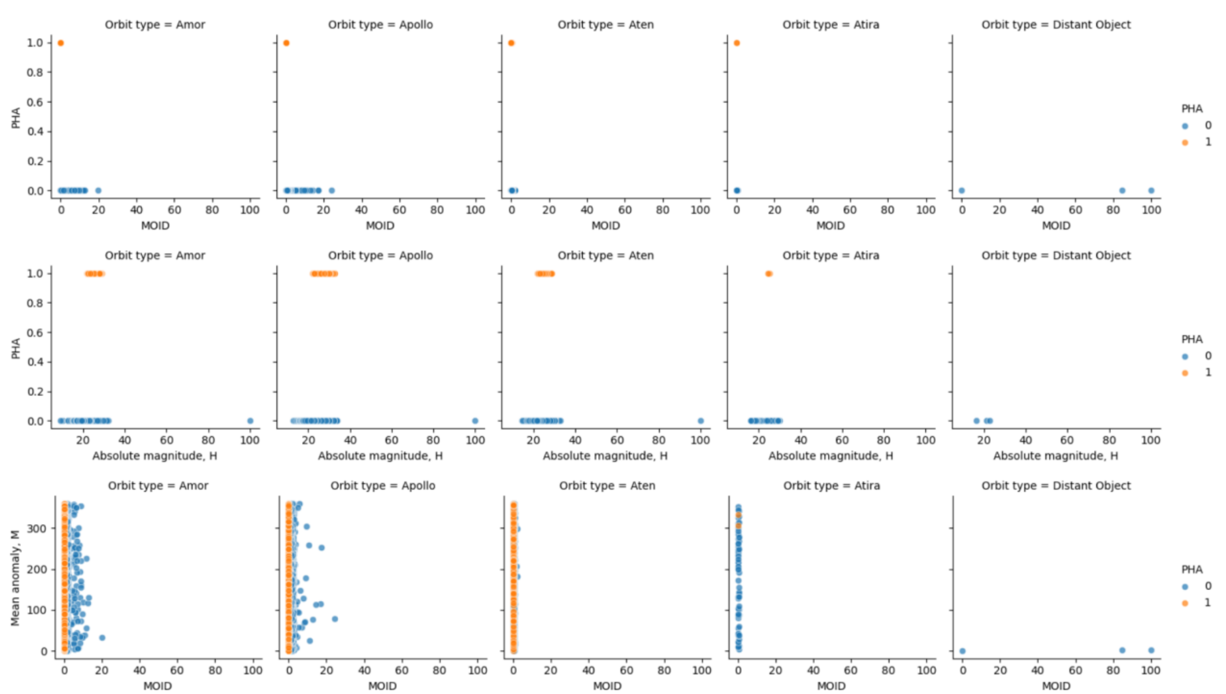


Fig. 9

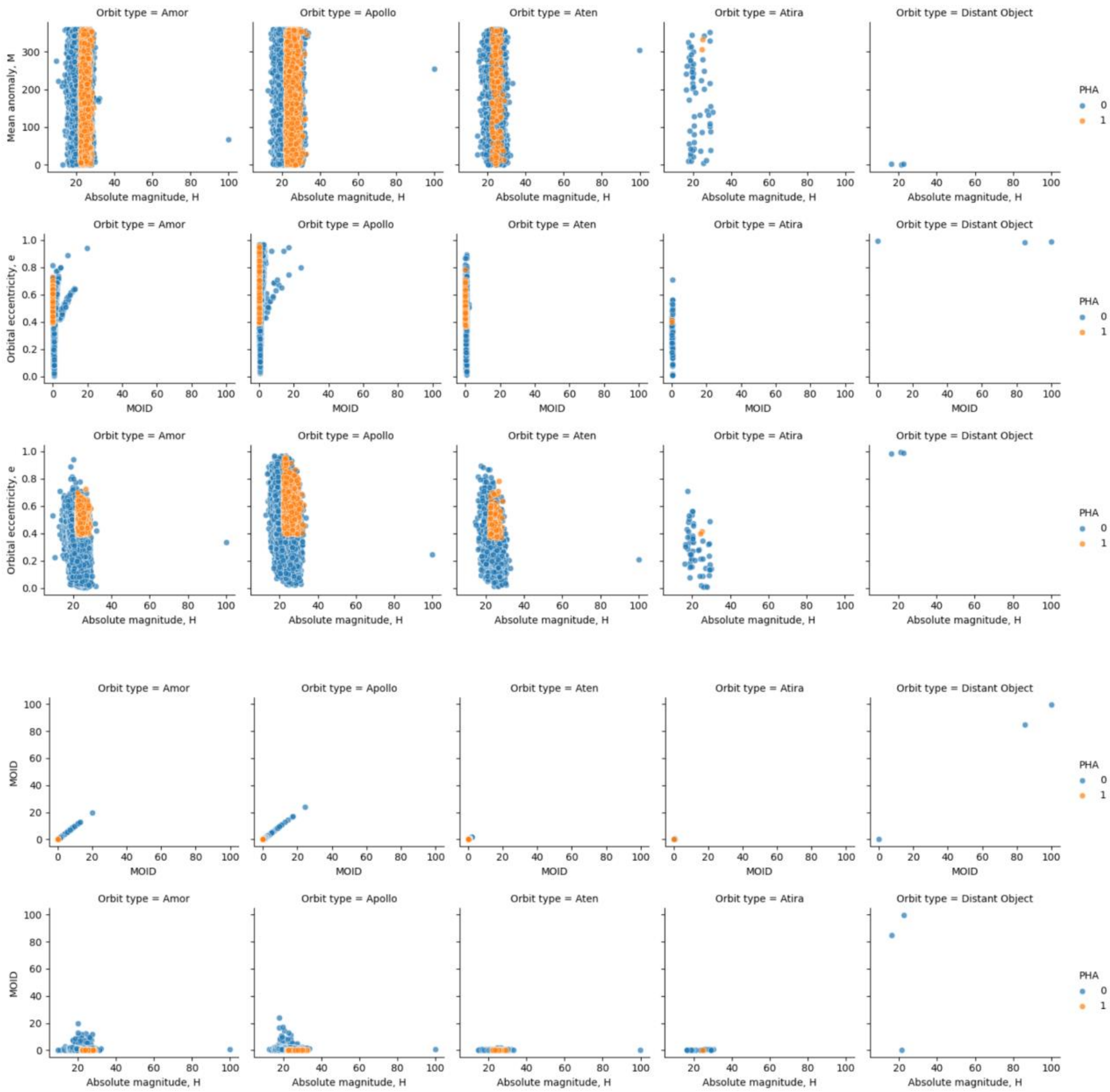


Fig. 10

To conclude, it is clear that the absolute magnitude and the MOID largely affect the PHA. For every asteroid that is potentially hazardous, the absolute magnitude is greater than 22 and the MOID is less than 0.05.

```
new_data[imp_parameters].describe()
```

	Slope parameter, G	Argument of perihelion	Absolute magnitude, H	Semilatus rectum distance (AU)	Mean daily motion, n (degrees/day)	Aphelion distance (AU)	PHA	Mean anomaly, M	Orbital eccentricity, e	MOID
count	33348.000000	33351.000000	33348.000000	33351.000000	33351.000000	33351.000000	33351.000000	33351.000000	33351.000000	3.335100e+04
mean	0.149979	182.599326	23.441543	0.650506	0.525355	2.611130	0.101736	178.042332	0.437671	5.825701e-01
std	0.001697	104.263415	2.990552	0.162290	0.282644	3.729409	0.302305	104.615752	0.176736	9.667485e-01
min	0.000000	0.010290	9.260000	0.069099	0.000189	0.653767	0.000000	0.005530	0.002356	7.744250e-11
25%	0.150000	93.288435	21.280000	0.543733	0.307089	1.674874	0.000000	87.266665	0.305382	1.760351e-01
50%	0.150000	184.606320	23.760000	0.660444	0.446295	2.443347	0.000000	176.235920	0.451947	4.380254e-01
75%	0.150000	272.601320	25.590000	0.768309	0.667421	3.358525	0.000000	267.790300	0.564722	8.051747e-01
max	0.150000	359.997940	99.990000	1.249472	3.141582	600.274566	1.000000	359.998910	0.995837	9.984212e+01

Fig. 11 Statistical description of data

The above figure shows the spread of the data across the selected parameters. The mean and the median for almost all the selected parameters are similar, this indicates the data is normally distributed.

Data Cleaning

The dataset used was in the raw format. There were many missing values, which were denoted as NaN.

```
[6]: neo.isnull().any()

[6]: Last observation                         False
     r.m.s. residual (")                      False
     Argument of perihelion                   False
     Other designations                        True
     Tp                                       False
     Orbit type                               False
     NEO flag                                 False
     One km NEO flag                           True
     Epoch of the orbit (Julian Date)         False
     Reference                                False
     Node                                     False
     Slope parameter, G                        True
     Name                                      True
     Perturbers 2                              True
     Absolute magnitude, H                     True
     Mean anomaly, M                          False
     Number of oppositions                    False
     Perturbers                                True
