Machine Learning Application to Classify Asteroids Based on Orbital Parameters

Article *in* Journal of Physics Conference Series · October 2024

DOI: 10.1088/1742-6596/2866/1/012047

Janaka Adassuriya
University of Colombo
42 PUBLICATIONS **60** CITATIONS

Shafiyyah Azzahra
Politeknik Kesehatan Makassar
1 PUBLICATION **0** CITATIONS

Trismidianto ,-
National Research and Innovation Agency
54 PUBLICATIONS **265** CITATIONS

Journal of Physics: Conference Series

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

Machine Learning Application to Classify Asteroids Based on Orbital Parameters

To cite this article: M Delina *et al* 2024 *J. Phys.: Conf. Ser.* **2866** 012047

View the article online for updates and enhancements.

Machine Learning Application to Classify Asteroids Based on Orbital Parameters

M Delina[1*], J Adassuriya[2], S A Azzahra[1], A M Hussaan[3], Trismidianto[4]
[1]Department of Physics, Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta 13220, Indonesia
[2]Astronomy and Space Science Unit, Department of Physics Colombo University
Sri Lanka
[3]Department of Computer Science, IQRA University
Karachi, Karachi City, Sindh, Pakistan
[4]Research Center for Climate and Atmospheric, National Research and Innovation Agency
Bandung 40135, Indonesia
Email*:* *mutia_delina@unj.ac.id

Abstract. A machine learning application was developed to detect Potentially Hazardous Asteroid and mitigate asteroid collision risk with the Earth by applying three classifiers: the K-Nearest Neighbors, Naïve Bayes, and Random Forest. The study determined the most effective classifier for developing an asteroid classification program based on orbital motion. The machine learning classifier was then evaluated by its precision, accuracy, F1-score, and recall in determining Potentially Hazardous Asteroids and non-Potentially Hazardous Asteroids. The result presented Random Forest as the most appropriate classifier with the highest accuracy score of 99.53%, followed by the Naive Bayes classifier with an accuracy score of 92.00%, and the KNN classifier with an accuracy score of 84.45%. The study provided information on the most accurate machine learning classifier with the impact parameters for asteroid classification in an early warning system. By improving an embedded real-time detection system for Potentially Hazardous Asteroids, the study contributes to more effective strategies for mitigating the risk of asteroid impacts and enhancing planetary defence.

Keywords: asteroids, machine learning, Potentially Hazardous Asteroid, K-Nearest Neighbors, Naïve Bayes, and Random Forest

1. Introduction

Our solar system has at least 25,000 identified Near Earth Asteroids (NEAs) [1], and over 2,200 of them are Potentially Hazardous Asteroids (PHA) which have the potential to collide with the Earth [2]. An Asteroid is potentially hazardous if the Minimum Orbital Interaction Distance (MOID) with the Earth is similar to or less than 0.05 AU. Besides, its absolute magnitude is equal to or less than 22 or brighter (indicating a diameter of approximately 140 meters or larger). The collision between PHA and the Earth has occurred before and will happen again someday [3]. For example, in February 2013, a 20-meter diameter Chelyabinsk meteor fell in Russia, injuring over 1,500 people and destroying many properties [4]. The collision emphasizes the critical need for effective PHA detection or classification program to mitigate the risk of future impacts [2]. Regular parameter tracking and updates are necessary, as gravitational interactions with other celestial bodies can alter the asteroids' orbit over time. Correspondingly, the orbit alteration potentially increases or decreases the asteroids' impact risk with

Earth [5]. Therefore, the orbital motion shows the asteroid's trajectory and determines whether it could cross the Earth's orbit and pose a threat [3].

Machine learning provides some methodologies to enhance asteroid classification accuracy and reliability based on their orbital parameters. For example, the Quantum Machine Learning (QML) methodology conducted the asteroids' dataset's quantum properties to increase the asteroid's classification accuracy (98.11 percent), with an average F1-score of 92.69 percent [5]. Another study applied Super-Vector Machine (SVMs), gradient boosting, and Multi-Layer Perceptron (MLP) in a program's classifier. The study evaluated 11 surveys and space missions [6]. However, until today there was no explanation of the best acceptable classifier for identifying asteroids based on the orbital motion.

