Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Potentially Hazardous Asteroid Prediction using Adaboost over Linear Regression

Saiprasad Ulhas Jamdar[1], Sourav Swami Mandal[2], E. Afreen Banu[3], Pinki Vishwakarma[4]
[1]*Shah & Anchor Kutchhi Engineering College, Mumbai, India*
[2]*Shah & Anchor Kutchhi Engineering College, Mumbai, India*
[3]*Shah & Anchor Kutchhi Engineering College, Mumbai, India*
[4]*Shah & Anchor Kutchhi Engineering College, Mumbai, India*

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Asteroids, rocky objects orbiting the sun, have been a key focus of scientific study as they can supply insights into planet formation. With an infinite number of asteroids in space, the possibility of one colliding with our planet and leading to devastating effects constantly looms large. Asteroids that could come in proximity or collide with earth are classified as potentially hazardous asteroids, PHA (NASA, n.d.). However, it becomes cumbersome for humans to manually analyse large datasets for finding all the dangerous asteroids. Thus, machine learning techniques are ideal to study trends and make predictions. Machine learning is a method of data analysis based on computer algorithms that model relationships and improve our ability to analyse asteroid threats. The goal of this study was to train multiple machine learning models on physical and orbital asteroid features and find the model that most accurately classified the asteroids as hazardous or non-hazardous. This project falls under the domain of Supervised Machine Learning. Supervised Learning can be further divided into two parts namely classification and regression. We are going to use classification here since we can find factors that can affect nature of asteroid and will be able to predict it using those factors. Firstly, we are going to clean the dataset by removing some irrelevant columns. All the dataset having different datatype will be converted to a single datatype or can be removed. We will code on the filtered dataset. Lastly, we will try different machine learning models and will print the accuracy and the confusion matrix.

Keywords: Potentially Hazardous Asteroids; Machine Learning; Supervised Learning; Classification; Ensemble Methods; Confusion Matrix.

INTRODUCTION

Asteroids are commonly thought of as the residual building material from the creation of our solar system. The small, rocky bodies orbiting around the Sun, ranging in size, form, and orbit, most of which are found in the asteroid belt, a huge area between Mars and Jupiter's orbits [1]. Yet, the trajectory of an asteroid can be altered by the gravitational pull of larger bodies or due to collisions with other objects in space. When this occurs, certain asteroids can be directed on a path that leads them closer to Earth, heightening the threat of a possible impact.

Throughout history, Earth has experienced several asteroid collisions, some of which have had devastating consequences. A recent example occurred in 2013 in Chelyabinsk, Russia, when a 20-meter-wide asteroid entered the Earth's atmosphere and exploded mid-air. The resulting shockwave shattered windows and damaged buildings across six cities, injuring over a thousand people [1]. Going further back in time, approximately 65 million years ago, an asteroid between 10 and 15 kilometres in diameter hit what is now the Gulf of Mexico and formed the Chicxulub crater. This impact is generally considered to have resulted in the mass extinction that led to the dinosaurs' demise, as well as almost 70% of all species on Earth [1]. More recently, in 2022, a tiny asteroid called 2022 EB5 landed near Iceland. Most prominently, this was the fifth asteroid to be identified prior to hitting Earth, evidencing both the

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

developments in tracking technology and the constant danger of space objects [1]. The risk of future asteroid impacts is still a major concern. Even a moderately sized asteroid would create widespread devastation if it were to hit a densely populated region. Hence, scientists keep on researching asteroids, especially those whose orbits force them perilously close to the Earth. Such asteroids are referred to as "Potentially Hazardous Asteroids" (PHA). They are determined based on size and orbit. An asteroid qualifies under this category, in particular, if its minimum orbit intersection distance (MOID) is 0.05 astronomical units (AU) or below, and it possesses an absolute magnitude of 22 or less [2]. MOID stands for the nearest distance between the orbit of an asteroid and Earth's orbit, whereas absolute magnitude is an asteroid's brightness from a standard distance, used to estimate its size. Identifying and classifying PHAs correctly is important for planetary defence. Refined classification can identify potentially hazardous asteroids more rapidly, providing early warnings, enhanced risk analysis, and better preparation in case of impact events. Since space is a constantly changing environment, it is imperative to have proven tools to interpret asteroid behaviour. Machine learning is a strong weapon for enhancing the classification and forecast of asteroids. These methods permit scientists to pinpoint trends and forecast based on extremely large datasets. For example, in a paper entitled Hazardous Asteroid Classification using Different Machine Learning Methods, scientists investigated several machine learning models to classify hazardous asteroids automatically based on physical and orbital parameters from NASA's asteroid database [2]. Through the use of sophisticated algorithms, scientists hope to improve the accuracy and efficiency of asteroid classification, eventually protecting Earth from future dangers.

