STA 160 Project

Niraj Bangari, Jayoung Kim, Melanie Lue, Cecilia Nguyen, and Marvin Pepito

I. INTRODUCTION / BACKGROUND

The dataset chosen involves the characteristics of asteroids near Earth as recorded by NASA. The data was collected through NeoWs (Near Earth Object Web Service), a web service for near earth Asteroid information. Our goal and questions regarding the dataset is to discover which characteristics mainly determine whether an asteroid was hazardous or not, where hazardous is defined as having the 'potential to make threatening close approaches to the Earth [1].

To have a solid understanding of the dataset, we will look at its characteristics. The asteroid dataset comes from the data portal of NASA, and it has 4687 rows and 40 columns, where each row corresponds to an observation for each category. In order to determine whether an asteroid was hazardous or not, we chose 'Hazardous' as a dependent variable which contains boolean values denoting whether the asteroids are hazardous or not. Due to the large number of independent variables, which can make analyzing the data confusing, we decided to choose specific columns with high correlations to an asteroid's hazardousness. The independent variables in this report are 'Orbit uncertainity,' 'Absolute magnitude,' and 'Miles per hour.' The first column, 'Orbit uncertainity,' denotes an uncertainity of the orbit of asteroids, which is affected by several parameters such as the number of observations, the time spanned by those observations, the quality of the observations, and the geometry of the observation. The second column, 'absolute magnitude,' indicates the absolute magnitude of an asteroid which is the visual magnitude an observer would record if the asteroid is placed 1 Astronomical Unit (AU) away. The third column, 'miles per hour,' provides the speed of asteroids in units of miles per hour. Although we have narrowed the scope of our analysis to these four columns, we note that there are other aspects of hazardousness of asteroids in the dataset that could also be explored.

III. METHODOLOGY

A. Exploring the Data

Exploratory data analysis was performed to clean and understand our data. This will give us a clearer picture of our variables in order to perform machine learning classification. As shown in Fig. 1, we observed that the frequency of hazardous asteroids in our data was 755 out of 4687 (16.1%) of the observations, and 3932 out of 4687 (83.9%) of the observations were not hazardous. The strength of correlations was analyzed between other variables and whether they were related to the asteroid's hazardousness; this was to choose variables which we think have strong correlations for identifying the asteroid's hazardousness and to also ensure that our chosen independent variables were not multicollinear (which can affect our machine learning models), as shown in Fig. 2,3, and 4. To view the distribution of our chosen variables, we implemented a boxplot, where we are able to see the interquartile range from the 25th to 75th percentile. To classify potential asteroids' hazardousness, supervised machine learning classifications to train a model were implemented. This is due to the fact that there are labeled inputs and outputs.

B. Cleaning up the data

We removed redundant variables, such as variables that were identical but measured in different units(MPH vs KPH), and then performed a correlation matrix to determine which variables we should focus on for our machine learning classification. According to Fig. 6, 'Hazardous' has strong negative correlation with 'Orbit Uncertainity' and 'Absolute Magnitude' and a strong positive correlation with 'Miles Per Hour'.