This research studied three machine learning classifiers: K-Nearest Neighbor (KNN), Random Forest, and Naïve Bayes. Here, the Random Forest combined several decision trees as a weak classifier [7][8]. It reduced a single tree error impact. Therefore, the classification stability was improved to process a final prediction [9]. In the classification task, the result was produced by the most trees [10]. On the other side, Naïve Bayes uses the Bayes rule as follows:

$$P(y|X) = \frac{P(y|X)P(y)}{P(x)} \tag{1}$$

where y and X are a class variable and a dependent feature vector with n-dimensions, respectively. The variable (y) has two outputs: PHA and non-PHA [11]. Naive Bayes at Eq. (1) is perceiving the label y probability, given an x input feature set. The Naïve Bayes algorithm assumes that input features x is independent of one another [12]. Meanwhile, the following step of KNNs' working mechanism include calculate the distance, discovering the closest neighbors, and labeling voting. k is the number of nearest neighbors in the KNN classifier, and it is a controlling variable for this method [13]. The classifiers may perform differently depending on how it interprets and processes the orbital data. By evaluating the performance of the classifiers in terms of precision, accuracy, F1-score, and recall, the study aims to identify the most effective approach for detecting PHAs and improving planetary defense strategies [14].

The study outcome contributed to the asteroid classification systems and enhanced risk mitigation to protect the Earth from the asteroid's collision.

2. Method

2.1. Datasets

The study collected 4,678 asteroid data through the National Aeronautics and Space Administration (NASA) open data portal. The asteroid dataset included 40 features, whereas only 13 features were related to the asteroid's orbital motion. Besides that, one target class was also applied to develop the asteroid classification program.

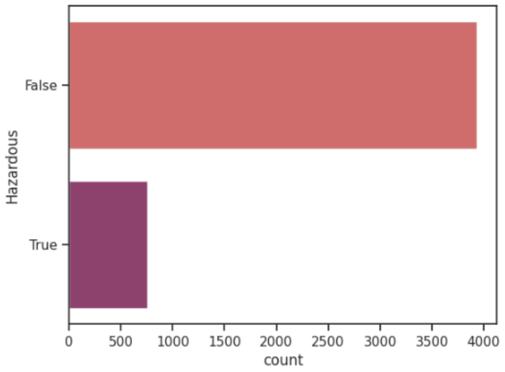

Figure 1. Data distribution

The difference between PHA and non-PHA data was big, as shown in Figure. 1, so the Synthetic Minority Over-Sampling Technique (SMOTE) method was applied to overcome the data imbalance issue.

By generating *n* synthetic samples, SMOTE increased the number of minority class instances [15]. Basically, the synthetic instance was created from the nearest neighbors and the minority instances. The instances x_i and $\widehat{x_\iota}$ were set to generate a synthetic instance as shown by the following formula.

$$x_{synthetic} = x_i + (\widehat{x_\iota} - x_i) \cdot \delta \tag{2}$$

where δ was a random number between 0 and 1, x_i was the minority class instance, and $\widehat{x_\iota}$ was a random number selected from the x_i nearest neighbor [16]. Based on the data distribution in Figure 1, the number of PHA class instances was smaller than the number of non-PHA class instances. Therefore, x_i and $\widehat{x_\iota}$ represented a PHA class instance and the nearest neighbor from the PHA class, respectively. Synthetic instances ($x_{synthetic}$) was generated by x_i and $\widehat{x_\iota}$ to balance the data points in each class.

2.2. Data Preparation

In the classification model development, the target label class was converted to a binary integer by a Label Encoder module. The module converted categorical columns into numerical columns [17]. Furthermore, Figure. 2 presented the data distribution after the SMOTE was applied to overcome the data imbalance. Here, SMOTE resampled 3932 PHA and 3932 non-PHA data.

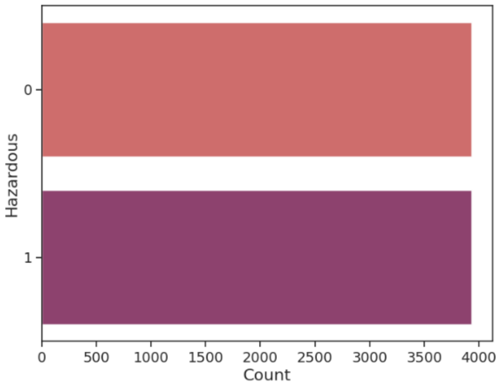

Figure 2. Data distribution after applying SMOTE

2.3. Model Training

The asteroid dataset was divided into a 7:3 ratio, 70% data training and 30% data testing. The study applied GridSearchCV to identify the hyperparameters of each model by evaluating all possible combinations based on cross-validation performance [18]. The optimal hyperparameters of KNN model were found to be: ‘metric’: ‘euclidean’, ‘n_neighbors’: ‘1’, and ‘weights’: ‘uniform’. The Random Forest model's optimal settings were ‘max_depth’: ‘None’, ‘min_samples_leaf’: ‘1’, ‘min_samples_split’: ‘5’, and ‘n_estimators’: ‘50’. The Naive Bayes model had a single hyperparameter, ‘var_smoothing’, set to ‘1×10^{-9}’. The model training phase was conducted using these selected hyperparameters to minimize error in prediction.

