Hazardous Asteroid Prediction

P13: Aakarsh Satish, Firasat Hussain Mohammed, Utkarsh Sharma
Department of Computer Science, North Carolina State University Raleigh NC 27695
(asatish2, fmohamm8, usharma3)@ncsu.edu

1 Background and Introduction

Asteroids, commonly referred to as "space rocks" or "minor planets," are celestial objects that orbit the Sun, primarily in the asteroid belt between the orbits of Mars and Jupiter. While most asteroids are small and benign, some have the potential to pose a significant threat to our planet. Understanding, tracking, and predicting the trajectories of these near-Earth objects (NEOs) is a crucial scientific endeavor.

1.1 The Asteroid Impact Threat

Throughout Earth's history, asteroid impacts have played a significant role in shaping our planet's geology and evolution. The most notable event is the extinction of the dinosaurs approximately 66 million years ago, believed to be caused by a massive asteroid impact. While such catastrophic events are rare, they underscore the importance of monitoring and predicting asteroid movements. A smaller impact could still have devastating regional or global consequences.

1.2 Relevant Papers

[1] Oscar Fuentes-Mu noz, Daniel J Scheeres, Davide Farnocchia, and Ryan S Park. The hazardous km-sized neos of the next thousands of years. The Astronomical Journal, 166(1):10, 2023.

[2] Leonid Sokolov, Nikita Petrov, Galina Kuteeva, and Andrey Vasilyev. Scattering of trajectories of hazardous asteroids. In AIP Conference Proceedings, volume 1959. AIP Publishing, 2018.

1.3 Significance of Asteroid Prediction

Asteroid prediction and monitoring are critical for several reasons:

- Early Warning: Accurate prediction models provide early warnings, enabling governments and space agencies to develop mitigation strategies in case of an impending impact threat.
- Planetary Defense: Understanding asteroid orbits and potential impact scenarios is the first step in developing planetary defense strategies to prevent catastrophic impacts.
- Scientific Research: The study of asteroids contributes to our understanding of the solar system's formation and evolution.
- Space Exploration: Identifying and characterizing NEOs also presents opportunities for future asteroid mining and exploration missions.

Given the potential consequences of an asteroid impact, this project focuses on improving our ability to predict, track, and mitigate the threat posed by these space objects, contributing to the safety and well-being of our planet.

2 Methodology

Our methodology for the "Hazardous Asteroid Prediction" project outlines steps and techniques to identify and classify hazardous asteroids using the NeoWs dataset. It's organized into key phases:

2.1 Data Collection and Preprocessing

In this initial phase, we gathered data from the NeoWs API, which provides comprehensive information about near-Earth asteroids. We have utilized the API's capabilities to search for asteroids based on their closest approach dates, retrieve specific asteroid information, and access the complete dataset of 4687 data points with 40 attributes. Effective data pre-processing was essential to ensure data quality and suitability for analysis. These are explained in detail under the data preprocessing sub-section in our "experiment's setup" section.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a critical step to gain insights into the characteristics of the data set. We performed the following:

1. We converted the 'Hazardous' column from boolean values to numerical values, where 'Hazardous' is mapped to 1 and 'Not Hazardous' to 0.
2. Dropped several columns that are considered redundant or unrelated to the analysis.
3. A correlation heatmap using Seaborn is portrayed to identify the correlation between different features in the dataset.