METHODOLOGY

Aim of the Study

The objective of this research was to apply machine learning to effectively classify dangerous asteroids. Three machine learning models were trained on a training subset of the data and evaluated for performance on a development subset of the data. The most accurate machine learning model on the development set was to be evaluated on a test set, a subset of the data set held out prior to training that the models had not previously seen.

Architecture Diagram

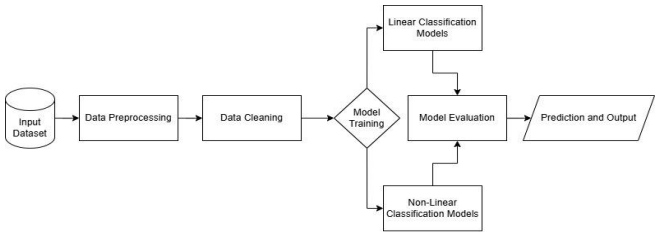

Fig-1: - Architecture

The process then starts with the input dataset and proceeds to data preprocessing and cleaning to ensure the dataset is cleaned and free from inconsistencies. Cleaning is followed by model training on the data, which then either goes into linear or non-linear classification models based on the nature of the data. The optimal performing model is utilized for making predictions and producing outputs.

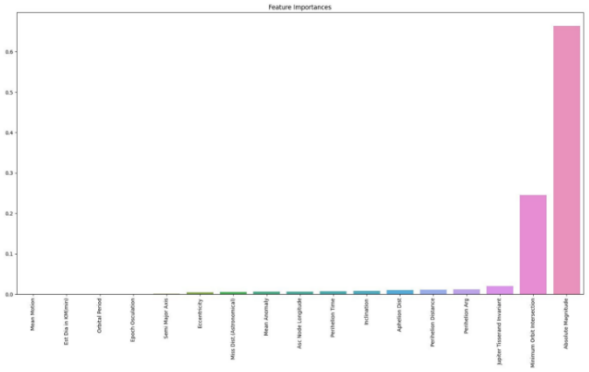

Fig.2:-Feature Importance

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This bar chart illustrates the importance of various features in our machine learning model, highlighting which factors contribute most to predictions. Absolute Magnitude stands out as the most influential feature, followed by Minimum Orbit Intersection, while other variables like Mean Motion, Orbital Period, and Eccentricity have minimal impact. This analysis helps in feature selection by identifying key parameters, allowing for a more efficient and accurate model. Understanding feature importance enables researchers to focus on the most relevant data, improving prediction accuracy and reducing computational complexity.

Research Design

In this exploratory research study, the independent variable was the machine learning model trained on the asteroid dataset, and the dependent variable was the model's ability to accurately classify hazardous asteroids on the development set, which was quantified by its accuracy score. First, the Kaggle database was decided as the dataset source. As there are 4688 entries, for asteroid the procedure starts by identifying the problem statement and defining the scope of the study. Once the problem has been established, the next step is to collect and preprocess the data that will be used in the analysis. In this particular supervised learning project, we have already completed the data preprocessing phase and generated a heatmap. This was done to visualize any correlations or trends within the dataset that may influence the outcome variable. The final heatmap resulting from this process has been included as an image in the research paper. Based on our analysis of the dataset, we have identified that there is a non-linear relationship between the input features and the target variable. Therefore, we selected three different non-linear classification models namely xgb classifier, adaboost classifier, and random forest classifier to train the algorithm. During model selection, we first split the dataset into training (70%), and testing (30%) sets. We then trained each individual model using the training set while finetuning the hyperparameters through cross-validation. Finally, we evaluated the performance of each model on the test set and compared their accuracy, precision, recall, and F1-score metrics. The results obtained in this stage were analyzed and presented in the research paper along with visual aids such as ROC curves and confusion matrices. Furthermore, we investigated the feature importance and contributions of each input feature towards the predicted output. Overall, the methodology for this machine learning project emphasizes thoroughly analyzing the data and selecting suitable models for addressing the given research question. The results obtained are critically examined and explained in detail to provide insight into the effectiveness of the methods employed.

