NASA Asteroids Classification

Jungwon Yoon
ojyoon@ucdavis.edu

Hyunjin Chang
hjchang@ucdavis.edu

Sung Won Lee
sgwlee@udavis.edu

Jong Wook Choe
wchoe@ucdavis.edu

Abstract

A potentially hazardous asteroid is with an orbit that can make close approaches to the Earth and is large enough to cause significant regional damage in the event of impact. By Identifying the potential hazardous asteroids, we can assess potential prevent the collision between the Earth and the asteroid. To achieve our goal to classify whether the asteroid is potentially hazardous or non-hazardous, we trained NASA Asteroids data with the Naive Bayes Classifier, Support Vector Machine, and Decision Tree. As a result, Decision tree modeling predicted hazardous asteroids with the best performance by achieving 99% accuracy.

1 Introduction

Throughout our analysis, we will be using a large data of Asteroids - NeoWs. We will discuss the factors that might be useful to identify an asteroid as potentially hazardous or not. Also we will explain the classification power of those remaining factors in the absence of absolute magnitude and minimum orbit intersection values.

Near Earth Objects (NEOs) are comets and asteroids that have been nudged by the gravitational attraction of nearby planets into orbits that allow them to enter the Earth's neighborhood. The scientific interest in comets and asteroids is due largely to their status as the relatively unchanged remnant debris from the solar system formation process some 4.6 billion years ago. As the primitive, leftover building blocks of the solar system formation process, comets and asteroids offer clues to the chemical mixture from which the planets formed some 4.6 billion years ago.

2 ASTEROID DATASET

2.1 Data Exploration

The data is about Asteroids - NeoWs. NeoWs (Near Earth Object Web Service) is a RESTful web service for near earth Asteroid information. With NeoWs a user can: search for Asteroids based on their closest approach date to Earth, lookup a specific Asteroid with its NASA JPL small body id, as well as browse the overall data-set.

Acknowledgements Data-set: All the data is from the (http://neo.jpl.nasa.gov/). This API is maintained by SpaceRocks Team: David Greenfield, Arezu Sarvestani, Jason English and Peter Baunach.

There are 4687 rows and 40 columns(features) in the dataset. The data does not contain null values.

2.2 Data Visualization

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Categorical Variables

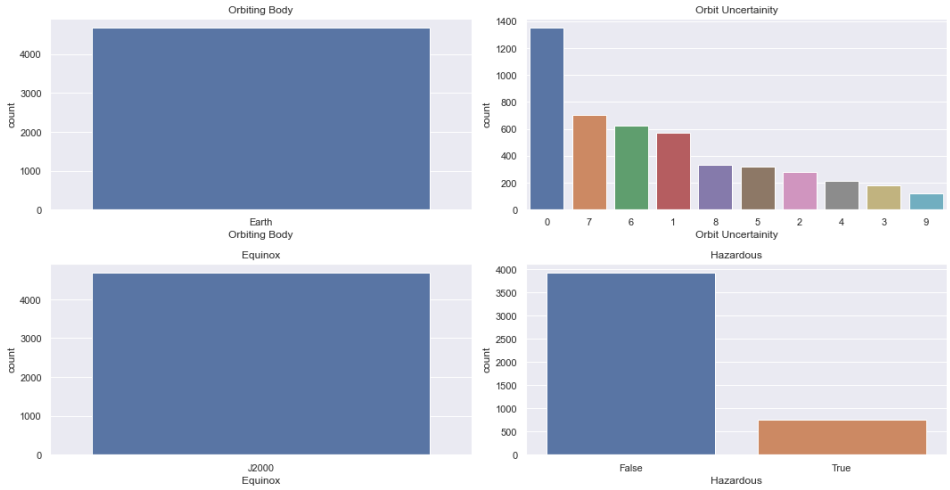

Figure 1: Categorical Variables

Orbiting Body - the orbiting body type of every asteroid recorded in the system is the Earth.

Orbit Uncertainty - About 28.9% recorded asteroid has uncertainity type 0, 14.9% has type 7, following 13.2% with 6.

Equinox - Every asteroid has the same type of Equinox, which is J200.

Hazardous - 83.9% of recorded asteroid found to be not Hazardous.