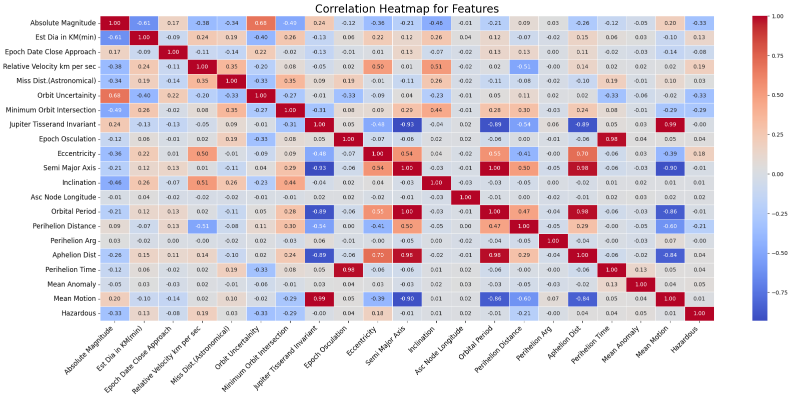


Figure 1: Correlation Heatmap

4. Box plots are generated to detect and visualize potential outliers in the data.
5. Implemented functions to apply the flooring and capping method to treat outliers in the numerical columns of the dataset. This method replaces extreme values with the lower or upper whiskers of the box plots.
6. Applied transformations to specific columns to reduce skewness in the data.
7. Visualization of the class distribution in the 'Hazardous' column is done and the problem of class imbalance is addressed using SMOTE - Synthetic Minority Oversampling Technique.

Conclusively, we computed statistics to understand data distributions and create visualizations such as boxplots, heatmaps, and correlation matrices to identify patterns and relationships in the data.

2.3 Model Development and Evaluation

This phase involves building predictive models to classify asteroids as either hazardous or non-hazardous. We have used several machine learning algorithms to create and evaluate models based on the preprocessed data.

1. We have imported various libraries for machine learning, including scikit-learn models, evaluation metrics, and tools for hyper-parameter tuning.
2. Evaluation metrics learned through our course like accuracy, recall, precision, and F1-score for a given model to create a confusion matrix.
3. Models such as Decision Trees, Random Forrest, KNN, and AdaBoost were implemented.
4. Hyper-parameter tuning was done to see any possible optimizations in the models for better performance. Evaluation metrics for the same were provided as well.

2.4 Model Validation

To assess the performance of our models, we loaded the dataset, preprocessed the data and then ran our model on this whole new data. The model evaluation summary is included in the report under the results section.

3 Experiment

In our project, "Hazardous Asteroid Prediction," our primary objective was to determine whether a particular asteroid is on a collision course with Earth and, based on this assessment, classify the asteroid as dangerous or not.

3.1 Dataset

NeoWs (Near Earth Object Web Service) is a RESTful web service for near-earth Asteroid information. With NeoWs a user can: search for Asteroids based on their closest approach date to Earth, look at a specific Asteroid with its NASA JPL small body ID, as well as browse the overall data set. The dataset has 40 attributes as such:

- Neo Reference ID
- Name
- Absolute Magnitude
- Est Dia in KM(min)
- 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)
- Close Approach Date
- Epoch Date Close Approach
- Relative Velocity km per sec
- Relative Velocity km per hr
- Miles per hour
- Miss Dist.(Astronomical)
- Miss Dist.(lunar)
- Miss Dist.(kilometers)
- Miss Dist.(miles)
- Orbiting Body
- Orbit ID
- Orbit Determination Date
- Orbit Uncertainty
- Minimum Orbit Intersection
- Jupiter Tisserand Invariant
- Epoch Osculation
- Eccentricity
- Semi Major Axis
- Inclination
- Asc Node Longitude
- Orbital Period
- Perihelion Distance
- Perihelion Arg
- Aphelion Dist
- Perihelion Time
- Mean Anomaly
- Mean Motion
- Equinox
- Hazardous

The dataset has 4687 records that have all the relevant data according to the attributes defined above. The Hazardous attribute is the class label for this dataset and the project. The NeoWs API empowers users with several capabilities, allowing them to search for asteroids based on their closest approach date to Earth, retrieve detailed information about a specific asteroid using its NASA JPL small body ID, and explore the comprehensive dataset. The data is sourced from (http://neo.jpl.nasa.gov/), and this API is diligently maintained by the SpaceRocks Team, including David Greenfield, Arezu Sarvestani, Jason English, and Peter Baunach.