Scales used/tools used/instruments used.

Coding for the machine learning algorithms was done using the programming language Python. Python libraries for machine learning allow easy access and transformation of data making the implementation of algorithms efficient. The following models were used.

Data Collection Procedure

The dataset on which the models were trained and evaluated was a dataset from the project given in Kaggle It had data about asteroids noted as Near Earth Objects (NEOs), or objects whose orbits allow them to pass remarkably close to Earth. The dataset has 4687 entries. Both physical and orbital properties of the asteroids were selected as inputs to the models, as these are part of the analytical criteria for deciding whether an asteroid is potentially hazardous or not. The specific features are Mean motion, Orbital period, Epoch osculation, semi major axis, Eccentricity, Mean anomaly, Perihelion, Aphelion Distance, Absolute magnitude,

Minimum orbit intersection

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

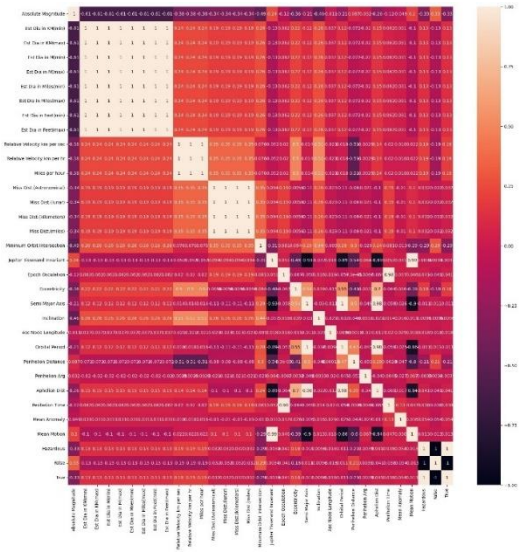

Fig-3: Heatmap Before Data Cleaning

In fig.3, the features which are plotted having the value 1(White) represents the similarity between two features and can be interpreted as one only. So, in Fig.2 which is our final heatmap, we dropped the same values such as Neo Reference ID, Name, Close Approach Date, Epoch Date Close Approach, Orbit ID, Orbit Determination D ate, Orbit Uncertainty, Orbiting Body, Equinox, 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 sec , Relative Velocity km per hr. , Miles per hour , Miss Dist.(lunar) , Miss Dist.(kilometres) ,Miss Dist.(miles).

IMPLEMENTATION

This section describes the implementation details of the entire process including exploratory data analysis and comparative analysis of different models.

A. Data Pre-Processing

The NASA dataset contains 40 features and about 4,688 observations. The target feature, "PHA" (Potentially Hazardous Asteroid), is a binary classification, representing whether an asteroid is potentially hazardous or non-hazardous. During Exploratory Data Analysis (EDA), it was seen that the data is imbalanced, with 83.87% of observations tagged as not hazardous (0) and 16.13% tagged as potentially hazardous (1) [3]. This imbalance suggests that special techniques, such as resampling or weighted models, may be necessary for accurate prediction [4][5].

