Applying Machine Learning Principles to Categorize Asteroids as Hazardous or Non-Hazardous

Harshil Sharma
Department of Computer Science
Symbiosis Institute of Technology
Pune, India
harshil.sharma.btech2020@sitpuneedu.in

Ojasv Issar
Department of Computer Science
Symbiosis Institute of Technology
Pune, India
ojasv.issar.btech2020@sitpune.edu.in

Lakshita Sharma
Department of Computer Science
Symbiosis Institute of Technology
Pune, India
lakshita.sharma.btech2020@sitpune.edu.in

Usha Jogalekar
Department of Computer Science
Symbiosis Institute of Technology
Pune, India
usha.jogalekar@sitpune.edu.in

Dr. Ketan Kotecha
Director
Symbiosis Institute of Technology
Pune, India
director@sitpune.edu.in

Arishi Agarwal
Department of Computer Science
Symbiosis Institute of Technology
Pune, India
arishi.agarwal.btech2020@sitpune.edu.in

***Abstract*— Asteroid impacts on Earth pose significant risks to our planet's safety and well-being. Identifying potentially dangerous asteroids among the multitude of celestial objects is a crucial task for planetary defense. In this research paper, we present a comprehensive analysis of a predictive modeling approach for classifying asteroids as potentially dangerous or not. The study encompasses a wide range of data preprocessing techniques, feature selection, and model fine-tuning, ultimately leading to the development of an accurate predictive model. The research leverages astronomical data, including orbital parameters, physical properties, and historical records, to build a predictive model capable of distinguishing between near-Earth asteroids that pose potential threats and those that do not. The data analysis includes exploratory data visualization, feature engineering, and several data transformation methods, such as standard scaling, min-max scaling, and robust scaling. Additionally, we employ the Synthetic Minority Over Sampling Technique for Nominal and Continuous (SMOTENC) to address class imbalance. The predictive modeling process involves the evaluation of various machine learning algorithms, including Random Forest, Extra Trees, XGBoost, and Multi-Layer Perceptron. We identify the best-performing model through cross-validation and hyperparameter tuning, while also assessing the model's ability to generalize and avoid overfitting. We measure the model's performance using classification metrics such as accuracy, F1-score, recall, precision, and ROC-AUC. The results show that the proposed predictive model, based on XGBoost and utilizing the Robust Scaler transformation, achieves high accuracy and F1-score in classifying potentially dangerous asteroids. It demonstrates robustness to class imbalance and performs well on both the training and test datasets, with an ROC-AUC score close to 1. These findings suggest the model's effectiveness in helping to identify potentially hazardous asteroids in a real-world planetary defense scenario. This research contributes to the field of space science and planetary defense by providing an advanced tool for automated asteroid classification, offering enhanced accuracy and reliability in identifying asteroids that may require further monitoring and mitigation strategies.**

Keywords—Potentially dangerous asteroids, predictive modeling, machine learning, feature selection, SMOTENC, planetary defense, astronomical data analysis

I. INTRODUCTION

The ever-present possibility of an asteroid impact on Earth remains a significant concern for the safety and security of our planet. While rare on human timescales, asteroid impacts have shaped Earth's geological history and have the potential to cause widespread destruction, injury, and loss of life. To mitigate these risks, planetary defense efforts have become increasingly vital, involving the identification and monitoring of Near-Earth Objects (NEOs), including potentially dangerous asteroids. Achieving this goal requires advanced predictive models capable of distinguishing between benign and hazardous celestial bodies.

This paper focuses on the development and evaluation of a predictive modeling approach for the classification of potentially dangerous asteroids, which may pose a threat to Earth's inhabitants. Potentially dangerous asteroids are those that come within proximity to our planet's orbit and have the potential to impact Earth. The objective is to build an automated system that can effectively classify NEOs into two main categories: "Potentially Hazardous" and "Not Hazardous."

Asteroid classification is an intricate task that involves analyzing a diverse range of astronomical and physical properties. Orbital parameters, physical characteristics, and historical records provide essential information for differentiating between these categories. Machine learning and data analysis techniques play a fundamental role in processing and interpreting these data to make accurate predictions.

