The International Journal of Engineering and Information Technology

journal homepage:www.ijeit.misuratau.edu.ly

MISURATA UNIVERSITY

Utilizing the H2O AutoML Approach to Predict Hazardous Near-Earth Objects

A. Al-Gaddari, M
University of Benghazi
marwa.abdulaziz@uob.edu.ly

B. Najem, T
University of Benghazi
Tarek.nagem@uob.edu.ly

Abstract— **The universe is replete with various types of objects that need to be studied from time to time, such as stars, asteroids, comets, and a multitude of small astronomical bodies. The small bodies surrounding the Earth are called Near-Earth Objects (NEOs) and may pose a danger to our planet. Therefore, analyzing the attributes and composition of NEOs by utilizing effective machine-learning approaches is considered a crucial mission to detect hazardous near-earth objects and help astrophysics and other scientists figure out the appropriate solutions before the occurrence of this phenomenon. This research focuses on leveraging the H2O AutoML prediction approach, in conjunction with pertinent data features, to accurately identify hazardous NEOs. The H2O AutoML approach has been applied and evaluated using different data-splitting techniques in three different experiments to reach and demonstrate superior performance, which was achieved with an impressive accuracy of 98.27%, precision of 98.37, recall of 98.17%, and F1-score of 98.26%.**

Index Terms— ***NASA, Near-Earth Object, Asteroid, H2O AutoML.***

I. INTRODUCTION

Since 4.6 billion years ago, after the formation of our solar system, the leftovers of that formation have made the solar system full of an enormous number of objects that contain rock, ice, and organic composition, which are now known as asteroids and comets. The current known asteroid count according to NASA is 1,281,591, and the unknown Near-Earth Objects data amount is increasing as well, which requires being studied and analyzed from time to time [1]. Most asteroids are discovered by satellites, probes, and telescopes with huge aperture lengths. Moreover, the main purpose of the large telescopes on Earth is to track main belt asteroids. In response to the 2014 report from the NASA Office of Inspector General, NASA reorganized its Near-Earth Object Observations Program and set up a Planetary Defense Coordination Office in January 2016 [2]. Nowadays, there are large NASA-funded observatories (including Pan-STARRS, the Catalina Sky Survey, NEOWISE, and, in the future, NEO Surveyor). NEOs, or near-Earth objects, are a popular term for objects close to Earth. These are asteroids and comets, and their orbits either approach or cross over that of the Earth. The majority of NEOs do not threaten Earth, but there is a slight possibility that a few could collide with it and do significant damage. Data analysis techniques such as data mining methods, machine learning algorithms, and deep learning models play a big role in analyzing and discovering the details and attributes of the NEOs, classifying the asteroids according to their danger, and predicting the dangerous events and impacts that may occur during specific periods caused by those asteroids in space [3]. The artificial neural network (ANN) was the most well-known machine learning technique in astronomy in the 2000s, and it has been presented since the middle of the 1980s [4]. Interestingly, the asteroid images and the features with the details can be analyzed using data mining methods, and the impacts of those asteroids can be predicted using machine learning techniques with acceptable performance. Even though the asteroids or NEO datasets are available and the models' performance is acceptable, certain tasks involved in machine learning, such as training models on large datasets, can be time-consuming. Therefore, using H2O AutoML can be beneficial as it allows a large number of models to be quickly built and evaluated, with many of the tedious and time-consuming tasks being automated, enabling us to focus on higher-level tasks. Moreover, the uncertainty surrounding the attributes and composition of asteroids poses a significant challenge in accurately predicting their behavior and hazard potential. However, the H2O AutoML approach can mitigate this challenge by employing the needed feature engineering to extract the significant features of near-earth objects. Additionally, it can be trained on large and small astronomical datasets to predict potentially hazardous near-earth objects.

Received 14 Apr , 2024; revised 06 May, 2024; accepted 15 Mar 2024.
Available online 08 Aug, 2024.

A. Problem Statement

Near-Earth objects (NEOs) pose a significant risk to our planet and man-made space objects. Understanding the characteristics and behavior of these objects is crucial for analysis and monitoring purposes. Collisions between NEOs and satellites can lead to a cascade of destructive events, such as communication disruptions and satellite failures. Additionally, the gravitational pull of large asteroids passing near Earth can cause orbital disturbances and unpredictable collisions. The primary concern lies in identifying hazardous NEOs that could potentially collide with Earth, leading to catastrophic damage and loss of life [5]. Mitigating this risk requires accurate detection and analysis of hazardous near-earth objects. To address this challenge, this research aims to leverage the H2O AutoML approach to analyze large and different volumes of data and discover patterns that enable precise identification of hazardous near-earth objects. By employing machine learning techniques and extensive data analysis, the proposed solution aims to improve our ability to detect and predict the behavior of these objects and contribute to effective risk mitigation strategies.

B. Aim of Research

The main aim of this proposed dissertation is to utilize the H2O AutoML approach to identify and predict near-earth objects that pose a potential threat to Earth with the highest possible accuracy.

C. Research Questions

- Can the H2O AutoML approach accurately predict potentially hazardous near-earth objects?
- Can H2O AutoML performance be better than the previously existing machine learning approaches?