3.2 Questions of interest

1. What characteristics of an asteroid from the dataset are more important than the others?
2. How do we predict future close approaches of potentially hazardous asteroids?
3. What data analyzing techniques can be used to get a better understanding of these asteroids?
4. How do the monitoring agencies define the attributes of an asteroid?

3.3 Data Preprocessing

1. Handling Missing Data: We had to carefully address missing values in our dataset. Depending on the extent of missing data, we had to employ appropriate imputation techniques or consider the removal of rows with significant missing information.
2. Duplicate Data: To maintain data integrity, we identified and eliminated duplicate entries. This ensured that our analyses were not skewed by redundant information.
3. Outlier Management: Outliers can significantly impact our project's results. We implemented robust statistical methods to detect and handle outliers, ensuring that they don't distort our analysis and modeling.

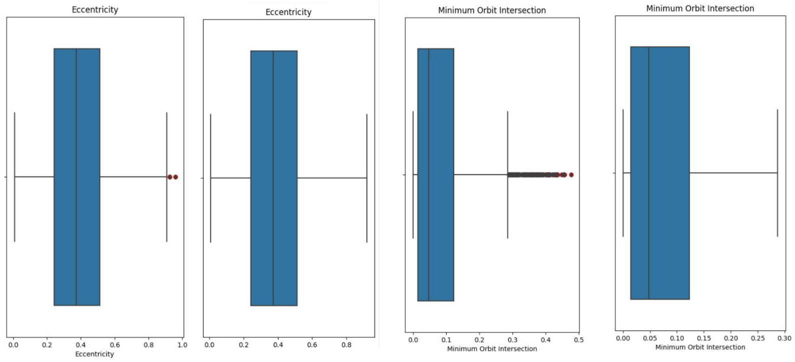


Figure 2: Treatment of Outliers

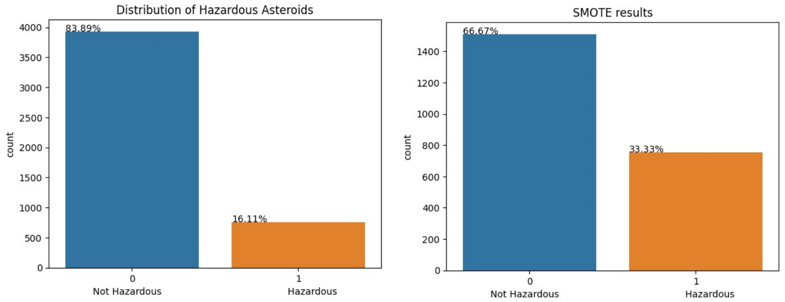


Figure 3: Class Imbalance

4. Dropping Columns: To reduce the dimensionality of the dataset, we have dropped several columns that were redundant. For instance, the estimated diameter in KM, miles, feet, and meters are the same with different units. So we decided to drop these columns to reduce the dimensionality before moving on to the feature selection step.
5. Feature Selection/Extraction: Given the complexity of our dataset, we utilized feature selection techniques to pinpoint and retain the most relevant features. This step reduces dimensionality and optimizes our modeling process. This was done based on a correlation heat map which we performed and noted down the major features which was impacting the hazardous feature.

6. Standardization and normalization are integral to our project, as they ensure that our numerical data maintains a consistent scale.
7. Class Imbalance: Our dataset had a class imbalance where hazardous asteroids were way less than the non-hazardous asteroids. To make a fair analysis and assessment we had to over-sample the minority class. We implemented SMOTE to achieve the same.