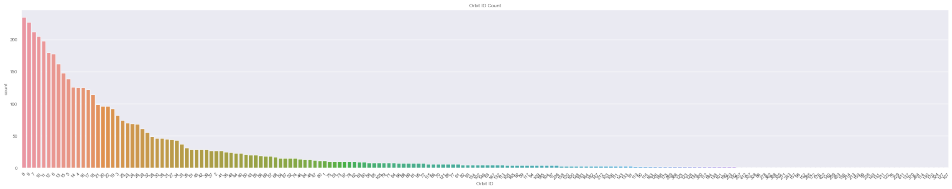

Figure 2: Orbit ID Graph

There are 188 unique Orbit IDs, and counts for each ID are following in the graph. And the most 10 common IDs are 8, 9, 7, 10, 11, 12, 6, 13, 15, 5

Distribution of Continuous Variables

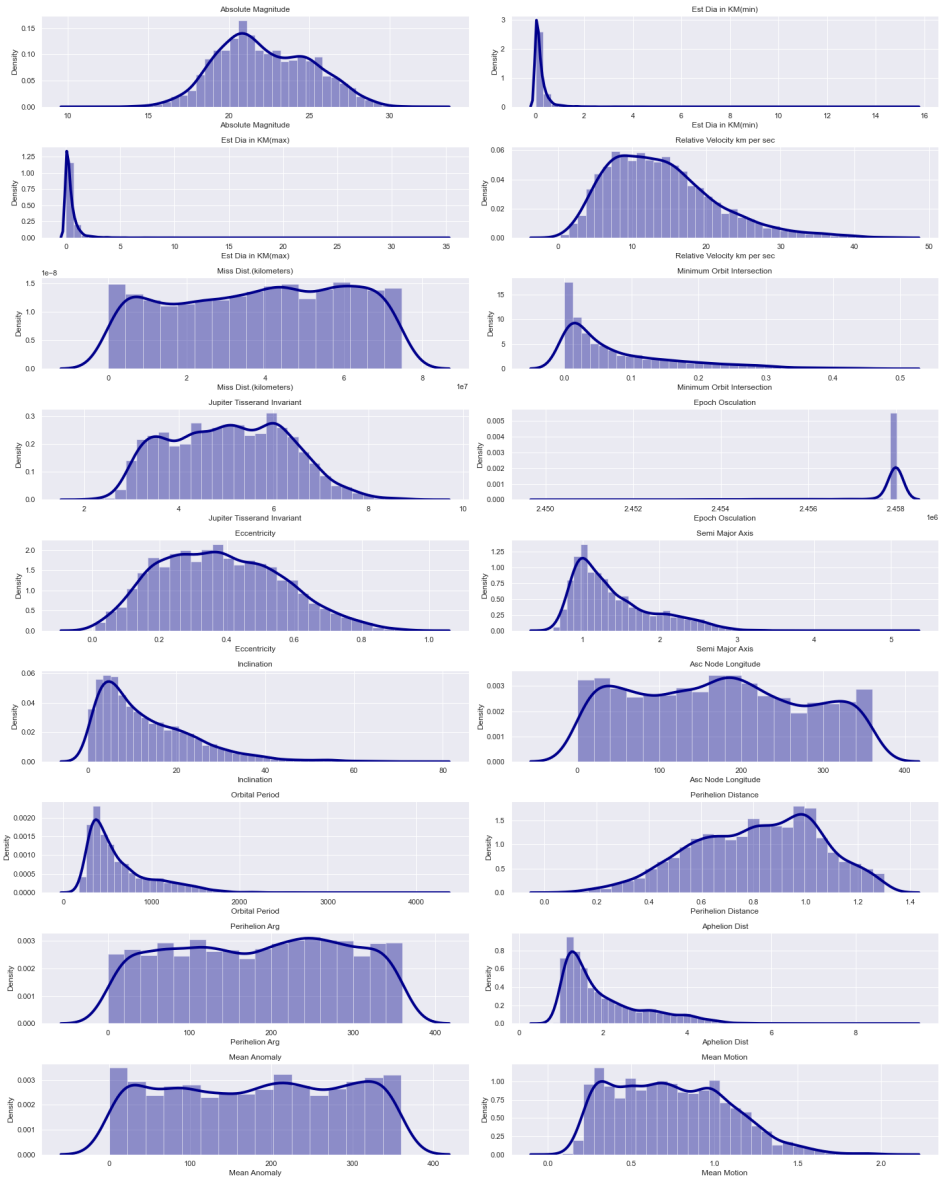

Figure 3: Distribution of Continuous Variables Graphs

Some features are right or left skewed like Est Dia in KM.

There are features that seem to be uniformly distributed such as Miss Dist Perihellon Arg.

Also, there are normally distributed features like Absolute Magnitude.