     Orbital period (years)                   False
     Uncertainty parameter                     True
     Number of observations                   False
     Arc years                                 True
     Semimajor axis, a (AU)                   False
     Orbital eccentricity, e                  False
     Inclination to the ecliptic              False
     Perihelion distance (AU)                 False
     Number                                    True
     Mean daily motion, n (degrees/day)       False
     Semilatus rectum distance (AU)           False
     Hex flags                                False
     Computer                                 False
     Synodic period (years)                   False
     Aphelion distance (AU)                   False
     Principal designation                    False
     dtype: bool
```

Fig. 12 dataset containing null values

As the above figure shows, there are few important columns that have the NaN values.
For some columns like the NEO flag, where the description of the data said that the column has binary values(either 1 or 0), for such cases we used the fillna() function and replaced all the blank values with 0.

```
neo['NEO flag'] = neo['NEO flag'].fillna(0)
```

```
neo['One km NEO flag'] =  neo['One km NEO flag'].fillna(0)
```

Fig. 13 data cleaning

Data Preprocessing

Data preprocessing includes addition of extra columns based on the available data, if required.
Our dataset required an additional column with a PHA flag.
We know that an asteroid is said to be potentially hazardous asteroid if the absolute magnitude >= 22 and the minimum orbital intersection distance(MOID)<=0.05AU. Since our dataset does not include any column with MOID value. Hence, we calculated the MOID value based on semi-major axis, eccentricity, inclination, argument of perihelion, longitude of ascending node, and mean anomaly. [8]

```
from scipy.optimize import minimize

def moid(row):
    G = row['Slope parameter, G']
    a_p = np.radians(row['Argument of perihelion'])
    H = row['Absolute magnitude, H']
    q = row['Semilatus rectum distance (AU)']
    n = np.radians(row['Mean daily motion, n (degrees/day)'])
    Q = row['Aphelion distance (AU)']
    e = row['Orbital eccentricity, e']

    def relative_distance(theta):
        r = q * (1 + e) / (1 + e * np.cos(theta))
        return r

    def distance_between_orbits(theta):
        r_asteroid = relative_distance(theta)
        r_earth = Q / (1 + e * np.cos(theta - a_p))

        return np.abs(r_earth - r_asteroid)

    result = minimize(distance_between_orbits, 0)
    min_distance = distance_between_orbits(result.x)

    return float(min_distance)


neo['MOID'] = neo.apply(moid, axis=1)

neo.head()
```

