Potentially Hazardous Asteroid Detection

Miftaun Noor 21241021
Munawar Mahtab Ansary 23341112
Avishek Paul 21301171
Mantaqa Abedin 23241130
Dewan Golam Mortoza 23141095

***Index Terms*—Asteroid, Machine Learning, Random Forest, K-Nearest Neighbour, Support Vector Machine.**

I. Introduction

In December 2018, an asteroid exploded in the upper atmosphere over the Bering Sea (western Pacific Ocean) with the explosive force of nearly 200 kilotons, or 10 times that of the Hiroshima bomb. This event, which was detected by various sensors and spotted by a Japanese weather satellite, demonstrates that Earth is frequently hit by objects, some of which could cause significant damage if they hit a populated area, as happened almost 6 years earlier over the Russian city of Chelyabinsk. [1]

Our aim is to train some machine learning models on a dataset of asteroids using supervised learning and finally come up with a trained model which can classify potential hazardous and non-hazardous asteroids with precision.

II. Methodology

A. Dataset description

The dataset we have found on asteroids is provided on ”Kaggle” titled ”NASA: Asteroids Classification”. [2] The dataset is provided by NEOWS (Near-Earth Object Web Service). All the data is collected from the website ”neo.jpl.nasa.gov”. [2]

The dataset is consisted of 4687 data points and 40 features. Among the features, 8 features are categorical in nature, the target feature termed ”Hazardous” has boolean values and all other features are numerical in nature. The features present in the dataset covers not only the information about the geometry of the asteroid, but also its path and speed.

B. Feature Engineering

The dataset has 6 irrelevant features to the target feature namely ’Neo Reference ID’, ’Name’ and ’Close Approach Date’, ’Epoch Date Close Approach’, ’Orbit ID’, ’Orbit Determination Date’ and 2 features namely ’Orbiting Body’ and ’Equinox’ which have only one unique value. These 8 features were the 8 categorical features mentioned earlier. We have omitted all of them, and thus are left with only numerical features. The target feature is encoded to 0s and 1s. There are no null values and duplicated values in the dataset.

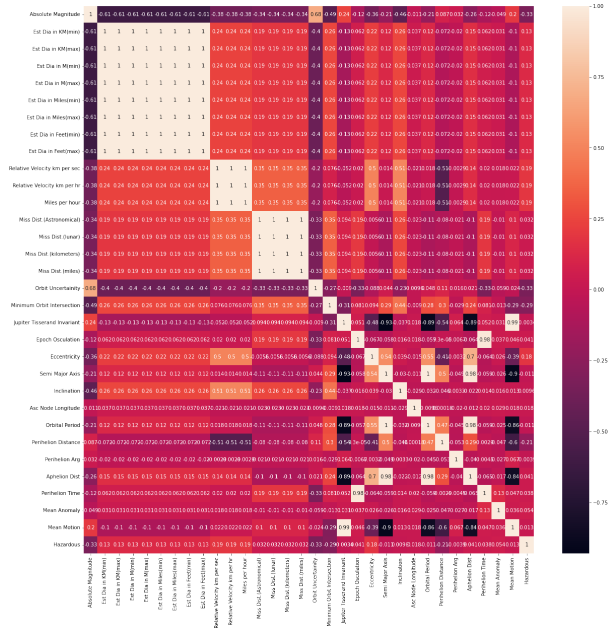
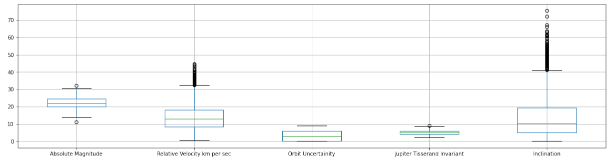


Fig. 1. Pearson Correlation values for the features.

From the pearson correlation values of the features, we have found the feature closely correlated to each other and omitted one feature out of each feature pair which has a absolute correlation value greater than or equal to 0.9. These features are ’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)’, ’Relative Velocity km per hr’, ’Miles per hour’, ’Miss Dist.(lunar)’, ’Miss Dist.(kilometers)’, ’Miss Dist.(miles)’, ’Mean Motion’, ’Perihelion Time’, ’Orbital Period’, ’Aphelion Dist’ and ’Semi Major Axis’. Thus we were left with 15 features, including the target feature.