Once the training phase was complete, the training models were evaluated by classification metrics (such as precision score, recall score, F1-score, and accuracy score). The accuracy score was the ratio between the true positive prediction total and the total data. Meanwhile, the precision score represented the correctly predicted positive out of all the positive predictions. The precision score represented the classifier's accuracy in determining the hazardous asteroid possibility. The recall score was the actual number of hazardous asteroids that were identified by the classifier. F1-score is the mean of precision and recall [14]. The processes of the study are illustrated in Figure 3.

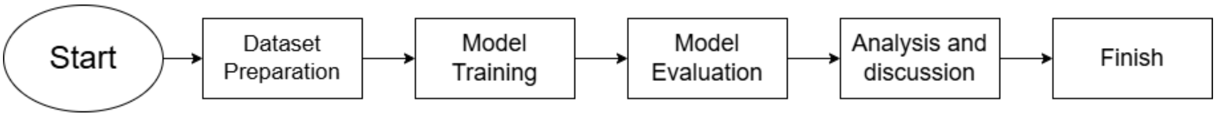

Figure 3. Research Flowchart

3. Result and Discussion

This study implemented an asteroid classification program to categorize PHA and non-PHA. Besides the study also determined the most appropriate machine learning classifiers (Random Forest, Naïve Bayes, and KNN) to classify the PHA. The classification metrics were applied to evaluate each classifier model. The models' evaluations result is as follow:

TABLE 1. Classification metrics of each method

	Accuracy	Precision	Recall	F1 Score
Random Forest	99.53%	100%	99.00%	100%
Naïve Bayes	92.00%	89.00%	95.00%	92.00%
KNN	84.45%	79.00%	94.00%	86.00%

Furthermore, the study applied the Receiver Operating Characteristic (ROC) curve to study each model performance. The ROC curve represented the true positive rate as opposed to the false positive rate for different classification thresholds in plot form. The ROC curve evaluated the model's capability to distinguish PHA (positive class) and non-PHA (negative class). The Area Under the Curve (AUC) represented the general model performance and measured the separability degree between each class [19].

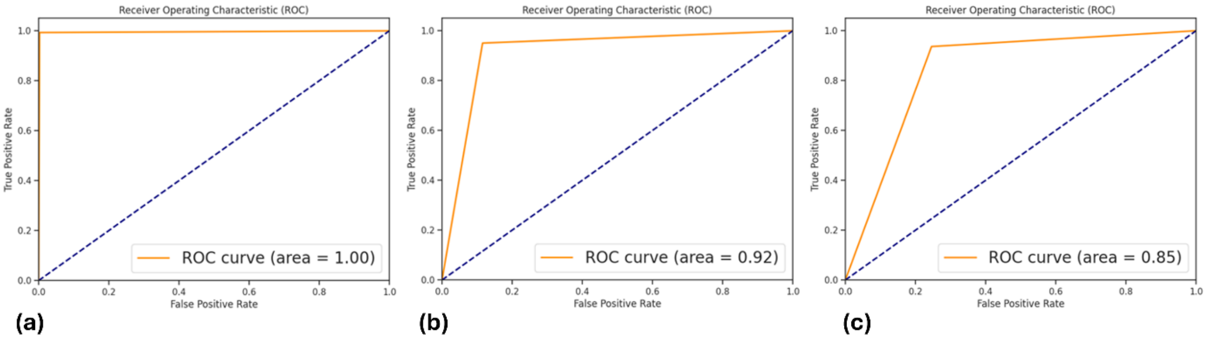

Figure 4. (a) Random Forest ROC curve, (b) Naïve Bayes ROC curve, (c) KNN ROC curve

Based on the classification metrics and ROC curve (Figure. 4), the Random Forest classifier outperformed the KNN and Naive Bayes classifiers in distinguishing PHA and non-PHA, as reflected by its accuracy score of 99.53% and AUC value of 1.00. The Random Forest performance can be attributed to several factors related to the nature of the dataset and the strengths of the algorithm. The orbital parameter of asteroids has a complex and non-linear relationship, thus the Random Forest particularly adept at capturing these non-linear patterns. Moreover, Random Forest can handle feature importance and selection [8], identifying the most relevant orbital parameters for classification which is beneficial when certain features are more informative in separating PHA from non-PHA. On the other hand, the Naïve Bayes classifier obtained an accuracy score of 92.00% with AUC value of 0.92, demonstrating a significant accomplishment, although the Naïve Bayes model was limited by asteroids' features were independent [12]. In contrast, the KNN classifier obtained an accuracy score of 84.45% and AUC value of 0.85. The KNN classifier's reliance on a fixed number of neighbours and distance metric might limit its ability to handle the complexity of the orbital parameters, leading to slightly lower performance [13].