II. LITERATURE REVIEW

The study of near-Earth objects (NEOs) and asteroids has gained widespread attention in recent years, thanks to the dissemination and availability of NEOs and asteroid data collections. In this context, artificial intelligence (AI) and machine learning have emerged as valuable tools for various tasks, particularly in the classification and prediction of NEOs, including the identification of hazardous objects. One area of research that has gained prominence is the classification of near-Earth objects based on extracted features from their images. Researchers have explored the use of machine learning algorithms to differentiate between asteroids and non-asteroids, leveraging the rich information present in these images. Additionally, the studies in the survey [6] have investigated the multi-classification of near-Earth objects based on their orbital class, a crucial task for predicting the hazards associated with these objects. Among the machine learning algorithms applied, binary classification models have been widely used to predict whether a near-Earth object is hazardous or non-hazardous. Notably, algorithms such as random forest and gradient boosting have demonstrated exceptional performance in numerous research projects. These algorithms have shown remarkable accuracy and have become the preferred choice in recent publications. The NEO data image is often collected from multiple resources, such as telescopes, radars, and satellites. Nevertheless, some of the data must be collected by taking images under the supervision of human experts during outer space missions to record essential and detailed information about asteroids, like the ATLAS [7] data and the Catalina Sky Survey [8]. Also, there are several projects and approaches for some international institutes that provide asteroids and near-earth object datasets with various file formats (i.e., jason,.txt, or CSV format), such as the NEOWISE Project data [9]and the data of the International Astronomical Union Minor Planet Centre (MPC) [10]. NASA gathers and keeps track of data related to NEO research. Concerning these objects' properties, orbits, and possible dangers, these datasets provide valuable information. Tracking and characterizing NEOs is the responsibility of NASA's Jet Propulsion Laboratory's (JPL) Near-Earth Object Program [11]. They find and track these objects using a variety of approaches to observation, such as space-based missions and ground-based telescopes. The data collected include measurements of an object's position, velocity, size, shape, rotation, and composition. NASA provides access to NEO data through various platforms and APIs (application programming interfaces). One such resource is the Asteroids-NeoWs API [12], available through the NASA Open Data Portal. This API allows users to retrieve information about NEOs, including their orbital elements, close-approach data, and physical parameters. Machine learning algorithms have been widely employed in the prediction of near-earth objects (NEOs) due to their ability to handle complex patterns and nonlinear relationships in the data. Several popular machine learning algorithms have been applied in NEO prediction studies. Decision trees, such as the Random Forest algorithm, are commonly used for classification tasks to determine whether an NEO is hazardous or not based on its features. Support Vector Machines (SVM) have also been utilized for classification, leveraging the separation of classes in a high-dimensional feature space. Additionally, neural networks, including deep learning architectures like convolutional neural networks (CNNs) have shown promise in capturing intricate relationships in NEO data for both classification and regression tasks. Other algorithms, such as Random forest, k-nearest neighbors (k-NN), Naive Bayes, and Gradient Boosting algorithms like XGBoost and Light GBM, have also been explored in the context of NEO prediction [13] [14] [15]. The choice of algorithm depends on the specific prediction task, the available data, and the desired performance metrics, and researchers continue to explore and compare the effectiveness of different machine learning techniques in accurately predicting NEOs and assessing their potential hazards.

A. Related Works

The previous studies in the literature review section have presented some works that performed classification and prediction tasks using machine learning and deep learning algorithms and discussed applying these algorithms to near-earth object datasets, which are taken from the NASA Jet Propulsion Laboratory "Small-Body Database" Search Engine with different sizes. This section will show the research studies that have used the same dataset that has been used for the main experiment of this research. Diya Khajuria and Amisha Sharma et al. [16] analyzed the NASA Nearest Earth Objects dataset through several plots and charts. The study discussed the performance of different machine learning classifiers such as the decision tree classifier, Logistic regression, and Random Forest Classifier for predicting hazardous or non-hazardous near-earth objects. Therefore, the authors split the data into 75% training data and 25% testing data. After comparing the algorithms' performance, the random forest algorithm outperforms other machine learning algorithms with 91.9% accuracy. As well as, Yao Wang [17] has applied seven machine learning algorithms to predict the hazardous near-earth objects from the NASA-Nearest Earth Objects dataset, and the best performance was for the Random Forest algorithm with an accuracy of 95%, the Voting algorithm with an accuracy of 94%, the Decision Tree with an accuracy of 93%, and the Gradient Boosting algorithm with an accuracy of 89%. In August 2023, the article [18] looked at 90,836 asteroids in a 70,000 km radius. Five algorithms—Lightgbm, Gradient Boosting, Ada Boost, Extra Tree, and Random Forest—were used to classify asteroids into high-risk and low-risk categories. It was found that the Random Forest method performed the best, while the AdaBoost approach performed the worst. In this manner, high-risk asteroids were predicted by the Random Forest method with 94% accuracy and the AdaBoost approach with 91% accuracy. The ELMs implemented in the study [19] include the standard ELM, the regularized ELM, and the weighted ELM with W1 and W2 versions. The" Nearest Earth Objects" dataset has been used to train and validate the hazardous object classification ELM models with only five features of the near-earth objects and a binary output indicating whether they were dangerous to Earth or not.