The dataset is found to be prone to outliers by using box- plotting. Thus, we need to scale the dataset using a Scaler which is not sensitive to outliers. We have chosen to use Robust Scaler.



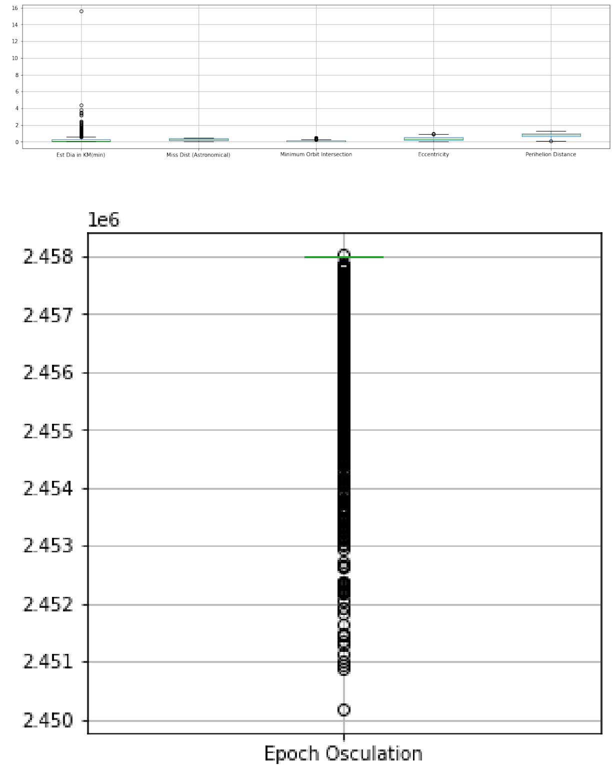


Fig. 2. Box plot of the features.

After scaling, we have removed the features with significant number of outliers namely 'Est Dia in KM(min)','Relative Velocity km per sec', 'Epoch Osculation' and 'Inclination' so that we can use outlier sensitive classifiers to train our model as well.

III. Model Application

As the Target Feature is categorical in nature, we would use Classifier ML Models. In addition, by pair plotting the values of the features, it is clear that the data points are linearly inseparable. Thus, we would use Non-linear Classification models.

We have concluded to use Random Forest Classifier, a tree based classifier, as it is a non-linear classifier not sensitive to outliers. In addition, as the features with significant number of outliers are removed now, we decided to use outlier sensitive non-linear classifiers i.e. K-Nearest Neighbour Classifier and Support Vector Classifier.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. For our dataset, we have used 'Gini Index' as a criterion for calculating information gain.

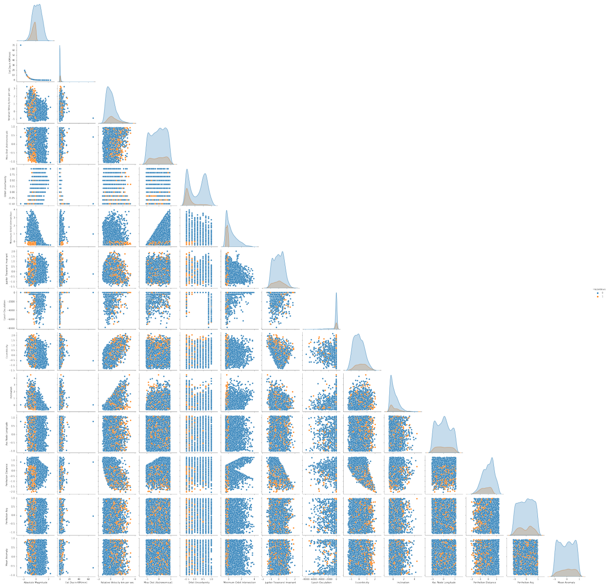


Fig. 3. Pair plot of the features.

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. We would be using distance as the weight function for the dataset.

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. As our data points are so much overlapped, we would use Radial Base Function(RBF) as the Kernel Function.