Fig. 14 data Preprocessing

[8] "Minimum Orbital Intersection Distance: An Asymptotic Approach | Astronomy & Astrophysics (A&A)."

Other nations	Tp	Orbit type	NEO flag	One km NEO flag	Epoch of the orbit (Julian Date)	Reference	Node	...	Perihelion distance (AU)	Number	Mean daily motion, n (degrees/day)	Semilatus rectum distance (AU)	Hex flags	Computer	Synodic period (years)	Aphelion distance (AU)	Principal designation	MOID
1956 PC	2.460446e+06	Amor	1.0	1.0	2460200.5	E2023-V42	304.28598	...	1.133254	(433)	0.559777	0.692869	1804	MPCLINUX	2.314555	1.782981	A898 PA	3.680317e-01
000 JW8	2.459956e+06	Amor	1.0	1.0	2460200.5	E2023-V08	183.85389	...	1.194321	(719)	0.230242	0.923801	1804	MPCLINUX	1.304809	4.078500	A911 TB	1.189120e-08
NaN	2.459867e+06	Amor	1.0	1.0	2460200.5	E2023-TI3	171.31940	...	1.082993	(1221)	0.370736	0.777411	1804	MPCLINUX	1.602948	2.755167	1932 EA1	1.095690e+00
NaN	2.460009e+06	Apollo	1.0	1.0	2460200.5	E2023-P11	87.95271	...	0.186626	(1566)	0.880499	0.170473	9803	MPCLINUX	9.377001	1.969530	1949 MA	8.927135e-01
NaN	2.460736e+06	Amor	1.0	1.0	2460200.5	E2023-V42	62.23069	...	1.127241	(1580)	0.302558	0.838124	1804	MPCLINUX	1.442951	3.267777	1950 KA	3.588154e-09

Fig. 15

Now that we have the MOID value, using the absolute magnitude values and the MOID value, we created a column with a PHA flag containing binary values(1 or 0) using the below code.

```
neo['PHA'] = np.where((neo['Absolute magnitude, H'] >= 22) & (neo['MOID'] <= 0.05), 1, 0)
```

```
neo.head()
```

Other designations	Tp	Orbit type	NEO flag	One km NEO flag	Epoch of the orbit (Julian Date)	Reference	Node	...	Number	Mean daily motion, n (degrees/day)	Semilatus rectum distance (AU)	Hex flags	Computer	Synodic period (years)	Aphelion distance (AU)	Principal designation	MOID	PHA
1956 PC	2.460446e+06	Amor	1.0	1.0	2460200.5	E2023-V42	304.28598	...	(433)	0.559777	0.692869	1804	MPCLINUX	2.314555	1.782981	A898 PA	3.680317e-01	0
2000 JW8	2.459956e+06	Amor	1.0	1.0	2460200.5	E2023-V08	183.85389	...	(719)	0.230242	0.923801	1804	MPCLINUX	1.304809	4.078500	A911 TB	1.189120e-08	0
NaN	2.459867e+06	Amor	1.0	1.0	2460200.5	E2023-TI3	171.31940	...	(1221)	0.370736	0.777411	1804	MPCLINUX	1.602948	2.755167	1932 EA1	1.095690e+00	0
NaN	2.460009e+06	Apollo	1.0	1.0	2460200.5	E2023-P11	87.95271	...	(1566)	0.880499	0.170473	9803	MPCLINUX	9.377001	1.969530	1949 MA	8.927135e-01	0
NaN	2.460736e+06	Amor	1.0	1.0	2460200.5	E2023-V42	62.23069	...	(1580)	0.302558	0.838124	1804	MPCLINUX	1.442951	3.267777	1950 KA	3.588154e-09	0

Fig. 16

Division of dataset

The dataset has been divided into training and testing dataset. We used the train_test_split() function to divide the dataset. The test_size used is 0.2, which implies that 80% of the data is the training data and the 20% is the testing data. We also set the random_state as 25, which will ensure the data in both the data set is picked randomly and will be constant throughout different models.

```
training_data, testing_data = train_test_split(df, test_size=0.2, random_state=25)

print(f"No. of training examples: {training_data.shape[0]}")
print(f"No. of testing examples: {testing_data.shape[0]}")
```

```
No. of training examples: 26721
No. of testing examples: 6681
```

Fig. 17 Dataset splitting

```python
# sns.boxplot(x="PHA", y="Absolute magnitude, H",
#             data=new_data)
sns.violinplot(data=new_data, x="PHA", y="Absolute magnitude, H", hue='PHA')
# plt.xticks(rotation=45)
```

```
<Axes: xlabel='PHA', ylabel='Absolute magnitude, H'>
```