The models were evaluated based on accuracy, geometric mean, and time of training. From the results, the weighted ELM in its W1 version obtained the best performance, with 80% accuracy, a 70% geometric mean, and 1.8 seconds of training time. Based on the results achieved, the viability of classifying whether objects are potentially hazardous for Earth is confirmed. Nevertheless, to improve the performance of ELM models, it is recommended to continue discovering other types of ELMs. From the scanning of those related works, all the reviewed studies used the same methodology, which is applying different machine learning algorithms to the main dataset to be compared with each other and highlighting the best machine learning model performance with the highest achieved accuracy. What represents a gap is that there are numerous machine-learning techniques, but the studies did not emphasize the technical reasons for choosing the machine-learning techniques that have been applied. Moreover, all the related works applied machine learning algorithms to the dataset without balancing the hazardous and non-hazardous classes of the data records. On the other side, all the features of the dataset are important; however, 'Name' and 'Object ID' were removed in all the related works as they do not impact the performance of the machine learning approaches. Additionally, it was noticeable that the ensemble machine-learning models demonstrated the best and most effective performance compared to individual algorithms.

III. DATA ANALYSIS

The selected dataset on near-Earth objects in this proposal research is available on Kaggle [20] and was collected from NASA JPL's Small-Body Database. This data has 90,837 records and 10 columns. The dataset features or columns are described as follows:

- Object ID: The unique identifier for each Near-Earth Object (NEO).
- Object Name: The name given to the NEO by the Minor Planet Centre (MPC).
- Orbiting Body: The celestial body around which the NEO is orbiting.
- Sentry Object: It is represented by a Boolean data type, which presents whether the asteroid is included in the sentry.
- Relative Velocity: The velocity of the NEO relative to Earth at the time of its closest approach.
- Miss Distance: The distance between the NEO and Earth at the time of its closest approach.
- Estimated Diameter (min): The minimum estimated diameter of the NEO, in meters.
- Estimated Diameter (max): The maximum estimated diameter of the NEO, in meters.
- Absolute Magnitude: The intrinsic brightness of the NEO on a logarithmic scale.
- Hazardous: The label column that shows whether the object is hazardous or not, which is represented in a Boolean data type.

  The dataset is imbalanced because it consists of only 8840 hazardous objects, and the majority of the near-earth objects within the data observations are non-hazardous objects' records. The label attribute is the hazardous feature column, and its data type is Boolean and contains two values. The '0' value represents an object classified as a non-hazardous object, and the '1' value represents an object classified as a hazardous object. "Figure 1" shows the distribution of the hazardous class in the dataset. The number of non-hazardous objects is more than 80000, which is precisely 81,996 objects.

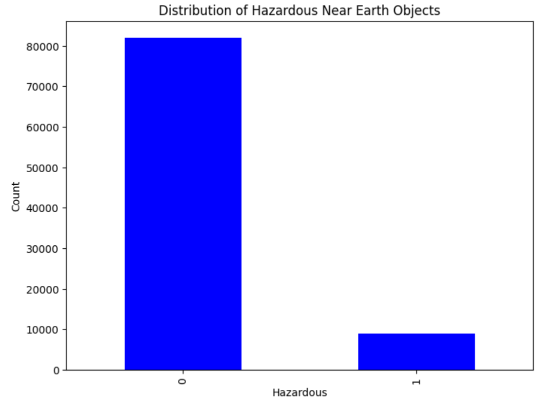

Figure 1. The Distribution of the Hazardous Class in the Dataset.

IV. H2O AUTOML APPROACH

The proposed approach in this research is the H2O AutoML algorithm that has been applied using Python code on the Google Colab online tool. H2O AutoML Approach is an automated machine-learning framework provided by H2O.ai. It aims to simplify developing high-performing machine learning models by automating various tasks, such as feature engineering, model selection, and hyperparameter tuning. The approach followed by H2O AutoML can be decomposed as follows:

A. *Data Preparation*

The first step in using H2O AutoML is to prepare the input data. This typically involves loading the dataset, handling missing values, encoding categorical variables, and splitting the data into training and validation sets.

B. *Model Building and Ensemble Construction*

Once the data is prepared and the AutoML configuration is set, H2O AutoML starts building and evaluating a diverse range of models. It automatically explores different algorithms, feature transformations, and hyperparameter settings to find the best model for the given task. H2O AutoML leverages H2O's distributed computing capabilities to train multiple models in parallel, making the process more efficient. H2O AutoML leverages the power of ensemble learning to improve model performance. It combines the predictions of multiple individual models to create a more accurate and robust ensemble model. The ensemble construction process involves selecting the best models based on their performance and combining their predictions using techniques like stacking or blending.

C. *Model Selection and Hyperparameter Optimization*

After building a collection of models, H2O AutoML ranks them based on the specified performance metric. The best-performing model, or the "leader" model, is selected based on this ranking. The leader model represents the most optimal model discovered by the AutoML process.