Time Series - Close Approach Date

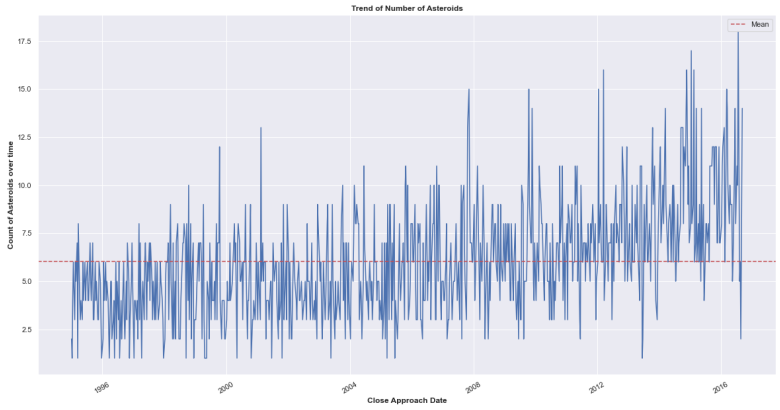

Figure 4: Time Series Trend of Asteroids Counts over Years

As year goes by, there seems to be more asteroids coming close to the Earth. In 2016, there has been about 17 asteroids approached close to the earth, with the mean of 6 asteroids over 20 years. Code reference (https://www.kaggle.com/code/shrutimehta/data-preprocessing-and-correlation)

