NASA Asteroids Classification

Iragavarapu Sai Pradeep AI21BTECH11013
Suraj Kumar AI21BTECH11029

Abstract

Asteroids, remnants of our solar system's formation, hold valuable clues about its history and composition. Asteroids exist in a very sparsely populated region of the solar system, often referred to as the asteroid belt, located between Mars and Jupiter. Asteroids have properties like diameter, relative velocity, composition, and orbital characteristics. All these properties help us know more about these celestial objects. This project aims to classify various asteroids into hazardous and non-hazardous, which provides further information about the potential effect of these asteroids on our planet. Various features affect this classification, few of them are size of asteroid, orbit characteristics like eccentricity inclination, relative velocity of asteroid and By analyzing these features and building a classification model, we can prioritize asteroids that warrant further observation and potential mitigation strategies, if necessary. This project contributes to planetary defense efforts by helping us identify and understand potential threats from asteroids.

1. Introduction

1.1. Background and Motivation

celestial objects like asteroids could cause a potential threat to our planet. Space organizations like ISRO, NASA, CNSA work hard to know about these asteroids to gain useful insights for future defense for the safety of mankind. Important features of these asteroid like size velocity eccentricity are under prime consideration while learning about these asteroids.

1.2. Objective

We aim to focus on important features of the asteroids and try to detect whether these celestial objects could cause hazard to our planet.

1.3. About the dataset

For our analysis, we considered the NASA: Asteroids Classification, in kaggle. This dataset has more than 30 features for about 4000 asteroids. The data present in this dataset is officially released by NASA. Our dataset can be found here. and can be found here

2. Exploratory Data Analysis(EDA)

Our initial venture into asteroid classification hinges on a thorough Exploratory Data Analysis (EDA). This stage involves meticulously examining the characteristics of known asteroids within NASA's dataset to uncover patterns and relationships that will guide our model building.

2.1. Data Cleaning and Preprocessing

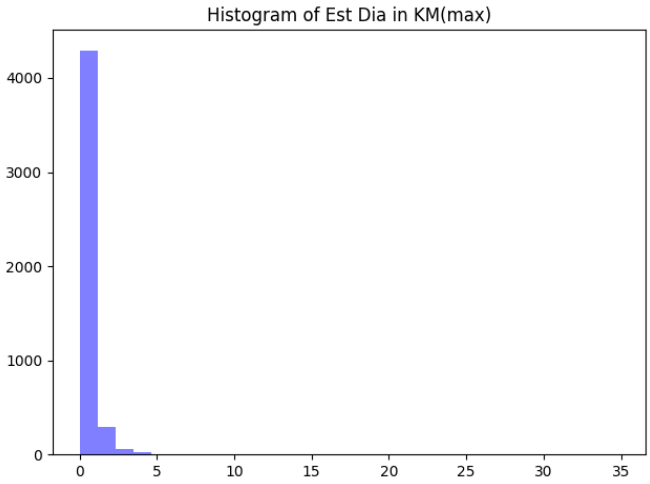
Before diving into analysis, we need to ensure the data's quality. This might involve handling missing values, identifying and correcting inconsistencies, and potentially transforming data formats for optimal analysis. We found that the dataset doesn't have any missing or undefined values, hence no handling is needed. We can see that, in the dataset we have multiple columns which depict the same quantity, but in different units. Columns like 'Est Dia in KM(min)','Est Dia in M(min)','Est Dia in Miles(min)', 'Est Dia in Feet(min)' convey the same meaning, i.e estimated diameter of asteroids in different distance metrics, similarly for 'Est Dia in KM(max)', 'Est Dia in M(max)', 'Est Dia in Miles(max)'and 'Est Dia in Feet(max)' convey the same parameter. Similarly 'Relative Velocity km per sec', 'Relative Velocity km per hr', 'Miles per hour' measures the asteroid relative velocity in different metrics. 'Miss Dist.(Astronomical)', 'Miss Dist.(lunar)', 'Miss Dist.(kilometers)', 'Miss Dist.(miles)' convey the distance by which the asteroid is missed from earth. For the further analysis and classification, we use one of the metrics of each of these features, for avoiding redundancy.

