Classification of Hazardous and Non-Hazardous Asteroids

Abdul Hadi Durrani, Nawall Aamer, Abdullah Awan, Soha Bilal

May 12, 2024

1 Introduction

This report presents a comprehensive analysis of hazardous and non-hazardous asteroid classification utilizing various machine learning techniques. Employing preprocessing, feature extraction via Recursive Feature Elimination (RFE) and Forward Feature Selection, and diverse classification algorithms, we aim to discern patterns distinguishing hazardous asteroids from benign ones. Additionally, we integrate Explainable AI models such as LIME and SHAP to enhance interpretability and transparency in classification outcomes. Furthermore, clustering algorithms are utilized to uncover hidden structures within the datasets. By elucidating the distinguishing features of hazardous asteroids, this study contributes to efforts in asteroid risk assessment and reinforces preparedness against potential cosmic threats.

2 Technologies Used

- Google COLAB
- VSCode
- Python3
- Pandas
- SKLearn
- TensorFlow
- SHAP
- LIME

3 About data and Pre-Processing

The dataset comprises 40 columns and 4688 rows, encompassing various attributes pertinent to asteroid characteristics and orbital parameters. Key features include Neo Reference ID, Name, Absolute Magnitude, and estimations of asteroid diameter in different units. Orbital parameters such as Relative Velocity, Minimum Orbit Intersection, and Orbital Period are also included. Of particular importance is the 'Hazardous' column, serving as the target class, indicating whether an asteroid is hazardous or not. It's worth noting the presence of an imbalance in the hazardous class, necessitating Pre-Processing techniques to address this issue.

3.1 Pre-Processing

In the preliminary preprocessing phase, various techniques were employed to ensure data quality and handle missing values. Object values were addressed by using Imputer for mode as well as one hot encoding, while missing numeric values were imputed using the mean and scaled for uniformity across features. A heatmap (**Figure 1**) was utilized to visualize the correlation among columns, aiding in

the identification of redundant or highly correlated features. Subsequently, recursive feature elimination and forward feature selection were applied to select the most relevant features for classification, yielding a subset comprising attributes such as 'Absolute Magnitude', 'Est Dia in KM(min)', 'Est Dia in KM(max)', 'Est Dia in M(max)', 'Est Dia in Miles(min)', 'Est Dia in Miles(max)', 'Est Dia in Feet(max)', 'Minimum Orbit Intersection', 'Inclination', and 'Perihelion Distance'. Furthermore, we visualized the imbalance in the hazardous class **Figure 2**, oversampling techniques were employed to mitigate class imbalance, ensuring robust model training and evaluation.

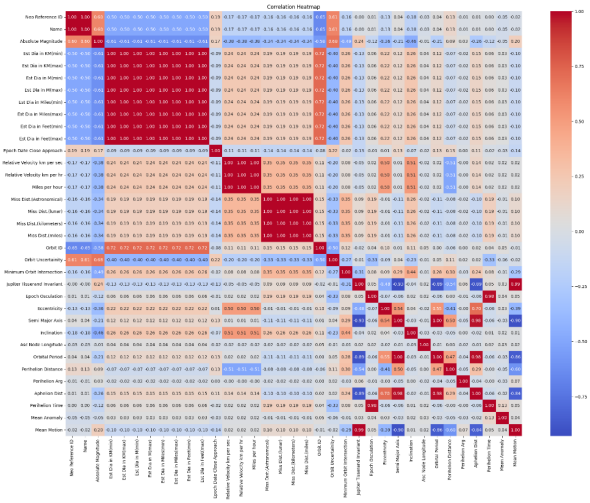

Figure 1:

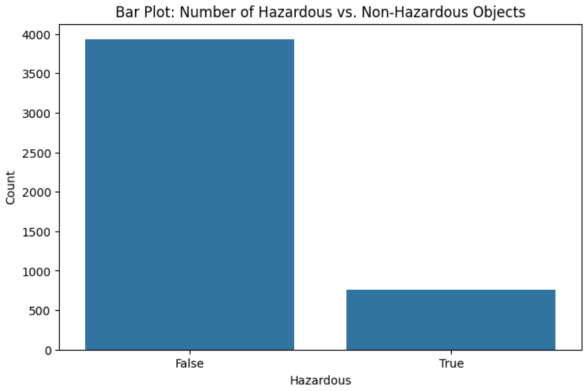

Figure 2:

4 Different Models

To discern the most effective classification model for our selected features, a variety of machine learning algorithms were evaluated. Utilizing pipelines and cross-validation techniques, the following classifiers were assessed:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Support Vector Classifier (SVC)

5. K-Nearest Neighbors Classifier (KNeighbors)
6. Gaussian Naive Bayes Classifier (GaussianNB)
7. AdaBoost Classifier
8. Gradient Boosting Classifier
9. Linear Regression

Each model was rigorously evaluated using cross-validation to ensure robustness and mitigate overfitting. Furthermore, to optimize model performance, grid search cross-validation was employed to tune hyperparameters for each classifier.