Boxplot of Continuous Variable by Hazardousness

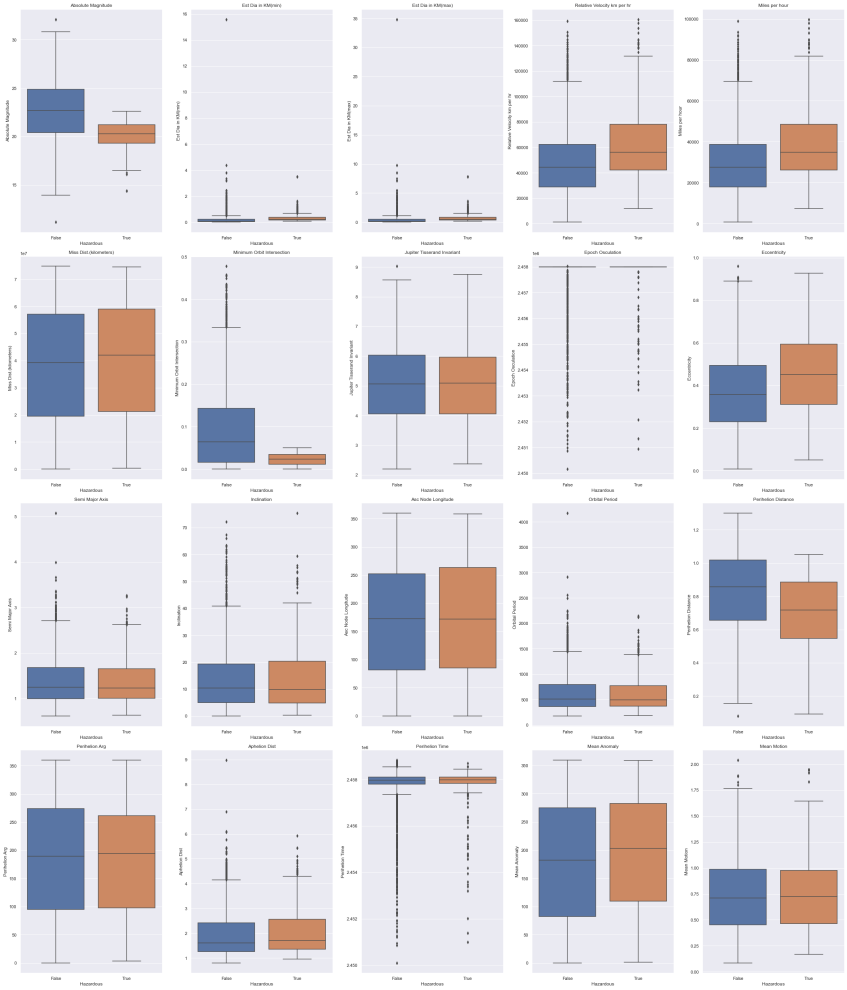

Figure 5: Boxplot for Harzardous Classification

Distribution of Continuous Variable by Hazardousness

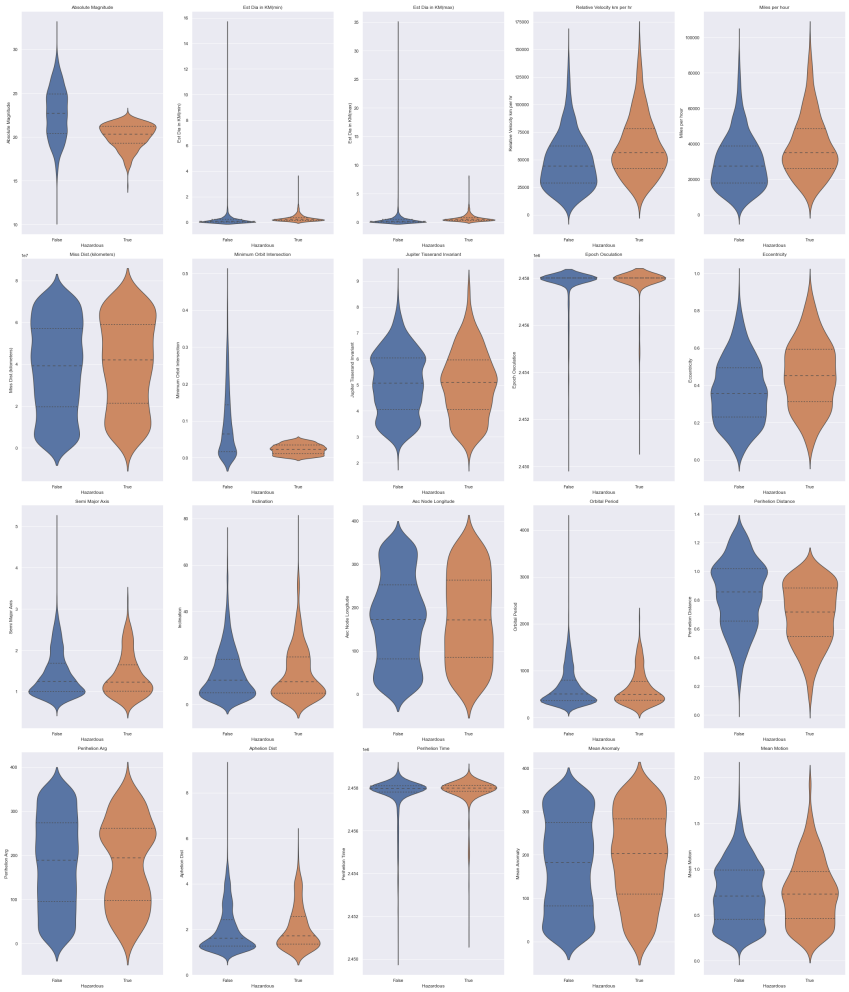

Figure 6: Violion Plot for Harzardous Classification

Based on the boxplot and the violin plot, a lot of features seem to have similar distribution whether it's hazardous or not. Some notable features are Absolute Magnitude, absolute magnitude of hazardous aestroids are very densely normally distributed and non hazardous one seems to be normally distributed but in a wider range. Also their means are pretty different, mean of absolute magnitude of non hazardous aestroid is about 23 whereas hazardous aestroids seem to have less magnitude, which is about 20.

We also see the significant distribution difference in minimum orbit intersection. And mean of minimum orbit intersection of hazardous aestroids is about 0.02 and non hazardous aestroids have its mean about 0.06 and it ranges widely, upto about 0.5.

Hence, Absolute Magnitude and Minimum Orbit Intersection might play an important role in classification.

Correlation between variables

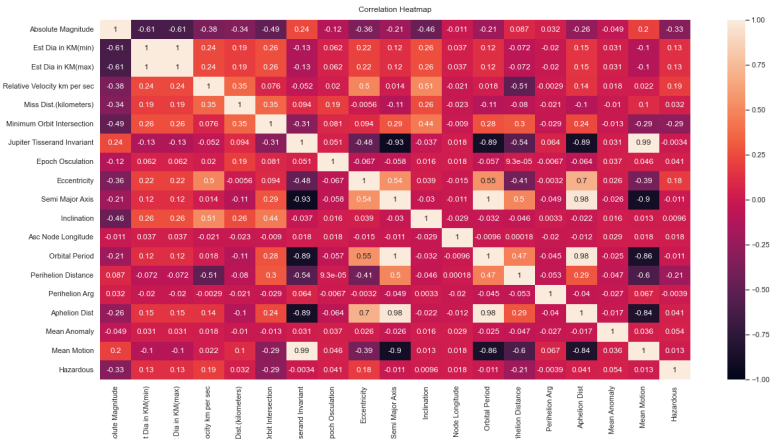

Figure 7: Correlation Heatmap

Negative correlations between two variable mean as one variable increases, the other variable decreases, and positive correlations mean as one variable increases, the other one increases as well. Based on our heatmap, there are some high correlation between variables. Some notable positively correlated variables are Mean Motion & Jupiter Tisserand Invariant with 99% correlated rate, Perihelion Time & Epoch Osculation, Aphelion Dist & Semi Major Axis, Aphelion Dist & Orbital Period with 98% rate. Some negatively correlated variables are Semi Major Axis & Jupiter Tisserand Invariant with the rate of 93%, Semi Major Axis & Mean Motion with the rate of 90%.