H2O AutoML considers a predetermined collection of machine learning algorithms, known as the model base, which generally consists of multiple types of models, including random forests, gradient boosting machines (GBMs), generalized linear models (GLMs), deep learning models, and more. Whereas the hyperparameters are settings or configurations that control the behavior of machine learning models, for instance, the learning rate, regularization strength, number of layers in a neural network, maximum depth of a decision tree, etc. During the model selection stage, H2O AutoML trains several models with different configurations and settings for each algorithm in the model base by using the k-fold cross-validation technique, where the training data is divided into k subsets or folds. It trains the models on k-1 folds and evaluates their performance on the remaining folds. Based on the performance metrics that H2O AutoML collects (such as accuracy, AUC, or RMSE) for every model and algorithm combination, H2O AutoML selects the best-performing models from the model base. After that, H2O AutoML automates the process of tuning hyperparameters to find the optimal combination that maximizes the model's performance by performing different strategies for hyperparameter optimization, such as grid search, random search, or Bayesian optimization, and evaluates models with different hyperparameter configurations using cross-validation and chooses the combination that obtains the best performance [21] [22] [23].

D. *Model Assessment and Interpretation*

Following the model creation and selection process, H2O AutoML summarizes the trained models' performance metrics. Using a holdout validation dataset might produce predictions that can be tested further. Additionally, H2O AutoML provides methods for analyzing and interpreting models, such as feature importance analyses and partial dependence graphs, to put more light on the behavior of the models.

E. *Model Deployment and Scoring*

Once the leader model is selected, it can be deployed for production use to make predictions on new, unseen data. H2O provides a simple interface to score new data using the deployed model, allowing you to apply the trained model to other real-world scenarios [24] [25].

H2O AutoML combines the power of automated feature engineering, model selection, and ensemble learning to create a comprehensive and efficient machine-learning workflow. It helps to reduce the manual effort and time required for developing high-quality machine learning models. ‘Figure 2’ shows the workflow of the H2O AutoML approach.

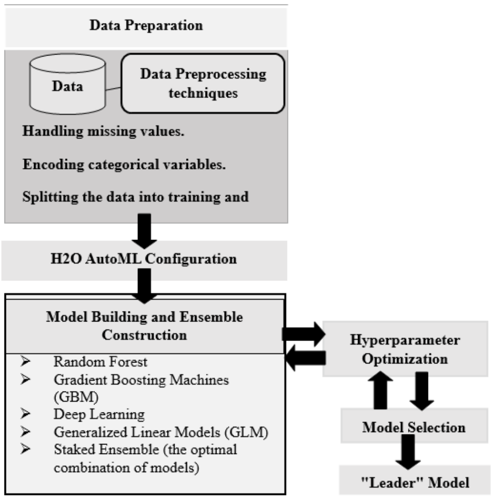
Data Preparation
Data
Data Preprocessing techniques
Handling missing values.
Encoding categorical variables.
Splitting the data into training and
H2O AutoML Configuration
Model Building and Ensemble Construction
Random Forest
Gradient Boosting Machines (GBM)
Deep Learning
Generalized Linear Models (GLM)
Staked Ensemble (the optimal combination of models)
Hyperparameter Optimization
Model Selection
"Leader" Model

Figure 2. The workflow of the H2O AutoML approach.

V. RESEARCH METHODOLOGY

The methodology employed in this research consists of a large dataset, which is 'Nearest Earth Objects'. The dataset is subjected to three experimental procedures to assess the performance of the H2O AutoML model and to pose unique challenges and opportunities for model training and evaluation.

- The first experiment involves utilizing the traditional data splitting approach, wherein the 'Nearest Earth Objects' dataset was divided into a training set and a testing set using varying ratios while adjusting for the dataset size as shown in “Figure 3”. This allows for effective evaluation and comparison of the model's performance under different data split configurations.
- The second experiment employs 10-fold cross-validation, a technique that systematically divides the 'Nearest Earth Objects' dataset into ten subsets (folds). Each fold is iteratively used as the testing set, while the remaining folds collectively form the training set as shown in “Figure 4”. This approach provides a robust assessment of the model's generalization capabilities.
- The third experiment is consequently the proposed solution that encompasses key components aimed at enhancing the accuracy and reliability of predicting hazardous near-earth objects. These components include utilizing a large dataset, implementing cross-validation for rigorous testing, and maintaining a balanced distribution of object records. This proposed solution strives to optimize the predictive performance and reliability of the model in accurately classifying hazardous near-earth objects.

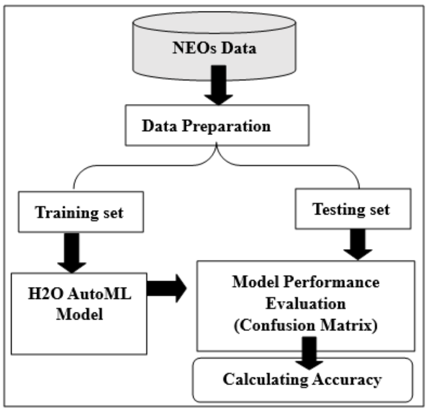

Figure 3. The workflow of The First Experiment.