The study begins with an extensive data analysis phase, encompassing exploratory data visualization and feature engineering. The dataset comprises a wealth of information, including orbital parameters such as aphelion distance (farthest point in the orbit) and perihelion distance (closest point in the orbit), absolute magnitude, epoch, and more. Preprocessing techniques are applied to prepare the data for modeling. Several data transformation methods, such as standard scaling, min-max scaling, and robust scaling, are assessed to identify the most suitable approach for achieving robust and reliable predictions.

To address the inherent class imbalance in asteroid datasets, the Synthetic Minority Over Sampling Technique for Nominal and Continuous (SMOTENC) is employed. SMOTENC is particularly valuable for generating synthetic

samples of the minority class, mitigating issues related to class imbalance while preserving the distribution of categorical and continuous features.

The heart of this research lies in the evaluation of various machine learning algorithms for asteroid classification. Models, including Random Forest, Extra Trees, XGBoost, and Multi-Layer Perceptron, are subjected to rigorous cross-validation and hyperparameter tuning processes. These steps help determine the model that offers the best balance between accuracy and generalization performance, ensuring the model's ability to make reliable predictions on both the training and test datasets.

Performance assessment is conducted using key classification metrics, including accuracy, F1-score, recall, precision, and the area under the ROC curve (ROC-AUC). Achieving high accuracy and precision, while maintaining a high recall rate, is crucial for identifying potentially dangerous asteroids with confidence.

This research contributes to the field of planetary defense by providing a state-of-the-art tool for automating the classification of near-Earth objects. With advancements in predictive modeling and machine learning, the potential exists to enhance the accuracy and reliability of asteroid classification, enabling early detection and more effective risk mitigation strategies. The findings and methodology presented in this research paper hold significant implications for our understanding and preparedness regarding potentially hazardous asteroids.

II. Literature Review

The potential of asteroid classification using Machine Learning (ML) techniques to improve our understanding of these celestial bodies has made it an important topic of research with implications for resource utilization, planetary defense, and space exploration [1]. An extensive summary of the most significant advancements in this topic is given in this literature review. ML techniques seek to simplify the complex task of asteroid categorization. In 2002, Bus and Binzel introduced the Bus-DeMeo taxonomy, which has been a significant contribution to this discipline [2]. This technique of classification makes it possible to distinguish between various types of asteroids by grouping them according to their spectral properties [3]. The automation of this taxonomy's application, made possible by ML techniques, has simplified and improved the accuracy of asteroids' classification [4]. Studying the makeup and formation of these celestial bodies would be affected by this.

Important parts of classifying asteroids are predicting their orbits and locating possible impactors on Earth. Our capacity to handle these issues has greatly increased because of machine learning approaches. By utilizing machine learning methods to precisely forecast asteroid orbits, Milani et al. (2005) reduced uncertainty and improved our capacity to evaluate possible impact occurrences [5]. This use of machine learning is crucial for both comprehending solar system dynamics and planetary defense. The availability of large datasets from missions such as Pan-STARRS and NEOWISE has further revolutionized the field of asteroid categorization in recent years [6]. These datasets are rich in information; they include time-series data and photos of asteroids. Researchers have turned to deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to efficiently extract meaningful information from these datasets [7]. The ability of these deep learning models to extract intricate features from time-series data and photos is astounding. CNNs were used by Deo et al. (2020) to categorize asteroids according to their visual features [8]. CNNs made it possible to automatically extract features from asteroidal photos, which eliminated the need for human feature engineering [9]. Their method demonstrated how deep learning can handle massive asteroid datasets and expedite the categorization procedure [10]. Deep learning was extended to time-series data for the classification of asteroids by Ge et al. in 2021 [11]. They employed RNNs for the analysis of time-series observations of an asteroid's brightness, known as asteroid light curves [12]. By efficiently identifying temporal patterns in the data, the RNNs enhanced the classification of asteroids according to their rotational properties [13]. This method shows how deep learning can be applied to asteroid classification in a variety of situations, not only visual data. In conclusion, by automating the classification process and improving our capacity to anticipate asteroid orbits, asteroid classification using machine learning has advanced significantly. Large-scale asteroid datasets have made it easier to use deep learning models, such as CNNs and RNNs, which enable the extraction of complicated features from a variety of data formats. Improved asteroid classification could result from further development of machine learning techniques in this area, which would provide a more thorough knowledge of these space objects and their effects on our solar system. The results of this study could have a significant impact on future asteroid resource utilization as well as planetary defense [14].

