NASA – NEAREST EARTH OBJECTS

CAPSTONE PROJECT II

SHIVANI GADGI
PGA18 WD

AGENDA

- Introduction
- Objective
- Data Set
- About The Data
- Visualizations
- Problem Statement
- Performance Metrics
- Conclusion

Introduction

- **Neartest Earth objects:**

Near-Earth Objects (NEOs) are a group of celestial objects, including asteroids and comets, whose orbits bring them into proximity with Earth's orbit.

In space, there are countless objects, and some are closer than we realize. Even though 70,000 kilometers may seem far to us, in space terms, it's actually a tiny distance. Surprisingly, this relatively short distance can potentially cause problems for various natural processes. That's why it's important to be aware of what's around us, especially objects like asteroids, as they have the potential to be harmful.

This dataset gathers information about asteroids certified by NASA, specifically those classified as the nearest to Earth. Understanding these objects helps us stay informed about potential risks and their proximity to our planet.

Asteroids:

• Asteroids are small, rocky objects that orbit the Sun, primarily found in the asteroid belt, which is located between the orbits of Mars and Jupiter.

•The potential harm from NEOs arises from the fact that their orbits bring them close to Earth, and in some cases, they can intersect Earth's orbit.

• The impact energy released in such an event could cause widespread damage, including massive explosions, fires, tsunamis, and potentially global environmental effects.

• The degree of harm depends on several factors, including the size, composition, speed, and angle of impact. Larger NEOs have the potential to cause more significant damage.

Learning Objective

DATA PREPARATION AND EXTRACTION

Importing the necessary libraries.

Numpy

Pandas

Matplotlib

Seaborn, etc.

DATA CLEANING AND TRANSFORMATION

Check for missing values

Outlier treatment

Encoding techniques (duplictes checking, etc.)

DATA EXPLORATION AND FEATURE SCALING

Check for missing values

Outlier treatment

Encoding techniques (duplicates checking, etc.)

TRAIN-TEST SPLITTING REGRESSION MODELS

Split the data in train and test.

Training the Linear regression model .

Objective

The objective is to use the features (est_diameter_min, est_diameter_max, relative_velocity, miss_distance, absolute_magnitude) to predict whether an asteroid is hazardous or not (binary classification)

Problem Statement

Developing a machine learning model to predict whether a Near-Earth Object (NEO) is hazardous based on its various characteristics. The model should be trained on the provided dataset, which includes information such as estimated diameter, relative velocity, miss distance, absolute magnitude, and a binary indicator of whether the object is hazardous.

Data Set

	A	B	C	D	E	F	G	H
1	id	name	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	absolute_magnitude	hazardous
2	2162635	162635 (2000 SS164)	1.198270801	2.679414966	13569.24922	54839744.08	16.73	FALSE
3	2277475	277475 (2005 WK4)	0.2658	0.594346868	73588.72666	61438126.52	20	TRUE
4	2512244	512244 (2015 YE18)	0.722029558	1.614507173	114258.6921	49798724.94	17.83	FALSE
5	3596030	(2012 BV13)	0.096506147	0.215794305	24764.30314	25434972.72	22.2	FALSE
6	3667127	(2014 GE35)	0.255008688	0.570216761	42737.73376	46275567	20.09	TRUE
7	5.4E+07	(2021 GY23)	0.036354232	0.081290534	34297.58778	40585691.23	24.32	FALSE
8	5.4E+07	(2021 PY40)	0.171614894	0.383742569	27529.47231	29069121.42	20.95	FALSE
9	5.4E+07	(2021 XD6)	0.005327887	0.011913517	57544.47008	55115019.26	28.49	FALSE
10	2088213	88213 (2001 AF2)	0.350392641	0.783501764	56625.21012	69035980.04	19.4	FALSE
11	3766065	(2016 YM)	0.105816886	0.23661375	48425.84033	38355261.56	22	FALSE
12	5.4E+07	(2020 OT6)	0.252670754	0.564988982	58430.6972	38337496.95	20.11	TRUE
13	5.4E+07	(2020 XW4)	0.152951935	0.342010925	64393.92832	71983105.31	21.2	FALSE
14	5.4E+07	(2021 AW1)	0.069912523	0.156329154	38018.61529	52093021.6	22.9	FALSE
15	5.4E+07	(2022 AM)	0.006145468	0.013741685	24323.04614	24617585.59	28.18	FALSE
16	2198752	198752 (2005 EA60)	0.290104841	0.648694146	10402.00218	60789296.03	19.81	FALSE
17	3069224	(2000 YT134)	0.483676488	1.081533507	74576.93076	59880809.9	18.7	FALSE
18	3739154	(2016 AF2)	0.006991252	0.015632915	75486.09085	71387057.95	27.9	FALSE
19	3795026	(2017 YU3)	0.04411182	0.098637028	70770.59114	27717237.02	23.9	FALSE
20	3797456	(2018 AN2)	0.029144391	0.065168838	42111.04408	39421282.19	24.8	FALSE
21	3825138	(2018 LC3)	0.46190746	1.032856481	104810.0937	18832837.96	18.8	FALSE
22	3835974	(2018 VK1)	0.009784868	0.02187963	49452.24058	57667088.95	27.17	FALSE
23	3842597	(2019 KN2)	0.010105434	0.022596438	27462.37801	38913011.95	27.1	FALSE
24	5.4E+07	(2020 RD4)	0.003051792	0.006824015	47644.44846	52948852.89	29.7	FALSE
25	2506491	506491 (2003 UW29)	0.201629919	0.450858206	115899.1805	15101017.14	20.6	TRUE
26	3329370	(2006 GB1)	0.050647146	0.113250461	33803.16656	70394867.73	23.6	FALSE
27	3623582	(2013 AU27)	0.4411182	0.986370281	69871.97731	58438716.67	18.9	FALSE
28	3768024	(2017 CQ)	0.175612319	0.392681082	71487.94083	21123317.71	20.9	FALSE
29	3781344	(2017 RV)	0.110803882	0.247765013	48655.30513	32797747.23	21.9	TRUE
30	5.4E+07	(2020 UF5)	0.031956189	0.07145621	17906.49109	66979305.81	24.6	FALSE