Addition to that based on low correlation some of the variables can be removed without hesitation.

Such as,

‘Est Dia in KM(min)’, ‘Est Dia in KM(max)’,

‘Est Dia in M(min)’, ‘Est Dia in M(max)’,

‘Est Dia in Miles(min)’, ‘Est Dia in Miles(max)’,

‘Est Dia in Feet(min)’, ‘Est Dia in Feet(max)’

Also,

‘Est Dia in M(min)’, ‘Est Dia in M(max)’,

‘Est Dia in Miles(min)’, ‘Est Dia in Miles(max)’,

‘Est Dia in Feet(min)’, ‘Est Dia in Feet(max)’.

As mentioned above the correlation matrix quickly shows redundancy, and we can delete the repeated information.

A similar explanation can be given for the features,

‘Relative Velocity km per sec’, ‘Relative Velocity km per hr’, ‘Miles per hour’,

and

‘Miss Dist.(Astronomical)’, ‘Miss Dist.(lunar)’, ‘Miss Dist.(kilometers)’, ‘Miss Dist.(miles)’

Out of the features mentioned above, we are going to keep ‘Relative Velocity km per sec’ and ‘Miss Dist.(Astronomical)’.

2.3 Data Description

After dropping irrelevant columns, we get 20 numeric features. The factors that might be useful to identify an asteroid as potentially hazardous are

1. *Name*: This feature denotes the name given to an asteroid.
2. *Absolute Magnitude*: The visual magnitude an observer would record if the asteroid were placed
3. *Est Dia in KM(min), Est Dia in KM(max)*: Min and Max of Estimated Diameters in km
4. *Relative Velocity km per sec*: This feature denotes the relative velocity of the asteroid in kilometre per second.
5. *Miss Distance in km*: Missed distance from an asteroid to the Earth.
6. *Orbit Uncertainity*: The orbital uncertainty of asteroids.
7. *Minimum Orbit Intersection*: A measure used in astronomy to assess potential close approaches and collision risks between astronomical objects.
8. *Jupiter Tisserand Invariant*: This feature denotes the Tisserand's parameter for the asteroid.
9. *Epoch Osculation*: The epoch of osculation of the asteroid.
10. *Eccentricity*: This feature denotes the value of eccentricity of the asteroid's orbit.
11. *Semi Major Axis*: This feature denotes the value of the Semi Major Axis of the asteroid's orbit.
12. *Asc Node Longitude*: Longitude of the ascending node
13. *Orbital Period*: This feature denotes the value of the orbital period of the asteroid.
14. *Perihelion Distance*: This feature denotes the value of the Perihelion distance of the asteroid
15. *Perihelion Arg*: Argument of Perihelion
16. *Aphelion Dist*: This feature denotes the value of Aphelion distance of the asteroid.
17. *Perihelion Time*: Perihelion is the point where a planet is closest to the sun in its orbit. Earth reaches its annual perihelion around two weeks after the winter solstice.
18. *Mean Anomaly*: The fraction of an elliptical orbit's period that has elapsed since the orbiting body passed periapsis.
19. *Mean Motion*: The angular speed required for a body to complete one orbit.
20. *Inclination*: The tilt of an object's orbit around a celestial body.

3 Methodology

3.1 Task

- Develop a model that predicts if an asteroid is going to be hazardous or not

To tackle this task we'll use the methods listed below, and extract the results from the method with best performance on unseen data.

- Naive Bayes Classifier
- SVM
- Decision Tree

3.2 Approach

Clustering of unlabeled data can be performed with the module sklearn.cluster.

Each clustering algorithm comes in two variants a class, that implements the fit method to learn the clusters on train data, and a function, that, given train data, returns an array of integer labels corresponding to the different clusters.

Among clustering algorithms we will discuss method that build nested clusters by merging or splitting them successively, Hierarchical clustering. Strategies for hierarchical clustering generally fall into two types:

Agglomerative: This is a "bottom-up" approach each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top-down" approach all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. It is a branching diagram that represents the relationships of similarity among a group of entities.