III. Methodology

The methodology adopted for this research encompasses several key stages, each designed to address the primary objectives of developing an accurate predictive model for classifying potentially dangerous asteroids. These stages include data preprocessing, feature selection, scaling, the selection of the best machine learning algorithm, and model fine-tuning. The overall workflow is structured as follows:

*A. **Data Pre-Processing***

The initial phase of this study involves the acquisition, exploration, and preparation of the dataset. Data preprocessing is an essential step to ensure that the input data is in a suitable format for machine learning. It includes the following steps:

- Data Acquisition: The dataset, consisting of information on Near-Earth Objects (NEOs) and their various

attributes, was collected from reliable sources, such as astronomical observatories and space agencies.

- Exploratory Data Analysis (EDA): EDA is performed to gain a comprehensive understanding of the dataset. This step includes data visualization, statistical analysis, and the identification of any missing values or outliers.
- Feature Engineering: Several new features are generated from existing data. For instance, converting orbital parameters from perihelion distance and aphelion distance to semi-major axis (a) and eccentricity (e) to simplify feature representation.

B. Feature Selection

Feature selection is crucial for creating a lean and efficient model. Techniques such as random forest feature importance, XGBoost feature importance, and recursive feature elimination are applied to determine which features are most relevant for asteroid classification. The most important features identified by these methods guide the selection process, reducing dimensionality while preserving predictive power.

C. Data Transformation and Scaling

Data transformation is applied to prepare the dataset for modeling. The selected numeric features are subjected to different scaling techniques to ensure that all features are on a similar scale. Three scaling methods are evaluated: Standard Scaler, Min-Max Scaler, and Robust Scaler. These techniques mitigate the effects of outliers and help models converge efficiently during training.

D. Machine Learning Algorithm Selection

A significant aspect of the research is the selection and evaluation of machine learning algorithms for asteroid classification. The following algorithms are assessed:

- Random Forest: A powerful ensemble learning method known for its ability to handle complex datasets and capture feature importance.
- Extra Trees: A variant of the random forest, Extra Trees further enhances the ensemble approach.
- XGBoost: An optimized gradient boosting algorithm, XGBoost is known for its high predictive accuracy and efficiency.
- Multi-Layer Perceptron (MLP): A neural network model that can capture complex patterns in the data.

Each algorithm is evaluated using cross-validation, focusing on the F1-score, a critical metric that balances precision and recall. Additionally, the ROC-AUC (Area Under the ROC Curve) is computed to assess the models' ability to discriminate between potentially dangerous and non-dangerous asteroids.

.

E. Synthetic Minority Over Sampling Technique for Nominal and Continuous (SMOTENC)

The dataset exhibits class imbalance, with most asteroids being non-dangerous. To address this, SMOTENC is employed to oversample the minority class (potentially dangerous asteroids). SMOTENC generates synthetic samples, mitigating class imbalance and improving model performance.

F. **Model Fine-Tuning**

The model with the best cross-validation performance is selected, and hyperparameter tuning is conducted using grid search. The goal is to identify the optimal hyperparameters for the selected model, striking a balance between performance and generalization.

G. **Model Evaluation**

The selected model is evaluated using the test dataset. Classification metrics, including accuracy, F1-score, recall, and precision, are computed. Additionally, the ROC-AUC metric is utilized to assess the model's ability to discriminate between classes.

The methodology outlined in the paper aims to provide a robust framework for automated classification of potentially dangerous asteroids, contributing to planetary defense efforts. By meticulously preparing and analyzing the data, selecting suitable features, scaling numeric attributes, and leveraging advanced machine learning algorithms, the research seeks to create a reliable predictive model capable of enhancing early detection and risk mitigation strategies associated with NEOs.

IV. DATA

The dataset used for this research was meticulously curated by the Jet Propulsion Laboratory (JPL) of the California Institute of Technology, an esteemed institution operating under the auspices of NASA. It was obtained from Kaggle and contains valuable information about Near-Earth Objects (NEOs), particularly asteroids. Each feature is carefully curated and serves a unique purpose in the analysis. These features include various identifiers such as SPK-ID and Object ID, as well as designations like primary and IAU names. Key attributes such as the orbit class, NEO (Near-Earth Object) flag, and PHA (Potentially Hazardous Asteroid) flag offer crucial insights into an object's nature and potential danger. Characteristics like absolute magnitude (H), diameter, albedo, and associated uncertainties are pivotal in understanding an object's size and reflectivity. Orbital parameters, including eccentricity, semi-major axis, perihelion distance, inclination, and others, provide essential data for analyzing an asteroid's path through space. Moreover, features like mean motion, time of perihelion passage, and orbital period add valuable information about an object's orbit. The dataset offers a comprehensive foundation for conducting in-depth analyses and modeling related to these celestial bodies, with applications in space research and planetary defense.