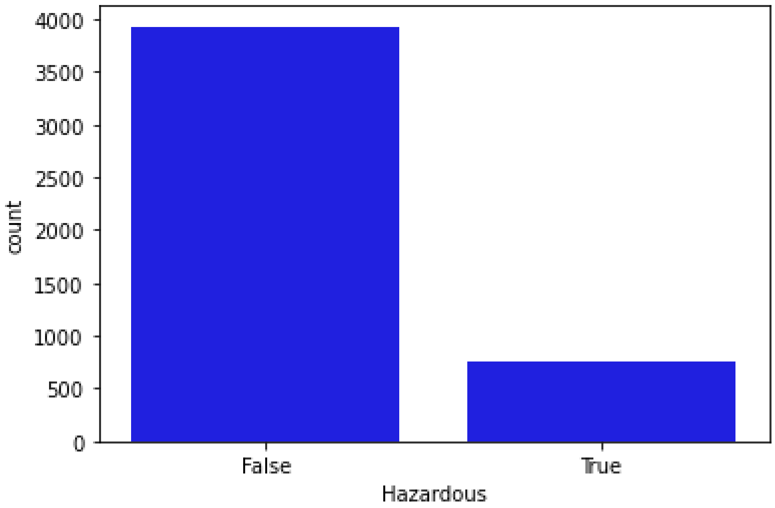

Figure 1. Frequency of hazardous asteroids

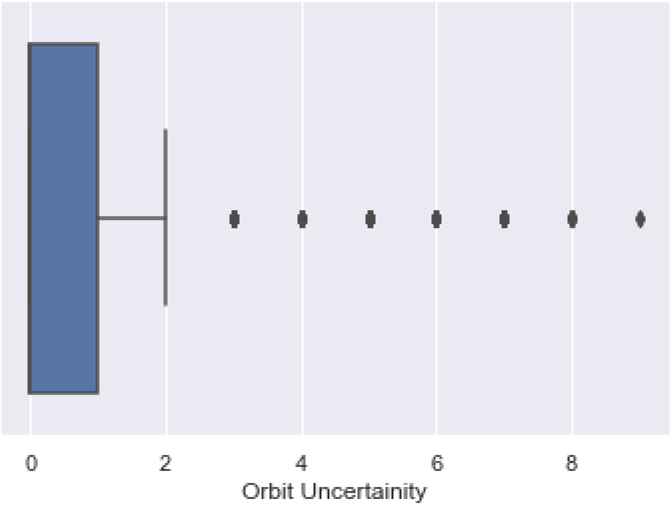

Figure 2. Boxplot of Orbit Uncertainity

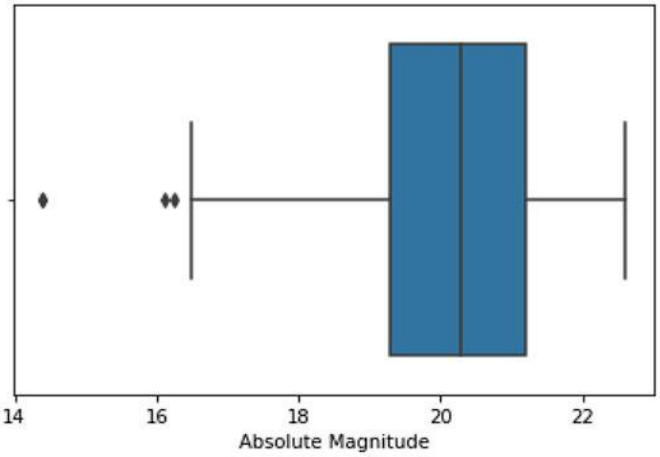

Figure 3. Boxplot of Absolute Magnitude

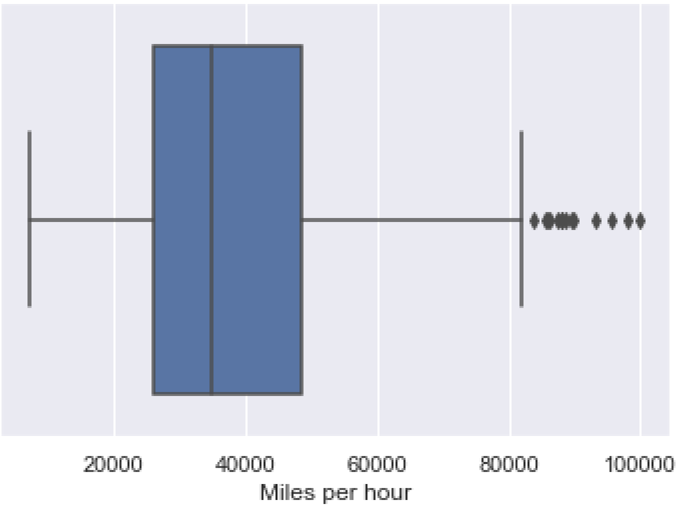

Figure 4. Boxplot of speed of asteroids in units of miles per hour

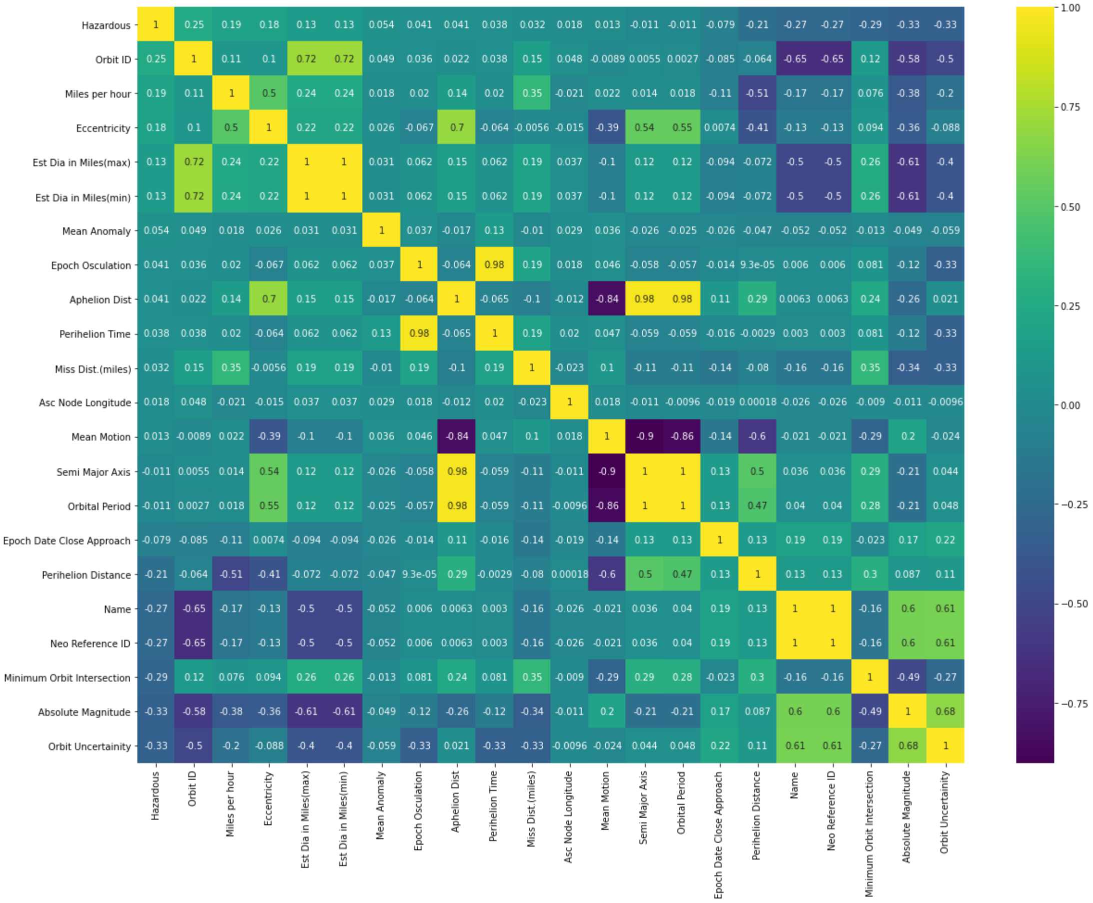

Figure 5. Correlation matrix

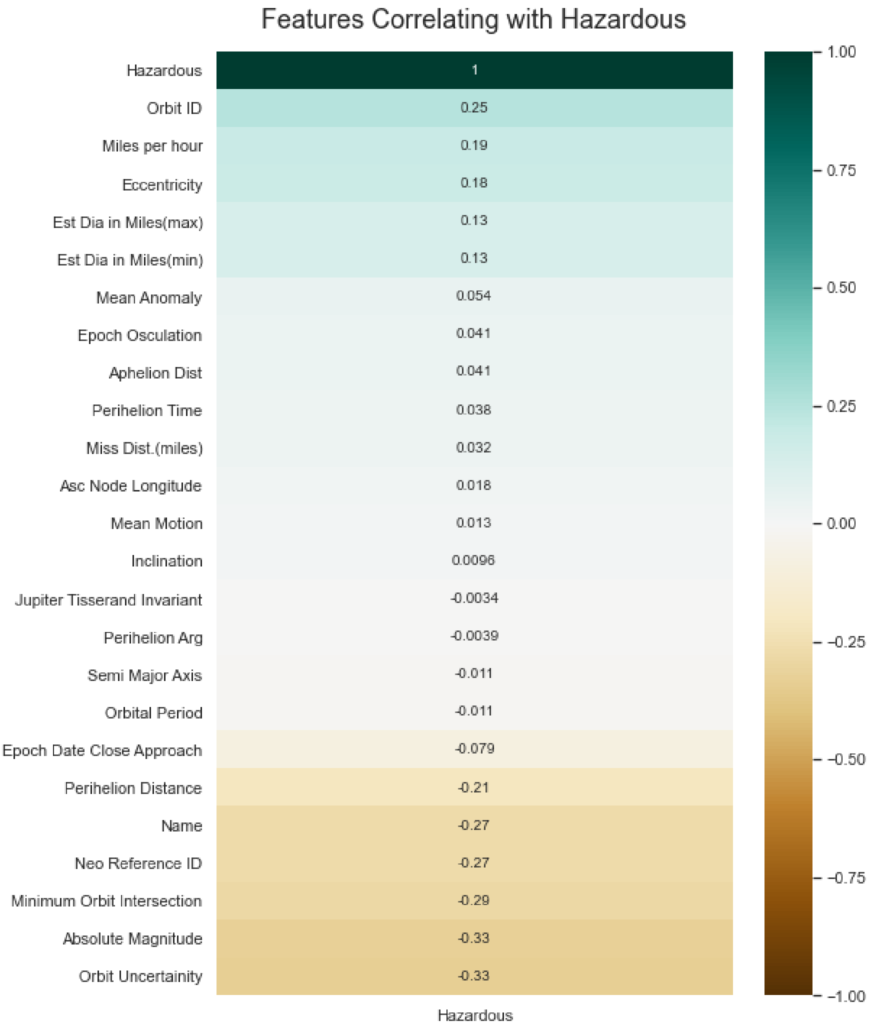

Figure 6. Heat Map of features correlating with Hazardous

IV. INTERPRETATION of EDA

A. Figure 1