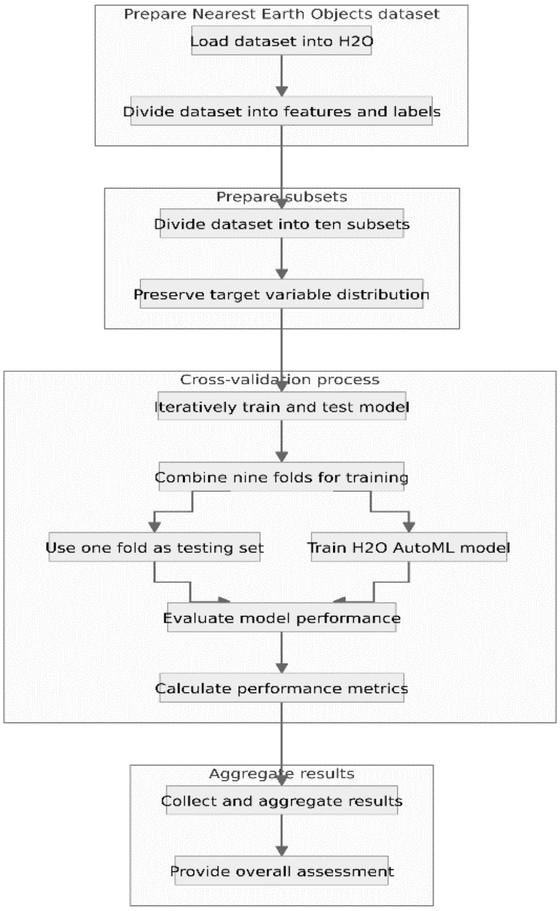
Prepare Nearest Earth Objects dataset
Load dataset into H2O
Divide dataset into features and labels
Prepare subsets
Divide dataset into ten subsets
Preserve target variable distribution
Cross-validation process
Iteratively train and test model
Combine nine folds for training
Use one fold as testing set
Train H2O AutoML model
Evaluate model performance
Calculate performance metrics
Aggregate results
Collect and aggregate results
Provide overall assessment

Figure 4. The workflow of The Second Experiment

VI. EVALUATION METHODS AND RESULTS

Evaluating the machine-learning model’s performance is an essential task to measure the efficiency that can be detected by approximating the correct model predictions. In this section, all of the metrics used to assess the model's performance have been explained and presented before using them in presenting the results as the chosen evaluation criteria.

A. Confusion Matrix

The confusion matrix is a fundamental tool used in the evaluation stage of classification models. It presents a tabular representation that summarizes the model's predictions in comparison to the actual ground truth labels. The matrix includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a detailed breakdown of the different types of classification errors made by the model. This breakdown aids in identifying specific areas of improvement and fine-tuning the model's performance. Moreover, the confusion matrix serves as the foundation for calculating key evaluation metrics to offer a quantitative assessment of the model's performance in predicting hazardous near-earth objects, such as the accuracy, precision, recall, and f1-score that have been used in the evaluation process.

B. The Accuracy

It measures the overall correctness of the model's predictions by calculating the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances. It can be described as the next equation:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

a) Sensitivity

It is also known as recall or true positive rate (TPR) and measures the proportion of actual positive instances (hazardous NEOs) correctly identified by the model. as described by the following equation.

$$Recall(R) = \frac{TP}{TP+FN}$$

b) Precision

It measures the proportion of predicted positive instances (hazardous NEOs) that are actually positive.

$$Precision(P) = \frac{TP}{TP+FP}$$

c) F1 Score

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives.

$$F1\ Score = (2 * \frac{Presision*Recall}{Precision+Recall})$$

A)

B) Results of Experiment 1: In the first experiment, the 'Nearest Earth Objects' dataset was divided into a training set and a testing set. Initially, the dataset was split using a 70% training set and a 30% testing set ratio. Subsequently, the ratios were adjusted to 75% training and 25% testing, followed by 80% training and 20% testing. The rows of the dataset were extracted and utilized to train the H2O AutoML model. The model was applied to the training set, undergoing a total runtime of 2 minutes for the training process. Following the completion of model training, predictions were generated from the confusion matrixes that are shown in Figure 5, Figure 6, and Figure 7.

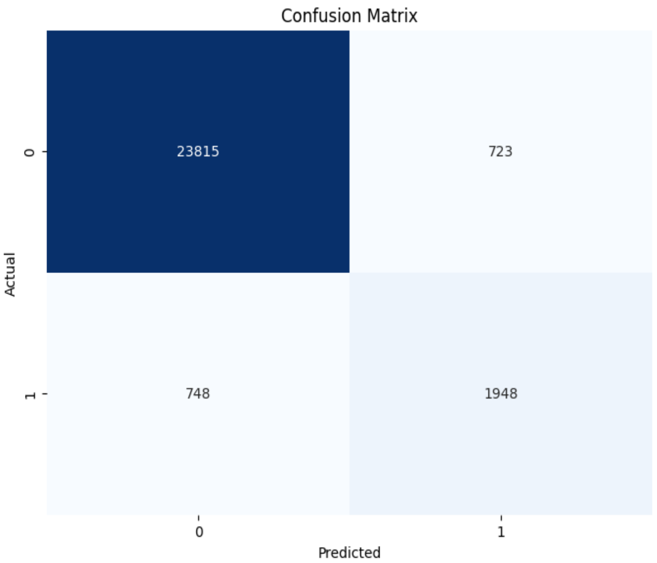

Figure 5. Confusion matrix for 30% testing set.