Moreover, the study visualized of two essential parameters of asteroid orbital motion such as the MOID and the absolute magnitude. The MOID measured the closest distance between the asteroids' osculating orbit and the Earth [20]. The absolute magnitude measured an asteroid's luminosity without being affected by the distance between the asteroid and the observer [21].

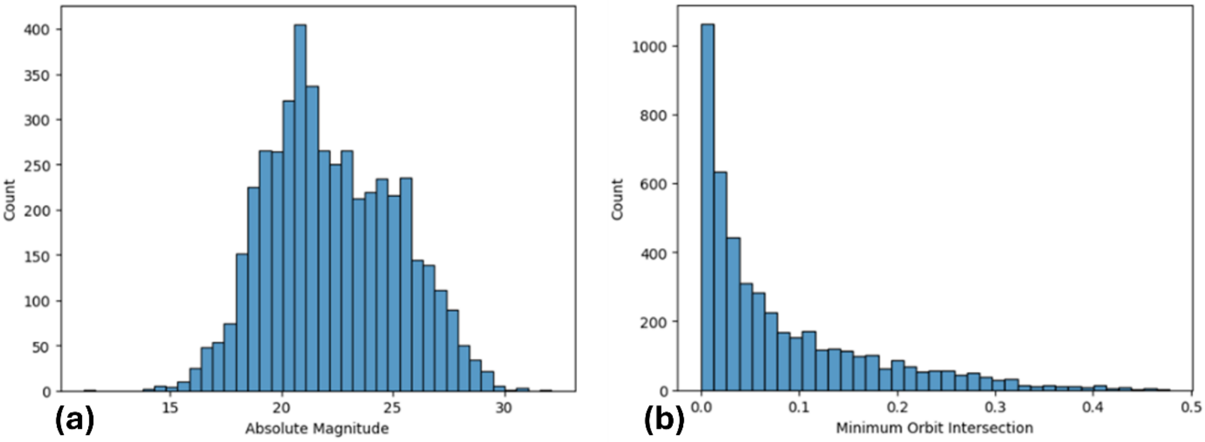

Figure 5. (a) Absolute magnitude data count, (b) MOID data count

Figure. 5 shows that asteroids' absolute magnitude in the dataset was 21.5. It indicated that the asteroids' brightness was relatively low. Meanwhile, its size was small. Furthermore, asteroids' MOID values were between 0.0 AU and 0.1 AU. It indicated that the asteroids potentially intersected with the Earth's orbit. The relationship between the asteroids' absolute magnitude, MOID, and potential hazards is shown in the following Figure. 6.

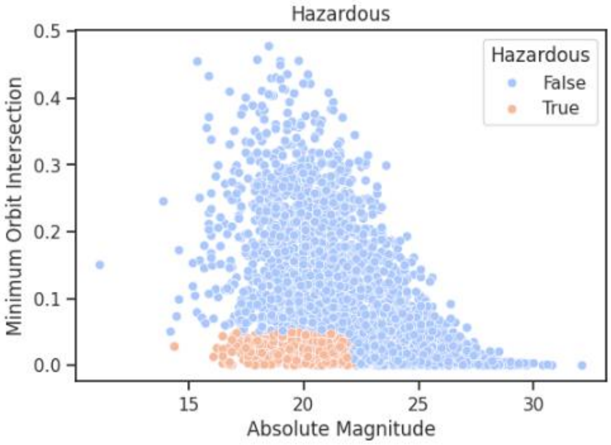

Figure 6. Visualization of absolute magnitude and MOID with hazardous

The PHA MOID value was less than 0.10 AU. Meanwhile the absolute magnitude value was less than 23. In the asteroid dataset, asteroids with an absolute magnitude below 23 have MOID values above 0.1 AU. It means that their orbit was not close to the Earth. Meanwhile, asteroids with MOID values below 0.1 AU and an absolute magnitude above 23 have a low brightness. It indicated a smaller size and a lower potential impact.