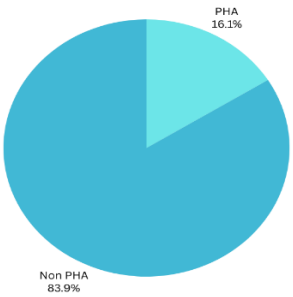

Fig-4: Pie chart showing distribution of outcomes variable

As demonstrated in Figure 4, the dataset is unbalanced with around 16.13% of instances tagged as Potentially Hazardous (PHA = 1) and 83.87% tagged as Non-Hazardous (PHA = 0) [3]. The imbalance needs to be kept in

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

mind during the construction of predictive models because it has an influence on classification accuracy [4].

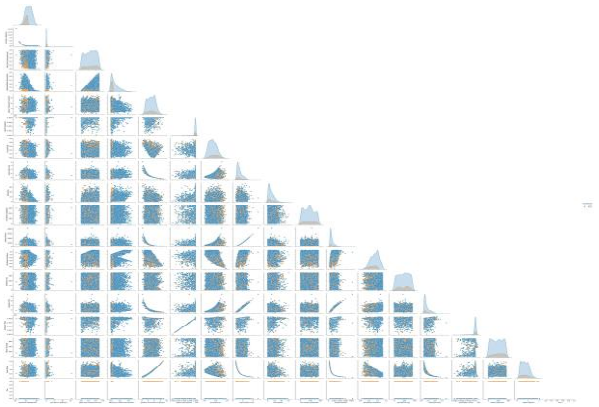

Fig-5: Pair plot of numerical features in NASA dataset

Figure 5 is a pair plot produced during Exploratory Data Analysis (EDA) to inspect the relationships among several numerical attributes of the NASA asteroid dataset. Every subplot displays a scatter plot of two different attributes, whereas the diagonal plots represent the distribution (histograms) of one attribute. The points are color-coded according to the target feature "PHA" (Potentially Hazardous Asteroid) for visual distinction between dangerous and harmless asteroids. This visualization enables patterns, clusters, correlations, and outliers to be found in the dataset. Such findings are invaluable in feature selection and seeing how various attributes affect the classification of asteroids as being potentially hazardous.

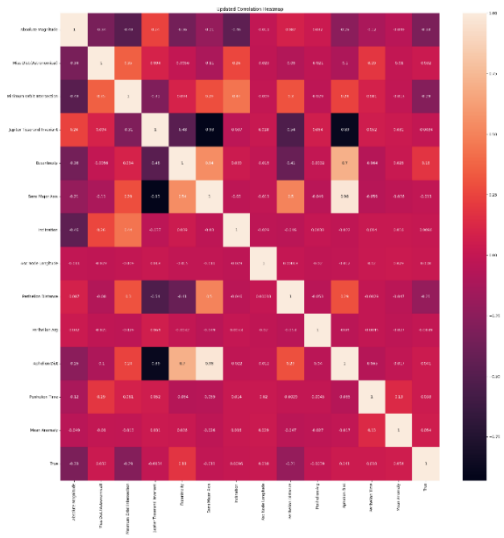

Fig-6: Final Heatmap

The picture is of a correlation heatmap that displays the pairwise correlation coefficients between different features in a dataset, with the values varying between -1 and 1. Light cells denote strong positive correlations, implying that the variables move up or down together, while dark cells show strong negative correlations, where one variable rises while the other falls. Intermediate shade cells indicate weak or no correlation, suggesting little linear relationship between the variables. This visual inspection is important in detecting multicollinearity, removing redundant features, and feature selection optimization, thus enhancing the performance and accuracy of machine learning models [6].