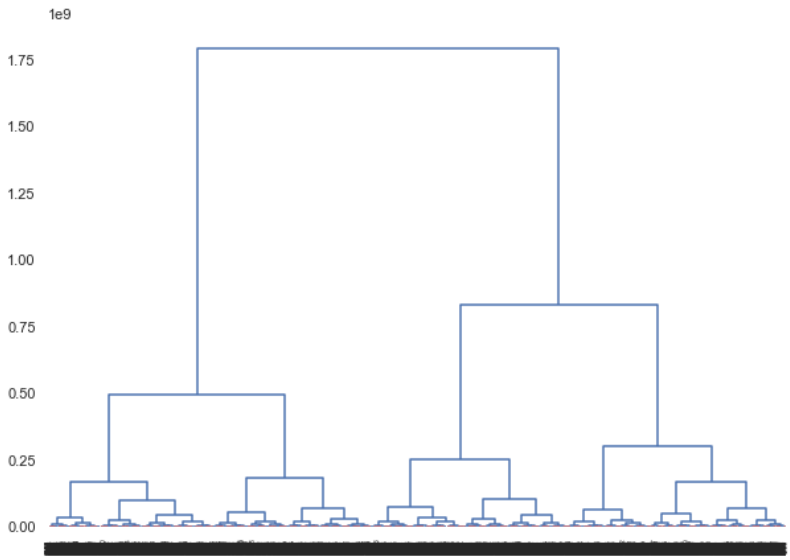

Figure 8: Dendrogram 1

Above dendrogram shows

1. Metrics calculation of the similarity between all objects
2. Clustering observations with adjacent distances
3. Updating affinity Matrix.

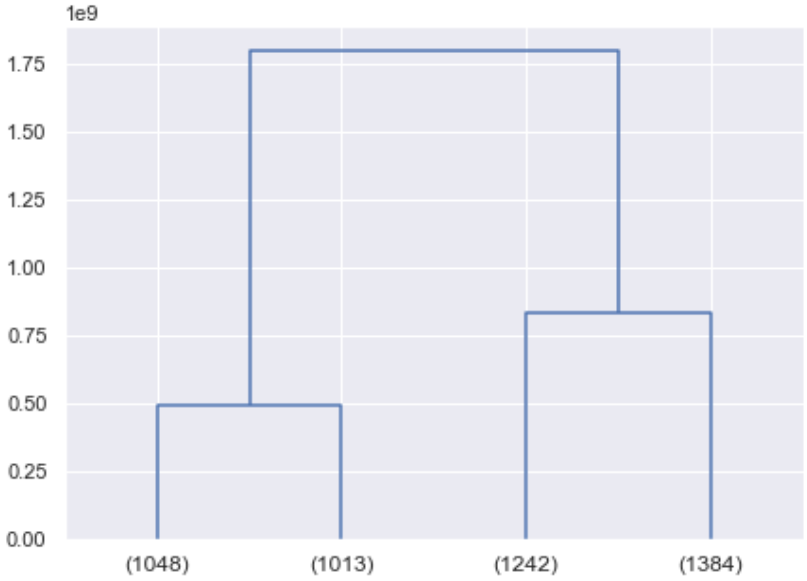

Figure 9: Dendrogram 2

Truncated and count each cluster we see the data is clustered into two group. Moreover, when using a dendrogram to display the result of a cluster analysis, it is always good to add the corresponding heatmap. It allows to visualise the structure of entities, and to understand if this structure is logical.

We build a dendrogram and heatmap by using the clustermap() function of seaborn library. Also we determined the distance calculation method, therefore we set the linkage method to use for calculating clusters with the method parameter.