Fig. 1 depicts the counts of the possible values for the Hazardous variable, the values being True and False. The majority of the asteroids in our dataset are not hazardous, with only 16% of the data having hazardous asteroids. Percentages do not tell the whole story however, with the 16% still amounting to a high count of 755 hazardous asteroids.

B. Figure 2 / Figure 3 / Figure 4

Fig. 2, 3, and 4 show the distribution of some of the variables (Orbit Uncertainity, Absolute Magnitude, and Miles Per Hour) that showed the strongest correlation with the Hazardous attribute, when the asteroid was indeed Hazardous (Hazardous=True). Fig. 2 shows the large majority of values for Orbit Uncertainity distributed around 0-1, with a few outliers, which likely had the greater impact on the Hazardous attribute. Fig. 3 shows that Absolute Magnitude was relatively normally distributed, with very few outliers, and the average value of this attribute for a hazardous asteroid was around 20. Fig. 4 for Miles Per Hour contains many more outliers, and is slightly positively skewed, with the average mph for a hazardous asteroid being around 30,000.

C. Figure 5

Fig. 5 is a correlation matrix showing the relationship between all variables. Within this matrix, the highest correlation generally occurs between naming variables and variables of distance. Fig. 6 continues the correlation exploration by singling out the attribute under focus.