Columns in the Dataset

- **Id**: A unique identifier for each
- **Asteroid.Name**: The name or designation of the asteroid.
- **Est_diameter_min**: Minimum estimated diameter of the asteroid.(Kilometer)
- **Est_diameter_max**: Maximum estimated diameter of the asteroid. .(Kilometer)
- **Relative_velocity:** The relative velocity of the object with respect to Earth, measured in kilometers per hour.
- **Miss_distance: T**he closest distance the object comes to Earth during its orbit, measured in kilometers.
- **Absolute_magnitude:** A measure of the object's brightness as seen from Earth, where a lower value indicates a brighter object.
- **Hazardous:** A binary indicator (TRUE or FALSE) denoting whether the asteroid is classified as hazardous

About The Data

- **Shape of the data :** (90836, 8)
- **Info :**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90836 entries, 0 to 90835
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   id                  90836 non-null  int64
 1   name                90836 non-null  object
 2   est_diameter_min    90836 non-null  float64
 3   est_diameter_max    90836 non-null  float64
 4   relative_velocity   90836 non-null  float64
 5   miss_distance       90836 non-null  float64
 6   absolute_magnitude  90836 non-null  float64
 7   hazardous           90836 non-null  bool
dtypes: bool(1), float64(5), int64(1), object(1)
memory usage: 4.9+ MB
```

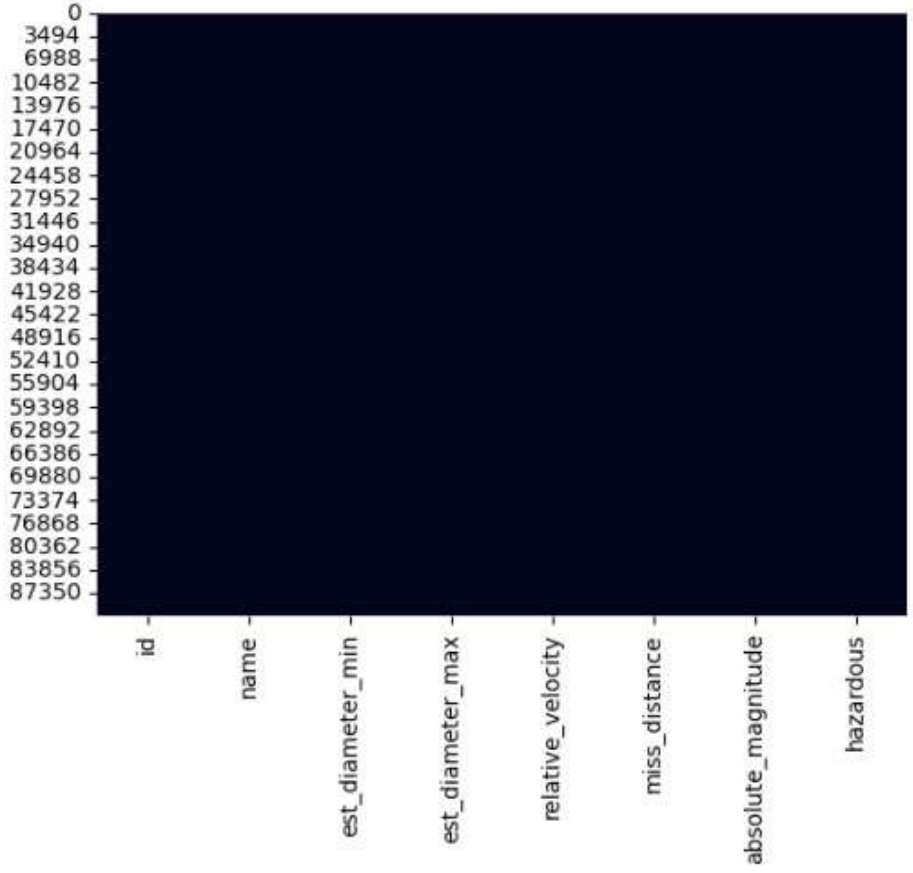
- **Describe:**

	id	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	absolute_magnitude
count	9.083600e+04	90836.000000	90836.000000	90836.000000	9.083600e+04	90836.000000
mean	1.438288e+07	0.127432	0.284947	48066.918918	3.706655e+07	23.527103
std	2.087202e+07	0.298511	0.667491	25293.296961	2.235204e+07	2.894086
min	2.000433e+06	0.000609	0.001362	203.346432	6.745533e+03	9.230000
25%	3.448110e+06	0.019256	0.043057	28619.020648	1.721082e+07	21.340000
50%	3.748362e+06	0.048368	0.108153	44190.117890	3.784658e+07	23.700000
75%	3.884023e+06	0.143402	0.320656	62923.604635	5.654900e+07	25.700000
max	5.427591e+07	37.892650	84.730541	236990.128100	7.479865e+07	33.200000

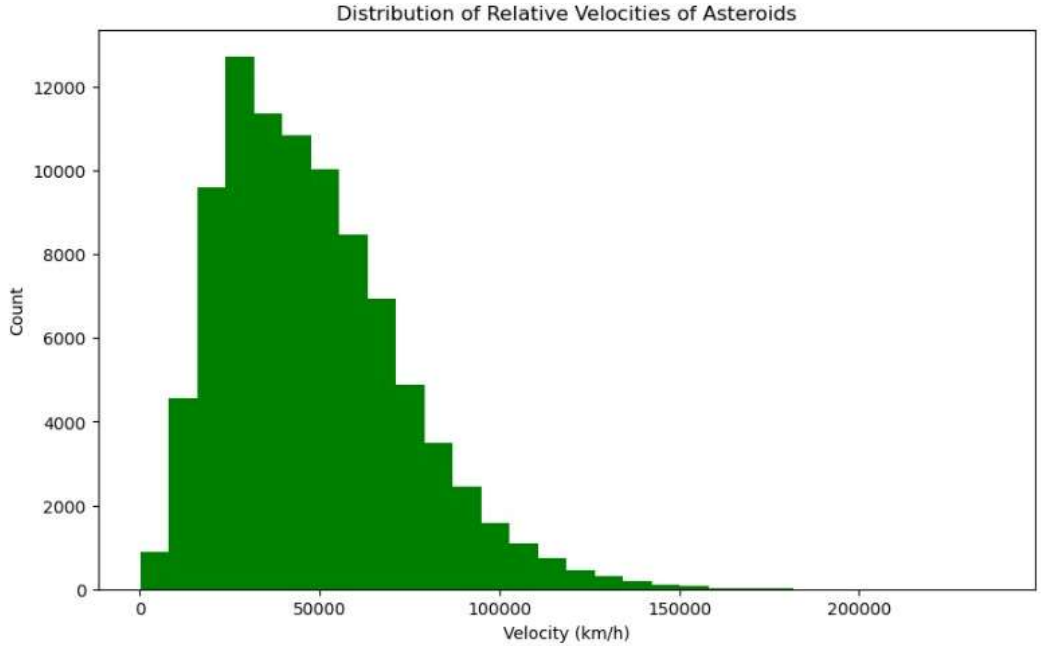
- **Null Values :**