4.1 Logistic Regression

The Logistic Regression model achieved a mean F1 score of approximately 0.86 with a standard deviation of 0.03, indicating good performance in classifying hazardous and non-hazardous asteroids.

4.2 Decision Tree Classifier

The Decision Tree Classifier outperformed other models with a mean F1 score of approximately 0.99 and a low standard deviation of 0.01, demonstrating excellent classification accuracy and consistency.

4.3 Random Forest Classifier

Similarly, the Random Forest Classifier yielded a high mean F1 score of approximately 0.99 with a small standard deviation of 0.01, showcasing robust performance and reliability in asteroid classification.

4.4 Support Vector Classifier (SVC)

The SVC model achieved a mean F1 score of approximately 0.90 with a moderate standard deviation of 0.02, indicating satisfactory performance in distinguishing between hazardous and non-hazardous asteroids.

4.5 K-Nearest Neighbors Classifier

The K-Nearest Neighbors Classifier obtained a mean F1 score of approximately 0.88 with a standard deviation of 0.03, demonstrating decent performance in asteroid classification but slightly lower than other models.

4.6 Gaussian Naive Bayes Classifier

The Gaussian Naive Bayes Classifier attained a mean F1 score of approximately 0.85 with a standard deviation of 0.03, indicating acceptable performance but lower than some other models evaluated.

4.7 AdaBoost Classifier

The AdaBoost Classifier achieved a high mean F1 score of approximately 0.99 with a small standard deviation of 0.01, demonstrating robust performance and consistency in asteroid classification.

4.8 Gradient Boosting Classifier

Lastly, the Gradient Boosting Classifier achieved a mean F1 score of approximately 0.99 with a standard deviation of 0.01, indicating excellent performance and consistency similar to other ensemble methods.

5 Explainable AI Models

we augmented our analysis with Explainable AI (XAI) models, specifically LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). These XAI techniques offer valuable insights into the underlying mechanisms driving classification outcomes, shedding light on the importance of features and their impact on model predictions. By employing LIME, we generated local explanations for individual predictions, providing intuitive insights into how specific features influence the model's decision-making process. Similarly, SHAP analysis facilitated a deeper understanding of feature importance by quantifying the contribution of each feature to model predictions across the entire dataset. Through the integration of LIME and SHAP, we enhance the transparency and trustworthiness of our classification models, empowering stakeholders with actionable insights into the factors driving asteroid categorization.

5.1 LIME (Local Interpretable Model-agnostic Explanations)

Lime analysis reveals key features contributing to the classification of hazardous and non-hazardous asteroids. Notably, a high Minimum Orbit Intersection (¿ 0.06) is negatively correlated with hazard classification, indicating that asteroids with greater orbital intersection tend to be classified as non-hazardous. Similarly, a high Perihelion Distance (¿ 0.95) exhibits a negative correlation with hazard classification. Conversely, features such as Absolute Magnitude (¡= 19.70), Est Dia in KM(min) (¿ 0.31), and Est Dia in KM(max) (¿ 0.68) positively contribute to hazard classification, suggesting that larger asteroid size and higher absolute magnitude are associated with hazardous classification. These insights provide valuable interpretability into the model's decision-making process, aiding in understanding the underlying factors influencing asteroid classification.

5.2 SHAP (SHapley Additive exPlanations)

The SHAP summary graph (**Figure 3**) illustrates the impact of different features on the model's output for classifying hazardous and non-hazardous asteroids. Features such as 'Absolute Magnitude' and 'Minimum Orbit Intersection' exhibit notable impacts on the model output. A lower absolute magnitude is associated with a higher likelihood of hazardous classification, as evidenced by its negative SHAP value of -5.7. Conversely, higher values of 'Minimum Orbit Intersection' are negatively correlated with hazard classification, as indicated by its negative SHAP value of -13. Additionally, features such as 'Est Dia in KM' and related metrics have relatively lower impacts on the model output, with SHAP values around 0.2. Features such as 'Inclination' and 'Perihelion Distance' also exhibit negative impacts on hazard classification, albeit to a lesser extent compared to 'Absolute Magnitude' and 'Minimum Orbit Intersection'.