V. Results & Discussions

A. Dataset Overview

The dataset comprises 958,524 rows and 40 attributes. During the preprocessing stage, various steps were taken to refine and enhance the dataset for analysis.

B. Attribute Evaluation and Correlation

Several columns were assessed, leading to the removal of the Root Mean Square (RMS) column due to its lack of meaningful information. A high correlation of approximately 0.95 was observed between Moid and Moid_ld, resulting in the decision to retain Moid and discard Moid_ld.

Furthermore, the examination of 'tp' and 'tp_cal' columns revealed their shared significance as both represented the "Time of Perihelion Passage (TDB)." It was confirmed that they were equivalent, and 'tp' was retained, while 'tp_cal' was dropped.

C. Handling Missing Data

Significant missing data was associated with the "pha" target variable, with 19921 out of 19922 rows having missed "pha" data. These rows were removed, rendering these observations unusable.

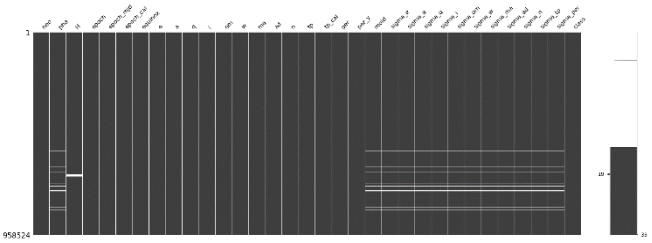

Fig. 1. Missing Data Matrix.

For other missing data, primarily in the "H" columns, imputation was employed using the mean value. For remaining sporadic missing data, these observations were dropped due to their insignificance relative to the dataset's size.

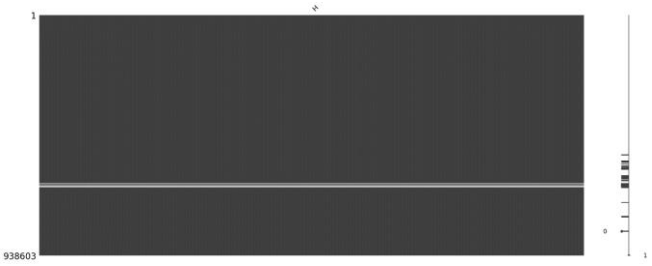

Fig. 2. Missing Data H Column

D. Observations From Column Characteristics

Descriptive statistics unveiled various characteristics of the data. For instance, the "sigma_tp" column displayed the largest standard deviation and the highest maximum value.

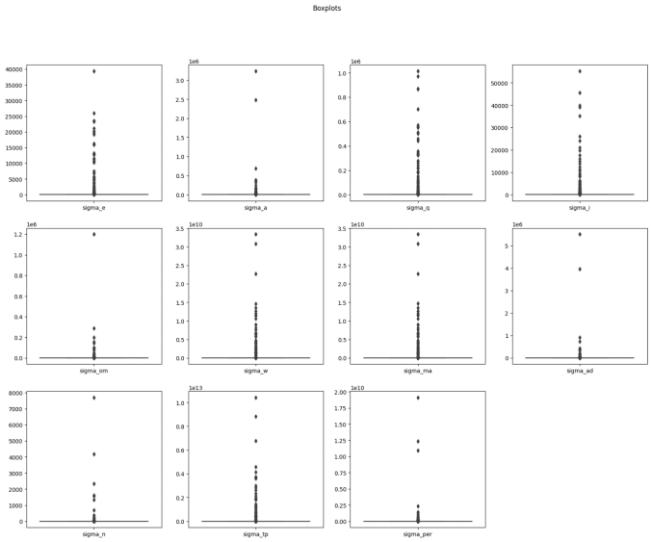

Fig. 3. Boxplot of Sigma Columns

The "H" (Absolute Magnitude) column had negative values, indicating a potential issue with data accuracy. Additionally, "epoch," "epoch_mjd," "epoch_cal," "e," "a," "q," and "i" columns demonstrated non-normal distributions, each with their unique shape and the presence of outliers.