```
id                  0
name                0
est_diameter_min    0
est_diameter_max    0
relative_velocity   0
miss_distance       0
absolute_magnitude  0
hazardous           0
dtype: int64
```

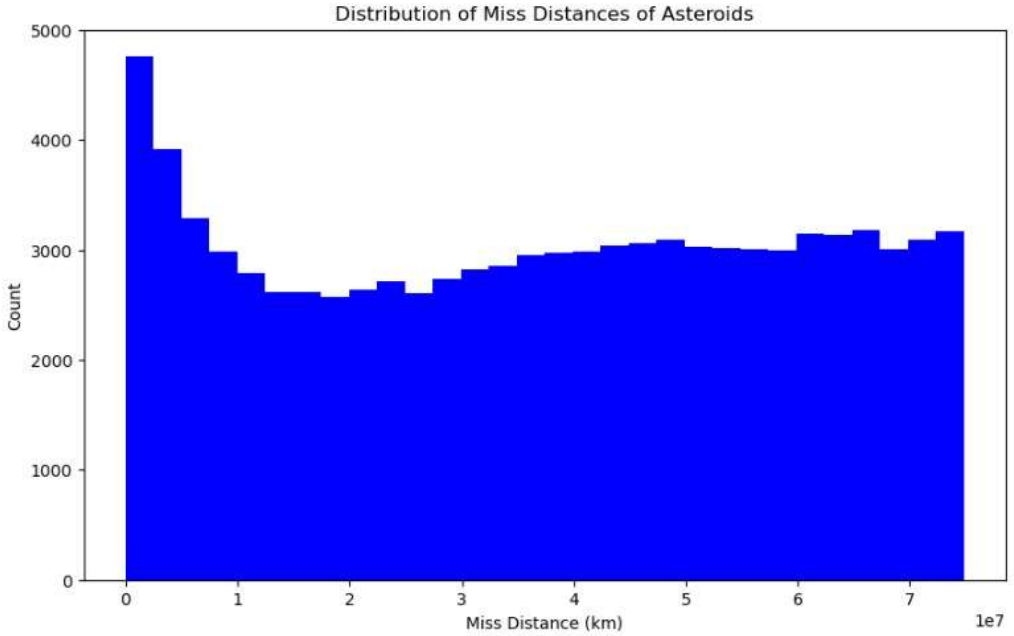
- **Heatmap to visualize null values :**



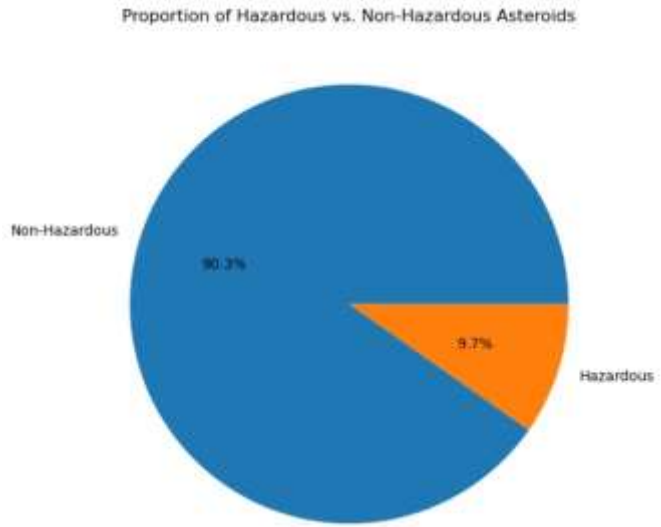
Visuals



This histogram illustrates the distribution of relative velocities of asteroids. The x-axis represents the velocity in kilometers per hour, and the y-axis shows the count of asteroids falling within each velocity range. This visualization provides insights into the speed of the asteroids as they approach Earth.

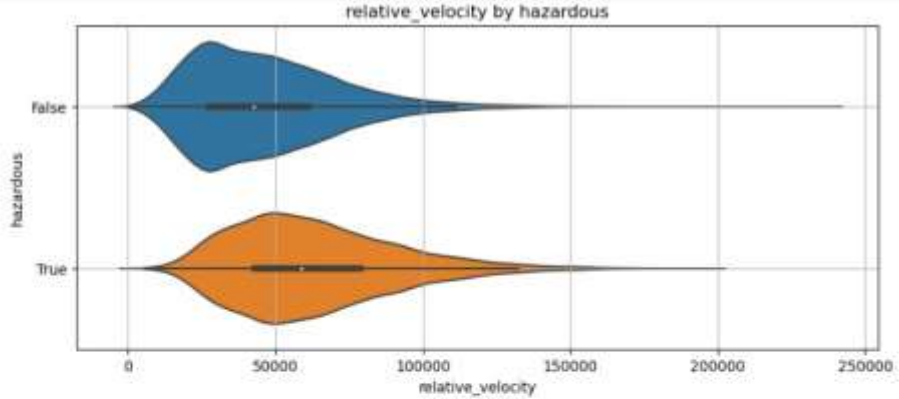


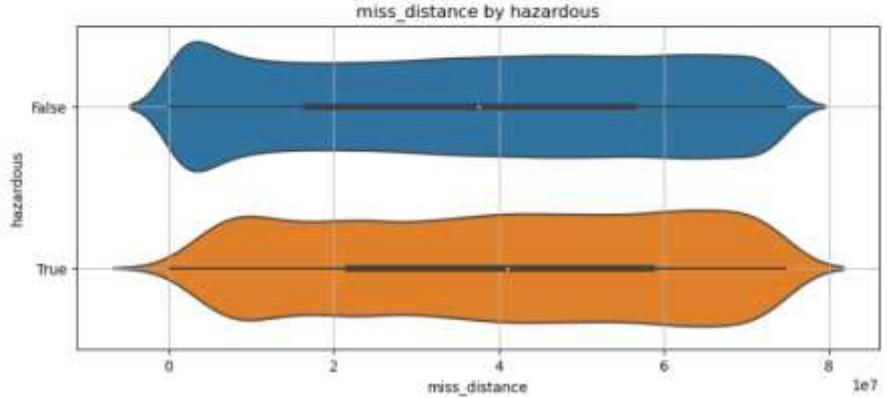
This histogram visualizes the distribution of miss distances of asteroids from Earth. The x-axis represents the miss distance in kilometers, and the y-axis shows the count of asteroids falling within each miss distance range. It provides insights into how closely these asteroids come to Earth.