2.2. Univariate Analysis

In this section we would examine and analyse individual features of the asteroids.

1. **Estimated Diameter of Asteroids (in KM):**
   insights: We can see that, in the dataset most of the asteroids have their diameter less than 5KM, and the number of asteroids are decreasing with the increase in

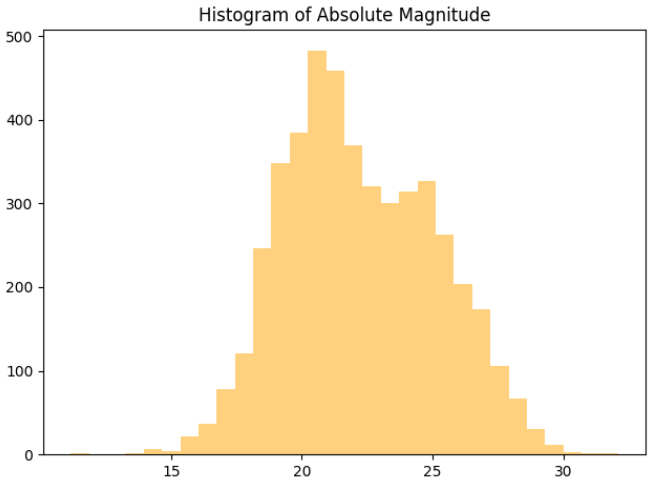
Figure 1. **histogram of Est Dia Max(in KM)**



the diameter of asteroids.

2. **Absolute Magnitude of Asteroids:**

Figure 2. **histogram of Absolute Magnitude**



insights: We can see that, in the dataset most of the asteroids have moderate brightness, with a few having small and large values.

3. **Inclination of Asteroids:**
   insights: A majority of the asteroids within the data exhibit inclinations concentrated in the range (0, 30 degrees), with a peak around 3.74 degrees. This indicates that most asteroids in dataset have orbits that lie close to the plane of the solar system (ecliptic plane).

4. **Orbital Period of Asteroids:**

Figure 3. **histogram of Inclination**

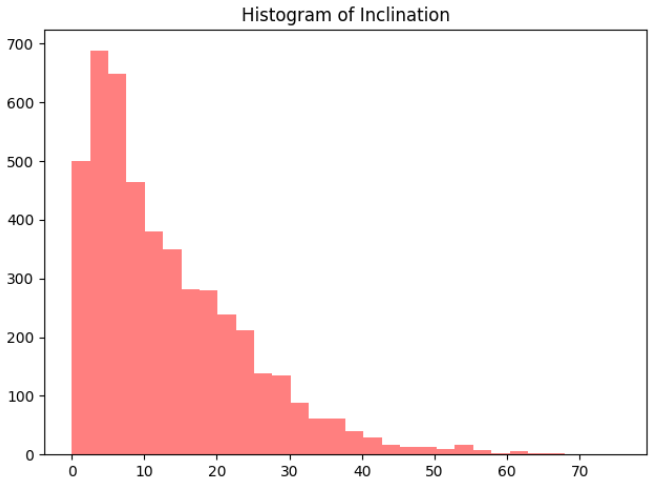
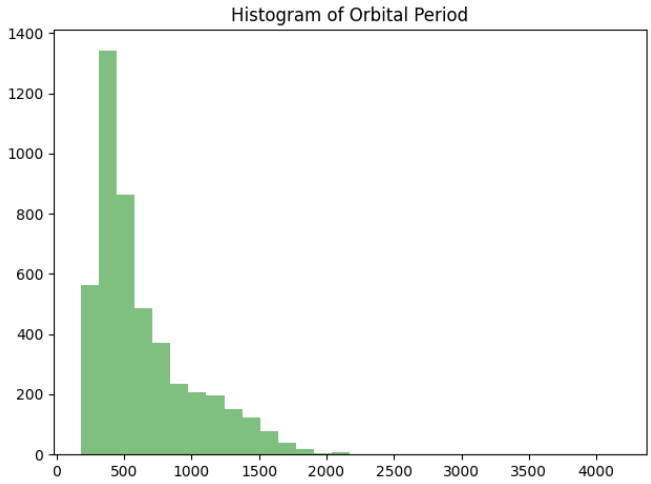


