Classifying the NEA’s and PHA’s and Predicting the Impact Collisions of the Asteroids

Jocelyn Baduria, Bhuvana Gopala Krishna Basapur, and Weifeng Ma, *SJSU*

**Abstract**—We used datasets from NASA JPL and CNEOS Sentry System, and we used Muller Loop Classification with five selected classifier models to identify the NEA’s and PHA’s and Muller Loop Regression with seven selected regression models to predict the impact collisions of the Asteroids.

**Index Terms**—NEA (Near Earth Asteroids), NEO (Near Earth Object), PHA (Potentially Hazardous Asteroids), NASA (National Aeronautics and Space Administration), JPL (Jet Propulsion Laboratory), CNEOS (Center for Near Earth Object Studies).

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

————————————————

* *Jocelyn Baduria; E-mail: jocelyn.baduria@sjsu.edu.*
* *Bhuvana Gopala Krishna Basapur; E-mail: bhuvanagopalakrishna.basapur@sjsu.edu*
* *Weifeng Ma; E-mail: weifeng.ma@sjsu.edu*
* *Code Repository:* [*UnderdogTeam github link*](https://github.com/jocelynbaduria/cmpe-257_Numeric_Project)*(See Appendix A)*

—————————— ◆ ——————————

# 1 Introduction

An Asteroid is a small rocky body that orbits the Sun. They are smaller than a planet, but larger than the pebble-size objects called meteoroids. Most asteroids in our solar system are found in the region between Mars and Jupiter (**asteroid belt)**.

They can also hang out in other locations around the solar system. For example, some asteroids orbit the Sun in a path that takes them near Earth.

Near-Earth Asteroids (**NEA’s**) are asteroids that have been nudged by the gravitational attraction of nearby planets into orbits that allow them to enter the Earth’s neighborhood. NEA’s are asteroids with perihelion distance less than 1.3 [**au**](https://cneos.jpl.nasa.gov/glossary/au.html). NEA’s are divided into groups (Atira, Aten, Apollo and Amor) according to their perihelion distance, aphelion distance and semi-major axes.

Potentially Hazardous Asteroids (**PHAs**) are asteroids with potential to make threatening close approaches to the Earth.

Though the probability of an asteroid hitting the earth is very low, the impact of collision has catastrophic effects (A huge asteroid wiped out the dinosaurs around 66 million years ago).

In this project we want to find the asteroids which are Earth-crossing NEA’s (with semi-major axis>1 and perihelion distance <1.017) and are Potentially Hazardous Asteroids (PHA) and predict the impact of collision of such asteroids with a danger level of (dia=1Km, neo=1, pha=1).

# 2 Data Collection

We obtained data from Jet Propulsion Laboratory, maintained by NASA for small body objects (Comets and Asteroids).

Data about the asteroids’ orbital and model parameters are from <https://ssd.jpl.nasa.gov/sbdb_query.cgi#x>.

Data about asteroids potential future Earth impact events are maintained by JPL Sentry system and is available from <https://cneos.jpl.nasa.gov/sentry/>.

The API from <https://ssd-api.jpl.nasa.gov/> is used to scrape or get our latent variables and related data. This API is provided by NASA.

We used data preparation, wrangling techniques to process the data so we can use it for the data exploratory visualization analysis.

## 2.1 NASA JPL DATA

For NASA JPL data, we collected data from NASA hosted public dataset in Jet Propulsion Laboratory.

Collected data was prepared using google colab tool to clean the dataset mainly focusing on NEO and PHA. The dataset collected has features spkid, fullname, pdes, name, e, a, q, i, om, w, ma, ad, n, moid, neo, pha. These abbreviations have been identified as the following by NASA JPL.

1. **Spkid** - object primary SPK-ID
2. **fullname**- object full name
3. **pdes** - object primary designation
4. **name** - object IAU name
5. **e** - eccentricity, is a measure of how far from circular each orbit is
6. **a** - semi-major axis (au), describes an object's distance from the Sun.
7. **q**- perihelion distance (au), means the closest point to the sun for the object, and "astronomical unit" or AU
8. **i** - inclination; angle with respect to x-y ecliptic plane (degree)
9. **om**- Longitude of the ascending node, is one of the orbital elements used to specify the orbit of an object in space
10. **w** - argument of perihelion
11. **ma** - Mean anomaly(deg) - The locations of the asteroid in a certain time period
12. **ad** - aphelion distance(au), the point in the orbit of a planet
13. **n** - Mean motion(deg/d)
14. mood- earth minimum orbit intersection distance (au)
15. **neo** - Near Earth Object
16. **pha** - Physically Hazardous Asteroid

The data collected is based on the latest dataset from NASA JPL. We only selected the needed features for this project.

For the data table summary, we did some feature engineering to remove some of the missing values for name, ma, ad, moid, neo and pha. After data preparation we performed exploratory data analysis mainly on neo and pha using Plotly express for interactive visualization of the NASA JPL data.

## 2.2 NASA JPL Visualization Exploratory Analysis

##### 

Fig. 1. NEA (Near Earth Asteroids) total counts 162.

We can see that there is a total count of 162 for near earth asteroids. We looked deeper and sorted the names with Near Earth Asteroids. Below is the list of names of Asteroids that are near to earth.