3.4 Data Mining

1. Decision tree: Decision tree models are used for classification and regression tasks where interpretability is essential. Decision tree models use entropy, Gini impurity, and information gain to partition data based on feature values, creating a structured set of decision rules. They are applied in classification and regression tasks, emphasizing interpretability and suitability for mixed data types. They excel in feature importance analysis, scalability, and rule extraction. Their flexibility in handling nonlinear relationships contributes to their value across domains. However, precautions against overfitting, such as pruning and ensemble methods, are commonly employed for enhanced generalization.
2. Random Forests: Random Forests, an ensemble of decision trees, mitigate overfitting and improve generalization. Crucial parameters like a number of estimators and class weight impact optimization. The model excels in interpretability, is insensitive to feature scaling, and accommodates numerical and categorical data. Challenges like computational complexity and susceptibility to noisy data exist. Nonetheless, Random Forests offers a robust and powerful solution, striking a balance between transparency and performance in machine learning applications. Their versatility and ability to handle diverse data types make them a favored choice in practical scenarios.
3. K-nearest Neighbours: The k-nearest neighbors (KNN) algorithm is a method for classifying data points, determining the likelihood of a data point belonging to a specific group by considering the group memberships of its closest neighbors. This algorithm falls under supervised machine learning and is applied to solve both classification and regression tasks.
4. AdaBoost: AdaBoost is an ensemble learning algorithm that combines multiple weak learners to form a robust and stronger model. The optimization of AdaBoost models involves adjusting key parameters, such as number of estimators, which governs the quantity of weak learners integrated into the ensemble. AdaBoost is distinguished by its capacity to enhance model accuracy through iterative adjustments, focusing on instances where misclassifications occur. Although it is susceptible to noisy data and outliers it gave us the best result out of all the models with an accuracy of 94 percent.

3.5 Interpretation and evaluation

In the context of our project - “Hazardous Asteroid Prediction”, prioritizing the recall value and accuracy was crucial for assessing the effectiveness of the predictive model.

Recall is vital for our hazardous asteroid prediction model, measuring its ability to identify hazardous instances accurately. A high recall minimizes the chance of missing dangerous asteroids, prioritizing sensitivity for comprehensive hazard identification. This focus on recall enhances system reliability, crucial for averting the severe consequences of false negatives—failure to predict a hazardous asteroid.

While accuracy is important, in hazardous asteroid predictions, a higher emphasis on recall is justified. Missing a hazardous asteroid (false negative) poses a greater risk than misclassifying a non-hazardous one (false positive).

In summary, for our project - hazardous asteroid predictions, the focus on recall and accuracy is important. A high recall value ensures a thorough identification of potential threats, while accuracy provides an overall measure of the model’s correctness. This balanced emphasis aims to create a prediction system that is both sensitive to the identification of hazardous asteroids and generally accurate in its assessments.

In our project recall and accuracy values were best given by the AdaBoost model. It gave us fewer False negative cases compared to other trained and tested models and also gave a high accuracy compared to the other models that we implemented.

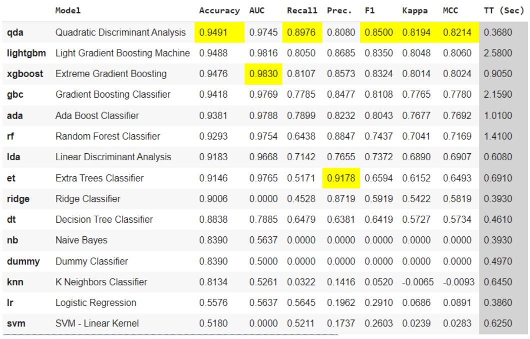
4 Results

In this section, we present a comprehensive overview of the performance of various machine learning models employed in our Hazardous Asteroid Prediction project. A visual representation of model summaries offers a quick snapshot of their respective performances, aiding in a comparative analysis.

Model Summaries: The model summaries visually depict the performance metrics of Decision Trees, KNN, Adaboost, and Random Forests. Notably, Adaboost emerges as the standout performer, showcasing superior accuracy compared to its counterparts. The summaries provide a good understanding of each model's strengths and weaknesses in the context of our specific problem statement.

Adaboost Confusion Matrix: For a detailed examination of Adaboost's performance, we present the confusion matrix, highlighting its ability to correctly classify hazardous and non-hazardous asteroids. The matrix reinforces the robustness of Adaboost, showcasing a high level of precision in its predictions.

