NASA Close Approach Analysis

Tzu Tang Liu*
tliu33@ncsu.edu
NC State University
Raleigh, NC State, USA

Prathamesh Thakur*
pthakur@ncsu.edu
NC State University
Raleigh, NC State, USA

Sankar Raghuthaman*
sraghut@ncsu.edu
NC State University
Raleigh, NC State, USA

Arjit Agarwal*
aagarw38@ncsu.edu
NC State University
Raleigh, NC State, USA

Abstract

Near-Earth objects (NEOs) represent a significant area of study in planetary science and risk assessment due to their potential threat to Earth. This project utilizes a comprehensive dataset from NASA's JPL, capturing detailed information on NEOs that approached Earth over a defined period. By leveraging real-time data from the CAD API, we can analyze the behavior of these celestial bodies and develop predictive models to assess their trajectories and impact risks.

Keywords

Data fetching, cleaning, EDA, Time series analysis, K means clustering, KNN

ACM Reference Format:
Tzu Tang Liu, Prathamesh Thakur, Sankar Raghuthaman, and Arjit Agarwal. 2024. NASA Close Approach Analysis. In . ACM, Raleigh, NC , USA, 5 pages.

1 Introduction and Background

1.1 Problem Statement

Near-Earth Objects (NEOs), such as comets and asteroids, are celestial bodies that approach Earth's orbit and present potential risks due to their proximity and impact potential. Among these, NASA defines Potentially Hazardous Objects (PHOs) as NEOs that come within 0.05 astronomical units (AU) of Earth and have an absolute magnitude (H) of 22.0 or less, indicating sufficient size to cause significant regional damage in the event of an impact.[1] This classification highlights the importance of understanding key features such as distance, velocity, and magnitude to assess the risk of NEOs effectively. Early identification of potentially hazardous NEOs is paramount. Building a preventive model to quickly and accurately identify high-risk asteroids before they become a direct threat is critical for taking timely mitigation actions. By leveraging NASA's PHO criteria and analyzing feature interdependencies, this project aims to develop a robust machine learning model capable of detecting hazardous objects early, enabling proactive measures to safeguard Earth.

1.2 Related Work

The quantification and classification of near-Earth objects (NEOs) are critical to assessing potential collision threats. Uchenna et al. (2023) present a comprehensive analysis of NEOs detected between 1990 and 2021, using observational data to quantify object populations. The authors discuss the distribution, orbital characteristics, and size classifications of these objects, providing a detailed statistical overview of NEO activity over the three-decade period [3]. Their study offers a foundation for ongoing research in asteroid detection and tracking, which is vital for planetary defense systems.

Building on this, Malakouti et al. (2023) leverage machine learning techniques to classify dangerous asteroids. The study implements various algorithms, including support vector machines (SVM), decision trees, and neural networks, to differentiate hazardous asteroids based on parameters such as size, velocity, and orbital trajectory. The researchers demonstrate that machine learning models can achieve high classification accuracy, highlighting the potential for automated systems in future asteroid monitoring programs [4]. This technical approach to asteroid classification emphasizes the integration of data-driven methodologies into astronomical research.

McLemore (2022) further extends the application of machine-learning by utilizing decision tree classifiers and k-nearest neighbors (k-NN) algorithms for hazard prediction in asteroids. The paper emphasizes the importance of feature selection in improving model performance and discusses how specific features, such as eccentricity and perihelion distance, influence the classification process. The study provides practical information on the implementation of machine learning models in real-world asteroid tracking systems [2]. This work underscores the growing role of computational methods in addressing astronomical challenges.

2 Method

2.1 Novel Aspect

This project offers a fresh approach to assessing near-Earth objects (NEOs) by combining real-time data from NASA's Close Approach Data API with machine learning models. Using K-means clustering, NEOs are grouped based on patterns in features like distance, velocity, and magnitude, while K-Nearest Neighbors (KNN) classifies new objects into these clusters.