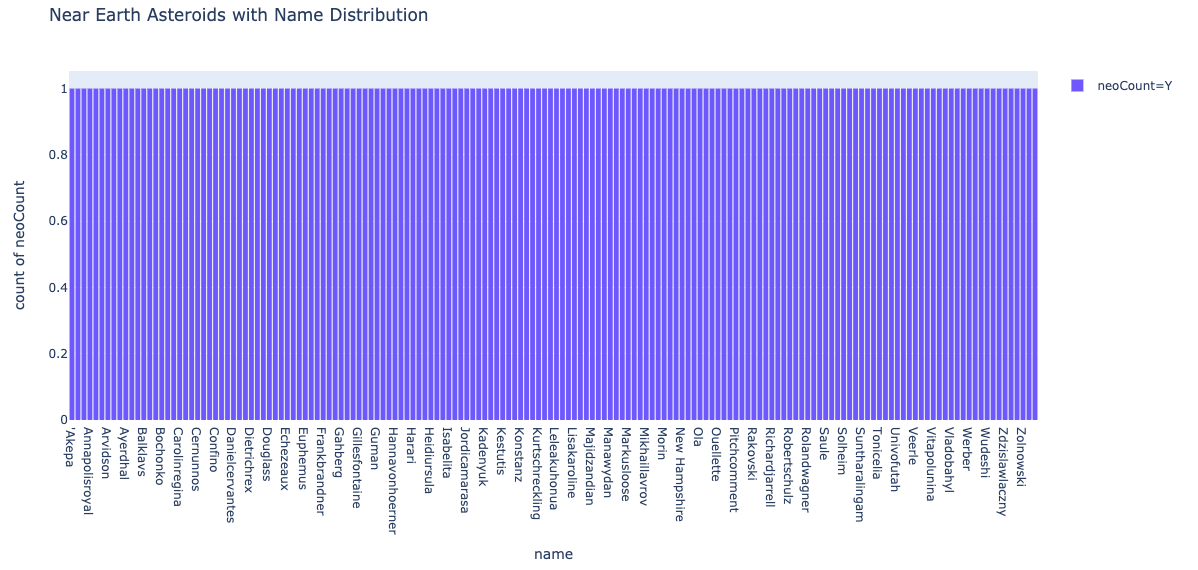


Fig. 2. NEA with name asteroids distributions

We also examined if these near-earth asteroids are potentially hazardous asteroids. There are a total 45 potentially hazardous asteroids.

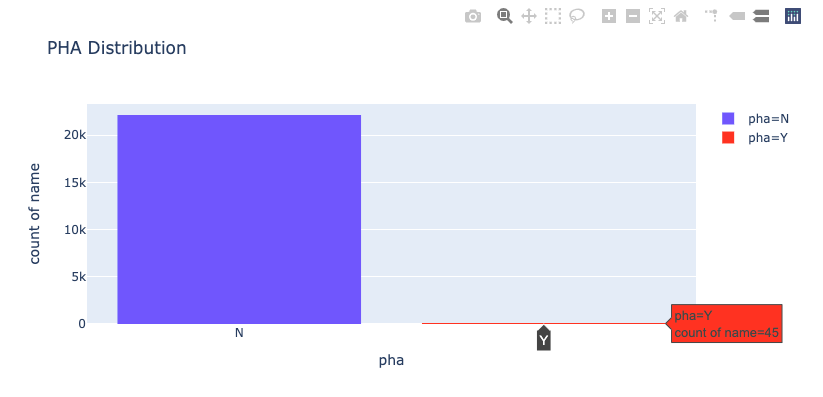


Fig. 3. PHA (Potentially Hazardous Asteroids) total counts 45.

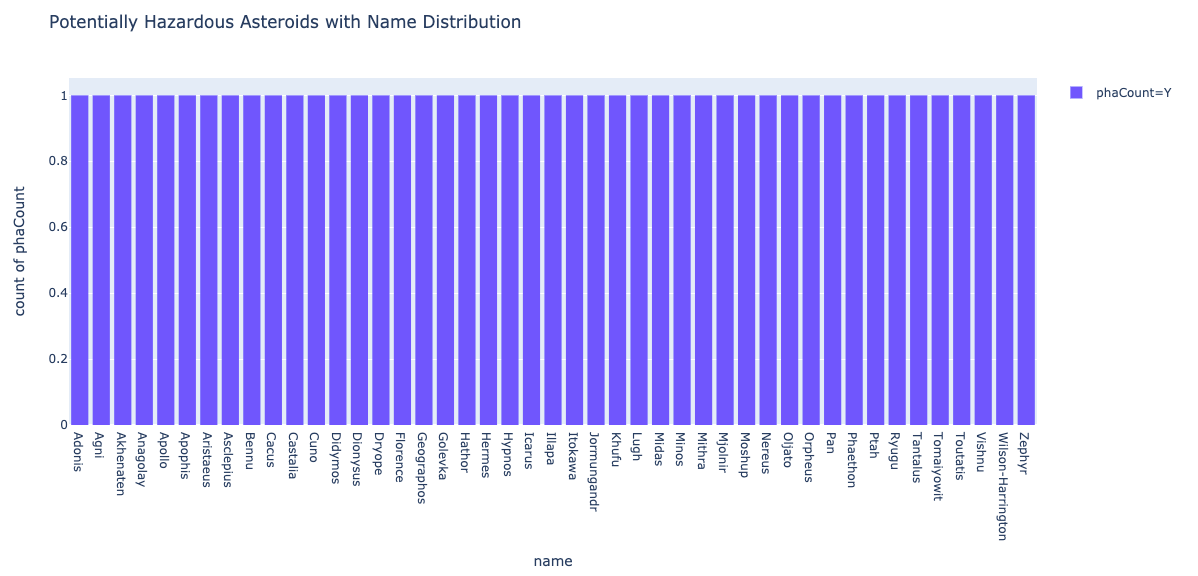


Fig. 4. PHA with name distributions

## 2. 2.3 CNEOS Sentry System Data

For CNEOS Sentry System data, we collected data from NASA hosted public dataset in CNEOS Jet Propulsion Laboratory.

Similarly, collected data was prepared using google colab tool to clean the dataset mainly focusing on Potential Impact and NEA. The dataset collected has features Object Designation, Year Range, Potential Impacts, Impact Probability (cumulative), Vinfinity, h, Estimated Diameter(km), Palermo Scale, Palermo Scale (max.). These abbreviations have been identified as the following by NASA JPL.