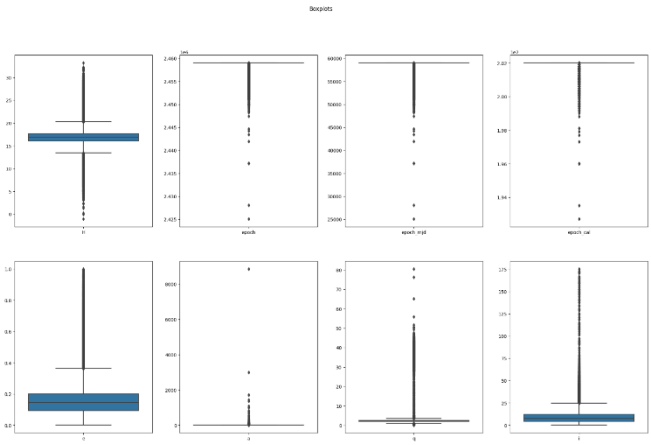

Fig. 4. Boxplot of Above given columns

E. Corrrelation & Feature Significance

Analysis revealed that the Longitude of the Ascending Node (om), Argument of Perihelium (w), and Mean Anomaly (ma) had similar distributions. On the feature importance front, Earth Minimum Orbit Intersection Distance (moid), Absolute Magnitude (H), and Epoch were found to be the most and least influential features, respectively, for classification according to XGBoost.

F. Categorical Variables and Relationships

Chi-squared tests indicated significant relationships between the 'class' and 'neo' attributes with the target attribute, with p-values below the significance level of 0.05.

G. Orbits And Distance Analysis

Eccentricity values were observed to significantly differentiate between potentially dangerous and non-

dangerous asteroids, with potentially dangerous asteroids having an average eccentricity of 0.879, while non-dangerous asteroids had an average eccentricity of 0.468.

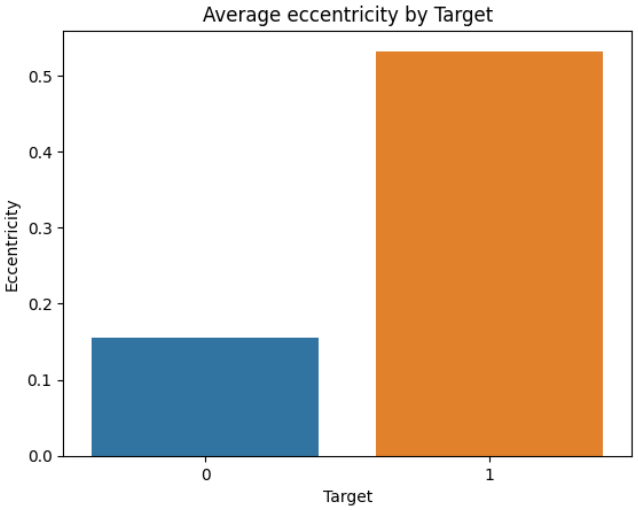

Fig. 5. Average Eccentricity by Target Barplot

Further analysis of Aphelium and Perihelium distances highlighted distinct characteristics between potentially dangerous and non-dangerous asteroids. Potentially dangerous asteroids had lower average Perihelium distances and higher Aphelium distances, indicating orbits that brought them closer to the sun.