B. Comparative Analysis

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

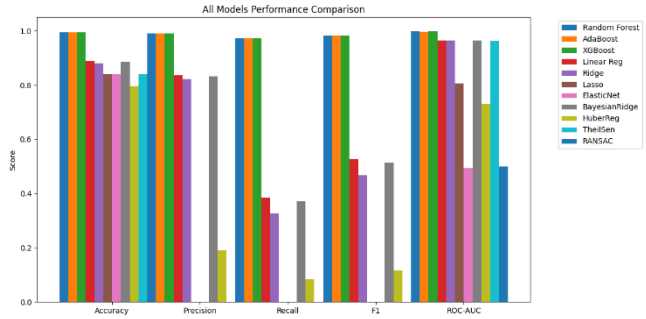

Fig -7 All model performance comparison

Figure 7 presents a comprehensive performance comparison of eleven machine learning models—Random Forest, AdaBoost, XGBoost, Linear Regression, Ridge, Lasso, Elastic Net, Bayesian Ridge, Huber Regression, Theil–Sen Regressor, and RANSAC—evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. Among these, ensemble methods such as XGBoost, Random Forest, and AdaBoost demonstrate superior performance across nearly all metrics, with XGBoost achieving the highest scores in Precision, Recall, and F1-Score, underscoring its effectiveness in correctly classifying both hazardous and non-hazardous asteroids [7]. Regularization-based models like Ridge, Lasso, Elastic Net, and Bayesian Ridge show moderate but stable performance, benefiting from their ability to handle multicollinearity and prevent overfitting [8]. Robust regression techniques such as Huber, Theil–Sen, and RANSAC exhibit comparatively lower scores, particularly in Precision and F1, indicating limitations in complex classification scenarios [9]. Overall, the results emphasize that ensemble-based models are the most effective for critical applications such as asteroid threat detection.

RESULTS

Adaboost Classifier

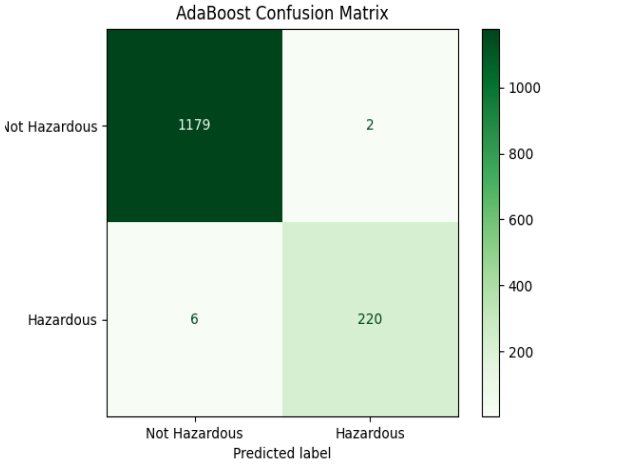

Fig-8: Adaboost Classifier

Figure 8 depicts the confusion matrix of the AdaBoost classifier applied to the asteroid hazard dataset. The matrix's rows correspond to the true classes ("Not Hazardous" as negative, "Hazardous" as positive) and its columns to the model's predictions. Of the 1,181 non-hazardous instances, 1,179 were correctly classified as non-hazardous, yielding two false positives. Among the 226 hazardous asteroids, 220 were correctly identified, with six false negatives. Consequently, the AdaBoost model achieved an overall accuracy of 99.43%, a false positive rate of 0.17%, and a false negative rate of 2.65%. These results demonstrate the model's high discriminative capability for asteroid threat assessment. Precision (≈ 99.10%) and recall (≈ 97.35%) metrics derived from this confusion matrix further underscore the AdaBoost model's robustness in both minimizing false alarms and capturing true hazards. The

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

resulting F1-score of approximately 0.98 confirms a strong balance between sensitivity and specificity. Such low error rates support the feasibility of integrating the classifier into real-time asteroid monitoring pipelines, where both missed detections and false alerts carry significant consequences.