1. **Object Designation** - Temporary designation or permanent number for this object
2. **Year Range** - Time span over which impacts have been detected
3. **Potential Impacts** - Number of dynamically distinct potential impacts that have been detected by Sentry
4. **Impact Probability (cumulative)** - Sum of the impact probabilities from all detected potential impacts.
5. **Vinfinity (km/s)** - Velocity of the asteroid relative to the Earth, assuming a massless Earth.
6. **H (mag)** - Absolute Magnitude, a measure of intrinsic brightness. It is the apparent magnitude of the object when it is 1 au from both the sun and the observer, and at full phase for the observer
7. **Estimated Diameter (km)** - Estimated diameter of the asteroid
8. **Palermo Scale (cum.)** - Cumulative hazard rating according to the Palermo technical impact hazard scale, based on the tabulated impact date, impact probability and impact energy
9. **Palermo Scale (max.)** - Maximum hazard rating according to the Palermo technical impact hazard scale, based on the tabulated impact date, impact probability and impact energy

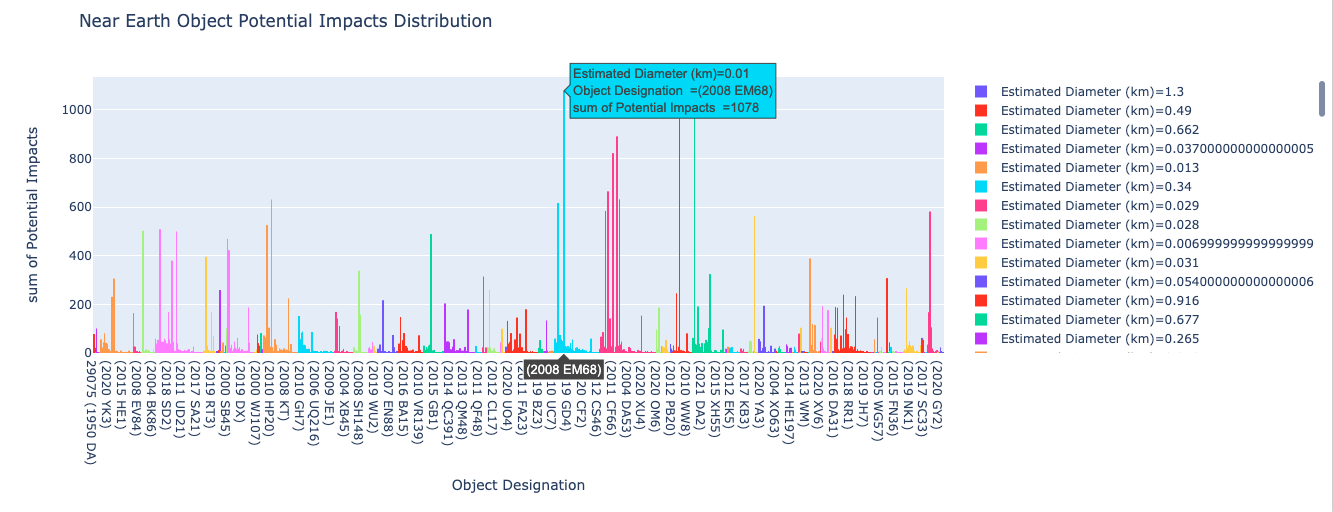
The data collected is based on the latest dataset using unconstrained settings for downloading. We only selected the needed features for this project.

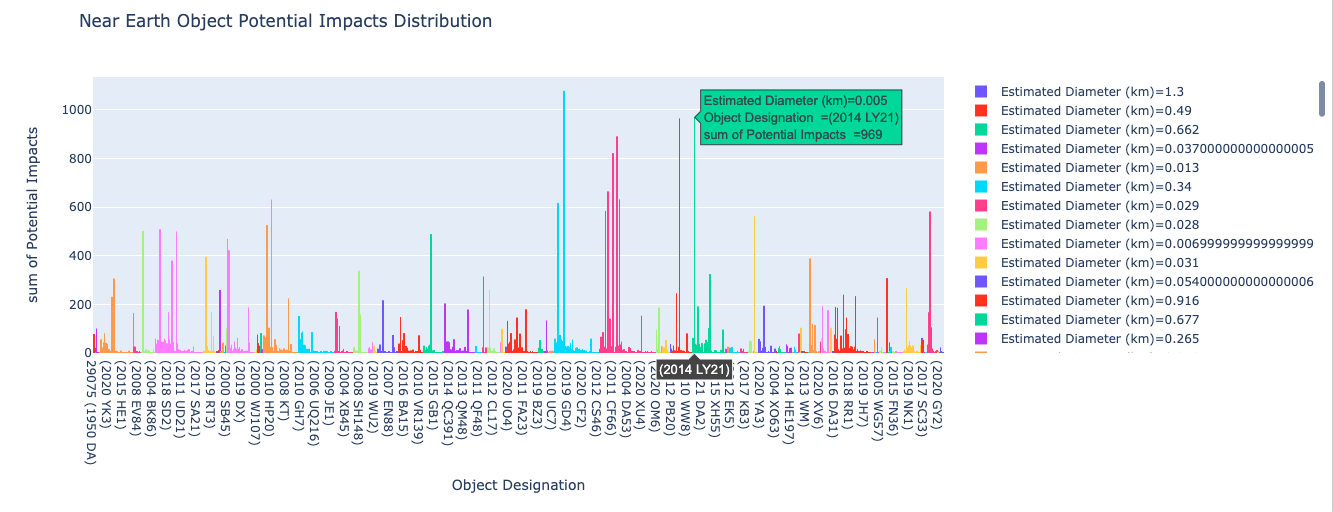
We renamed some of the features for clearer and easy identification of the features for data analysis.