```
hazardous
False    81996
True      8840
Name: count, dtype: int64
```

This pie chart represents the proportion of hazardous and non-hazardous asteroids in the dataset. It gives a visual overview of the safety concern by showing what percentage of asteroids are classified as hazardous. The legend provides clarity on the color-coding.





This violin plots is to visualize the distribution of each numeric feature with respect to the 'hazardous' status of the asteroids. Each plot will display a violin shape for each category of 'hazardous' (True or False). The width of the violin at different points represents the density of data, and the plot allows you to observe how the distribution of numeric values varies for hazardous and non-hazardous asteroids.

Model Performance Overview

Model	Accuracy	Classification Report			
XGBoost	90.12	precision	recall	f1-score	support
		Class 0	0.91	0.99	0.95	16373
		Class 1	0.50	0.09	0.15	1795
		accuracy			0.90	18168
		macro avg	0.70	0.54	0.55	18168
		weighted avg	0.87	0.90	0.87	18168
KNN	87.51	precision	recall	f1-score	support
		Class 0	0.90	0.96	0.93	16373
		Class 1	0.16	0.06	0.09	1795
		accuracy			0.88	18168
		macro avg	0.53	0.51	0.51	18168
		weighted avg	0.83	0.88	0.85	18168
RandomForest	89.49	precision	recall	f1-score	support
		Class 0	0.93	0.95	0.94	16373
		Class 1	0.46	0.38	0.42	1795
		accuracy			0.89	18168
		macro avg	0.70	0.67	0.68	18168
		weighted avg	0.89	0.89	0.89	18168
DecisionTree	88.96	precision	recall	f1-score	support
		Class 0	0.94	0.94	0.94	16373
		Class 1	0.44	0.44	0.44	1795
		accuracy			0.89	18168
		macro avg	0.69	0.69	0.69	18168
		weighted avg	0.89	0.89	0.89	18168

Conclusion

- All models perform well in predicting non-hazardous objects (Class 0) with high accuracy, precision, recall, and F1-score.
- Hazardous object (Class 1) prediction is more challenging for all models, with lower precision, recall, and F1-score.
- RandomForest and DecisionTree models show a better balance between the two classes compared to XGBoost and KNN.
- XGBoost appears to be the best-performing model overall, considering accuracy and other classification metrics.

After Hyper Tunning The XG-Boost Model

Tuned XGBClassifier Accuracy: 90.2

- After tuning the XGBoost classifier, the accuracy improved slightly from 90.12% to 90.2%. The tuning process likely resulted in fine-tuning the model's hyperparameters, contributing to a marginal increase in overall performance on the test data.

Model Performance after Oversampling and Undersampling

	Model	Original Data Accuracy	ROS Accuracy	RUS Accuracy
0	Decision Tree	0.890026	1.000000	1.000000
1	Random Forest	0.893989	1.000000	1.000000
2	XGBoost	0.901200	0.566844	0.566856
3	KNN	0.890357	0.549830	0.549397

- Oversampling and undersampling led to perfect training accuracy for Decision Tree and Random Forest but did not necessarily improve their performance on the original test data.
- XGBoost and KNN did not achieve perfect training accuracy after oversampling and undersampling, and their performance on the original test data was not significantly improved.

Thank you