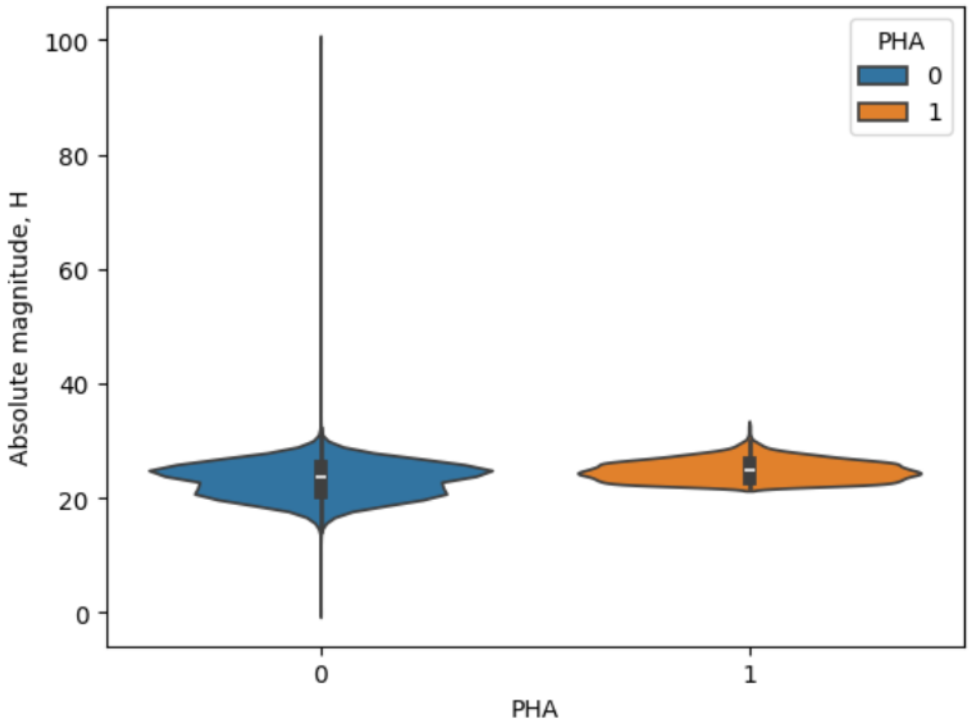


Fig. 18 violin plot(absolute magnitude vs PHA flag)

The violin plot between the absolute magnitude and the PHA clearly shows that the asteroids that if the absolute magnitude greater than or equal to 22 then the asteroids are potentially hazardous.
On the other hand the asteroids that have absolute magnitude lower than 22 are non hazardous.

```
sns.lineplot(data = new_data, x = "MOID", y = "Absolute magnitude, H", hue = "One km NEO flag")
```

```
<Axes: xlabel='MOID', ylabel='Absolute magnitude, H'>
```

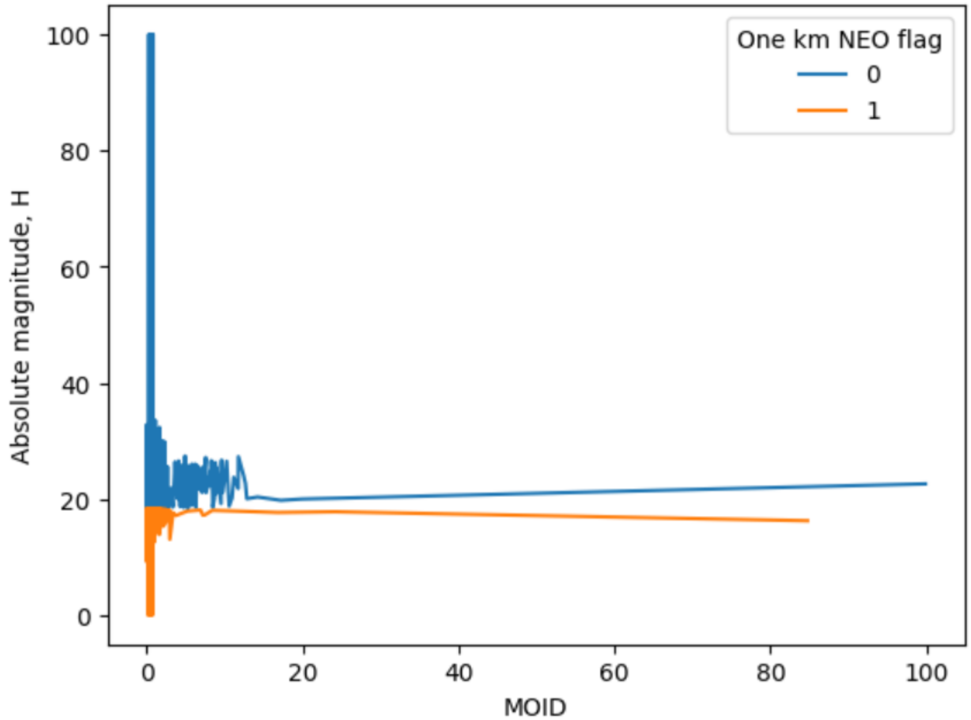


Fig. 19 line plot(absolute magnitude vs MOID)

This line plot has MOID on the x-axis and absolute magnitude on the y-axis.
The graph shows that most of the near-earth objects that are 1 km away from the earth have MOID ranging from 0-1 and absolute magnitude ranging from 1-20 AU.
We can also say that the absolute magnitude is slightly increasing with increasing MOID for near-earth objects that are more than 1 km away.

```
sns.catplot(data = new_data, y = "MOID", x = "PHA")
plt.gca().set_ylim([0, 5])
```

```
(0.0, 5.0)
```