4. Conclusion

The study implemented an asteroid classification program based on the orbital motion. It also determined the most appropriate classifiers: Naive Bayes, Random Forest, or KNN. The Random Forest, Naive Bayes, and the KNN classifier obtained an accuracy score of 99.53%, 92.00%, and 84.45%, respectively. Even though each classifier has its strengths, the data showed that the Random Forrest was better than the other two classifiers because it performed the most effective data interactions and disparity. The asteroid's orbital parameter was essential to classify the PHA. Furthermore, there were two most essential parameters in PHA Classification: MOID and Absolute Magnitude. Finally, the research output helped us to monitor the PHA easily. For further research, real-time PHA detection is essential for an early warning system to protect the Earth from PHA.

References

[1] Harris, A. W., & Chodas, P. W. (2021). The Population of Near-Earth Asteroids Revisited and Updated. *Icarus*.

[2] Gregg, C. R., & Wiegert, P. A. (2022). A dedicated Lunar Trojan Asteroid Survey with small ground-based telescopes. *Monthly Notices of the Royal Astronomical Society, 511*(4), 5396-5404.

[3] Li, M., Huang, Y., & Gong, S. (2019). Assessing The Risk of Potentially Hazardous Asteroids Through Mean Motion Resonance Analyses. *Astrophysics and Space Science, 1-12*.

[4] Ikenaga, T., Sugimoto, Y., Ceriotti, M., Yoshikawa, M., Yanagisawa, T., Ikeda, H., ... & Utashima, M. (2019). A concept of hazardous NEO detection and impact warning system. *Acta Astronautica, 156*, 284-296.

[5] Bhavsar, R., Jadav, N. K., Bodkhe, U., & Gupta, R. (2023). Classification of Potentially Hazardous Asteroids Using Supervised Quantum Machine Learning. *IEEE, 11*.

[6] Klimczak, H., Oszkiewicz, D., Carry, B., Penttila, A., Kotlowski, W., Kryszczynska, A., & Wilawer, E. (2022). Comparison of Machine Learning Algorithms used to Classify The Asteroids Observed by all-sky Surveys. *Astronomy & Astrophysics, 667*, 1-15. doi:10.1051/0004-6361/202243889

[7] Bahel, V., Pillai, S., & Malhotra, M. (2020). A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance. *IEEE*, 495-498.

[8] Rahman, S., Shamrat, J. M., Tasnim, Z., Roy, J., & Hossain, S. A. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 8*(11), 419-422.

[9] Liu, Q., Cheng, W., Yan, G., Zhao, Y., & Liu, J. (2019). A Machine Learning Approach to Crater Classification from Topographic Data. *MDPI, 11*, 1-30. doi:10.3390/rs11212594

[10] Ahmed, R., Tahid, S. T., Mitu, N. A., Kundu, P., & Yeasmin, S. (2020). A Comprehensive Analysis on Undergraduate Student Academic Performance using Feature Selection Techniques on Classification Algorithms. *IEEE*.

[11] Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Elsevier*, 422-431.

[12] Chase, R. J., Harrison, D. R., Burkee, A., Lackman, G. M., & McGovern, A. (2022). A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning. *American Meteological Society, 37*, 1509-1529. doi:10.1175/WAF-D-22-0070.1

[13] Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2019). A Comparative Assessment of Credit Risk Model Based on Machine Learning. *Elsevier*, 141-149

[14] Smirnov, E. (2024). A comparative analysis of machine learning classifiers in the classification of resonant asteroids. *Elsevier*, 1-28.

[15] Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32-64.

[16] Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1), 54.

[17] Jiang, D., Lin, W., & Raghavan, N. (2020). A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques. *IEEE, 8*.

[18] Madhav, Kohli, S., Rawat, H., & Joshi, P. (2021). A Brief Study on Random Forest Using Python. *International journal of advances in engineering and management (IJAEM), 3(6),* 2063-2069. doi: 10.35629/5252-030620632069

[19] Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., & Manuel, D. G. (2020). A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making, 4*(20), 1-12. doi:10.1186/s12911-019-1014-6

[20] Egal, A., Wiegert, P., Brown, P. G., Spurný, P., Borovička, J., & Valsecchi, G. B. (2021). A dynamical analysis of the Taurid Complex: evidence for past orbital convergences. *Monthly Notices of the Royal Astronomical Society*, *507*(2), 2568-2591.

[21] Veres, P., Jedicke, R., Fitzsimmons, A., Denneau, L., Granvik, M., Bolin, B., . . . Waters, C. (2015). Absolute magnitudes and slope parameters for 250,000 asteroids observed by Pan-STARRS PS1 - preliminary results. *Astrophysics*, 1-37.