Figure 4. **histogram of Orbital period**



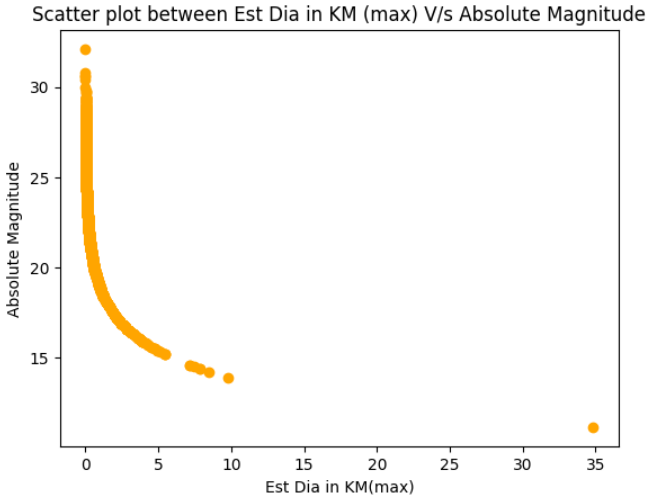
insights: The number of asteroids is first increasing, reaching the maximum, and decreasing with the increase in the orbital period.

2.3. Bivariate Analysis:

Moving beyond individual features, we now explore relationships between pairs of variables.

1. **Estimated Diameter in KM(max) and Absolute Magnitude**
   insights: From the scatter plot we can see that, as estimated diameter increases, the absolute magnitude decreases. From the plot, absolute magnitude seems to be negatively quadratic related to estimated diameter. This aligns with the expected trend, as larger objects tend to reflect more sunlight and appear brighter, have relatively less absolute magnitude. The logarithmic rela-

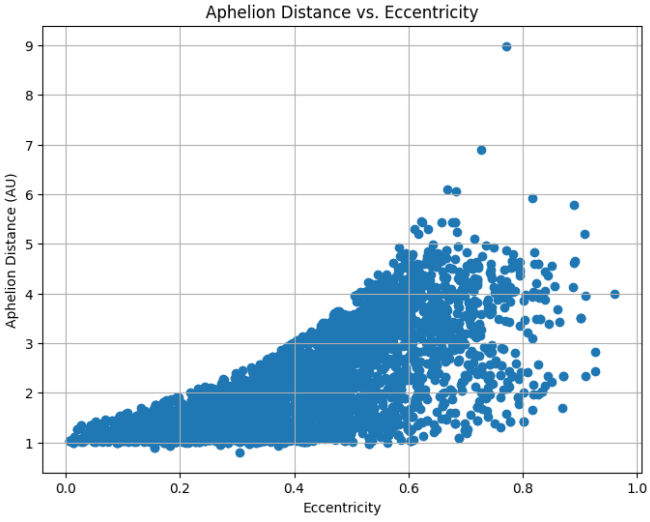
Figure 5. **scatter plot b/w Est Dia in KM(max) and Absolute Magnitude**



tionship between log of estimated diameter and absolute magnitude has a perfect correlation(correlation=1), further reinforcing our intuition.

2. **Aphelion Distance and Eccentricity**

Figure 6. **scatter plot between aphelion distance and eccentricity**



insights: We can see that, as the eccentricity is increasing, the aphelion distance is increasing, and also the variance in the aphelion distance is increasing. The observed correlation between eccentricity, aphelion distance, and its variance might be indicative of the influence of eccentricity on orbital characteristics.