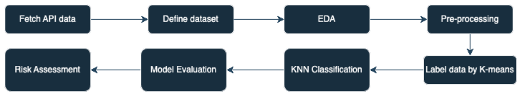

Figure 1: Work Flow of Analyze NEOs dataset

Unlike traditional methods that rely on fixed thresholds, this approach dynamically identifies potentially hazardous objects (PHOs) and provides actionable insights for planetary defense and mission planning. By mapping clusters to threat levels, the project supports both risk mitigation and the optimization of space exploration strategies.

2.2 Rationale

The vast and continuously growing dataset of NEOs requires efficient tools to analyze and classify their potential threat levels. Traditional methods often rely on fixed thresholds or manual evaluation, which may not adapt well to the dynamic nature of the data.

This project leverages machine learning to address these challenges. By using K-means clustering, we can uncover patterns and group NEOs based on their intrinsic characteristics, such as distance and velocity, which are difficult to interpret manually. Additionally, the use of K-Nearest Neighbors (KNN) allows for real-time classification of new NEOs, ensuring that any emerging risks are quickly identified.

The rationale lies in the need for a scalable, adaptive, and data-driven solution that not only identifies potentially hazardous objects (PHOs) but also provides insights that can guide defense strategies and optimize space exploration. This approach ensures a more effective and proactive response to the growing challenge posed by NEOs.

2.3 Approach

- **Extraction**: Data was read from a csv file and then was merged with the data collected from the JPL API.
- **Pre-processing**: For preprocessing of the data, duplicates and missing values were removed, and the numeric attributes of the data were normalized. The date and time were also extracted as separate columns from the data and the columns which were not useful for analysis were dropped. This was followed by an EDA of all the numeric attributes, which included their distributions.
- **K-means**: After wrangling of the data, the K-Means clustering algorithm was used to create two clusters based on the numeric attributes, one for potentially hazardous objects and the other for non-hazardous objects. The silhouette score was used to find the optimal number of clusters.
- **KNN**: Apart from K-Means, the KNN algorithm was also used for classification based on number of nearest neighbors. Again, two classes were used for the classification. The KNN algorithm classified the test data with 99% accuracy.
- **Risk assessment**: This step involved making use of the results generated from the analysis. It used domain knowledge to check the functioning and accuracy of the clustering algorithm.

3 Plan and Experiment

3.1 Dataset

The dataset used in this research is sourced from NASA's Center for Near Earth Object Studies (CNEOS), specifically focusing on near-Earth objects' close approaches during 2023-2024. This comprehensive dataset contains several critical parameters that characterize NEOs and their potential hazard classification.The dataset encompasses multiple features that are instrumental in determining an object's hazard potential.

Non numerical features include:

- Date-Time Measurements:
  - 'jd' (Julian Date): Exact time of closest approach.
  - 'cd' (Calendar Date): Human-readable date and time of closest approach.
  - t_sigma_f (Time Uncertainty): Uncertainty in the time of closest approach.
- Identifiers:
  - des (Designation): Unique identifier for each NEO.
  - orbit_id (Orbit Identifier): Identifier for the orbit calculation.

The primary numerical features include:

- Distance Measurements:
  - 'dist': The nominal approach distance
  - 'dist_min': Minimum approach distance
  - 'dist_max': Maximum approach distance
- Velocity Parameters:
  - 'v_rel': Relative velocity
  - 'v_inf': Infinity velocity
- Physical Characteristics:
  - 'h': Absolute magnitude, which serves as a proxy for the object's size

The dataset also includes temporal information through the 'cd' (close approach date) field, which has been preprocessed to separate into distinct date and time components for more in depth analysis. This temporal aspect is important for tracking and predicting NEO approaches.

A significant characteristic of this dataset is its handling of potentially hazardous objects (PHOs). The classification criteria align with NASA's established standards, where objects are considered potentially hazardous if they have a Minimum Orbit Intersection Distance (MOID) of 0.05 astronomical units or less, and an absolute magnitude of 22.0 or less. These thresholds are particularly important as they represent objects large enough to cause significant damage in the event of an Earth impact.

The dataset's quality is ensured using proper preprocessing steps, including handling missing values and applying appropriate normalization techniques. The numerical features are standardized using MinMaxScaler to ensure consistent scaling across all parameters. This preprocessing is essential for the following machine learning applications and ensures reliable model training and evaluation.