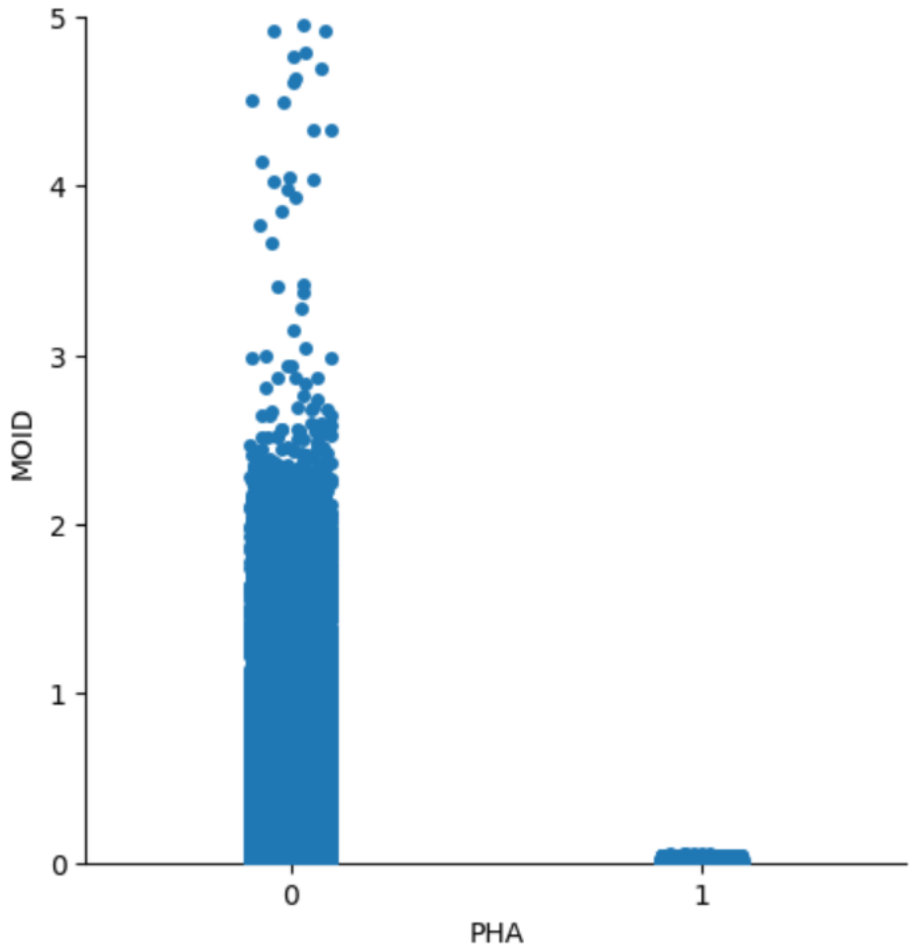


Fig. 20 Categorical plot(MOID vs PHA flag)

The above shown is a categorical plot with PHA on the x-axis and MOID on the y-axis.
We can see that the potentially hazardous asteroids strictly have MOID value less than 0.1.
Whereas for the asteroids which are not hazardous, the MOID values ranges from 0 to more than 5.
For non hazardous asteroids, although the MOID range starts from 0, the condition for an asteroid to be considered as PHA does not totally depend on the MOID value, other factor such as absolute magnitude sums up to form a condition to distinguish an asteroid as PHA.

Different Models (Evaluation)

To analyze which model provides a better accuracy in predicting the potentially hazardous asteroids, we used a different set of models to understand the accuracy. Few of the models that we used are: Random forest classification, gradient boosting classifier, (SVC)support vector classifier, ANN(Artificial Neural Networks), Linear regression.[9]