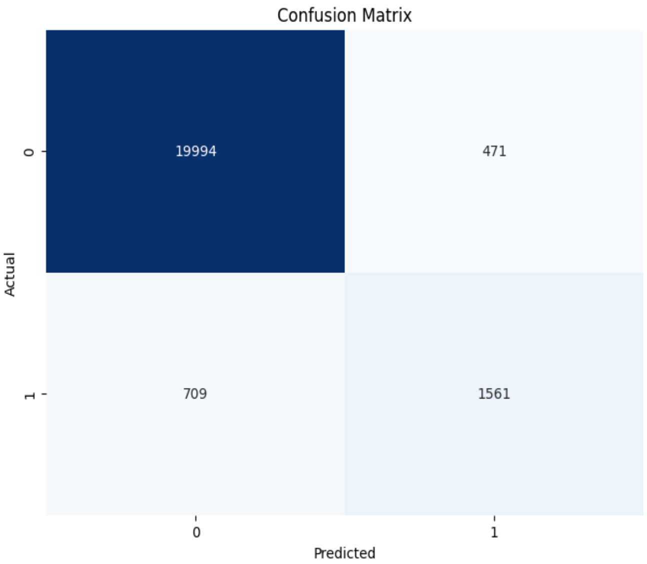

Figure 6. Confusion matrix for 25% testing set.

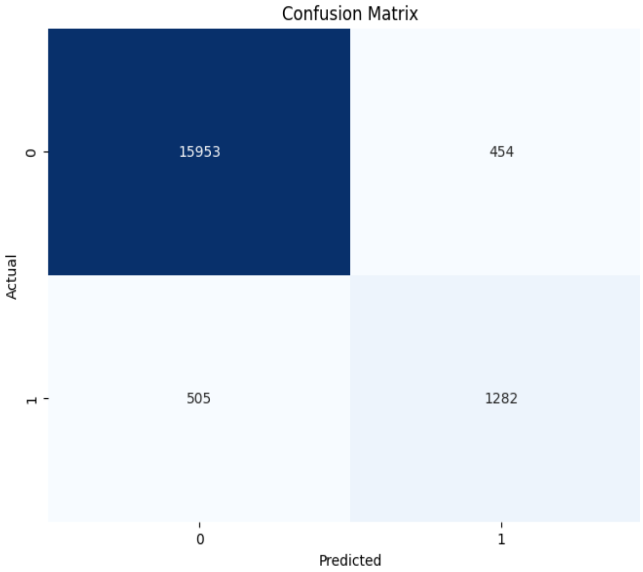

Figure 7. Confusion matrix for 20% testing set.

C) *Results of Experiment 2:* To assess the presence of overfitting and address the challenge of model generalization, a 10-fold cross-validation technique was employed as a pivotal component of the experimental methodology. By utilizing this approach, the performance of the model was evaluated across multiple iterations, ensuring robustness and reliability in the assessment of its accuracy and generalization capabilities.

D) *Discussion*

Accuracy is a commonly used evaluation metric to measure the performance of machine learning models. However, it's important to consider other factors, such as dataset size and class imbalance, when assessing the effectiveness of these models, especially in the context of detecting hazardous near-earth objects. Due to the rarity of hazardous near-earth objects compared to non-hazardous ones, the task of detection is challenging the effectiveness of these models, and the dataset tends to be imbalanced. In the initial experiment, where the data testing was split into a training set and a testing set, the achieved accuracy values were 94.59%, 94.80%, and 94.72%, which indicate excellent accuracy. To ensure the robustness and reliability of the H2O AutoML model's performance assessment and its generalization capabilities, 10-fold cross-validation was applied to the dataset.

This approach yielded an accuracy of 96.83%, demonstrating the reliability and precision of the H2O AutoML approach in predicting hazardous near-earth objects. To further explore the model's abilities, a solution has been extracted in a subsequent experiment.

E) *Results of Experiment 3:*

a third experiment was conducted, utilizing an under-sampling strategy to address the issue of imbalanced hazardous and non-hazardous object records to obtain reliable performance accuracy and a strong predictive model that could predict any near-earth object, whether it was hazardous or not.

A balancing technique was employed as a proposed solution to address potential class imbalance issues in the datasets. To assess the model's performance and its ability to generalize, 10-fold cross-validation was employed as a data-splitting technique. “Figure 8” shows the new version of the dataset after using the under-sampling technique.

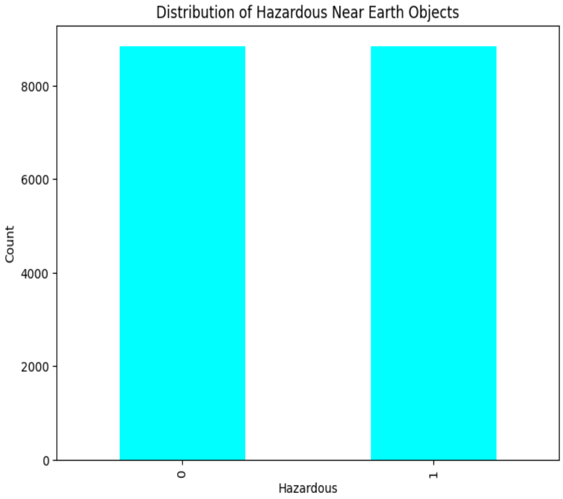

Figure 8. Distribution of Hazardous Attribute for New Version of NASA-Nearest Earth Objects Dataset.