The dataset's concise nature and inclusion of multiple orbital and physical parameters make it particularly suitable for developing and testing machine learning models for NEO hazard classification. The

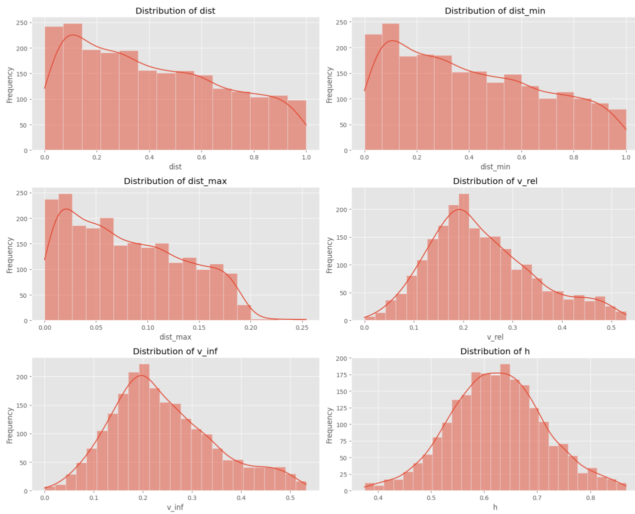

Figure 2: Feature distribution of numerical features

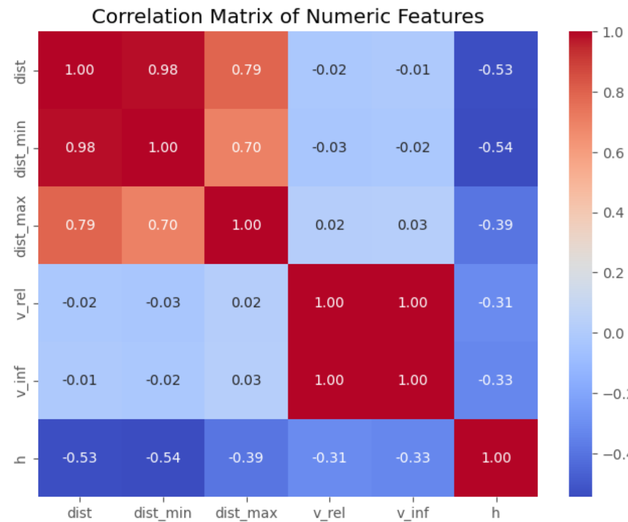

Figure 3: Correlation matrix of numerical features

data's structure allows for both supervised and unsupervised learning approaches, enabling robust analysis of NEO characteristics and their potential threats.

3.2 Hypothesis

This project investigates many key hypotheses regarding Near-Earth Objects (NEOs) and their classification using machine learning approaches. These hypotheses have been formulated prior to conducting experiments and are designed to yield meaningful insights regardless of the experimental outcomes.

- Primary Hypotheses
  - The combination of distance measurements (dist, dist_min, dist_max) and physical characteristics (absolute magnitude) can effectively distinguish between hazardous and non-hazardous NEOs with an accuracy exceeding random classification
  - Clustering algorithms can identify natural groupings in NEO characteristics that align with NASA's established hazard criteria (MOID ≤ 0.05 AU and absolute magnitude ≤ 22.0), potentially revealing additional classification patterns beyond traditional thresholds.
- Secondary Hypotheses
  - Velocity parameters (v_rel and v_inf) have a significant correlation with an object's hazard classification, independent of its physical size and distance measurements.
  - The temporal distribution of NEO approaches (based on approach dates) exhibits patterns that could enhance hazard prediction accuracy when combined with physical and orbital parameters.

These hypotheses are specifically designed to address gaps in current NEO classification methods and explore potential improvements in hazard assessment. Each hypothesis is testable through concrete experiments and statistical analysis, with outcomes that would be valuable regardless of whether they support or reject the initial assumptions. The investigation of these hypotheses could potentially reveal new insights about NEO characteristics and improve our ability to identify potentially hazardous objects.