2.4. Dimetionality Reduction:

Since our dataset has a lots of features, we wish to use dimensionality reduction techniques for better classification for preventing overfitting and model complexity.

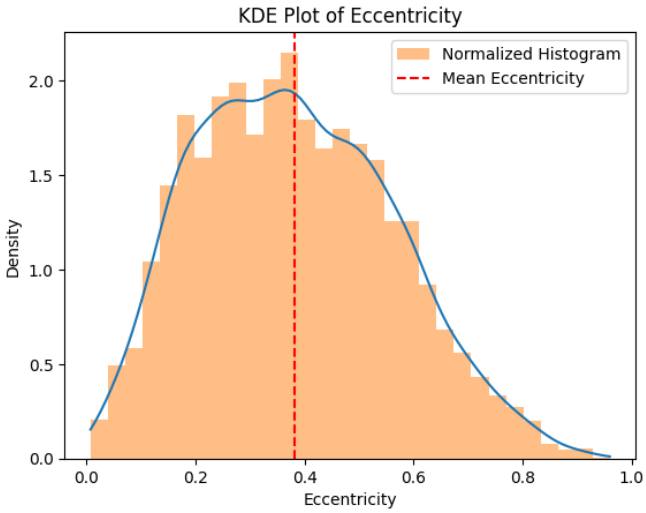
1. Principal Component Analysis(PCA): We employed PCA for dimensionality reduction on our dataset.
2. PCA has been performed to retain components that collectively explain 99% of the total variance in the dataset, resulting in a selection of 14 components.

2.5. Data Visualization Analysis:

We now explore more robust visualization methods for gaining further insights into the data.

1. **KDE on Eccentricity of the asteroids**

Figure 7. **histogram and KDE plot of asteroids eccentricity**



insights: The KDE plot of eccentricity suggests the presence of more asteroids in the moderate range of eccentricities, having the majority of them in (0.2,0.6) range and peak value around 0.4. This suggests that most asteroids in your data exhibit moderately elliptical orbits, neither perfectly circular nor highly elongated.

2. **Variation of eccentricity with Aphelion Distance and Perihelion Distance**
insights: From the plot, we can say, for a given Perihelion Distance, with decrease in Aphelion Distance, the eccentricity also decreases, and vice versa. This suggests a tendency for asteroids with similar perihelion distances to have more circular orbits (lower eccentricity) when their aphelion distance is also smaller.

3. **Correlation Heat Map**
insights: As described earlier, and from the correlation heat ma, we can say that may parameters depicts same

Figure 8. **Aphelion Distance V/s Perihelion Distance**

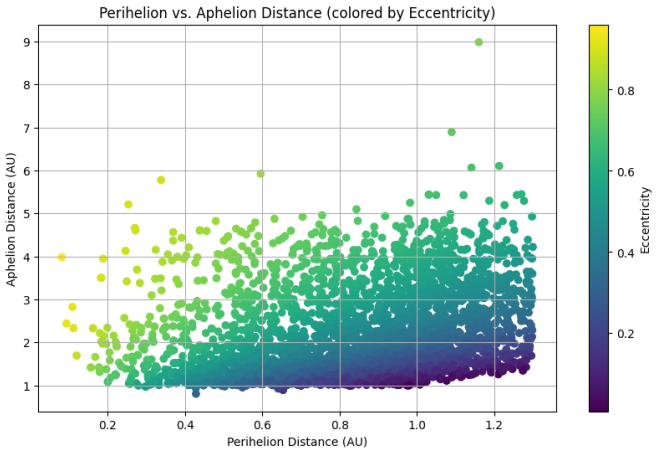
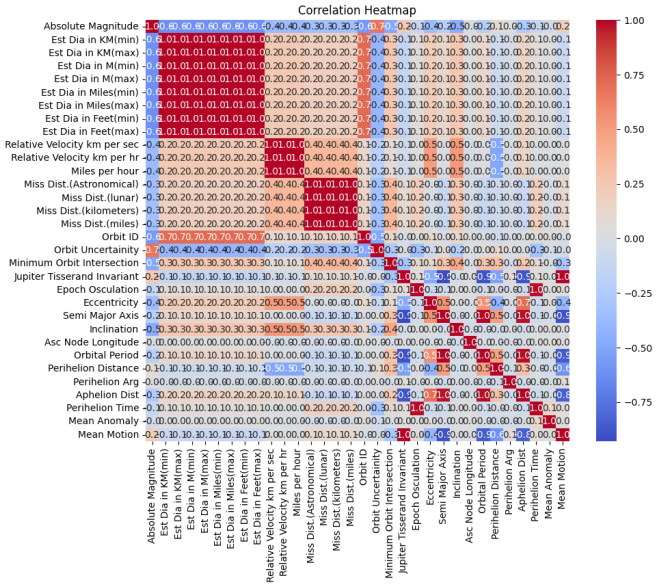