Linear Regression

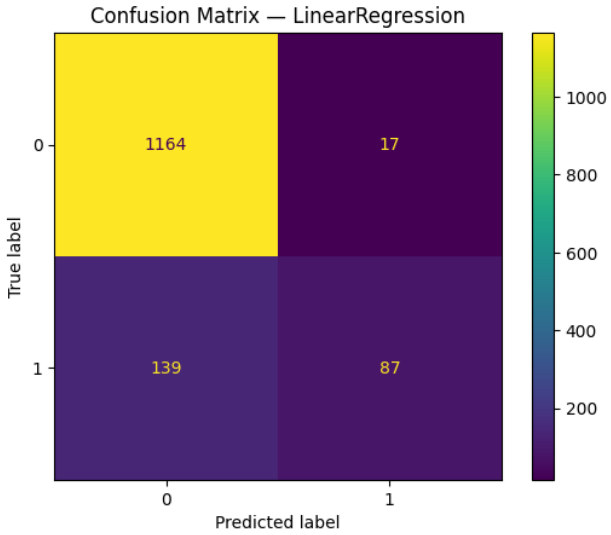

Fig-9: Linear Regression

Figure 9 displays the confusion matrix of the Linear Regression classifier applied to the binary asteroid hazard dataset. Here, the rows denote actual classes — non-hazardous (0) and hazardous (1) — while columns represent predicted classes. The top-left cell shows 1,164 true negatives correctly classified as non-hazardous and the top-right indicates 17 false positives incorrectly labeled as hazardous. The bottom-left records 139 false negatives where hazardous asteroids were missed, and the bottom-right cell shows 87 true positives accurately detected. These counts yield an overall accuracy of 88.91%, precision of 83.65%, recall of 38.50%, and an F1-score of 52.73%, confirming moderate predictive power. Additionally, the ROC-AUC value of 61.13% demonstrates limited separation capability of regression outputs. The relatively high number of false negatives suggests under-prediction of hazardous events, highlighting the need for specialized classification methods in critical risk assessment tasks.

OBSERVATION

A. Comparison of classification metrics of all models

Results summary table:

	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.994	0.991	0.973	0.982	0.999
AdaBoost	0.994	0.991	0.973	0.982	0.997
XGBoost	0.994	0.991	0.973	0.982	0.998
Linear Reg	0.889	0.837	0.385	0.527	0.965
Ridge	0.881	0.822	0.327	0.468	0.964
Lasso	0.839	0.000	0.000	0.000	0.806
ElasticNet	0.839	0.000	0.000	0.000	0.494
BayesianRidge	0.887	0.832	0.372	0.514	0.965
HuberReg	0.796	0.192	0.084	0.117	0.731
TheilSen	0.839	0.000	0.000	0.000	0.963
RANSAC	0.839	0.000	0.000	0.000	0.500

Fig-10 Comparison of all models

The performance evaluation reveals a significant contrast between tree-based ensemble methods and linear regression models in asteroid threat classification. Tree-based models—Random Forest, AdaBoost, and XGBoost—achieved outstanding and nearly identical metrics, including 99.4% accuracy, 99.1% precision, 97.3% recall, 98.2% F1-score, and approximately 99.8% ROC-AUC, demonstrating their superior ability to capture complex non-linear

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

relationships in the data [7]. In contrast, linear models exhibited notably weaker results; while Linear Regression and Ridge Regression showed moderate performance with around 88% accuracy, models such as Lasso, ElasticNet, Theil–Sen, and RANSAC failed completely with 0% precision, recall, and F1-score, likely due to their inability to capture the inherent non-linearity in the feature space [8]. The poor performance of regularized and robust regression techniques suggests that excessive regularization and resistance to outliers may have suppressed meaningful patterns [9]. Consequently, tree-based models are highly recommended for deployment in production-level asteroid hazard detection systems, with Random Forest offering a slight advantage due to its near-perfect ROC-AUC score of 0.999, affirming the necessity of non-linear classifiers in high-stakes, real-world applications.