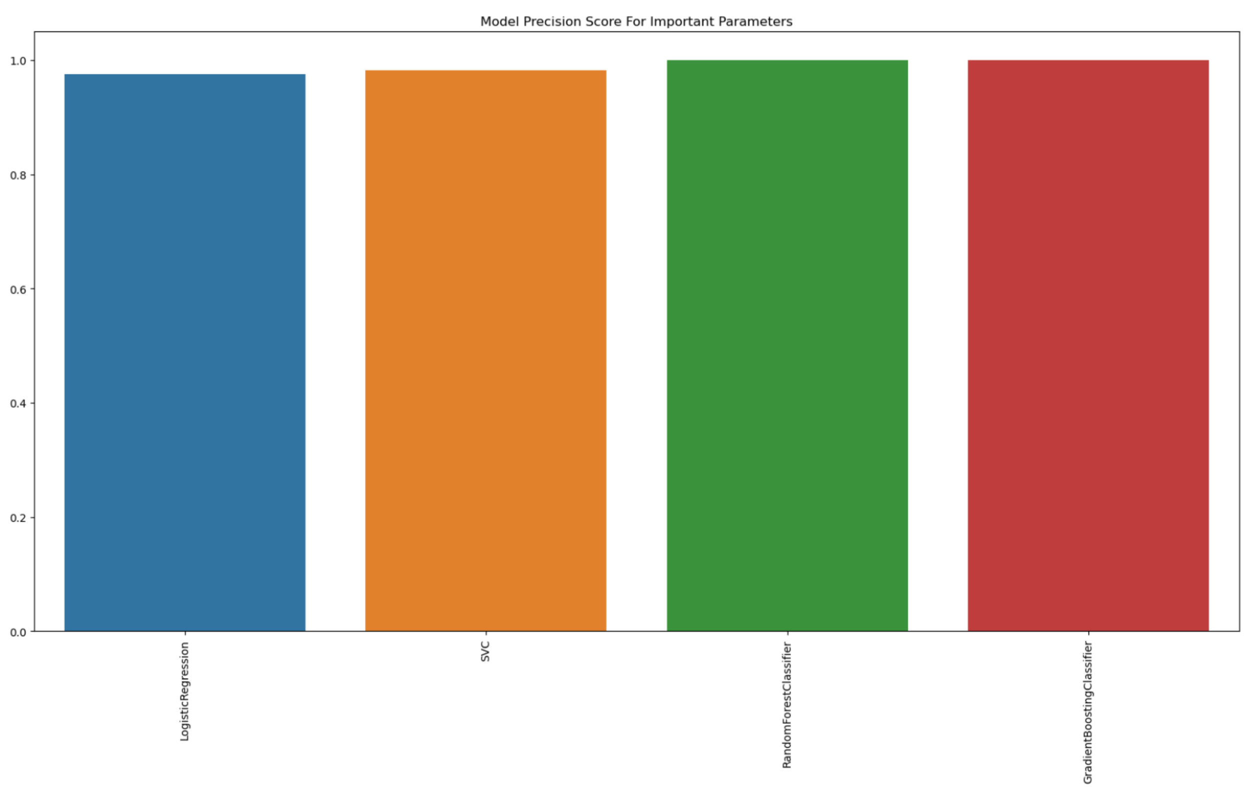
After trying the above models, the accuracy that we got for each model is shown below. These accuracies provide us an insight to which model best suits to predict the potentially hazardous asteroids.

```
model_scores_params_imp = model_fit_score(models, new_data[imp_parameters])
model_scores_params_imp.sort_values('Score', ascending = False)
```

	Score
GradientBoostingClassifier	1.000000
RandomForestClassifier	0.999960
SVC	0.982929
LogisticRegression	0.975853

Fig. 21 Different Model accuracy

As seen from the table, some of the models seem to give overfitting issues. Models like random forest classifier and gradient boosting are giving the accuracy of around 99-100%.



[9] "Algorithm Finds a Potentially Hazardous Asteroid Missed by NASA."

Fig. 22 Plot for Different Model accuracy

We used the random forest classifier model to analyze the important features. From the figure below, we can see that the two most important features are MOID and absolute magnitude, followed by argument of perihelion and orbital eccentricity.

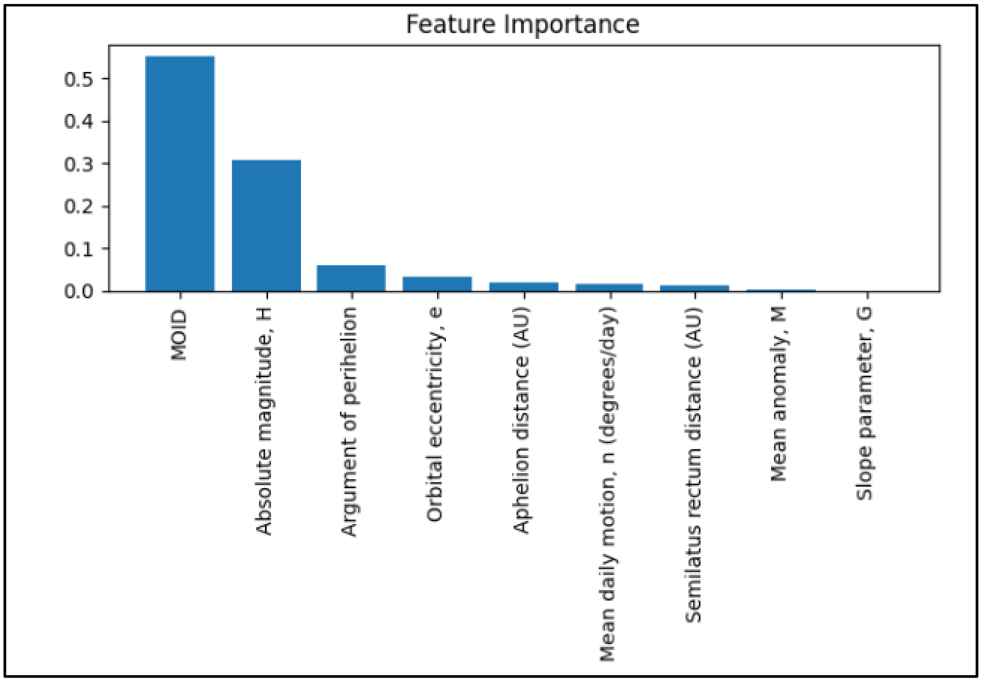


Fig. 23 Feature Importance

Random Forest Classifier

This is a machine learning technique used in regression and classification tasks. It combines the output of multiple decision trees to reach a single result.
For training this model, we split our dataset into training and testing data with test_size as 20% and random_state as 42.

```
Accuracy: 1.0

Confusion Matrix:
 [[5994    0]
 [   0  677]]
```

Fig. 24 Accuracy for random forest classifier

Classification Report				
	Precision	*recall*	*f1-score*	*support*
0	*1.00*	*1.00*	*1.00*	*5994*
1	*1.00*	*1.00*	*1.00*	*677*
accuracy			*1.00*	*6671*
macro average	*1.00*	*1.00*	*1.00*	*6671*
weighted average	*1.00*	*1.00*	*1.00*	*6671*

Table(2) Classification report for random forest classifier

As we can see the accuracy for the random forest classifier model is around 100% and the f1-score is 1.00 with the precision of 1.00. We understand this model is overfitted. Hence we tried other models.