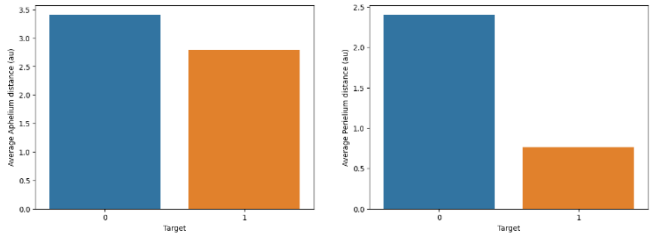

Fig. 6. Barplot of Aphelium and Perielium Distance by target class

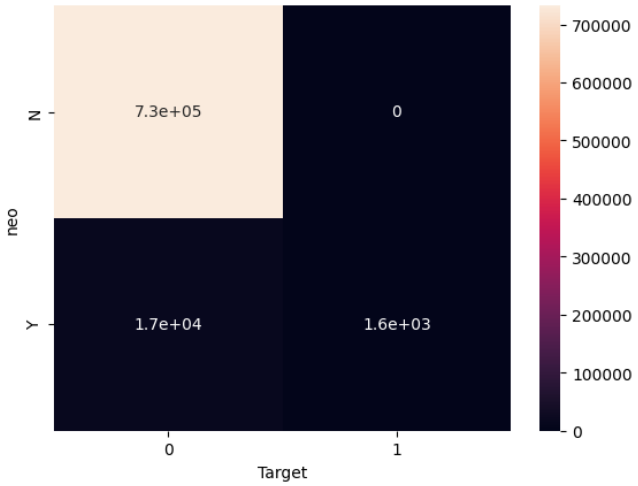

Fig. 7. Neo Flag Plot

H. Feature Selection And Model Scalling

Feature selection using Recursive Feature Elimination identified 30 important features for the XGBoost algorithm.

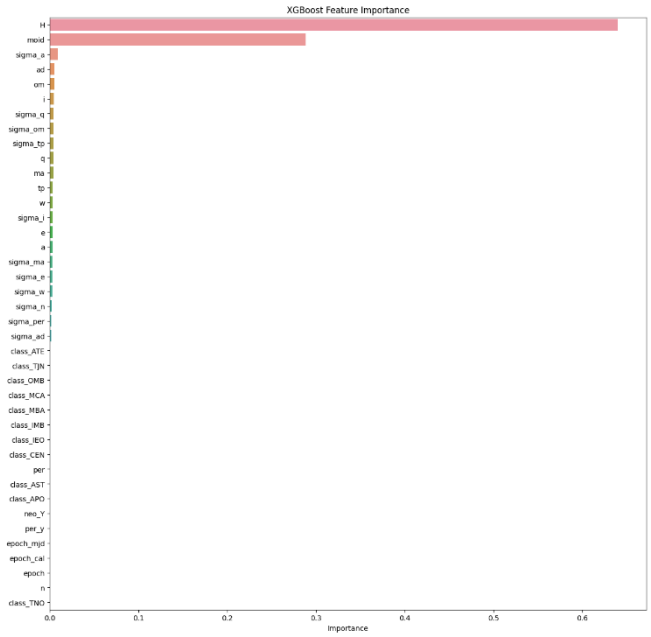

Fig. 8. Feature Selection Using XGBoost Algorithm

Three scaling techniques were employed:

- Standard Scaler: Achieved a mean accuracy of 0.9844 for RandomForest, 0.8277 for ExtraTree, 0.9844 for XGBoost, and 0.6408 for MLP.

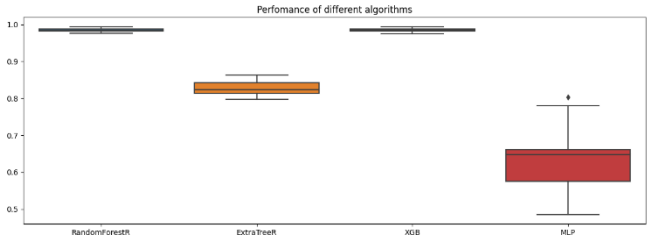

Fig. 9. Standard Scalar Model Evaluation

- Min-Max Scaler: Achieved mean accuracy scores of 0.9835 for RandomForest, 0.7997 for ExtraTree, 0.9856 for XGBoost, and 0.3480 for MLP.

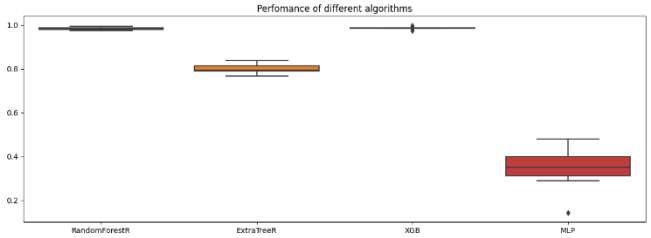

Fig. 10. Min Max Scalar Model Evaluation

- Robust Scaler: Achieved mean accuracy scores of 0.9848 for RandomForest, 0.8232 for ExtraTree, 0.9856 for XGBoost, and 0.3831 for MLP.