For the data summary we did feature engineering and we have to remove the columns that have no sense for analysis like columns with values that were all zeros Torino Scale (max.) and no column name Unnamed: 10. Also, we remove some missing values in column Vinfinity. Then we performed the exploratory data analysis mainly on Potential Impacts, Estimated Diameter with respect to NEA using Plotly express for interactive visualization of the NASA JPL data.

## 2.4 CNEOS Sentry System Data Visualization Exploratory Analysis

The highest contributor of Near-Earth Asteroids is the 2008 EM68 followed by 2014 LY21 with a total of 969.

Fig. 5. Top contributor of NEA with potential impacts crossing to earth is 2008 EM68. Total potential impacts are 1078 with an estimated diameter of 0.01.

Fig. 6. Second contributor of NEA with potential impacts crossing to earth is 2014 LY21 with a total of 969 with an estimated diameter of 0.005.

We also consider the Asteroids that have the highest cumulative potential impact in crossing near the earth. This asteroid with object designation as (2010 RF12) has an impact cumulative probability of 0.047 with a total 59 potential impact of crossing to earth.

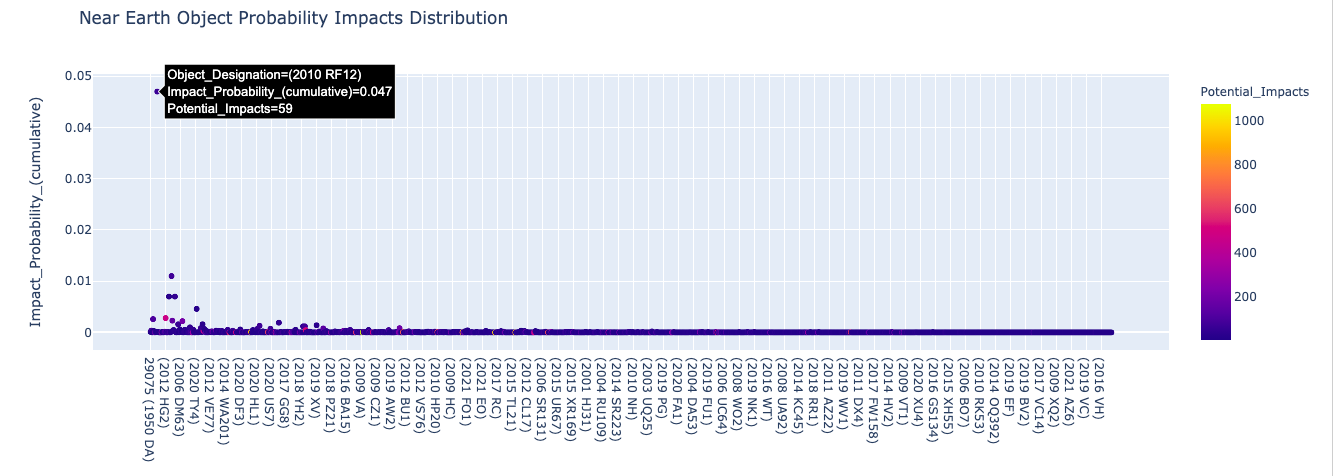


Fig. 7. (2010 RF12) with highest cumulative probability of 0.047.

Considering the asteroids with highest probability we also check if (2010 RF12) has a bigger diameter. We want to understand if this will cross the earth how big is the impact based on estimated diameter.

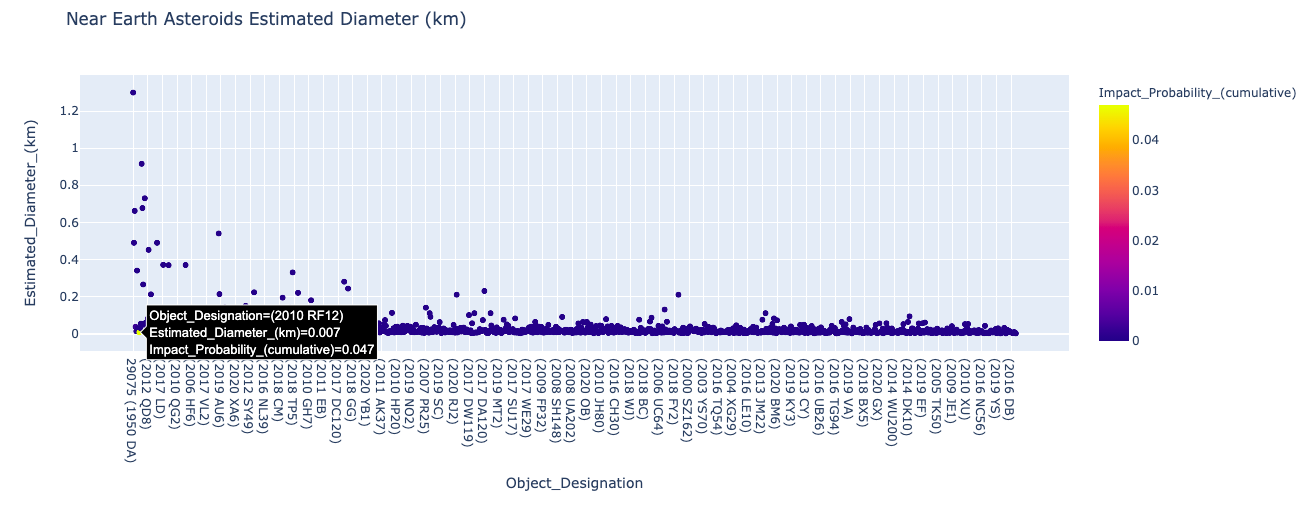


Fig. 8. (2010 RF12) NEA’s has a diameter of 0.047.

The biggest diameter of asteroids has a diameter of 1.3km and this asteroid has a probability of 0.00012 which is lower than the ((2010 RF12).