Gradient Boosting Classifier

This is a machine learning technique used in regression and classification tasks.
Here, we trained our model at different learning rates. For the learning rate 0.05, we got an accuracy of about 89.8%. For the learning rate 0.25, we got an accuracy of about 99.3%.

```
Learning rate:  0.05
Accuracy score (training): 0.898
Accuracy score (validation): 0.899
Learning rate:  0.075
Accuracy score (training): 0.898
Accuracy score (validation): 0.899
Learning rate:  0.1
Accuracy score (training): 0.898
Accuracy score (validation): 0.899
Learning rate:  0.25
Accuracy score (training): 0.993
Accuracy score (validation): 0.991
```

Fig. 25 Accuracy for gradient boosting classifier at different learning rate

For the learning rate 0.5, we got an accuracy of about 99.97%.

```
Accuracy: 0.9997001948733323

Confusion Matrix:
 [[5994    0]
 [   2  675]]
```

Fig. 26 Accuracy for gradient boosting classifier at 0.5 learning rate

Classification Report				
	Precision	*recall*	*f1-score*	*support*
0	*1.00*	*1.00*	*1.00*	*5994*
1	*1.00*	*1.00*	*1.00*	*677*
accuracy			*1.00*	*6671*
macro average	*1.00*	*1.00*	*1.00*	*6671*
weighted average	*1.00*	*1.00*	*1.00*	*6671*

Table(3) Classification report for gradient boosting classifier

Support Vector Classifier

For support vector classification, we trained the model in two kernels.
RBF kernel is one of the kernels used. For this kernel we split our dataset into training and testing data with test_size as 20% and random_state as 42.

```
Accuracy: 0.903912456903013

Confusion Matrix:
 [[5968   26]
 [ 615   62]]
```

Fig. 27 Accuracy for support vector classifier(RBF kernel)

Classification Report				
	Precision	recall	f1-score	support
0	0.91	1.00	0.95	5994
1	0.70	0.09	0.16	677
accuracy			0.90	6671
macro average	0.81	0.54	0.56	6671
weighted average	0.89	0.90	0.87	6671

Table(4) Classification report for support vector classifier(linear kernel)

As we can see the accuracy for the SVC model in RBF kernel is around 90.39% and the f1-score is 0.95 with the precision of 0.91.

The other kernel that we used is the linear kernel. For this kernel we split our dataset into training and testing data with test_size as 20% and random_state as 42.

```
Accuracy: 0.9829111077799431

Confusion Matrix:
 [[5936   58]
 [  56  621]]
```

Fig. 28 Accuracy for support vector classifier(Linear kernel)

Classification Report				
	Precision	*recall*	*f1-score*	*support*
0	*0.99*	*0.99*	*0.99*	*5994*
1	*0.91*	*0.92*	*0.92*	*677*
accuracy			*0.98*	*6671*
macro average	*0.95*	*0.95*	*0.95*	*6671*
weighted average	*0.98*	*0.98*	*0.98*	*6671*

Table(5) Classification report for support vector classifier(linear kernel)

As we can see the accuracy for the SVC model in linear kernel is around 98.29% and the f1-score is 0.99 with the precision of 0.99.

Logistic Regression Model

Logistic regression is a supervised machine learning algorithm that is mainly used for binary classification.

For this model, we split our dataset into training and testing data with test_size as 20% and random_state as 42.

```
Accuracy: 0.9833608154699446

Confusion Matrix:
 [[5947   47]
 [  64  613]]
```

Fig. 29 Accuracy for logistic regression model

Classification Report				
	Precision	recall	f1-score	support
0	0.99	0.99	0.99	5994
1	0.93	0.91	0.92	677
accuracy			0.98	6671
macro average	0.96	0.95	0.95	6671
weighted average	0.98	0.98	0.98	6671

Table(6) Classification Report for logistic regression model

As we can see the accuracy for the linear regression model is around 98.33% and the f1-score is 0.99 with the precision of 0.99.

Artificial Neural network (ANN)

ANN is a biologically inspired computer program designed to simulate the way in which the human brain processes information.

For this model, we split our dataset into training and testing data with test_size as 20% and random_state as 42.

For this model, we are executing 25 epochs in a batch of 16. Using this parameters, we executed the ANN model and below are the results.

```
1668/1668 [==============================] - 6s 3ms/step - loss: 0.5516 - accuracy: 0.8754
Epoch 2/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.2739 - accuracy: 0.8983
Epoch 3/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.2160 - accuracy: 0.9054
Epoch 4/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.1828 - accuracy: 0.9141
Epoch 5/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.1580 - accuracy: 0.9271
Epoch 6/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.1376 - accuracy: 0.9356
Epoch 7/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.1240 - accuracy: 0.9425
Epoch 8/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.1124 - accuracy: 0.9484
Epoch 9/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.1034 - accuracy: 0.9524
Epoch 10/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0969 - accuracy: 0.9567
Epoch 11/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0883 - accuracy: 0.9605
Epoch 12/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0883 - accuracy: 0.9590
Epoch 13/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0813 - accuracy: 0.9641
Epoch 14/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0782 - accuracy: 0.9654
Epoch 15/25
1668/1668 [==============================] - 4s 3ms/step - loss: 0.0753 - accuracy: 0.9671
Epoch 16/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0756 - accuracy: 0.9665
Epoch 17/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0782 - accuracy: 0.9651
Epoch 18/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0707 - accuracy: 0.9687
Epoch 19/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0732 - accuracy: 0.9677
Epoch 20/25
1668/1668 [==============================] - 4s 3ms/step - loss: 0.0678 - accuracy: 0.9696
Epoch 21/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0714 - accuracy: 0.9695
Epoch 22/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0658 - accuracy: 0.9722
Epoch 23/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0677 - accuracy: 0.9702
Epoch 24/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0667 - accuracy: 0.9713
Epoch 25/25
1668/1668 [==============================] - 5s 3ms/step - loss: 0.0657 - accuracy: 0.9711
209/209 [==============================] - 1s 3ms/step - loss: 0.0497 - accuracy: 0.9837

Accuracy of test: 98.37
```