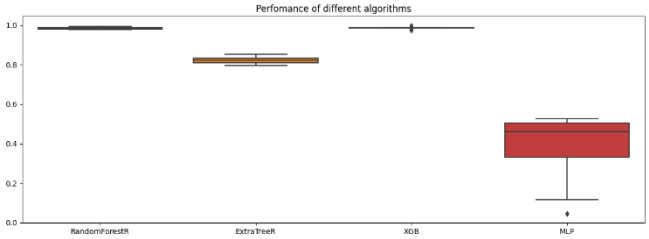

Fig. 11. Robust Scalar Model Evaluation

I. Model Fine Tuning and Evaluation

The XGBoost algorithm was fine-tuned with optimal parameters.

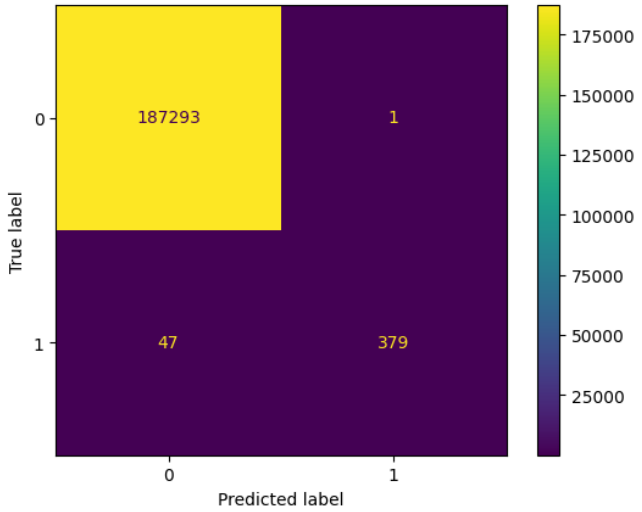

Fig. 12. Confusion Matrix

- *TP*: Number of Positive observations that was correclty predicted (379).
- *TN*: Number of Negative observations that was correctly predicted (187293).
- *FP*: Number of Negative observations that was predicted as positive (1).
- *FN*: Number of Positive observations that was predicted as negative (47).

The model's classification performance demonstrated high accuracy (99.97%), F1-score (0.94), precision (99.74%), recall (88.97%), and an exceptional ROC-AUC score of approximately 0.9999.

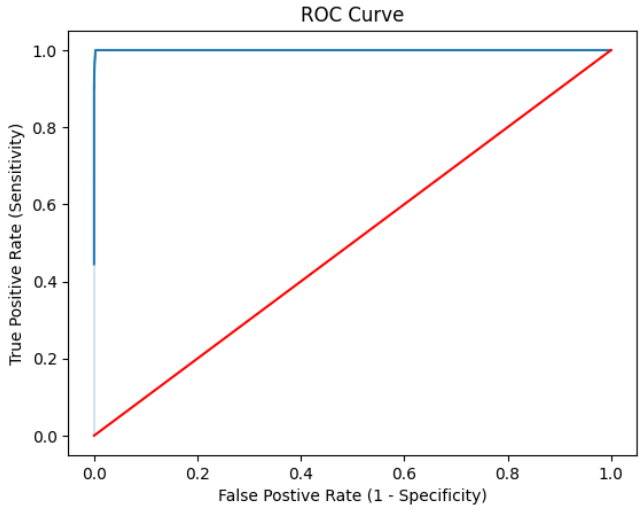

Fig. 13. ROC Curve

VI. CONCLUSION

In conclusion, this research embarked on a comprehensive analysis of an extensive dataset containing information about asteroids, particularly focusing on the identification of potentially hazardous asteroids. The primary aim was to develop a robust machine learning model capable of precisely classifying these celestial bodies while unearthing significant insights into their defining characteristics. Throughout the course of this study, several pivotal revelations emerged. The initial data exploration phase revealed a large dataset of 958,524 observations across 40 columns, necessitating thorough treatment of missing data and non-contributing features to prepare the dataset for modeling. Implementing varied techniques for feature selection and data transformation ultimately led to the adoption of the robust scaler, significantly impacting the model's performance. Constructing machine learning models, notably the Random Forest and XGBoost, demonstrated exceptional accuracy in discerning potentially hazardous from non-hazardous asteroids. While attempts to refine the model included experimentation with different data transformation techniques, model tuning, and the application of the Synthetic Minority Over Sampling Technique for Nominal and Continuous variables, SMOTENC did not significantly enhance the model's performance. The final model choice, XGBoost coupled with the robust scaler, yielded outstanding results. Notable insights gleaned from this endeavor revealed that potentially hazardous asteroids typically exhibit higher eccentricity, indicating more elliptical orbits, lower perihelium distances bringing them closer to the Sun, and a correlation between higher absolute magnitude and hazardous classification. Looking ahead, potential future avenues for this project encompass continued feature engineering, with the goal of identifying attributes to enhance model performance, focusing on improving recall without compromising precision, and establishing real-time monitoring systems for asteroid characteristics. This undertaking achieved its goals by developing an immensely accurate machine learning model for identifying potentially hazardous asteroids while providing crucial insights that could significantly impact the realm of planetary defense and space research. Through ongoing refinement and vigilance, this project stands poised to contribute to Earth's safety by aiding in the detection and mitigation of potential cosmic threats.