IV. RESULTS

A. Random Forest Classifier

From our test we get an accuracy score of 99.72f1 score of 0.9914 for Random Forest Classifier. In addition, there were 2 false positives and 2 false negative for this model.

B. K Nearest Neighbour

From our test we get an accuracy score of 94.67score of 0.826 for K Nearest Neighbour. In addition, there were 21 false positives and 54 false negatives for this model.

C. Support Vector Machine

From our test we get an accuracy score of 97.73score of 0.9304 for Support Vector Machine. In addition, there were 14 false positives and 18 false negatives for this model.

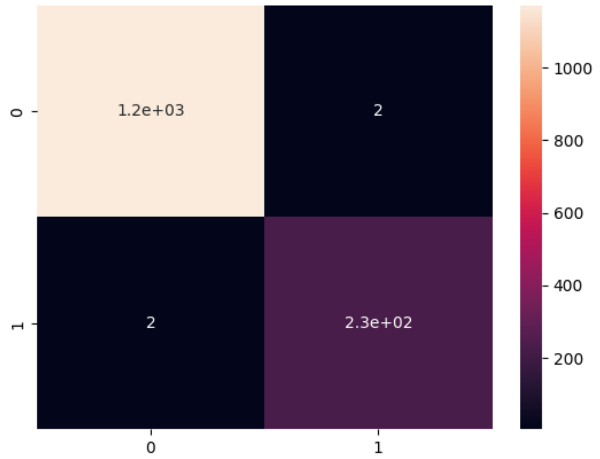


Fig. 4. Confusion Matrix of Random forest.

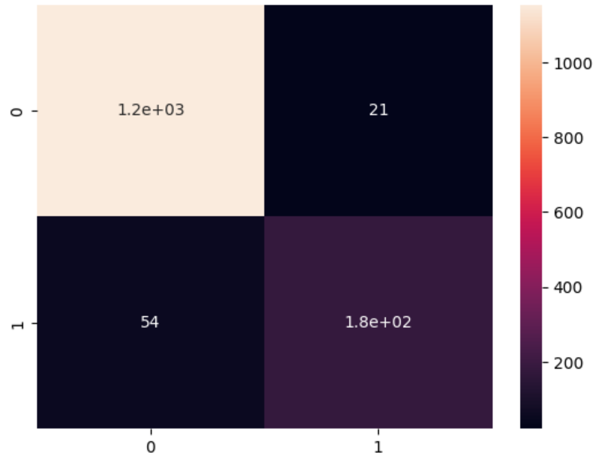


Fig. 5. Confusion Matrix of KNN.

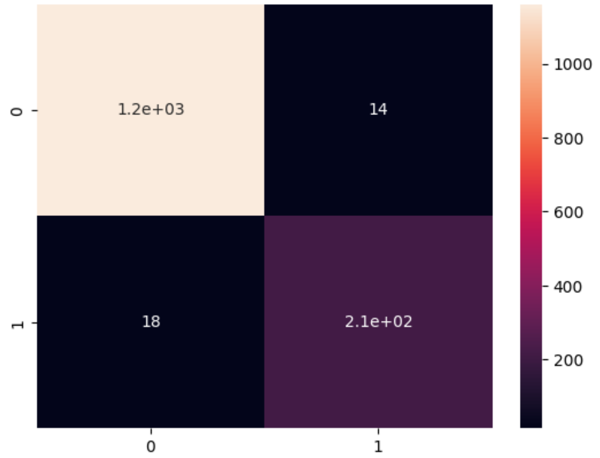


Fig. 6. Confusion Matrix of SVM.

V. Conclusion

Among the three Classifiers used on the dataset, Random Forest Classifier has the highest value of accuracy score and f1 score and the lowest value of false positives and false negatives. Thereby, it is concluded that for this dataset, Random Forest Classifier produced the best outcome.

References

[1] National Academies of Sciences, Engineering, and Medicine, 2019. Finding Hazardous Asteroids Using Infrared and Visible Wavelength Telescopes. Washington, DC: The National Academies Press. https://doi.org/10.17226/25476.

[2] NASA: Asteroids Classification. (2018, March 1). Kaggle. https://www.kaggle.com/datasets/shrutimehta/nasa- asteroids-classification