Fig. 30 Accuracy for ANN

From the above figures, we can see that the accuracy keeps getting better with each epoch. And at the end of the 25th epoch, the accuracy of the ANN model is around 98.37%

Limitations and Future work

Limitations

Data quality and the quantity could be one of the limitations. So there is limited or incomplete data available on the Internet and this can hinder the accuracy of our predictive models. So inconsistency, missing values or the bias within the data set could affect the performance of our machine learning algorithms.
There could be uncertainties in the orbital trajectories that are available. Sometimes it might be difficult to predict the nature of an asteroid.
In the near future there can be many more asteroids that can be detected. Currently our model just has been trained based on the data available, but I feel the limitations can be like if a new asteroid is added to the data, it might affect our prediction or it might affect our algorithm.

Future Work

Looking at the limitations of this project, the future work that we were looking forward to are:
The data that we are using currently has a lot of limitations like the missing values in consistencies. So we would like to have a structured data and work with the data to train our models and predict the potentially hazardous asteroids
We can improve the accuracy of our models by fine tuning the hyper parameters
We would like to collaborate with different space agencies and research groups who are already working on the predictive models to have a better understanding and improve our predictive model. [10][11]

Conclusion

To sum up, this effort has proved helpful in predicting Potentially Hazardous Asteroids (PHAs) by examining asteroid data. After a thorough investigation of several models, the Support Vector Classifier (SVC) proved to be the most effective and accurate mode with an accuracy of 98.33%l for predicting PHAs. Its strong performance highlights its potential as a trustworthy instrument for locating and evaluating these potentially dangerous celestial bodies. This achievement signifies a significant step forward in our efforts toward planetary defense and underscores the importance of continued research and development in this critical field.

---

[10] Ranaweera and Fernando, “Prediction of Potentially Hazardous Asteroids Using Deep Learning.”
[11] “New Algorithm Spots Its First ‘Potentially Hazardous’ near-Earth Asteroid — and It’s 600 Feet Long - CBS News.”

Citations

1. “Algorithm Finds a Potentially Hazardous Asteroid Missed by NASA.” Accessed December 19, 2023. https://www.freethink.com/space/potentially-hazardous-asteroids-heliolinc3d.
2. Arnold, J. O., C. D. Burkhard, J. L. Dotson, D. K. Prabhu, D. L. Mathias, M. J. Aftosmis, Ethiraj Venkatapathy, D. D. Morrison, D. W. G. Sears, and M. J. Berger. “Analysis of Potentially Hazardous Asteroids.” June 29, 2015. https://ntrs.nasa.gov/citations/20160000309.
3. “Asteroids - NASA Science.” Accessed December 19, 2023. https://science.nasa.gov/solar-system/asteroids/.
4. Eyes on Asteroids - NASA/JPL. “Eyes on Asteroids - NASA/JPL.” Accessed December 19, 2023. https://eyes.nasa.gov/apps/asteroids.
5. “Glossary.” Accessed December 19, 2023. https://cneos.jpl.nasa.gov/glossary/.
6. “Minimum Orbital Intersection Distance: An Asymptotic Approach | Astronomy & Astrophysics (A&A).” Accessed December 19, 2023. https://www.aanda.org/articles/aa/full_html/2020/01/aa36502-19/aa36502-19.html.
7. “NEO Basics.” Accessed December 19, 2023. https://cneos.jpl.nasa.gov/about/neo_groups.html.
8. “New Algorithm Spots Its First ‘Potentially Hazardous’ near-Earth Asteroid — and It’s 600 Feet Long - CBS News,” August 2, 2023. https://www.cbsnews.com/news/new-algorithm-spots-potentially-hazardous-near-earth-asteroid-heliolinc3d-rubin-observatory/.
9. Ranaweera, Rkmt Nishavi, and Tgi Fernando. “Prediction of Potentially Hazardous Asteroids Using Deep Learning.” In *2022 2nd International Conference on Advanced Research in Computing (ICARC)*, 31–36, 2022. https://doi.org/10.1109/ICARC54489.2022.9753945.
10. Science. “Asteroid Impacts:10 Biggest Known Hits,” February 15, 2013. https://www.nationalgeographic.com/science/article/130214-biggest-asteroid-impacts-meteorites-space-2012da14.
11. The Planetary Society. “Notable Asteroid Impacts in Earth’s History.” Accessed December 19, 2023. https://www.planetary.org/notable-asteroid-impacts-in-earths-history.

Video Link

https://www.youtube.com/watch?v=i0-NbNCweI0