Figure 9. **heat map correlation**



quantity in different metrics(these variables have correlation as 1.0). The high correlation between Orbital Period and Aphelion Distance, in accordance with the expected trend.

3. Classification task

After gaining handful insights about our data, we now focus at the classification task. We aim to classify Near Earth Objects (NEOs) as hazardous or non-hazardous. For this task, we applied several classification methods including Logistic Regression, Logistic Regression with L2 penalty (Ridge Regression), Support Vector Machine (SVM), Random Forest, and Decision Tree.

For making our classification task fit for foreign(unknown) data, we split the dataset into training and testing sets for our classification. The table below summarizes the dataset split:

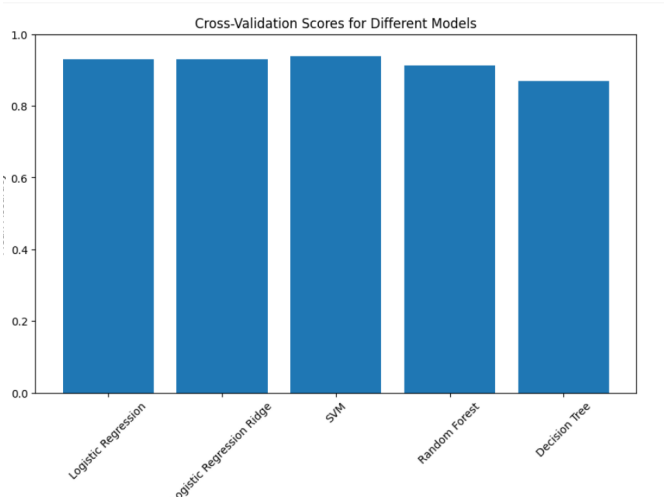
Table 1. Dataset Split for NEO Classification

	Total Samples	Hazardous NEOs	Non-Hazardous NEO
Training Set	3037	605	2432
Testing Set	1650	150	1500

4. Result

After defining our task, we further trained and evaluated different classification methods for selecting the best model for this task. We conducted 10-fold cross-validation for each classification model to evaluate their performance. The Figure below summarizes the mean accuracy obtained from cross-validation:

Figure 10. Mean Accuracy Results from 10-fold Cross-Validation



Based on the mean accuracy results, we observe that SVM achieved the highest mean accuracy of 0.93, indicating that it performed the best among the models tested.