In summary, the Adaboost model emerged as the top performer in our Hazardous Asteroid Prediction project, boasting an impressive accuracy of 94 percent. The model's consistency and adaptability were further exemplified through the examination of its confusion matrix and learning curve, reinforcing its efficacy in the accurate identification of hazardous asteroids.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
qda	Quadratic Discriminant Analysis	0.9491	0.9745	0.8976	0.8080	0.8500	0.8194	0.8214	0.3680
lightgbm	Light Gradient Boosting Machine	0.9488	0.9816	0.8050	0.8685	0.8350	0.8048	0.8060	2.5800
xgboost	Extreme Gradient Boosting	0.9476	0.9830	0.8107	0.8573	0.8324	0.8014	0.8024	0.9050
gbc	Gradient Boosting Classifier	0.9418	0.9769	0.7785	0.8477	0.8108	0.7765	0.7780	2.1590
ada	Ada Boost Classifier	0.9381	0.9788	0.7899	0.8232	0.8043	0.7677	0.7692	1.0100
rf	Random Forest Classifier	0.9293	0.9754	0.6438	0.8847	0.7437	0.7041	0.7169	1.4100
lda	Linear Discriminant Analysis	0.9183	0.9668	0.7142	0.7655	0.7372	0.6890	0.6907	0.6080
et	Extra Trees Classifier	0.9146	0.9765	0.5171	0.9178	0.6594	0.6152	0.6493	0.6910
ridge	Ridge Classifier	0.9006	0.0000	0.4528	0.8719	0.5919	0.5422	0.5819	0.3930
dt	Decision Tree Classifier	0.8838	0.7885	0.6479	0.6381	0.6419	0.5727	0.5734	0.4610
nb	Naive Bayes	0.8390	0.5637	0.0000	0.0000	0.0000	0.0000	0.0000	0.3930
dummy	Dummy Classifier	0.8390	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4970
knn	K Neighbors Classifier	0.8134	0.5261	0.0322	0.1416	0.0520	-0.0065	-0.0093	0.6450
lr	Logistic Regression	0.5576	0.5637	0.5645	0.1962	0.2910	0.0686	0.0891	0.3860
svm	SVM - Linear Kernel	0.5180	0.0000	0.5211	0.1737	0.2603	0.0239	0.0283	0.6250

Figure 4: Models Summaries

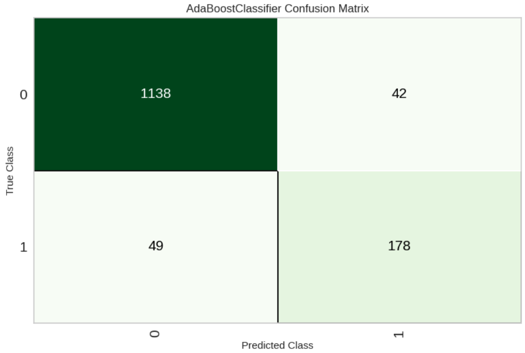


Figure 5: Confusion Matrix of AdaBoost

5 Conclusions

In our pursuit of enhancing Hazardous Asteroid prediction, we undertook a comprehensive approach, beginning with meticulous preprocessing techniques. Outlier removal, handling missing data, column selection, and feature engineering were employed to refine our dataset. To address the class imbalance, SMOTE was implemented, ensuring a more representative model.