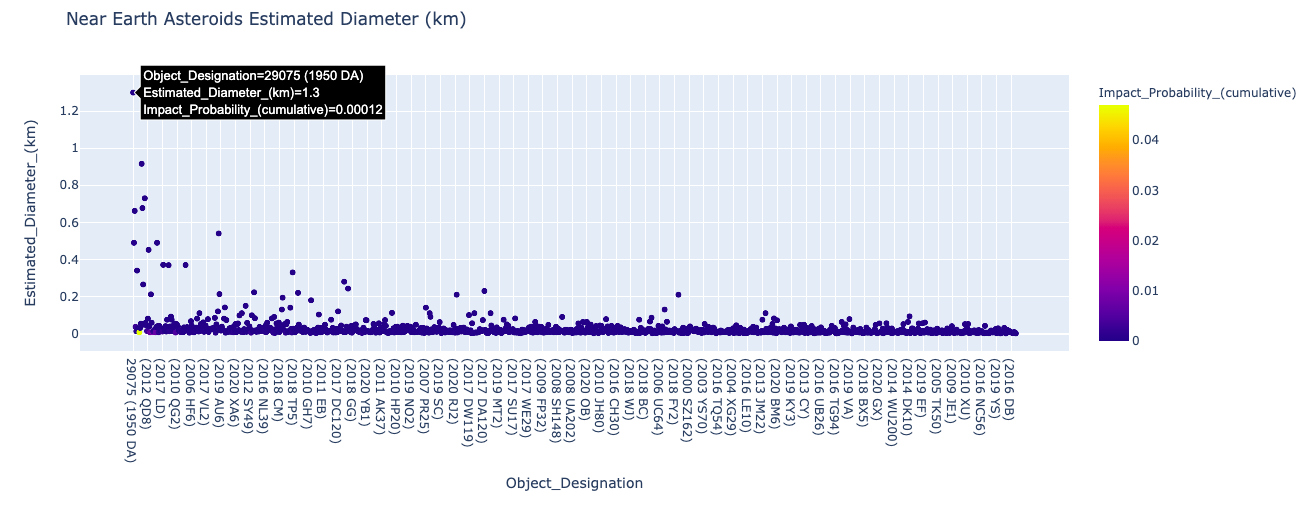


Fig. 9. (1950 AD) has a diameter of 1.3km

## 2. 2.5 Feature Importance

Using our first data enrichment with Random Forest Gini Importance and Permutation Importance we identified these features in our NASA JPL dataset as highly important. The full\_name, pdes and the spkid.