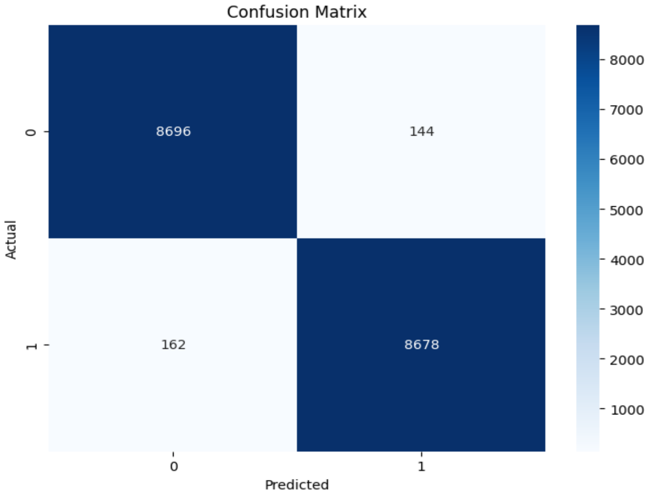

Figure 9. The Average Confusion Matrix of Using 10-Fold Cross-Validation with Balanced NASA-Nearest Earth Objects Dataset.

Based on the outcomes and the average confusion matrix shown in “Figure 9” of the H2O AutoML approach's 10-fold cross-validation, the model's overall accuracy was 98.27%. Furthermore, 98.37%, 98.17%, and 98.27 were reported for the precision, recall, and F1-score, respectively, as presented in “Table 1”. The performance metrics suggest that there is a high degree of accuracy and effectiveness in predicting near-earth objects, both hazardous and non-hazardous.

VII. H2O AUTOML PERFORMANCE COMPARED TO RELATED WORKS

The “NASA-Nearest Earth Objects” dataset has been used in several papers and articles, as discussed in the related works subsection of the literature review. In the research [16], the dataset was split into a 75% training set and a 25% testing set, and the random forest algorithm was applied and achieved an overall performance accuracy of 91.9%. However, when using the same data split, the H2O AutoML model surpassed this performance with an accuracy of 94.80%. Similarly, in another study [17], the random forest algorithm was used and achieved an accuracy of 95% by using cross-validation. In a separate investigation of the work [18], the accuracy of the random forest was 94% when the dataset was split into an 80% training set and a 20% testing set. Nevertheless, in both cases, the H2O AutoML model outperformed the random forest, achieving an accuracy of 96.80% with cross-validation and 95.0% with an 80% training set and a 20% testing set. All the comparisons of the H2O AutoML model with the related works are presented in "Table 2.".

Table 1. All of The Research Experiments Results.

Experiment 1				
Data-Splitting Ratio	**Accuracy**	**precision**	**recall**	**F1-Sore**
Training set: 70% Testing set: 30%	94.59%	72.93%	72.25%	72.59%
Training set: 75% Testing set: 25%	94.80%	76.82%	68.76%	72.57%
Training set: 80% Testing set: 20%	94.72%	73.84%	71.74	72.77%
Experiment 2				
10-Fold Cross-Validation	96.83%	85.78%	80.36%	83.25%
Experiment 3 (Proposed Solution)				
Applying H2O AutoML to Balanced NASA-Nearest Earth Objects Dataset Using 10-Fold Cross-Validation	**98. 27%**	**98.37%**	**98.17%**	**98.27%**

Table 2. H2O AutoML Performance Comparisons with Related Works.

The Approach	**Data-Splitting Technique**	**Ref. /Year**	**Accuracy**
Random Forests	75% Training set 25% Testing set	[16]/ 2023	91.1%
	Cross Validation	[17]/2023	95.00%
	80% Training set. 20% Testing set	[18]/2023	94.00%
H2O AutoML Approach	75% Training set 25% Testing set	The Proposed Research /2024	94.80%
	Cross Validation		96.83%
	80% Training set 20% Testing set		94.72%