3.3 Experimental Design

The experimental methodology is structured to systematically evaluate our hypotheses about NEO classification and hazard assessment through a series of controlled machine learning experiments.

- Data Splitting Strategy
  - The dataset is split into training (80%) and test (20%) sets using a random state of 42 to ensure reproducibility
  - Cross-validation will be implemented during model training to ensure robust performance evaluation.
- Data Preprocessing Pipeline
  - A systematic preprocessing pipeline is implemented using scikit-learn's Pipeline and ColumnTransformee.
  - Numeric features undergo MinMaxScaler normalization for the following parameters: dist, dist_min, dist_max, v_rel, v_inf, h

Experimental Approach
For Hypothesis 1:

- Implementation of K-Nearest Neighbors (KNN) classifier with the following parameters:
  - n_neighbors = 6
  - Feature set: normalized distance, velocity, and magnitude parameters.
  - Evaluation metrics: accuracy, precision, recall, and F1-score

For Hypothesis 2:

- Application of K-means clustering with:
  - Optimal cluster determination using silhouette score analysis
  - Principal Component Analysis (PCA) for dimensionality reduction and visualization.

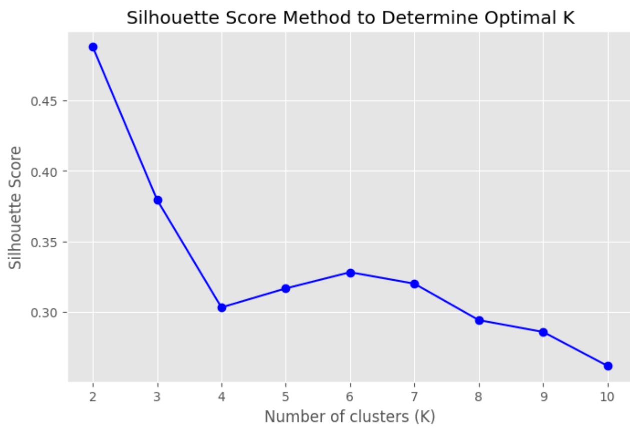

Figure 4: Silhouette score graph to select ideal cluster size

- Validation through cluster centroid analysis

For Hypothesis 3:

- Correlation analysis between velocity parameters and hazard classification
- Statistical significance testing of velocity features
- Feature importance analysis using PCA components

For Hypothesis 4:

- Temporal pattern analysis using the preprocessed date information.
- Time series analysis of approach frequencies.
- Integration with physical parameters for enhanced prediction.

The model performance is compared against two baselines:

- NASA's traditional classification criteria (MOID $\leq$ 0.05 AU and absolute magnitude $\leq$ 22.0)
- Random classification baseline

A significant challenge in this experimental design is the class imbalance inherent in NEO datasets, where hazardous objects are typically less common. This is addressed through:

- Careful selection of evaluation metrics that account for class imbalance.
- Implementation of clustering before classification to identify natural groupings.
- Use of PCA to reduce dimensionality while preserving important feature relationships.

To ensure experiment reproducibility, the following elements are standardized:

- Fixed random seed (42) for all random operations
- Documented preprocessing pipeline
- Specified hyperparameter values
- Consistent evaluation metrics across all experiments

This experimental design provides a robust framework for testing our hypotheses while maintaining scientific rigor and reproducibility. The multi-faceted evaluation approach ensures that results will be meaningful regardless of the specific outcomes.

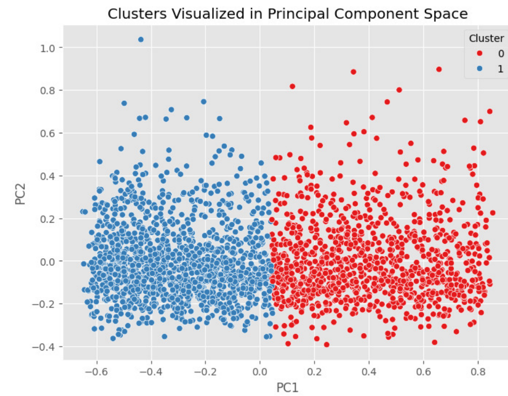

Figure 5: K-Means clusters visualized using PCA

	dist	dist_min	dist_max	v_rel	v_inf	h
PC-1	0.696989	0.687449	0.137747	-0.004901	-0.000804	-0.150403
PC-2	-0.017731	-0.029499	0.010329	0.687941	0.686196	-0.233625