Table 2. Model Results on Training and Testing Sets

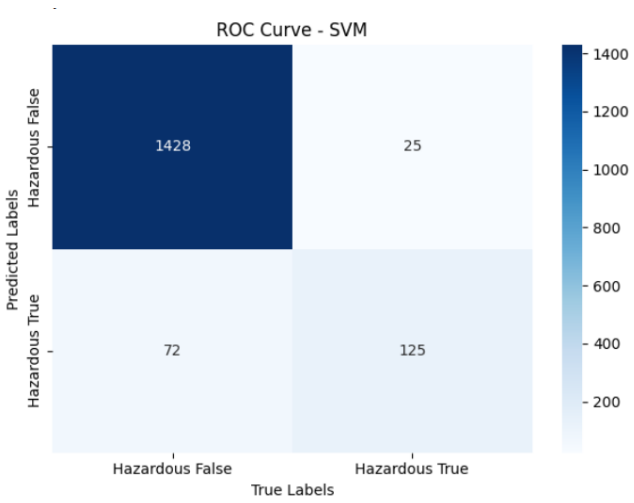
Model	Training Accuracy	Test Accuracy
Logistic Regression	0.9338	0.9321
Ridge Regression	0.9338	0.9321
SVM	0.9602	0.9412
Random Forest	1.0000	0.9315
Decision Tree	1.0000	0.8824

The table above presents the accuracy results for each model on both the training and testing sets. Based on these

results, the SVM model achieved the highest accuracy on the test set.

After evaluating the Support Vector Machine (SVM) model on the testing set, we obtained the following confusion matrix: After testing our model on 1650 data points,

Figure 11. Confusion metrix for SVM over test Data



we observed the following:

- Out of 1500 samples evaluated as non-hazardous, 1428 were correctly classified as non-hazardous (True Negatives).
- 72 samples were incorrectly classified as hazardous when they were actually non-hazardous (False Positives).
- 25 samples were incorrectly classified as non-hazardous when they were actually hazardous (False Negatives).
- 125 samples were correctly classified as hazardous (True Positives).

This breakdown provides insights into the performance of our model in distinguishing between hazardous and non-hazardous NEOs.

Some Insights of performance of each model

1. Logistic Regression achieved a training accuracy of 0.9338 and a test accuracy of 0.9321. It demonstrated stable performance with a mean accuracy of 0.92 during 10-fold cross-validation.
2. Ridge Regression, similar to Logistic Regression, achieved a training accuracy of 0.9338 and a test accuracy of 0.9321. During 10-fold cross-validation, it exhibited a mean accuracy of 0.92.
3. SVM outperformed other models with a training accuracy of 0.9602 and a test accuracy of 0.9412. It demonstrated the highest mean accuracy of 0.93 during 10-fold cross-validation, indicating its effectiveness in distinguishing between hazardous and non-hazardous NEOs.
4. Random Forest achieved a perfect training accuracy of 1.0 but exhibited a slightly lower test accuracy of 0.9315. During 10-fold cross-validation, it maintained a mean accuracy of 0.91.
5. Decision Tree achieved a training accuracy of 1.0 and a test accuracy of 0.8824. It exhibited a mean accuracy of 0.86 during 10-fold cross-validation, indicating slightly lower performance compared to other models.

5. Conclusion

In conclusion, in this paper we aimed to classify Near Earth Objects (NEOs) as hazardous or non-hazardous using various classification methods, including Logistic Regression, Ridge Regression, Support Vector Machine (SVM), Random Forest, and Decision Tree. Also we explored various trends in the dataset variables, which helped us gain useful insights about the data. In process of classifying, we split the dataset into training and testing sets, with 3037 samples used for training and 1650 samples for testing. After evaluating each model's performance using 10-fold cross-validation, we found that the SVM model achieved the highest mean accuracy of 0.93, indicating its effectiveness in distinguishing between hazardous and non-hazardous NEOs.

Further analysis of the SVM model on the testing set revealed a confusion matrix that provided insights into its performance. Out of 1650 data points tested, the SVM model correctly classified 1428 non-hazardous NEOs and 125 hazardous NEOs. However, it incorrectly classified 72 non-hazardous NEOs as hazardous and 25 hazardous NEOs as non-hazardous.

Overall, while the SVM model demonstrated strong performance with the highest mean accuracy.

In short, in our project, we intended to gain insights on our data, and classified the Near Earth Objects, which further helped us assess the potential impact risk posed by NEOs to Earth.

6. Resources

1. Codes for the project can be found here