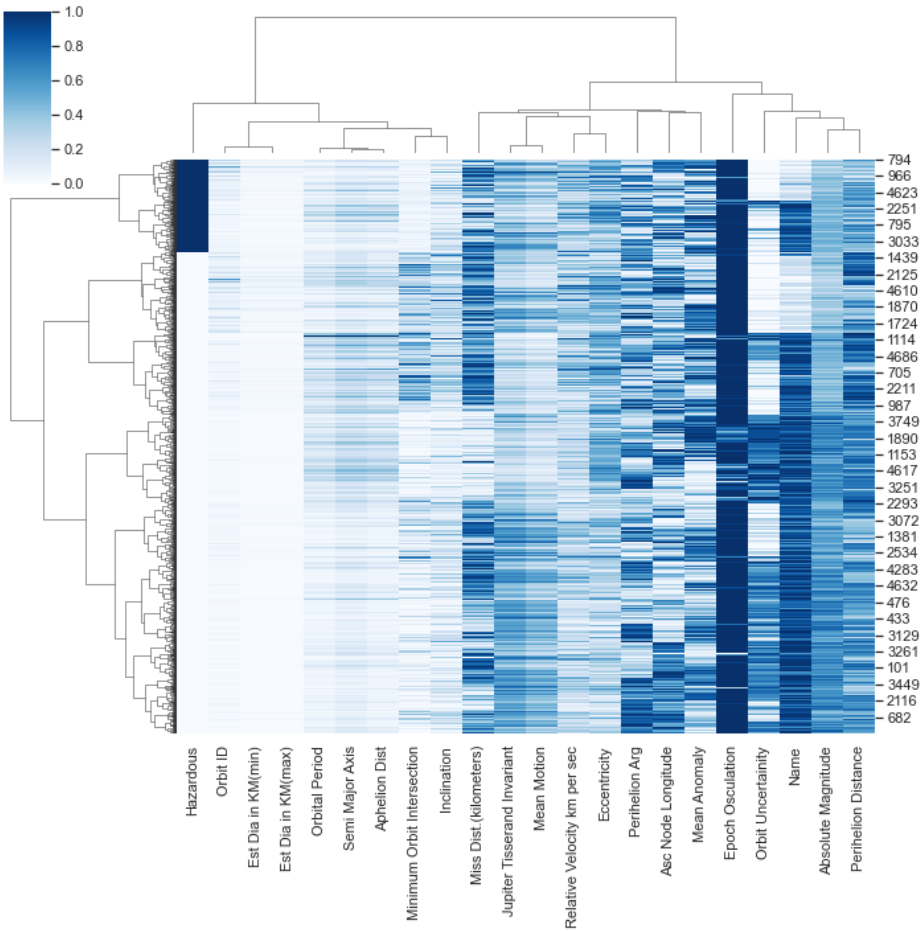

Figure 10: Dendrogram with heatmap

The dendrograms along the sides show how the variables and the rows are independently clustered. The heat map tells the data value for each row and column. Any patterns in the heat map indicate an association between the rows and the columns. Moreover, a rectangular area of about the same color suggests a group of rows that is correlated for the corresponding group of columns.

For instance, left corner ‘hazardous’ feature are sharing same tree with ‘Est Dia in M(min)’, ‘Est Dia in M(max)’, ‘Est Dia in Miles(min)’, ‘Est Dia in Miles(max)’, ’Orbital Period’, and ’Aphelion Dist’. On the other hand, ’Name’, ’Perihelion Arg’, ’Absolute Magnitude’, and etc sharing another. This tells that not every variables affect whether asteroid is going to be hazardous or not.

4 Analysis

Let us divide the data into an 80:20 ratio as training and test set respectively.

Generated train set contains 3749 data and has 610 cases labelled as 1, which means that if a model predicts all values as 0, then the accuracy will be 83.72. This will be considered as baseline accuracy for the train set.

Similarly, the baseline accuracy for the test set will be 84.54.

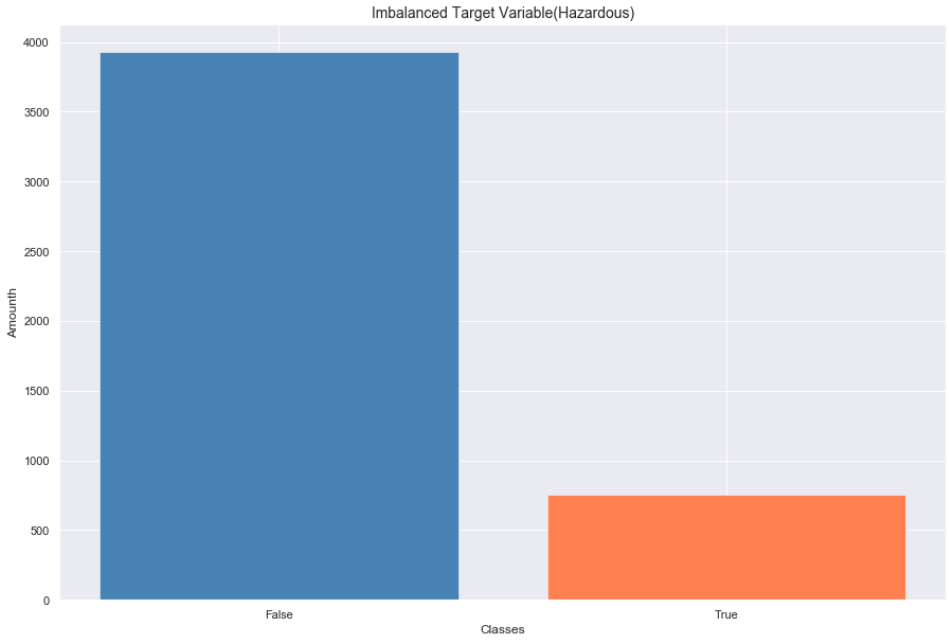

Figure 11: Without SMOTE algorithm

However this is a imbalanced data set. 83.89 are the case of not hazardous and only 16.10 are the case of hazardous.

This means even if we have a broken model that predicts all values as not hazardous, then the accuracy will be 83.89. So, we can not rely on accuracy to evaluate a machine learning classifier trained on this data set.