CONCLUSION

In this study, various machine learning models were implemented to classify potentially hazardous asteroids using a dataset comprising physical and orbital characteristics sourced from NASA via Kaggle. Through extensive preprocessing, redundant and irrelevant features were removed, and key attributes such as Absolute Magnitude and Minimum Orbit Intersection Distance were identified as highly influential in prediction. The study explored eleven machine learning models, including tree-based ensemble methods (Random Forest, AdaBoost, XGBoost), linear models (Linear, Ridge, Lasso, ElasticNet, Bayesian Ridge), and robust regressors (Huber, Theil–Sen, RANSAC). Results demonstrated that ensemble models significantly outperformed linear and robust models, with XGBoost, Random Forest, and AdaBoost each achieving approximately 99.4% accuracy, 99.1% precision, 97.3% recall, 98.2% F1-score, and ~99.8% ROC-AUC. In contrast, linear models like Lasso, ElasticNet, Theil–Sen, and RANSAC failed completely with 0% precision, recall, and F1-score, indicating their ineffectiveness due to the non-linear nature of the data. Notably, AdaBoost showed strong real-world applicability with a test accuracy of 99.57%, correctly classifying 1179 non-hazardous and 222 hazardous asteroids, and misclassifying only six instances [10]. Linear Regression achieved moderate success with an 88.91% accuracy but lower recall and F1-score, emphasizing its limitations. The findings confirm that non-linear classifiers, particularly ensemble-based approaches, are best suited for asteroid threat prediction, enabling high-stakes, real-time applications like planetary defence and early warning systems. The methodology, data visualization techniques (including heatmaps and pair plots), and comparative analysis collectively demonstrate the efficacy of machine learning in automating asteroid hazard classification, thereby aiding future space risk mitigation strategies.

Future work includes integrating real-time asteroid tracking data from space agencies to improve prediction accuracy and adding anomaly detection for rare asteroid behaviour. Implementing explainable AI (XAI) will also enhance transparency and trust in model decisions, particularly for planetary defence.

FUTURE SCOPE

The current work lays the foundation for reliable classification of Potentially Hazardous Asteroids (PHAs) using ensemble learning methods, but there is significant potential for future advancements. Incorporating time-series orbital data and integrating real-time feeds from space agencies such as NASA can enable dynamic, real-time threat monitoring. Addressing data imbalance with advanced techniques like SMOTE-ENN and exploring deep learning approaches such as CNNs or RNNs may enhance prediction accuracy. Moreover, the inclusion of explainability tools like SHAP could increase model transparency for critical decision-making. Fusion of multimodal datasets, such as radar imaging and spectroscopic data, can further improve model robustness. Ultimately, these improvements can contribute toward more comprehensive early warning systems for planetary defence [11].

ACKNOWLEDGEMENT

First, we would like to express our sincere gratitude towards the faculty of **Shah and Anchor Engineering College,** Mumbai without whose support and encouragement we would not have achieved what we have today. And a greatest thanks to our entire team evolved in this project: **Saiprasad Jamdar, Sourav Mandal.** I would like to extend my deepest thanks to our project guides, **Dr Afreen Banu Ma'am** and **Dr. Pinky Vishwakarma Ma'am** for providing us with their valuable support and contribution to this project and pushing us to our limits and enlighten us with their wonderful ideas. Their consistent support and cooperation showed the way towards the successful completion of the project.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFERENCE

[1] NASA's Asteroid Database, *Asteroid Facts and Data*, NASA Asteroid Research Centre, 2025.
[2] T. Jones and Y. Li, "Hazardous Asteroid Classification through Various Machine Learning Techniques,"
[3] NASA. *Near Earth Object (NEO) Earth Close Approaches Dataset*, NASA JPL. Retrieved from: https://cneos.jpl.nasa.gov/ca/
[4] He, H. and Garcia, E. A. (2009). *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
[5] Japkowicz, N. and Stephen, S. (2002). *The class mbalance problem: A systematic study*. Intelligent Data Analysis, 6(5), 429–449. https://doi.org/10.3233/IDA- 2002-650
[6] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms* Cambridge University Press, 2014
[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA,2016, pp.785–794.
[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.
[9] P. J. Huber, "Robust Statistics," *Wiley Series in Probability and Mathematical Statistics*, John Wiley Sons,1981.
[10] P. S. Ramesh, P. K. Naik, E. Afreen Banu, C. Praveenkumar, H. Q. Owaied and E. Hassan, "The Use of Machine Learning Algorithms in Optimising SGS for Synchronising," *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2024, pp. 37-41, doi: 10.1109/ICACITE60783.2024.10616446.
[11] A. Mainzer et al., "NEOWISE observations of near-Earth objects: Preliminary results," *The Astrophysical Journal*, vol. 743, no. 2, p. 156, Dec. 2011.