VIII. CONCLUSION

Nowadays, machine-learning algorithms are playing a noticeable role in several fields. Studying the universe's nature and discovering any hazards that may cause damage or loss of life on our planet was the motivation for this research. The study of near-Earth objects (NEOs) and asteroids has gained widespread attention in recent years, thanks to the dissemination and availability of NEOs and asteroid data collections. In this context, artificial intelligence (AI) and machine learning have emerged as valuable tools for various tasks, particularly in the classification and prediction of NEOs, including the identification of hazardous objects. One area of research that has gained prominence is the classification of near-Earth objects based on extracted features from their images. Recent research and studies have focused on detecting potentially hazardous objects, where some algorithms have shown effective results. The workflow of the H2O AutoML approach that has been utilized in this research encompasses multiple stages, beginning with data preparation to train the H2O AutoML model and culminating in the deployment of a significant model. Leveraging the power of Python, the H2O AutoML library, and the computational resources provided by Google Colab, three experiments were successfully implemented. The strengths of H2O AutoML, such as its automated machine learning capabilities, extensive algorithm selection, and hyperparameter optimization, enabled efficient model selection and improved predictive performance. Two experiments were conducted to assess the model's performance. In the first experiment, the traditional data-splitting technique was employed to apply and test the H2O AutoML model on the dataset. This experiment aimed to evaluate the model's performance in different data scenarios. The second experiment focused on improving the model's performance using 10-fold cross-validation. This technique was applied to enhance the model's efficiency and generalization abilities. The results of this experiment revealed that utilizing 10-fold cross-validation led to improved performance, suggesting that this approach is beneficial for near-earth object prediction. Also, the H2O AutoML model performance in the first two experiments outperformed the other approaches that have been utilized in the related works, with an accuracy of approximately 95% and 96.83%, respectively. Building on these findings, a third experiment was conducted to address the issue of imbalanced hazardous and non-hazardous object records in the dataset. An under-sampling technique was implemented as a proposed solution to balance the two classes. Subsequently, the H2O AutoML model was applied using the 10-fold cross-validation technique. The results of the third experiment demonstrated a high level of accuracy, with an overall accuracy rate of 98.27%. Additionally, precision, recall, and F1-score were reported as 98.37%, 98.17%, and 98.27%, respectively. These performance metrics indicate the model's effectiveness in correctly classifying hazardous and non-hazardous near-earth objects.

REFERENCES

[1] "nasa.gov," NASA, 2023. [Online]. Available: https://solarsystem.nasa.gov/. [Accessed 2024].

[2] P. K. Martin, "NASA's Efforts to Identify Near-Earth Objects and Mitigate Hazards," NASA, SEPTEMBER 15, 2014.

[3] N. M. B. a. R. J. BRUNNER, "DATA MINING AND MACHINE LEARNING IN ASTRONOMY," *International Journal of Modern Physics DC World Scientific Publishing Company,* August 11, 2010.

[4] W. J. a. R. Rosner, "OPTIMIZATION ALGORITHMS: SIMULATED ANNEALING AND NEURAL NETWORK PROCESSING," Harvard-Smithsonian Center for Astrophysics, (1986), pp. ApJ 310, 473.

[5] "Near-Earth Object Observations Program," NASA, [Online]. Available: https://www.nasa.gov/planetarydefense/neoo. [Accessed 2024].

[6] S. A. R. C. D. M. H. a. W. B. V. Carruba, "Machine Learning applied to asteroid dynamics," *arXiv,* Jun 2022.

[7] "The ATLAS Project. ATLAS," 2020. [Online]. Available: https://fallingstar.com/home.php.

[8] "The University of Arizona. Catalina Sky Survey," 2020. [Online]. Available: https://catalina.lpl.arizona.edu/.

[9] "California Institute of Technology. The NEOWISE Project.," 2020. [Online]. Available: https://neowise.ipac.caltech.edu/. [Accessed 8 September 2023].

[10] "The International Astronomical Union Minor Planet Centre," [Online]. Available: https://www.minorplanetcenter.net/data.

[11] "NASA JPL Small-Body Database," [Online]. Available: https://ssd.jpl.nasa.gov/tools/sbdb_lookup.html#/. [Accessed 3 January 2024].

[12] "Asteroids - NeoWs API | NASA Open Data Portal.," [Online]. Available: https://data.nasa.gov/Space-Science/Asteroids-NeoWs-API/73uw-d9i8/about_data.

[13] Breiman, "Random Forests. Machine Learning," 2001, pp. 45 (1) 5-32.

[14] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* pp. 46(3), 175-185., (1992).

[15] T. Chen, "XGBoost: A Scalable Tree Boosting System.," *The Journal of Machine Learning Research,* vol. 17, no. (1), p. 1–5., (2018).

[16] A. S. N. S. M. M. Diya Khajuria, "Classification and Comparative Analysis of Earth's Nearest Objects using Machine Learning Models," in *International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023.

[17] Y. Wang, "Comparison of Machine Learning Strategies in Hazardous Asteroids Prediction," *Highlights in Science, Engineering and Technology,* vol. Volume 39, (2023).

[18] M. B. M. A. A. S. Seyed Matin Malakouti, "Machine learning techniques for classifying dangerous asteroids," *ScienceDirect*, 2023.

[19] E. M. F. ´. u. D. Z.-B. X. A. L.-C. ´. e. J. P. R. M. I. S. Roberto Ahumada-Garc´ıa, "Extreme Learning Machine (ELM) for Detection of Hazardous Near Earth Objects," in *42nd IEEE International Conference of the Chilean Computer Science Society*, 2023.

[20] S. VANI, "Kaggle," 2022. [Online]. Available: https://www.kaggle.com/datasets/sameepvani/nasa-nearest-earth-objects. [Accessed 2024].

[21] E. L. a. S. Poirier, "H2O AutoML: Scalable Automatic Machine Learning," in *7th ICML Workshop on Automated Machine Learning*, (2020).

[22] "H2O.ai. (n.d.). H2O AutoML.," [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html.

[23] H. (n.d.)., "H2O AutoML: Scalable Automatic Machine Learning," [Online]. Available: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html.

[24] "H2O.ai, "H2O AutoML," [Online]. Available: https://h2o.ai/platform/h2o-automl/.

[25] A. W. J. G. K. H. C. B. B. R. F. Anh Truong, " Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools," *arXiv,* 3 Sep 2019.