The next phase involved dimensionality reduction using Principal Component Analysis (PCA), streamlining the dataset for more efficient model training. Subsequently, four diverse machine

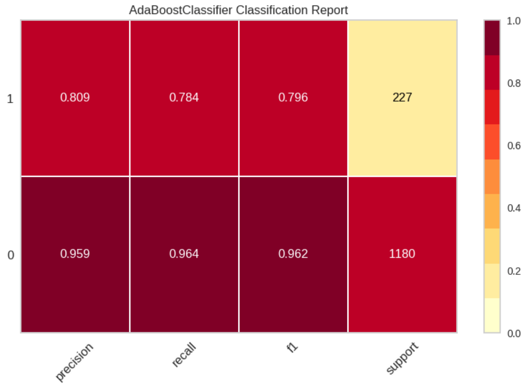


Figure 6: Classification Report of AdaBoost

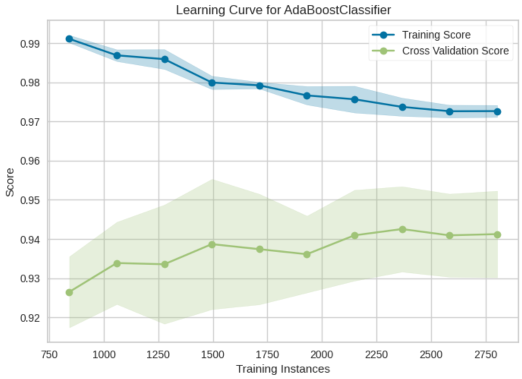


Figure 7: Learning Curve of AdaBoost

learning models—Decision Trees, KNN, AdaBoost, and Random Forests—were implemented to discern their effectiveness in predicting hazardous asteroids.

Upon rigorous evaluation, AdaBoost emerged as the top-performing model, showcasing an impressive 94 percent accuracy and 0.7899 recall rates. This robust performance underscores AdaBoost's effectiveness in our specific problem domain. Notably, the model's ensemble learning approach contributed to its ability to provide accurate predictions. The AUC (Area under Curve) of AdaBoost can be seen as shown in Figure-9.

Our approach, blending preprocessing, dimensionality reduction, and model selection, resulted in a successful Hazardous Asteroid prediction framework. AdaBoost, with high accuracy and recall scores, emerges as the optimal model. This comprehensive method not only advances asteroid prediction understanding but also holds practical implications for space exploration and planetary defense. The attained 94 percent accuracy underscores the methodology's effectiveness, paving the way for robust celestial body detection. The high AUC value signifies that the model effectively balances true positive rates with false positive rates, contributing to its overall robust performance.