D. Figure 6

Fig. 6 shows the correlation of all attributes with ‘Hazardous.’ We used this chart to choose one attribute that was highly positively correlated in comparison to the other features, MPH, and two negatively correlated attributes, Absolute Magnitude and Orbit Uncertainity.

V. INTERPRETATION OF CLASSIFIERS

We have implemented classification of our asteroid data using different classification algorithms, some linearand some non-linear. We will now compare the performance of these algorithms to decide which method would be most reliable for classifying new unseen data observations.

For comparing the accuracy of these classifiers when used to make predictions for other asteroid data, we split our data set into 70% for the training subset and 30% for the test subset.

First, we have our Multi-Layer Perceptron neural network classifier, where we use 5-fold cross validation to tune our parameters. We tune over hidden layer size, activation function, solver, max iteration and alpha (or L2 penalty).

We obtain a classification accuracy score of 0.8400000000000001 from the following cross-validated parameters; the optimal activation function was relu, L2-penalty or alpha is 0.0005, 10 hidden layers, 300 max iterations, and the adam solver.

For our K-Nearest Neighbors Classifier, we tune the parameters using 5-fold cross validation for number of neighbors, weights and algorithm used to compute nearest neighbors.

We obtain a classification accuracy score of 0.8288151658767774 with the tuned parameters of 9 nearest neighbors, automatically selected algorithm, and uniform weights.

Next, our Logistic Regression classifier tuned with 5-fold cross validation gives us a classification accuracy score of 0.8387372013651877.

For our Support Vector Machine (SVM) classifier, instead of using a linear version of the classifier, we implement the SVM with an radial basis function (RBF) kernel.

Using this kernel parameter, we obtain a classification accuracy on the test set of 0.8369843527738264.

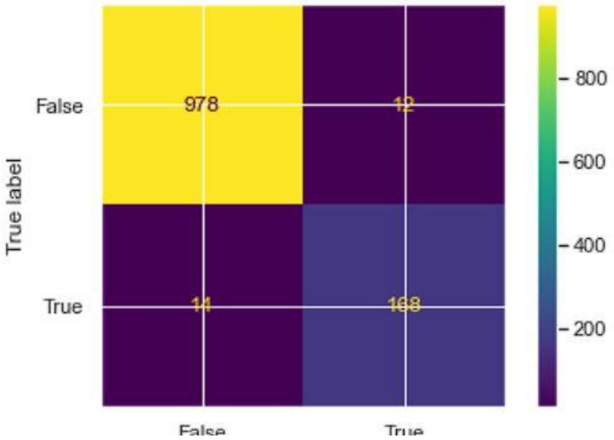

Figure 7. Confusion matrix of Naive Bayes Model

We also test a Gaussian based Naive Bayes classifier using 5-fold cross validation. This obtains a classification accuracy on the test set of 0.9759241706161138, which is our highest accuracy. The confusion matrix for this model is shown in Fig. 7 above, showing only 23 incorrect classifications.

This Naive Bayes classification produces the best performance amongst all our classifiers. The Naive Bayes near perfect classification accuracy may be attributed to the assumptions that this methodology employs. Since this classifier uses Bayes' theorem with strong (naive) assumptions about the independence of features considered in the classification, the 10 percent advantage in accuracy performance can be explained. The features we consider that are negatively correlated with hazardous asteroids such as orbit uncertainity and absolute magnitude may not be independent of each other, which contradicts this methodology's assumptions.

For this reason, we consider the neural network classifier via Multi-Layer Perceptron to be our strongest overall classification method, with an accuracy score of 0.8400000000000001. Since this is a neural network approach with multiple parameters to tune regarding the network's architecture and solving functions, we do not hold the same naive assumptions as the Bayes classifier and thus can more comfortably accept its classification accuracy and performance for future unseen data.

VI. References

[1] NASA. (n.d.). Neo basics. CNEOS. Retrieved May 7, 2022, from https://cneos.jpl.nasa.gov/about/neo_groups.html