Figure 6: Weights for the features according to PCA

4 Result

As discussed in the previous section, we used K-Means for clustering the Near Earth Objects data according to the numeric features based on distance, velocity and absolute magnitude. The K Nearest Neighbours algorithm was also used to judge the cluster divisions.

For the KNN algorithm, it performed with 99 percent accuracy, meaning the trends of the data were accurately captured and it was able to classify the data point correctly according to the values of its six nearest neighbors.

For the K Nearest Neighbors algorithm, the silhouette score was first computed for the range of number of clusters from 2 to 10. The silhouette score estimates the intra cluster density and inter cluster separability. A good silhouette score closer to 1 is considered best for the algorithm. The maximum silhouette score was for 2 clusters, which was close to 0.5. Hence, the number of clusters chosen were 2 and the class labels were considered either hazardous or not hazardous. Then, based on the numeric features, the model classified the points.

The formulation of these hypotheses aligns with both the practical requirements of NEO monitoring and the scientific rigor needed for meaningful research conclusions. They avoid the common pitfall of simply assuming the effectiveness of the proposed approach and instead focus on testing specific, measurable relationships between variables.

Figure 5 shows the cluster distribution for two principal components. Since the number of features used for clustering were six, visualization for the same wasn't possible, so PCA was used for dimensionality reduction and visualization.

Figure 6 shows the weights assigned according to the principal components for all features. As observed, the maximum weights are given to the distance features. This suggests that distance plays an important role in the clustering.

Based on the domain knowledge discussed in the hypothesis section, it is understood that the object is considered hazardous

	des	orbit_id	Date	Time	cd	dist	dist_min	dist_max	v_rel	v_inf	t_sigma_f	h
2982	2006 WB	31.0	2024-11-26	18:01:00	2024-Nov-26 18:01	0.005958	0.005958	0.005958	4.203735	4.095962	< 00:01	22.800
2983	2024 WD3	1.0	2024-11-27	03:53:00	2024-Nov-27 03:53	0.011574	0.011540	0.011608	5.361120	5.318005	< 00:01	28.124
2984	2024 WQ2	1.0	2024-11-27	08:01:00	2024-Nov-27 08:01	0.013900	0.013853	0.013948	17.308632	17.297554	00:02	24.580
2985	2024 WF3	1.0	2024-11-28	11:15:00	2024-Nov-28 11:15	0.027802	0.027698	0.027905	6.897005	6.883095	00:03	27.330
2986	2024 WT	4.0	2024-11-28	12:44:00	2024-Nov-28 12:44	0.005981	0.005973	0.005988	3.841196	3.723411	00:01	27.093
2987	2018 DC4	2.0	2024-11-29	07:28:00	2024-Nov-29 07:28	0.048082	0.042169	0.058864	4.486259	4.473889	1_03:07	27.300
2988	2024 WZ	2.0	2024-11-29	22:49:00	2024-Nov-29 22:49	0.009674	0.009653	0.009696	9.914152	9.886332	00:04	26.906
2989	2019 JN2	9.0	2024-11-30	07:39:00	2024-Nov-30 07:39	0.023970	0.023420	0.024542	7.713242	7.698817	01:04	25.720
2990	2024 WS	2.0	2024-11-30	15:35:00	2024-Nov-30 15:35	0.015734	0.015635	0.015833	10.519232	10.503121	00:06	26.533
2991	2021 XZ	1.0	2024-12-02	09:14:00	2024-Dec-02 09:14	0.033762	0.020087	0.047621	7.397151	7.386475	1_22:20	28.210
2994	2024 WO1	1.0	2024-12-04	20:03:00	2024-Dec-04 20:03	0.030042	0.029555	0.030529	4.332503	4.311983	00:02	26.953
2995	2021 WA5	4.0	2024-12-05	00:02:00	2024-Dec-05 00:02	0.025125	0.000035	0.058893	5.919745	5.901804	6_08:03	27.290
2996	2024 WB1	2.0	2024-12-05	23:40:00	2024-Dec-05 23:40	0.048972	0.048630	0.049314	5.804202	5.794820	00:04	25.668
2997	2024 WG3	1.0	2024-12-07	03:56:00	2024-Dec-07 03:56	0.046978	0.046756	0.047201	6.455722	6.446931	00:01	24.764
2998	2024 UU3	5.0	2024-12-08	20:47:00	2024-Dec-08 20:47	0.043335	0.043247	0.043422	4.801197	4.788373	00:03	24.826