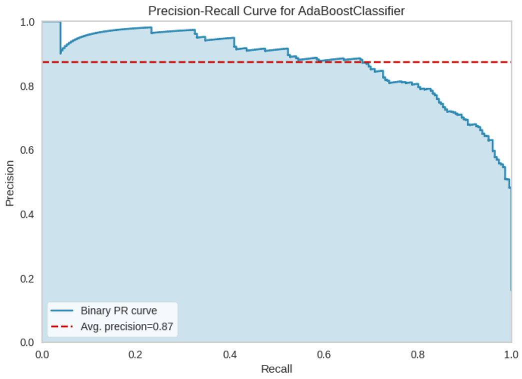


Figure 8: Precision-Recall Curve for AdaBoost

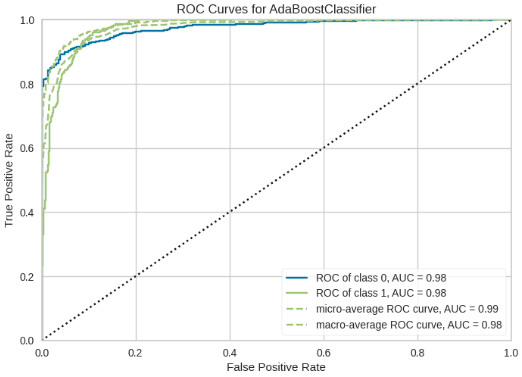


Figure 9: Area under curve of AdaBoost

References

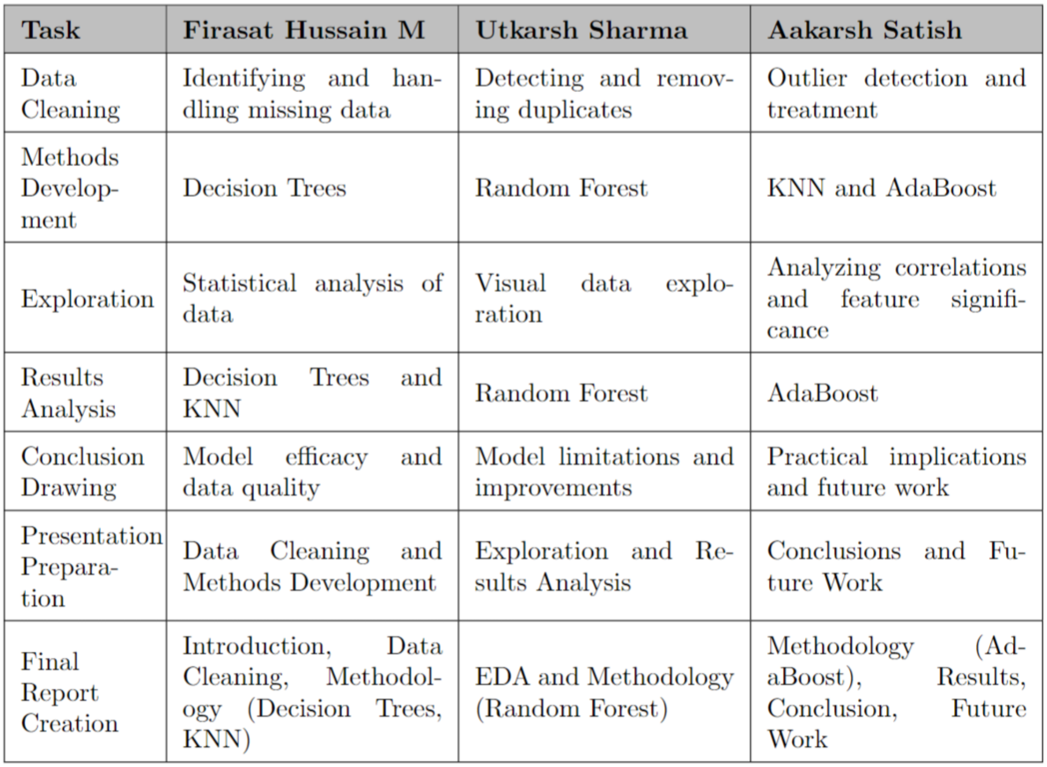
[1] G Alekhya, J Aakanksha, et al. Hazardous asteroid prediction using machine learning. In *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pages 1–6. IEEE, 2023.

[2] Rushir Bhavsar, Nilesh Kumar Jadav, Umesh Bodkhe, Rajesh Gupta, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. Classification of potentially hazardous asteroids using supervised quantum machine learning. *IEEE Access*, 2023.

[3] Muhammad Farae, Cameron Woo, and Anka Hu. An improved approach to orbital determination and prediction of near-earth asteroids: Computer simulation, modeling and test measurements. *arXiv preprint arXiv:2109.07397*, 2021.

GitHub Repo Link: https://github.ncsu.edu/asatish2/engr-ALDA-FALL2023-P13

Equal Work Distribution



Task	Firasat Hussain M	Utkarsh Sharma	Aakarsh Satish
Data Cleaning	Identifying and handling missing data	Detecting and removing duplicates	Outlier detection and treatment
Methods Development	Decision Trees	Random Forest	KNN and AdaBoost
Exploration	Statistical analysis of data	Visual data exploration	Analyzing correlations and feature significance
Results Analysis	Decision Trees and KNN	Random Forest	AdaBoost
Conclusion Drawing	Model efficacy and data quality	Model limitations and improvements	Practical implications and future work
Presentation Preparation	Data Cleaning and Methods Development	Exploration and Results Analysis	Conclusions and Future Work
Final Report Creation	Introduction, Data Cleaning, Methodology (Decision Trees, KNN)	EDA and Methodology (Random Forest)	Methodology (AdaBoost), Results, Conclusion, Future Work

Table 1: Equal Work Distribution Among Team Members