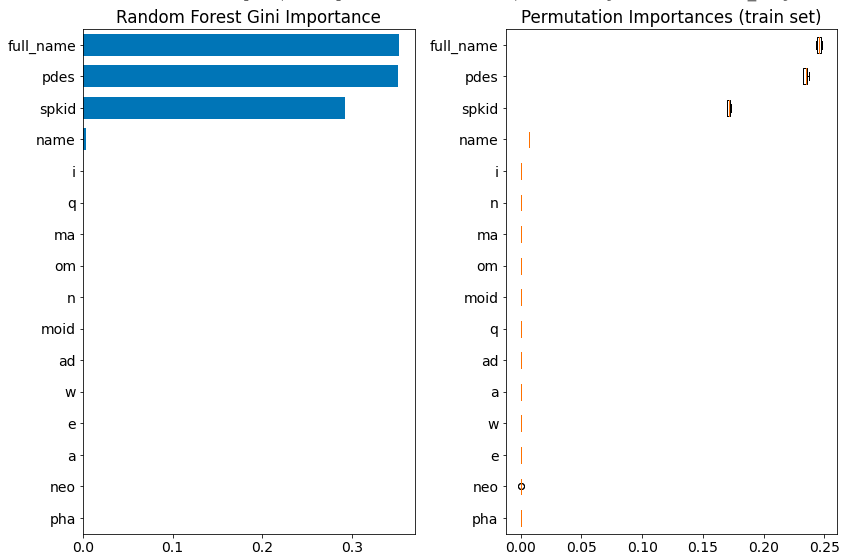


Fig. 10. Feature Importance

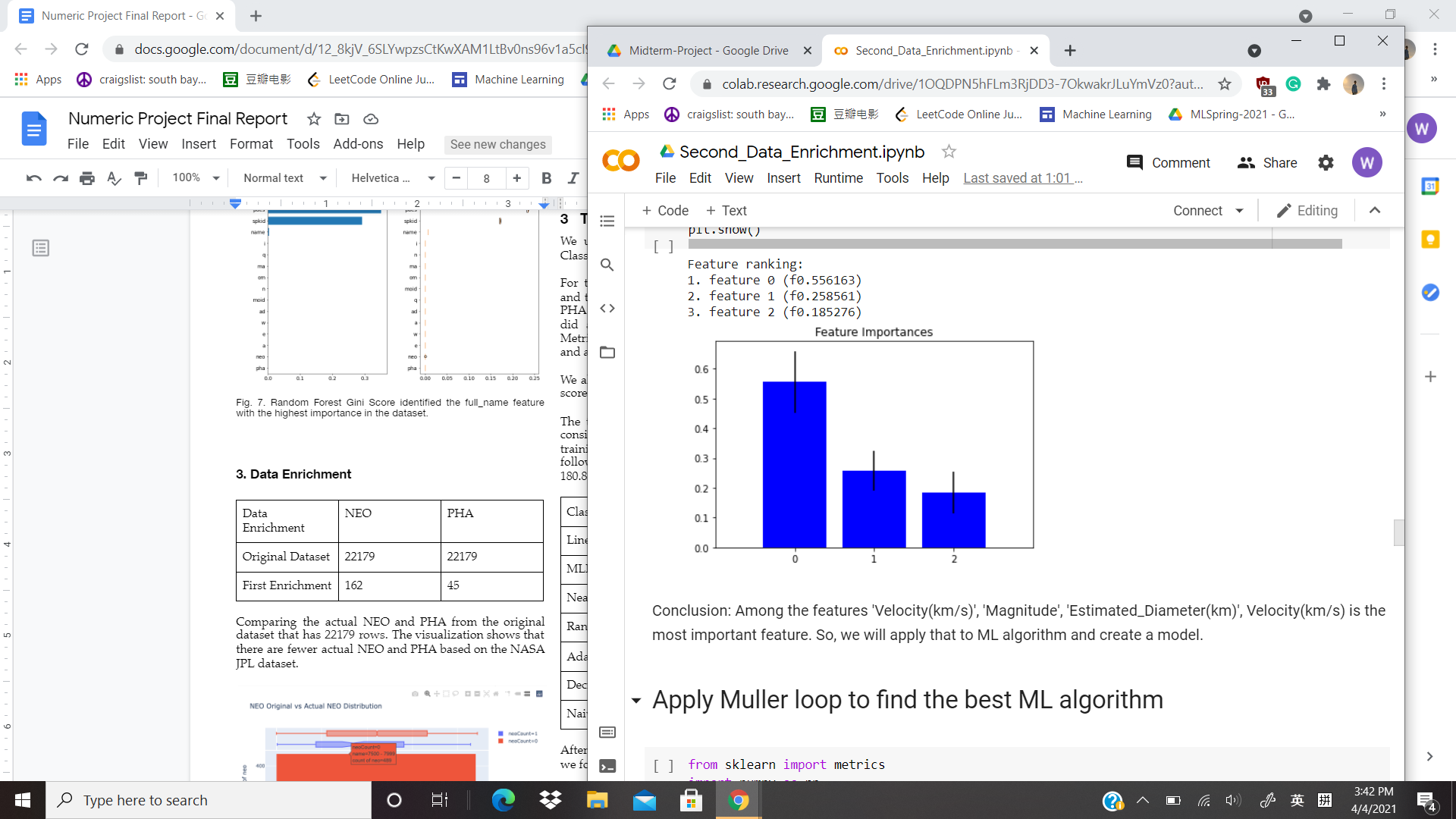


Fig. 11. Feature Importance on CNEOS Sentry

Among the feature’s velocity, diameter, and magnitude, the feature ‘Estimated\_Diameter’ has the highest score. It is the most important feature relative to the impact level, the potential impacts.

## 3. Data Enrichment

**3.1 First Data Enrichment**

|  |  |  |
| --- | --- | --- |
| Data Enrichment | NEO | PHA |
| Original Dataset | 22179 | 22179 |
| First Enrichment | 162 | 45 |

Comparing the actual NEO and PHA from the original dataset that has 22179 rows. The visualization shows that there are fewer actual NEO and PHA based on the NASA JPL dataset.

Latent variables created for data enrichment are neoCount, phaCount.

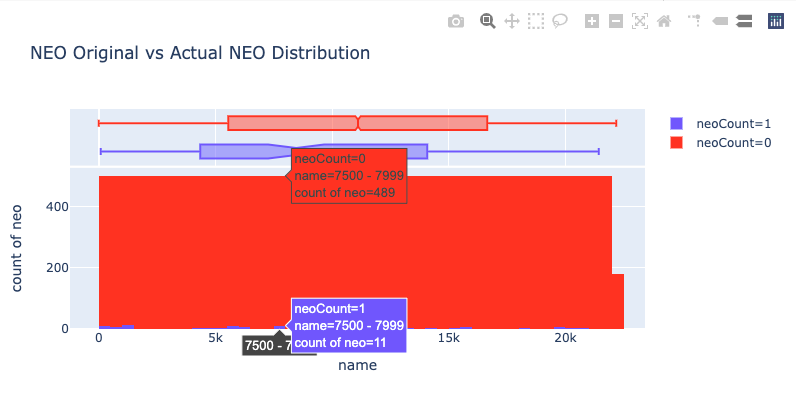


Fig. 12. Actual NEO vs NEO Original dataset

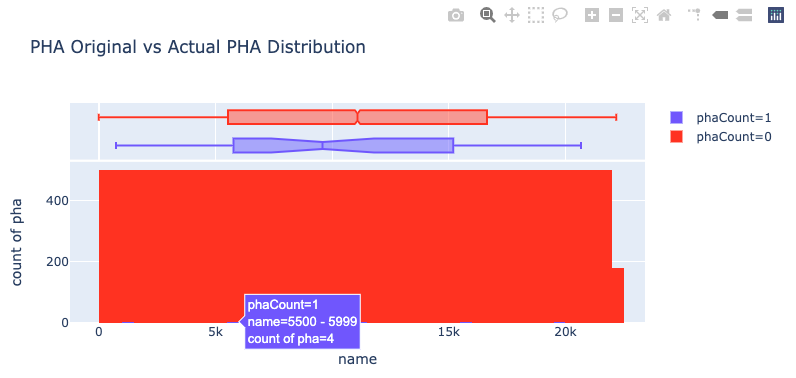


Fig. 13. Actual PHA vs PHA Original dataset

**3.2 Second Data Enrichment**

To find the potential impact, we first needed to analyze the CNEOS Sentry System Dataset. We applied the muller loop to find the best ML algorithm using the original and the second latent variable.

Table

Description automatically generated

Fig. 14. Muller Loop Regression Original Data

Table

Description automatically generated

Fig. 15. Muller Loop Regression with First Latent variable

We find that the accuracy improves after applying the second latent variables.

* Original Accuracy - 60.70%
* First Latent Variable - 66.14%

**Third Data Enrichment**

To find the potential impact, we added the third latent variables (close Approach) to the second latent variables (orbital parameters). We applied the muller loop Regression and find the best ML algorithm with the accuracy of 94%.