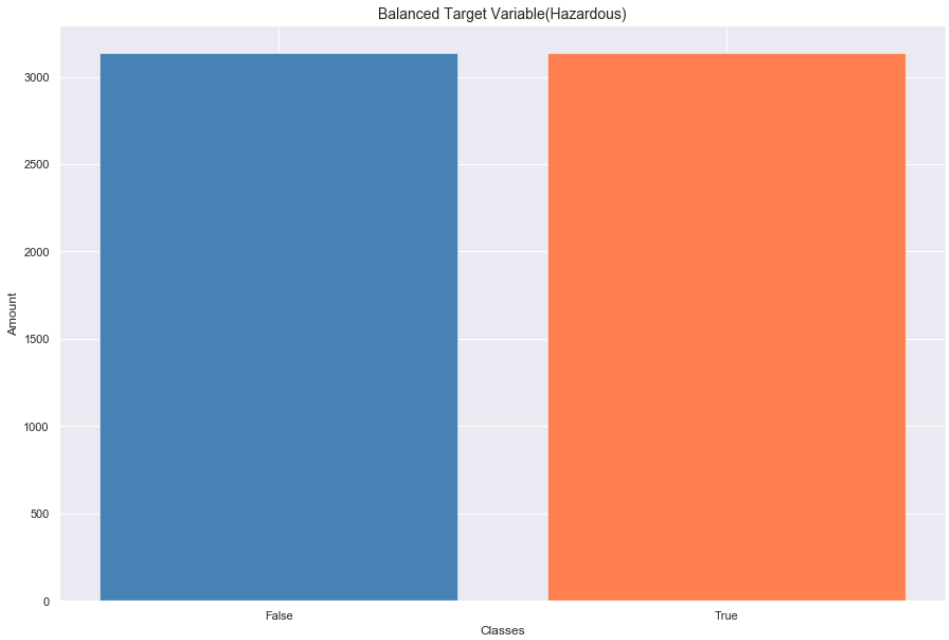

Figure 12: Using SMOTE algorithm

To handle the imbalanced distribution of the target value, SMOTE(Synthetic Minority Over-sampling Technique) had been applied for the oversampling method while training the model. The figure above shows the changes after applying SMOTE.

There are mainly three methods to resolve the unbalance issue: undersampling, oversampling, and the combination of those two methods. In this dataset, we chose to proceed with oversampling method since undersampling eliminates the data, it may lead to loss of information of the data from the majority class which is deleted. Therefore, among various oversampling method, we used SMOTE algorithm, which generates the data for the minority class by using *k-nn* algorithm.

- Naive Bayes Classifier
  The accuracy of Naive Bayes :
  - Train set : 0.5826
  - Test set : 0.5767

- SVM
  The accuracy of Support Vector Machine :
  - Train set : 0.5068
  - Test set : 0.4936

- Decision tree
  The accuracy of Decision Tree :
  - Train set : 1.0
  - Test set : 0.9936

As a result, by comparing three different models, we could observe that decision tree resulted the best performance on predicting whether the asteroids are hazardous or not. From the accuracy of the train set of Decision Tree algorithm, we could observe a overfitting problem. However, it occurs because we have used SMOTE oversampling method to solve the unbalance of the data set. Thus, although we see overfitting, we can say that the test result is reliable.

5 Conclusion

From the Asteroids - NeoWs data set we had examined trends, the relation between variables, and asteroid classification for predicting an asteroid hazard. Starting with various plots we discovered certain variables share similar distribution. Also, showed how data is imbalanced. Before analysis based on the correlation heatmap, we removed some variables that are overlapping or minority features.

After that, we used a dendrogram to see the relation between variables. Assuming all observations start in one and splits into different clusters, they fall into two clusters. Our target feature ‘hazardous’ was related to ‘Est Dia in M(min)’, ‘Est Dia in M(max)’, ‘Est Dia in Miles(min)’, ‘Est Dia in Miles(max)’, ’Orbital Period’, and ’Aphelion Dist’. Others grouped with others.

Lastly, solved the imbalanced distribution of the target value using SMOTE. Next, we used Naive Bayes Classifier, SVM, and Decision tree modeling to predict asteroid hazards. As a result, using Decision tree modeling we could perform the best prediction with 99% accuracy.

6 Reference

“NASA: Asteroids Classification.” Kaggle, 1 Mar. 2018,
https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification.

“Center for Neo Studies.” NASA, NASA, https://cneos.jpl.nasa.gov/.

"Data Preprocessing and Correlation"
https://www.kaggle.com/code/shrutimehta/data-preprocessing-and-correlation