REFERENCES

[1] "Asteroid Spectral Types," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Asteroid_spectral_types.

[2] Massachusetts Institute of Technology, "Asteroid Spectrum Classification Using Bus-DeMeo Taxonomy," [Online]. Available: http://smass.mit.edu/busdemeoclass.html.

[3] "Research Note," Astronomy & Astrophysics (A&A), [Online]. Available: https://www.aanda.org/articles/aa/full/2007/20/aa5064-06/aa5064-06.right.html.

[4] "Asteroid Spectral Types," Simple English Wikipedia, [Online]. Available: https://simple.wikipedia.org/wiki/Asteroid_spectral_types.

[5] "Spectroscopy of X-Type Asteroids," IOPscience, [Online]. Available: https://iopscience.iop.org/article/10.1086/424856/pdf.

[6] A. Milani, et al., "Multiple Solutions for Asteroid Orbits: Computational Procedure and Applications," [Online]. Available: https://www.researchgate.net/profile/A-Milani/publication/44392221_Solutions-for-asteroid-orbits-Computational-procedure-and-applications/links/550065120cf2d61f820d6f79/Solutions-for-asteroid-orbits-Computational-procedure-and-applications.pdf.

[7] A. Milani, "From Astrometry to Celestial Mechanics: Orbit Determination with Very Large Data Sets," [Online]. Available: https://copernico.dm.unipi.it/~milani/preprints/badhof.pdf.

[8] A. Milani, et al., "Unbiased Orbit Determination for the Next Generation Asteroid/Comet Surveys," [Online]. Available: http://adams.dm.unipi.it/~gronchi/PDF/acm05_milani.pdf.

[9] M. E. Sansaturio, A. Milani, and A. Valsecchi, "The Asteroid Identification Problem," [Online]. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/8F6515D97F4DA96CD787BD6DDC0DECAB/S0252921100072730a.pdf/div-class-title-the-asteroid-identification-problem-div.pdf.

[10] "Use of the Semilinear Method to Predict the Impact Corridor for Near-Earth Objects," Springer, [Online]. Available: https://link.springer.com/article/10.1007/s10569-020-09959-3.

[11] "Near-Earth Observations (NEO) Program," NASA Science, [Online]. Available: https://science.nasa.gov/planetarydefense-neoo.

[12] "The NEOWISE Project," Caltech, [Online]. Available: https://neowise.ipac.caltech.edu/news/.

[13] "Pan-STARRS," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Pan-STARRS.

[14] "Near-Earth Observations (NEO) Program," NASA, [Online]. Available: https://www.nasa.gov/planetary-defense-neoo/.

[15] "Predicting Asteroid Types: Importance of Individual and Ensemble Models," Frontiers, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fspas.2021.767885/full.

[16] "Deep Convolutional Neural Networks with Transfer Learning for Asteroid Taxonomy," Springer, [Online]. Available: https://link.springer.com/article/10.1007/s00138-020-01069-2.

[17] "Asteroid Families Classification: Exploiting Very Large Data Sets," arXiv.org, [Online]. Available: https://arxiv.org/pdf/1312.7702.pdf.

[18] "Machine Learning Applied to Asteroid Dynamics," arXiv.org, [Online]. Available: https://arxiv.org/abs/2110.06611.

[19] "SS-RNN: A Strengthened Skip Algorithm for Data Classification Based on Spectroscopy of Asteroids," Frontiers, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2021.746181/full.)