Table

Description automatically generated

Fig. 16. Muller Loop Regression with Second Latent variable

A picture containing text, receipt

Description automatically generated

Fig. 17. Regression Analysis after adding Orbital Parameters

# 4 Methodology

We used the Muller Loop Ensemble method both for Classification and Regression model.

For the Classification, we separated the training dataset and test dataset with a test size of 33%. We used NEO and PHA as our features to be analyzed in classification. We did a label encoding process for these two features. Metric used for classification is precision, recall, f1-score and accuracy.

We also implemented the interactive holoviz graph for f1 score and accuracy for classification.

The training time using the Muller loop classification is considered in the evaluation training process. The highest training time is Linear SVM model with 507.23 secs followed by MLP (MultiLayer Perceptron) Neural Net 180.81 secs.

|  |  |
| --- | --- |
| Classification Model | Training Time |
| Linear SVM | 507.23 seconds |
| MLP Neural Net | 180.81 seconds |
| Nearest Neighbors | 34.86 seconds |
| Random Forest | 4.79 seconds |
| Ada Boost | 1.52 seconds |
| Decision Tree | 1.36 seconds |
| Naive Bayes | 0.24 seconds |

Fig. 18. Muller Loop Classification Training Time

After finishing running the training and test set samples, we found the top four classifier models.

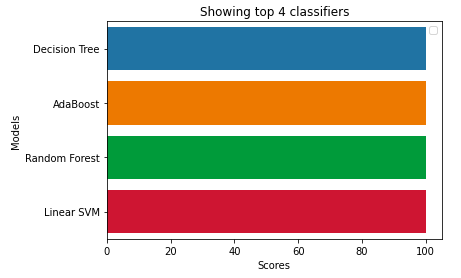


Fig. 19. Using Muller Loop: Identified Top 4 Classifier

The model trained was saved also in pickle file named classification\_model.sav

Secondly, we did the same process for Regression, we split the training dataset and use test size 30%. In regression we want to predict the potential impact of collision of NEA and PHA and will cause danger to the earth.

We did a standard scaling on the dataset for normalization of data range. Metrics used for regression are mae, mse and Rsquared. After the training is done, we found that ADABoost has the highest accuracy around 68.54% followed by Random Forest 65.29%. Below is the boxplot of metrics score with MAE as the best metric.

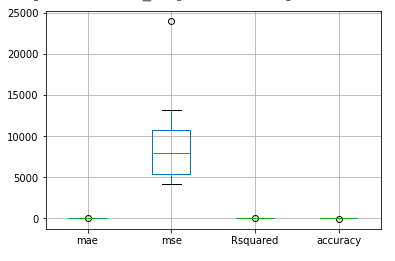


Fig.20. Boxplot MAE has the best metric score in Regression

Then we finally saved the trained model using the pickle file with a name mullerloop\_regression\_model.sav.

# 5 Models

We experimented with different ensemble models using the following two methods:

1. Muller Loop Classification
2. Muller Loop Regression

## 4. 5.1 Muller Loop Classification

### Classifying the NEA and PHA

Using muller loop classification and confusion matrix accuracy we predicted the actual NEA as dangerous correctly. The 344479 true value samples were predicted correctly and the test 8210 samples for prediction was correctly predicted.

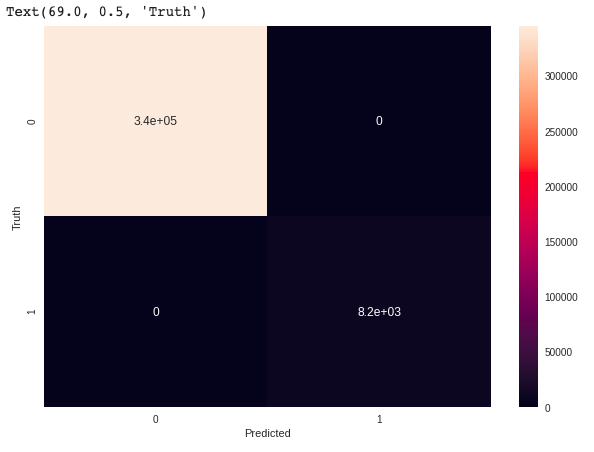


Fig. 21. Muller Loop Classification Confusion Matrix

# 5.2 Muller Loop Regression

Using Random Forest Regression as the best model, prediction shows that the accuracy met is 54.54% with the lowest MAE and MSE and highest R squared score in comparison with other regression models.

The Random Forest regression model does not form into a fit line model. The data was scattered, that's why the error (MAE and MSE) is higher and got only 55% accuracy.

### Predict Potential Impact

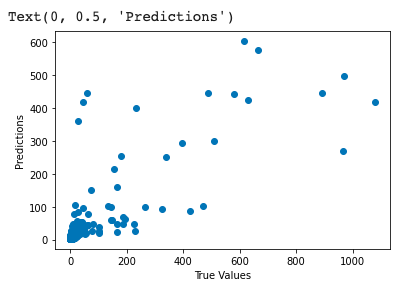


Fig. 22. Prediction Potential Impact with R squared 0.55.

# 6 Comparison Metrics

We track the following metrics for evaluation:

Classification

1. Precision
2. Recall
3. F1 score

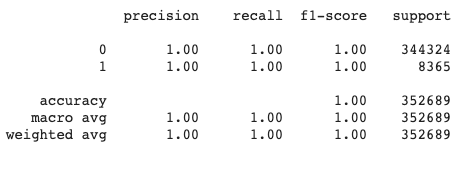


Fig. 23. Precision, Recall, F1 Score

Regression After First Latent Variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **R Squared** | **Accuracy** |
| Nearest Neighbor | 28.53 | 7742.62 | 0.28 | 27.52 |
| Linear SVR | 29.20 | 10609.91 | 0.01 | 0.67 |
| Gaussian Process | 34.30 | 9149.52 | 0.14 | 14.35 |
| Decision Tree | 22.78 | 4664.91 | 0.56 | 56.33 |
| Random Forest | 20.57 | 3616.52 | 0.66 | 66.14 |
| MLP | 35.72 | 5727.09 | 0.46 | 46.39 |
| AdaBoost | 47.40 | 4896.66 | 0.54 | 54.16 |

### 

Fig. 24. Table Regression Metric

# 7 Conclusions

From the first data enrichment, we concluded based on the visualization that there are 162 NEO (Near Earth Object) and 45 PHA (Potentially Hazardous Asteroids) and from the data transformations itself of neo to neoCount and pha to phaCount. We were able to identify the names and the actual NEO and PHA and that there are fewer of them in the NASA JPL dataset. The risk of danger is smaller based on the quantity compared to 22179 data.

After second enrichment, the asteroids named "2020 HO5", "2021 EZ4" have the highest potential impacts to the Earth, they both reach over 50 in the potential impacts level.

After the third data enrichment the accuracy of prediction improves from 60% to 66% and to 94% but using different regression (Random Forest – First latent variable, MLP – Second Latent Variable).

We also did the other modeling using Muller Loop Regression for identifying the best model. We found that Random Forest is the best model regression for our dataset.

Comparing the Classification and Regression method, the classification has a higher accuracy which is 100% in identifying the NEA and PHA compared with the regression method which has only 60% accuracy.

The probability of an asteroid hitting the earth is very low. The impact of collision of asteroids on earth is converted to kinetic energy. The kinetic energy released when the asteroids which might approach earth and collide in the next 60days is 1.62e+13 Joules and 5.38e+14 Joules.

The energy released is almost equivalent to 1/5 of yield of the Little Boy atomic bomb dropped on Hiroshima in World War II.

After all the study we did using the data enrichment and Muller Regression. Our team able to predict these two asteroids name ("2020 HO5", "2021 EZ4") with the highest potential impact of having collisions.

7 End Sections

## Appendices

Appendix A

**Final Colab Link**

[**https://colab.research.google.com/drive/1bD3mmmR-Tw9vMRdmSpvnzwcb3lnzu-IC?usp=sharing**](https://colab.research.google.com/drive/1bD3mmmR-Tw9vMRdmSpvnzwcb3lnzu-IC?usp=sharing)

**Github Link**

[**https://github.com/jocelynbaduria/cmpe-257\_Numeric\_Project**](https://github.com/jocelynbaduria/cmpe-257_Numeric_Project)

**Document Final Report**

[**https://drive.google.com/file/d/1QUtcbq75nYwTmu7vVsLkYQ6AKeHK2OX\_/view?usp=sharing**](https://drive.google.com/file/d/1QUtcbq75nYwTmu7vVsLkYQ6AKeHK2OX_/view?usp=sharing)

**Data Links:**

NASA JPL Dataset

<https://drive.google.com/file/d/1fIXV2Degdd4yYLPRLtPdpDnasptCJ438/view?usp=sharing>

Neo JPL dataset

<https://drive.google.com/file/d/1PE0YpkaenGjVb613kyfWeDdzNgTNrXX1/view?usp=sharing>

Latent variables/data

<https://drive.google.com/file/d/1SqUPabns-VO7tmP2fCyJRAqdjfsSqsps/view?usp=sharing>

<https://drive.google.com/file/d/1B-GwsTw6SG0H9DiZ-SiZdO5oyj7mSQAE/view?usp=sharing>

Cleaned NASA JPL Dataset

<https://drive.google.com/file/d/1d2tUMvxC5y6BRWZNl5cXP-5pcaJq1iBR/view?usp=sharing>

Cneos Sentry System Dataset

<https://drive.google.com/file/d/1OZ_ZjSXrl-8xfq_yFV4UqT83XObLdvo4/view?usp=sharing>

Cleaned Cenos Sentry System Dataset

<https://drive.google.com/file/d/1-4QUX0S_7iwLU3j15jtBFJG0xNFPlJLb/view?usp=sharing>

Cneos Close approach Data

<https://drive.google.com/file/d/1Q2ut-V9LagYTzR0R34sdltsT7hOvvUi8/view?usp=sharing>

Cleaned Cenos Close approach Data

<https://drive.google.com/file/d/1UL4dcYg-mwVIddM5bpURdW3cNtGLZkeY/view?usp=sharing>

Muller Loop regression analysis with model

<https://drive.google.com/file/d/1M2ACFz2esasvWyfYYTbEgRUn4DIj82yQ/view?usp=sharing>

NASA JPL first data enrichment dataset

<https://drive.google.com/file/d/1aqdT37rH-gkaio-0f821OzGOtwXAEBt0/view?usp=sharing>

Test Nea/the near-earth object

<https://drive.google.com/file/d/1D0W5ZsdTUguCtZu4sS7upzhqu7l9AV9q/view?usp=sharing>

Temporary Transformed data

<https://drive.google.com/file/d/1mOBr6xA37ssoajn4wCPjcoaTgnrd-JG1/view?usp=sharing>

## Acknowledgments

We would like to acknowledge our Professor Dr. Ali Arsanjani for guidance throughout this Spring2021 semester.

## References

1. NASA JPL

<https://ssd.jpl.nasa.gov/sbdb_query.cgi#x>

1. CNEOS Sentry

<https://cneos.jpl.nasa.gov/ca/>

1. Contributors

National Research Council,” Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies”.

<https://www.nap.edu/catalog/12842/defending-planet-earth-near-earth-object-surveys-and-hazard-mitigation>

1. Gizachew Tiruneh, Ph. D Dept of Political Science University of Central Arkansas. “Explaining Planetary-Rotation Periods Using an Inductive Method”, <https://arxiv.org/abs/0906.3531v3>, June 15, 2011.
2. API Link

<https://ssd-api.jpl.nasa.gov/sbdb.api>