Figure 7: Entries from the data which are considered hazardous according to domain knowledge

	dist	dist_min	dist_max	v_rel	v_inf	h	des	orbit_id	Date	Time	cd	t_sigma_f	Cluster
76	0.98029	0.973631	0.187358	0.136008	0.142598	0.585918	2024 WB1	2.0	2024-12-05	23:40:00	2024-Dec-05 23:40	00:04	0
92	0.545575	0.545892	0.10355	0.150146	0.156478	0.511353	2012 UK171	51.0	2025-01-11	00:40:00	2025-Jan-11 00:40	< 00:01	0
287	0.861897	0.862121	0.163587	0.262141	0.267869	0.574595	2018 XU3	11.0	2024-12-11	08:27:00	2024-Dec-11 08:27	< 00:01	0
430	0.923007	0.56674	0.242854	0.240064	0.245956	0.768536	2017 YD2	2.0	2024-12-27	22:01:00	2024-Dec-27 22:01	00:16	0
621	0.867303	0.865782	0.164944	0.110588	0.117283	0.535204	2024 UU3	5.0	2024-12-08	20:47:00	2024-Dec-08 20:47	00:03	0
750	0.568944	0.56615	0.172848	0.324864	0.330084	0.386677	2022 CE2	8.0	2025-01-16	20:29:00	2025-Jan-16 20:29	3_21:31	0
780	0.479211	0.468527	0.093127	0.184391	0.190488	0.58905	2019 JN2	9.0	2024-11-30	07:39:00	2024-Nov-30 07:39	01:04	0
845	0.65639	0.587051	0.137818	0.211683	0.217704	0.666145	2020 XY4	8.0	2024-12-19	23:45:00	2024-Dec-19 23:45	04:23	0
928	0.852412	0.811719	0.171423	0.264328	0.270039	0.505933	2023 OS	6.0	2025-01-15	14:50:00	2025-Jan-15 14:50	00:25	0
979	0.962448	0.84417	0.223683	0.102606	0.109373	0.684214	2018 DC4	2.0	2024-11-29	07:28:00	2024-Nov-29 07:28	1_03:07	0
1169	0.589578	0.452091	0.138272	0.219051	0.224998	0.608324	2024 BM1	3.0	2025-01-08	22:23:00	2025-Jan-08 22:23	06:14	0
1173	0.730058	0.721907	0.140162	0.174152	0.180438	0.608324	2024 BM1	3.0	2024-01-08	00:16:00	2024-Jan-08 00:16	< 00:01	0
1391	0.555999	0.554249	0.10592	0.163704	0.16997	0.686021	2024 WF3	1.0	2024-11-28	11:15:00	2024-Nov-28 11:15	00:03	0
2064	0.675446	0.401761	0.180916	0.17638	0.182632	0.739023	2021 XZ	1.0	2024-12-02	09:14:00	2024-Dec-02 09:14	1_22:20	0
2086	0.600893	0.591445	0.1159	0.098709	0.105301	0.663314	2024 WO1	1.0	2024-12-04	20:03:00	2024-Dec-04 20:03	00:02	0
2538	0.940327	0.936082	0.179317	0.15252	0.159	0.53147	2024 WG3	1.0	2024-12-07	03:56:00	2024-Dec-07 03:56	00:01	0

Figure 8: Entries from the data which are considered hazardous according to K-Means

	des	orbit_id	Date	Time	cd	dist	dist_min	dist_max	v_rel	v_inf	t_sigma_f	h
2995	2021 WA5	4.0	2024-12-05	00:02:00	2024-Dec-05 00:02	0.025125	0.000035	0.058893	5.919745	5.901804	6_08:03	27.29
3002	2022 YO1	2.0	2024-12-17	15:40:00	2024-Dec-17 15:40	0.005058	0.000036	0.013485	14.333602	14.296800	15:24	30.01

Figure 9: Human explorable NEOs

if the MOID (distmin attribute) is less than 0.05 and the absolute magnitude(diameter/h) is greater than or equal to 22. Based on this domain knowledge, we narrowed down the rows from the dataset.

Figure 7 shows an excerpt of the rows from the dataset which are considered hazardous according to domain knowledge. After this, we compared this table with the objects that were classified as hazardous based on the clustering(Figure 8). 16 of the objects were classified as hazardous in both the results.

The result is turning out to be so due to the fact that the data on hazardous objects is quite less. The objects which fit into the hazardous criteria make up a mere 26 rows out of 3002 rows. For more accurate results, more data should be used for hazardous object types. The API is constantly updating the data with new findings, so the results will depend on the nature of the new data.

Human exploration of planets is well known, but humans have also explored NEOs to mine material samples. The most recent example of this was carried out by JAXA, which brought back samples from the asteroid Ryugu in 2020. The mission took well over 5 years to complete. By using the data for the asteroid, the mapping of potential NEOs which are explorable for humans was also found out.

Figure 9 shows the NEOs which can be potentially explored by humans. Based on the condition that the minimum distance should be less than 0.0006 au, the NEOs narrowed down can be potentially explored by humans for understanding different space materials.

5 Conclusion

The universe is ever expanding and there are constantly new discoveries being made in this domain. While NEOs have gained attention since the 1980s, recently, more efforts have been made to monitor these objects to understand not only potential risk which these might pose, but also from an exploratory perspective. This project helped us understand how data is recorded for space objects, and how information from analyzing it can be used to set up countermeasures or explore. This domain is ever expanding and new objects are always being discovered. We also learnt that apart from this application, there are various web apis which provide a wealth of data to work with. We hope to use them for the next application.

References

[1] Center for Near Earth Object Studies (CNEOS). 2024. *About NEO Groups.* Retrieved November 25, 2024 from https://cneos.jpl.nasa.gov/about/neo_groups.html

[2] Trotter McLemore. 2022. *Classifying Hazardous and Non-Hazardous Asteroids Using Machine Learning.* Retrieved October 19, 2024 from https://nhsjs.com/2022/classifying-hazardous-and-non-hazardous-asteroids-using-machine-learning/#google_vignette

[3] Aernan Andofa Ojima Ocheni Aneke Ukpabi Wash Albert Madike Emeka Michael Justus Chukwunonyerem Okolo Obinna Okoro Emmanuel Chineche Okpara Richard Tobechi Vincent Osinachi Ezinwa Ekechukwu Uzoamaka Obasi Rita Chioma Ezenwukwa Nnenna Dorathy Osuji Emmanuel Uchenna, Ugwu Monday. 2023. Quantification of Near-Earth Objects Between 1990-2021: An Overview. *American Journal of Astronomy and Astrophysics*, Article 5 (june 2023). To appear.

[4] Amir Abolfazl Suratgar Seyed Matin Malakouti, Mohammad Bagher Menhaj. 2023. *Machine learning techniques for classifying dangerous asteroids.* Retrieved October 19, 2024 from https://www.sciencedirect.com/science/article/pii/S2215016123003345

6 Meeting Attendance

The meetings were scheduled every 2 weeks on Saturdays.

- October 12 2024: All 4 present
- October 26 2024: All 4 present
- November 9 2024: All 4 present
- November 23 2024: All 4 present
- November 24 2024: All 4 present
- November 25 2024: All 4 present

Other than that, majority of our communication requirements were satisfied through text channels.