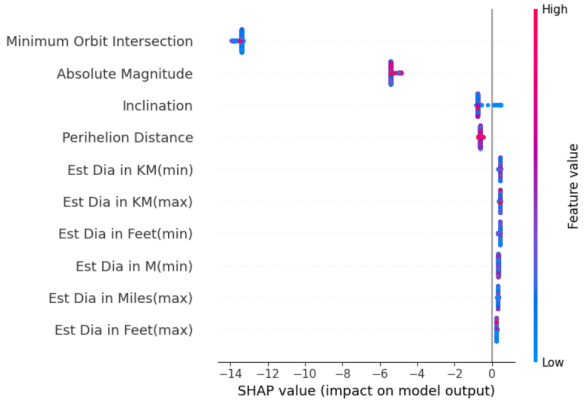

Figure 3:

6 Neural Network ANN

Utilizing TensorFlow, we constructed a neural network model for classifying hazardous and non-hazardous asteroids. Trained over 100 epochs with a dropout rate of 0.5, the model achieved a validation accuracy of 0.76, indicating its potential for accurate classification. However, a detailed examination of the classification report uncovered imbalances in precision and recall, particularly for hazardous asteroids. While the model exhibits high recall for hazardous asteroids (0.99), precision is lower (0.68), suggesting a need for further refinement to reduce false positives. Despite this, the model demonstrates promise in asteroid classification tasks, with opportunities for optimization to enhance its performance further.

7 Clustering

Incorporating clustering algorithms such as KMeans(**Figure 4**), DBScan(**Figure 5**), Agglomerative(**Figure 6**), and Hierarchical(**Figure 7**), we sought to glean additional insights into the inherent structures within our asteroid dataset. However, despite rigorous experimentation, these clustering techniques did not yield significant or interpretable results. The absence of discernible clusters suggests a high degree of heterogeneity or noise within the dataset, rendering traditional clustering approaches ineffective in extracting meaningful patterns or groupings. Nonetheless, while clustering did not provide direct insights into distinguishing hazardous and non-hazardous asteroids, its exploration remains valuable in comprehensively assessing the dataset's characteristics and guiding future analyses. We used "Minimum Orbit Intersection" and "Perihelion Distance" for clustering and cluster were made in regards to hazardous and non-hazardous.

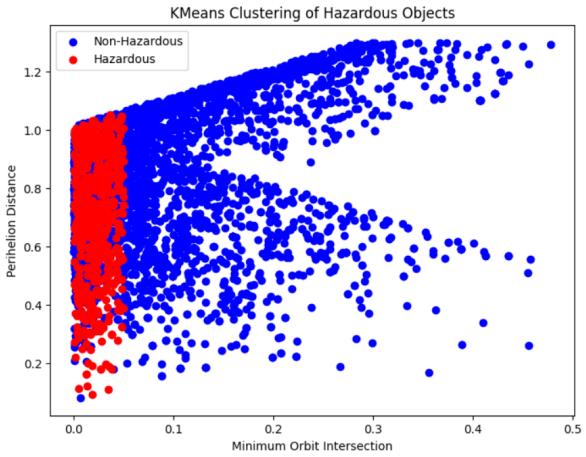

Figure 4:

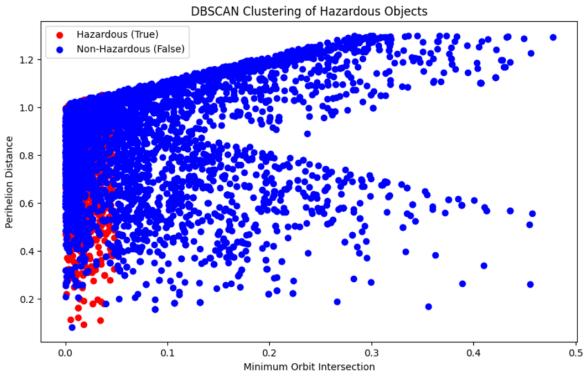

Figure 5:

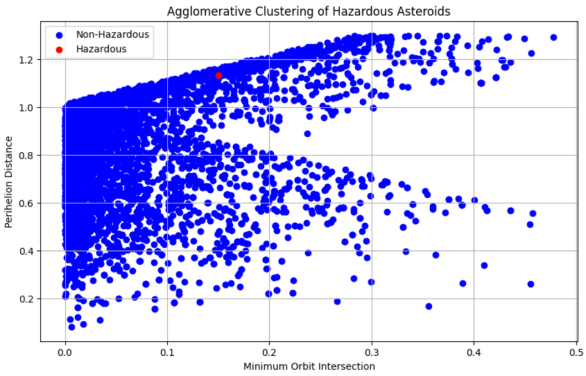

Figure 6:

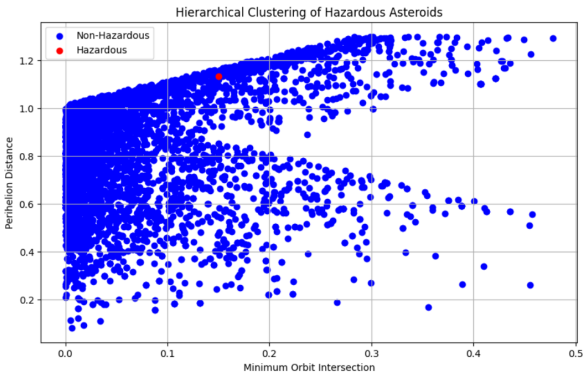

Figure 7: