Introduction

Space, in its boundless vastness, is a realm of both breathtaking beauty and profound mystery. It is home to countless celestial marvels—glittering stars that light up the night sky, majestic planets with swirling atmospheres, and awe-inspiring nebulae that serve as the cradles of new stars. Galaxies billions of light-years away form intricate cosmic tapestries that stretch across the universe, each a testament to the grandeur and scale of space. Yet, amid this cosmic splendor, space also conceals entities that, despite their often diminutive size compared to the vastness of the universe, hold the potential to bring about unimaginable destruction.

Among these entities are Near Earth Objects (NEOs), which include asteroids and comets with orbits that bring them perilously close to Earth. These celestial bodies, while fascinating, are not merely objects of scientific curiosity; they are potential harbingers of disaster. Throughout Earth's history, such impacts have caused mass extinctions and significant geological changes, highlighting the importance of monitoring these cosmic neighbors.

NASA, the National Aeronautics and Space Administration, is the United States government agency responsible for the nation's civilian space program and for aeronautics and aerospace research. Since its establishment in 1958, NASA has been at the forefront of space exploration, conducting pioneering missions to understand our solar system and beyond. One of NASA's critical areas of focus is the study and monitoring of near-Earth objects (NEOs), which are asteroids and comets that come within 1.3 astronomical units of Earth. NASA's Planetary Defense Coordination Office (PDCO) is tasked with detecting, tracking, and characterizing NEOs, as well as developing strategies to mitigate any potential impact threats.

As of 2024, data on NEOs has been continuously updated and expanded, encompassing a wide range of observations from various space missions and telescopes. This dataset, which spans from 1910 to 2024, offers a comprehensive overview of NEO encounters with Earth, providing valuable information for researchers, astronomers, and policymakers.

Following is an analysis of this dataset which was created using NASA’s open source API, and made available (with great thanks) in https://www.kaggle.com by IvanSher. The dataset is titled “nearest-earth-objects (1910-2024)”.

The dataset contains entries of 338199 asteroids. This dataset originally contained 9 features, but the feature labelled *orbiting_body* has been dropped as it contained only 1 unique value i.e. "Earth". Thus, the dataset now contains 8 features and their following information are mentioned respectively.

Column Name	Unit	Description	Data type
neo_id	nil	Unique identifier assigned to each near-Earth object	Integer
name	nil	Name or designation of the near-Earth object	String
absolute_magnitude	watts	The absolute magnitude (brightness) of the NEO, which indicates its size	Float
estimated_diameter_min	kilometers	The estimated minimum diameter of the NEO	Float
estimated_diameter_max	kilometers	The estimated maximum diameter of the NEO	Float
relative_velocity	kilometers per hour	The speed at which the NEO is moving relative to Earth, measured	Float
miss_distance	kilometers	The closest distance that the NEO will approach Earth, measured	Float
is_hazardous	nil	Indicates whether the NEO is potentially hazardous to Earth	Boolean

Basic Statistics (Part 1)

	neo_id	name	absolute_magnitude	estimated_diameter_min	estimated_diameter_max	relative_velocity	miss_distance	is_hazardous
count	3.381990e+05	338199	338171.000000	338171.000000	338171.000000	338199.000000	3.381990e+05	338199
unique	NaN	33514	NaN	NaN	NaN	NaN	NaN	2
top	NaN	277810 (2006 FV35)	NaN	NaN	NaN	NaN	NaN	False
freq	NaN	211	NaN	NaN	NaN	NaN	NaN	295037
mean	1.759939e+07	NaN	22.932525	0.157812	0.352878	51060.662908	4.153535e+07	NaN
std	2.287225e+07	NaN	2.911216	0.313885	0.701869	26399.238435	2.077399e+07	NaN
min	2.000433e+06	NaN	9.250000	0.000511	0.001143	203.346433	6.745533e+03	NaN
25%	3.373980e+06	NaN	20.740000	0.025384	0.056760	30712.031471	2.494540e+07	NaN
50%	3.742127e+06	NaN	22.800000	0.073207	0.163697	47560.465474	4.332674e+07	NaN
75%	5.405374e+07	NaN	25.100000	0.189041	0.422708	66673.820614	5.933961e+07	NaN
max	5.446281e+07	NaN	33.580000	37.545248	83.953727	291781.106613	7.479865e+07	NaN

A quick initial summary statistic of the dataset reveals the following facts:

Total number of rows is 338,199 but there exists only 33,514 unique asteroids (on the basis of *name* feature) in the dataset. The NaN in other columns suggest that there may be more than one value of other features for a single named asteroid. The most occurring asteroid is 277810 (2006 FV35) with a frequency of 211. On an average, asteroids had 22.93 watts absolute magnitude, 0.157812 km and 0.352878 km estimated minimum and maximum diameter respectively, 51060.66 km/h relative velocity and 4.15e+07 km miss distance.

The large differences between the maximum and minimum values for each of the numeric features (excluding *neo_id*) implies that all of them have high variances which suggests that standardization may be required for comparing the different features with one another.

The mean values for estimated diameters are close to each other (approximately 0.158 and 0.353, respectively), indicating a narrow range of uncertainty in the estimated sizes. However, the maximum values differ more significantly (37.55 for *estimated_diameter_max*).

The *miss_distance* column has a mean of 41,533,535 km, with a wide range from 6,745.54 km to 74,798,567 km. This suggests that some NEOs pass very close to Earth, while others are much farther away.

The majority of asteroids in the dataset, with a frequency of 295,037, are non-hazardous, indicating that while near-Earth objects are carefully monitored, most do not pose a significant threat. However, this distribution creates a data imbalance, which can potentially mislead algorithms during training and testing, impacting the accuracy of predictive models.

These statistics highlight the diversity in size, speed, and proximity of NEOs, with a small proportion of them being potentially hazardous. The missing values in some columns might require special attention during analysis.

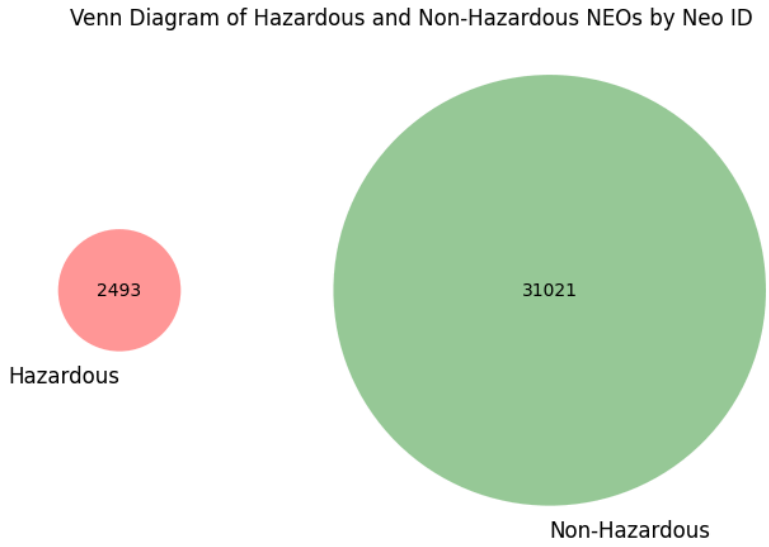
Exploratory Data Analysis (Part 1)

There are no missing values but there are null values present in some features as listed below.

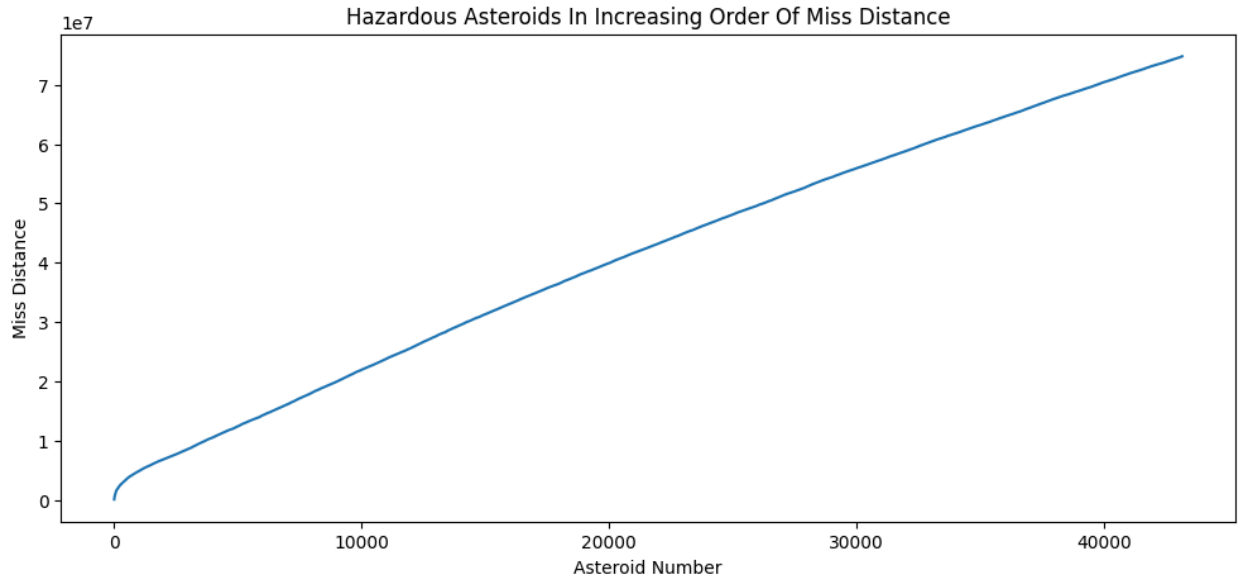
Features	Total null values present
absolute_magnitude	28
estimated_diameter_min	28
estimated_diameter_max	28

Fortunately, these missing values correspond to the same asteroid entries, making it straightforward to handle them. By dropping the rows containing these null values, we can ensure that the dataset remains consistent without affecting other data points.

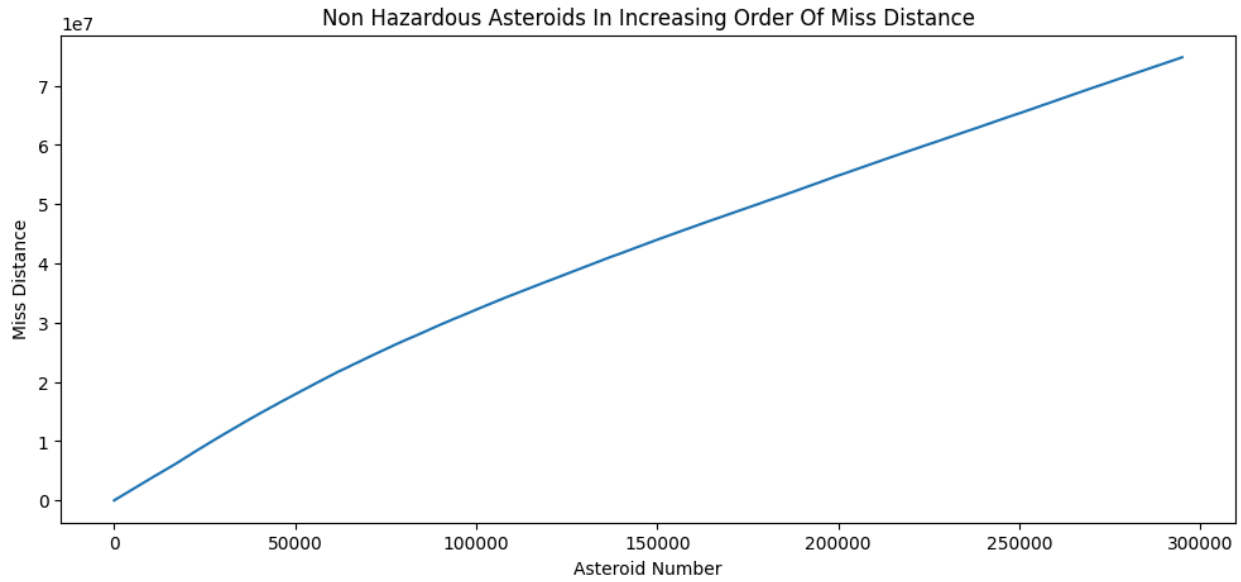
In the dataset, there does not exist a single asteroid which is both hazardous and non hazardous. To reach this conclusion we create 2 sets, one containing unique *neo_id* of asteroids which are hazardous and the other containing the non hazardous ones. We later create a venn diagram using the length of the 2 sets and their intersection, only to find that there is no intersection between the 2 sets.



The maximum miss distance for a potentially hazardous asteroid is 74796655.68197332 km.

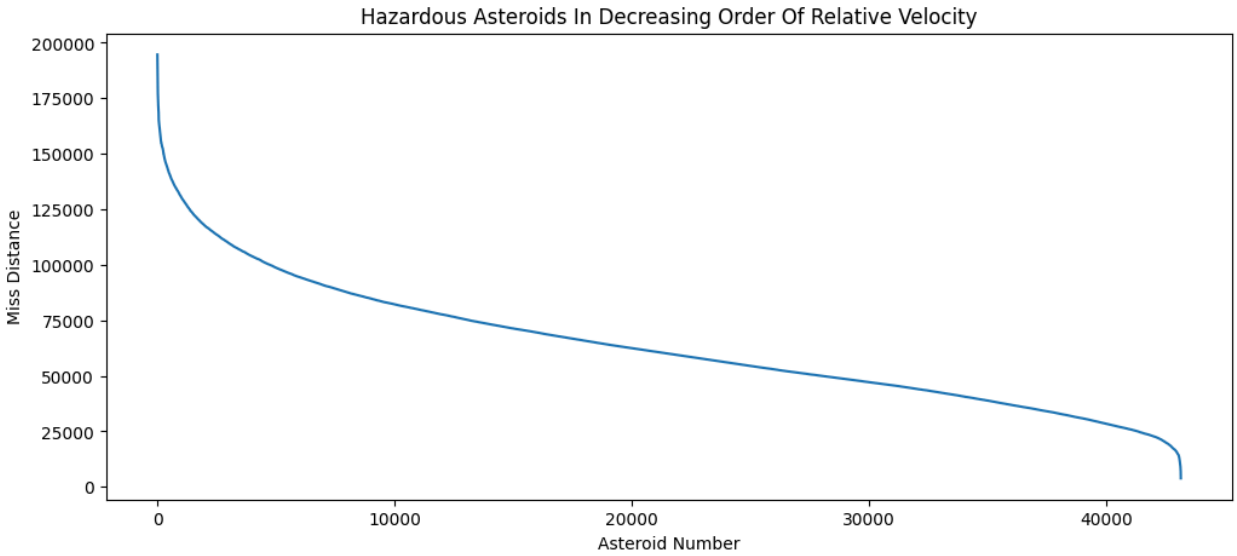


The minimum miss distance for a potentially non hazardous asteroid is 6745.532515957 km.

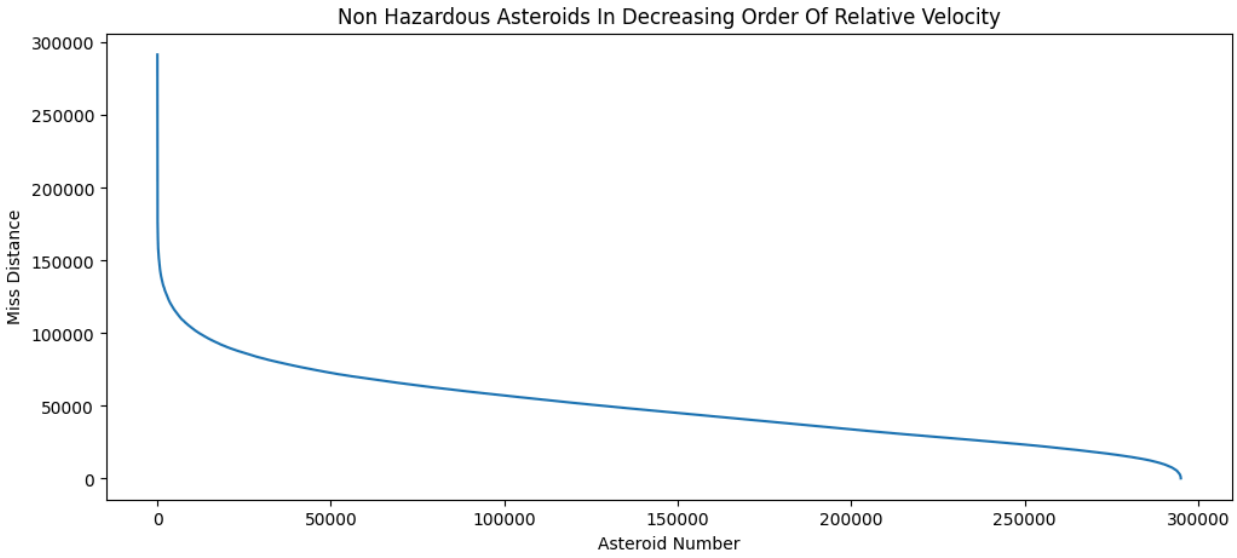


If we consider that hazardousness of an asteroid depends solely on miss distance, then any distance lower or equal than the maximum distance of hazardous asteroids should be considered hazardous. But it is not so, we can see that the minimum value of non hazardous is much lower than the max hazardous distance, which is a contradiction.
Thus using the above observations, we conclude that the hazardous feature does not solely depend on miss distance.

The minimum relative velocity for a hazardous asteroid is 3888.6028125806 km/hr.



The maximum relative velocity for a non hazardous asteroid is 291781.1066131202 km/hr.



If we consider that hazardousness of an asteroid depends solely on relative velocity, then any velocity higher or equal than the minimum velocity of hazardous asteroids should be considered hazardous. But it is not so, we can see that the maximum value of non hazardous is much higher than the min hazardous relative velocity, which is a contradiction.

Thus using the above observations, we conclude that the hazardous feature does not solely depend on relative velocity.

Data Remodelling

As mentioned before, this dataset contains multiple rows containing the same asteroid name, neo id and physical properties i.e. there are multiple recordings of the same asteroid but at different relative velocities, which in turn also differ the miss distance at each recording. The initial venn diagram depicts that no same asteroid is both hazardous and non hazardous, this proves that even though an asteroid may have multiple recordings with varying miss distance and relative velocities, they still fall under the same hazardous/non-hazardous class.

For better understanding of every asteroid, we take individual readings of each asteroid and remodel the data to display only 33514 unique rows. Following is the format of the remodeled dataset:

Column Name	Unit	Description	Data type
neo_id	nil	Unique identifier assigned to each near-Earth object	Integer
name	nil	Name or designation of the near-Earth object	String
absolute_magnitude	watts	The absolute magnitude (brightness) of the NEO, which indicates its size	Float
estimated_diameter_min	kilometers	The estimated minimum diameter of the NEO	Float
estimated_diameter_max	kilometers	The estimated maximum diameter of the NEO	Float
minimum_relative_velocity	kilometers per hour	Minimum speed at which the NEO is moving relative to Earth	Float

maximum_relative_velocity	kilometers per hour	Maximum speed at which the NEO is moving relative to Earth	Float
average_relative_velocity	kilometers per hour	Average speed at which the NEO is moving relative to Earth	Float
minimum_miss_distance	kilometers	Minimum closest distance that the NEO will approach Earth	Float
maximum_miss_distance	kilometers	Maximum closest distance that the NEO will approach Earth	Float
average_miss_distance	kilometers	Average closest distance that the NEO will approach Earth	Float
min_miss_distance_classification	nil	The minimum distance at which a NEO will approach Earth	String
is_hazardous	nil	Indicates whether the NEO is potentially hazardous to Earth	Boolean

The miss distance classifications is based upon the following percentile range of *minimum_miss_distance*:
[Here we consider the minimum miss distance for classification as smaller the asteroid distance will be, more will be the probability of that asteroid being hazardous.]

Percentile Range	**Class**
<=25%	Near
>25% and <=50%	Near-Midway
>50% and <=75%	Midway
>75%	Far

Basic Statistics (Part 2)

	neo_id	name	absolute_magnitude	estimated_diameter_min	estimated_diameter_max	minimum_relative_velocity	maximum_relative_velocity
count	3.351100e+04	33511	33511.000000	33511.000000	33511.000000	33511.000000	33511.000000
unique	NaN	33511	NaN	NaN	NaN	NaN	NaN
top	NaN	(2024 NP3)	NaN	NaN	NaN	NaN	NaN
freq	NaN	1	NaN	NaN	NaN	NaN	NaN
mean	2.306254e+07	NaN	23.427700	0.138637	0.310002	37265.089469	63843.182920
std	2.474539e+07	NaN	2.919482	0.354701	0.793136	21590.382682	30808.061447
min	2.000433e+06	NaN	9.250000	0.000511	0.001143	203.346433	1418.218469
25%	3.577491e+06	NaN	21.220000	0.020163	0.045086	21739.359535	40031.091350
50%	3.788052e+06	NaN	23.700000	0.048368	0.108153	33201.685280	62992.086620
75%	5.420045e+07	NaN	25.600000	0.151550	0.338875	48822.221309	82653.513747
max	5.446281e+07	NaN	33.580000	37.545248	83.953727	230051.873217	291781.106613

average_relative_velocity	minimum_miss_distance	maximum_miss_distance	average_miss_distance	min_miss_distance_classification	is_hazardous
33511.000000	3.351100e+04	3.351100e+04	3.351100e+04	33511	33511
NaN	NaN	NaN	NaN	4	2
NaN	NaN	NaN	NaN	Near-Midway	False
NaN	NaN	NaN	NaN	16755	31018
48305.115170	1.617521e+07	5.045405e+07	3.340955e+07	NaN	NaN
23017.282140	1.605806e+07	2.630486e+07	1.737756e+07	NaN	NaN
1418.218469	6.745533e+03	9.316925e+03	9.316925e+03	NaN	NaN
31889.933517	3.779695e+06	2.643296e+07	2.114211e+07	NaN	NaN
45096.713390	1.019570e+07	6.410632e+07	3.674749e+07	NaN	NaN
60926.819396	2.434135e+07	7.207361e+07	4.578375e+07	NaN	NaN
260830.496284	7.478833e+07	7.479865e+07	7.478833e+07	NaN	NaN

Following are the observations that can be derived from the above summary statistic for the remodeled data:

Total number of rows is 33514 which is equal to the total number of unique asteroids in the original dataset. Some columns, like neo_id, name, minimum_relative_velocity, maximum_relative_velocity, average_relative_velocity, and miss distances (minimum, maximum, average), are complete with 33,511 non-null values. However, columns such as *absolute_magnitude, estimated_diameter_min*, and *estimated_diameter_max* have slightly fewer non-null entries (33,511), indicating some missing data.

The *absolute_magnitude* has a mean of around 23.43, with a standard deviation of 2.92, indicating moderate variability in the brightness of the Near-Earth Objects (NEOs). The *estimated_diameter_min* and *estimated_diameter_max* have means of approximately 0.14 and 0.31, respectively, indicating that these objects are relatively small in size on average. The relative velocities (minimum, maximum, average) show significant variation, with mean values ranging from about 37,265 km/h to 63,843 km/h.

The *minimum_miss_distance*, *maximum_miss_distance*, and *average_miss_distance* have mean values around 16,175,210 km, 50,454,050 km, and 33,409,550 km, respectively. These distances suggest that while some NEOs come relatively close to Earth, most stay at considerable distances.
The *min_miss_distance_classification* column has 4 unique categories, with the most frequent category being Near-Midway occurring 16,755 times. The *is_hazardous* column is binary (True/False), with False being the most frequent, indicating that most of the NEOs in this dataset are not considered hazardous.

The standard deviation in the *maximum_relative_velocity* (23,017.28 km/h) and the *maximum_miss_distance* (2.63e+07 km) suggests a wide range of values, indicating a significant diversity in the velocities and distances of these objects from Earth. The NEOs have a wide range of miss distances, with most maintaining a safe distance from Earth.

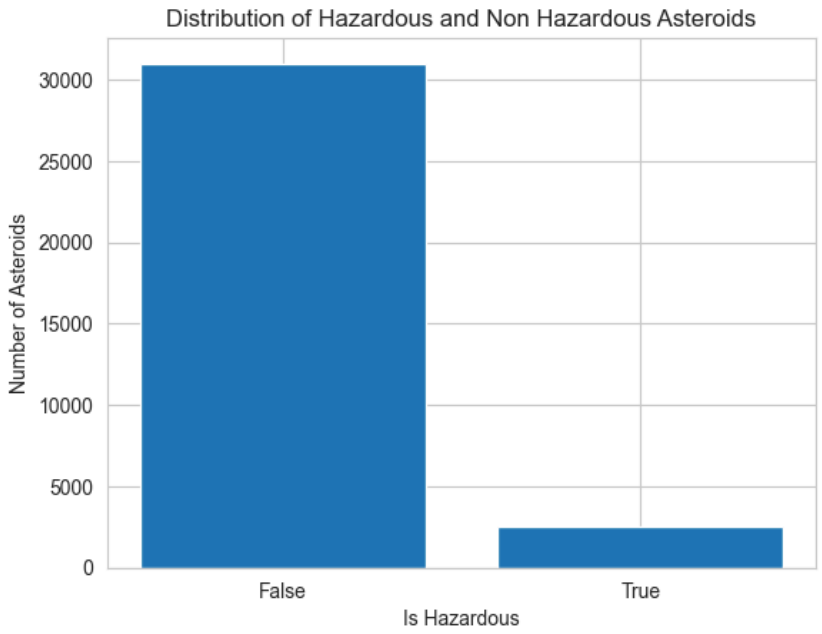
Exploratory Data Analysis (Part 2)

There are no missing values but there are null values present in some features as listed below.

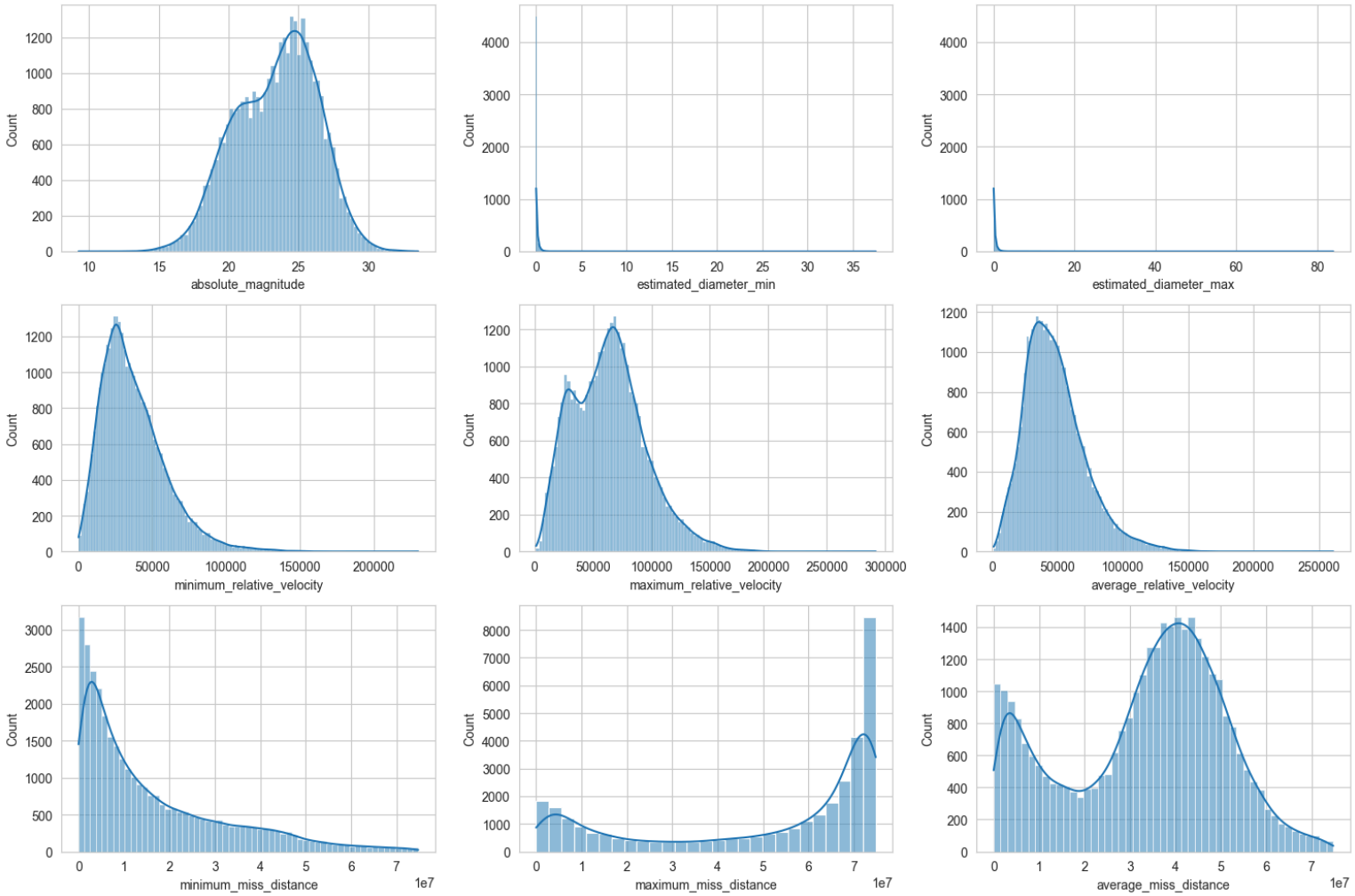
Features	Total null values present
absolute_magnitude	3
estimated_diameter_min	3
estimated_diameter_max	3

Fortunately, these missing values correspond to the same asteroid entries, making it straightforward to handle them. By dropping the rows containing these null values, we can ensure that the dataset remains consistent without affecting other data points.

The remodeled dataset contains 31,018 non hazardous asteroids and 2493 hazardous asteroids. Following is the graph depicting the distribution of hazardous and non hazardous asteroid:



A histogram distribution is then created for all the numeric features against *is_hazardous*:

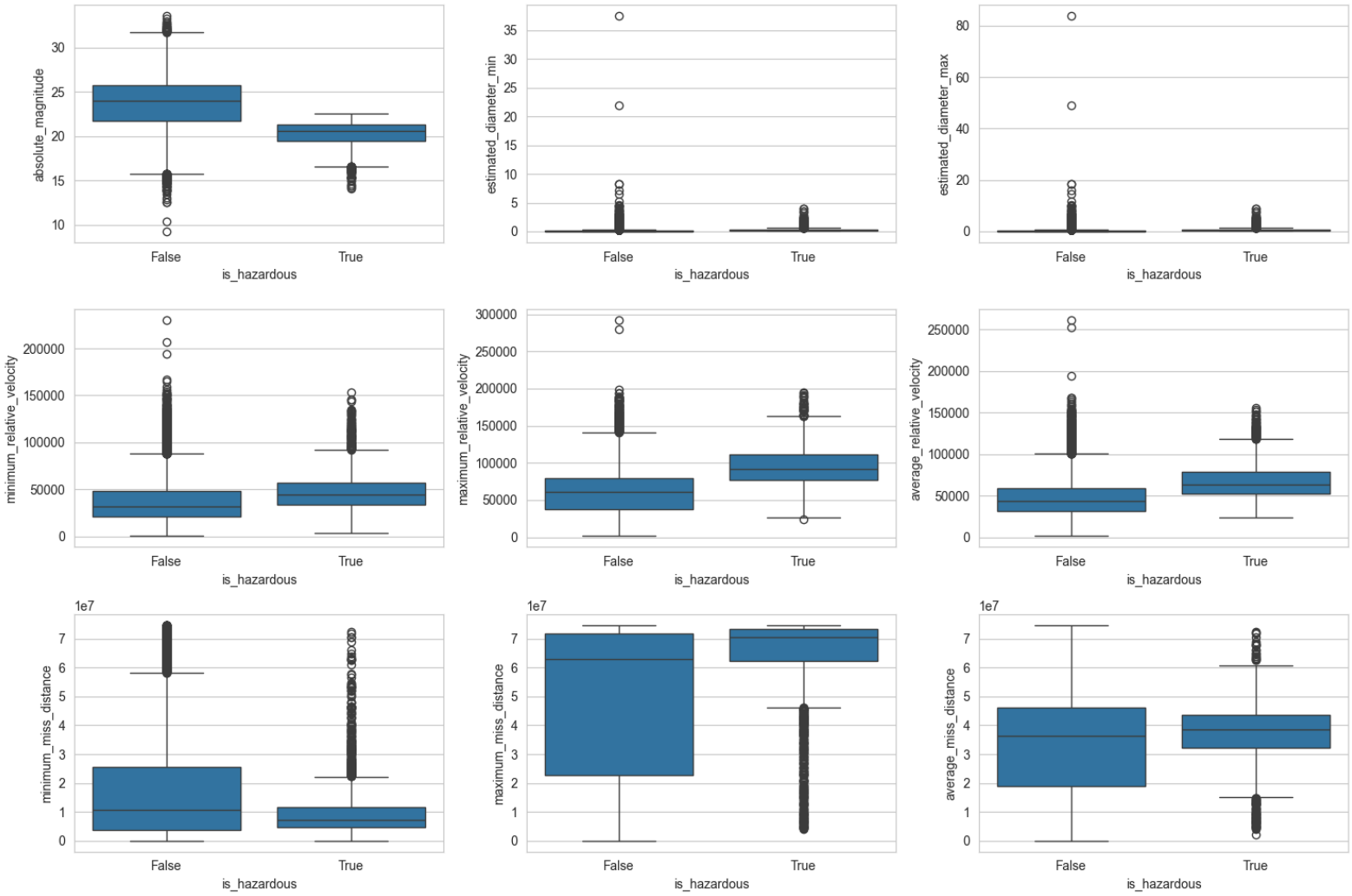


Conclusions derived from the above histogram exploration are:

1. The distribution of absolute magnitude is roughly normal with a peak around 25 which suggests that most objects have an absolute magnitude around 25.
2. The distribution of estimated diameter min is heavily skewed to the right with most values close to zero which indicates that the majority of objects have a very small minimum estimated diameter.
3. Similarly, the distribution of estimated diameter max is heavily skewed to the right with most values close to zero which indicates that most objects have a very small maximum estimated diameter.
4. The distribution of minimum relative velocity is right-skewed, peaking at low values and tapering off as the velocity increases which indicates that most objects have a relatively low minimum relative velocity.
5. The distribution of maximum relative velocity has a peak at lower values and decreases as the velocity increases, with some minor secondary peaks which suggests that most objects have a low maximum relative velocity, but there are some with higher velocities.
6. The distribution of average relative velocity is similar to the minimum and maximum relative velocity, with a peak at lower values and tapering off as the velocity increases. So, most objects have a low average relative velocity.

7. The distribution of minimum miss distance is right-skewed, with a peak at very low values and tapering off as the distance increases which indicates that most objects have a very small minimum miss distance from Earth.
8. The distribution of maximum miss distance has a peak at the higher end, suggesting a large number of objects have a maximum miss distance around 70 million units. There is a clear bimodal distribution, indicating two distinct groups of objects with different maximum miss distances.
9. The distribution of average miss distance is bimodal, with peaks around 10 million and 40-50 million units which indicates that there are two distinct groups of objects with different average miss distances.

Further we create box plot distributions for all of the numeric features against *is_hazardous*:



Conclusions derived from the above boxplot exploration are:

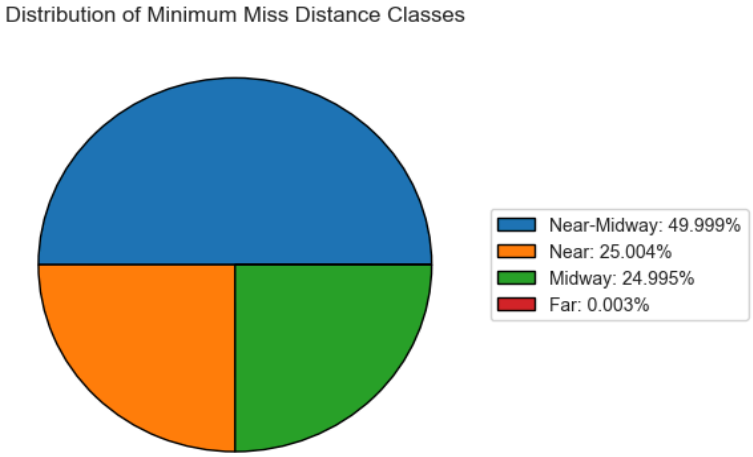
1. Hazardous objects tend to have lower (brighter) absolute magnitudes compared to non-hazardous objects. Lower median absolute magnitude for hazardous objects indicates they are generally brighter.

2. The minimum estimated diameter for both hazardous and non-hazardous objects is generally very small, with some outliers and there is no significant difference between the two categories.
3. Similar to the minimum diameter, the maximum estimated diameter for both categories is generally small, with some outliers and there is no significant difference between the two categories.
4. In the minimum relative velocity distribution, both categories have similar medians and ranges, with hazardous objects having slightly higher outliers.
5. In the maximum relative velocity distribution, both categories have similar medians and ranges, with hazardous objects having slightly higher outliers.
6. In the average relative velocity distribution, both categories have similar medians and ranges, with hazardous objects having slightly higher outliers.
7. Hazardous objects have a smaller minimum miss distance compared to non-hazardous objects. Lower median minimum miss distance for hazardous objects indicates they come closer to Earth.
8. Non-hazardous objects have a larger maximum miss distance compared to hazardous objects. Higher median maximum miss distance for non-hazardous objects indicates they stay farther away from Earth.
9. Hazardous objects have a smaller average miss distance compared to non-hazardous objects. Lower median average miss distance for hazardous objects indicates they come closer to Earth on average.

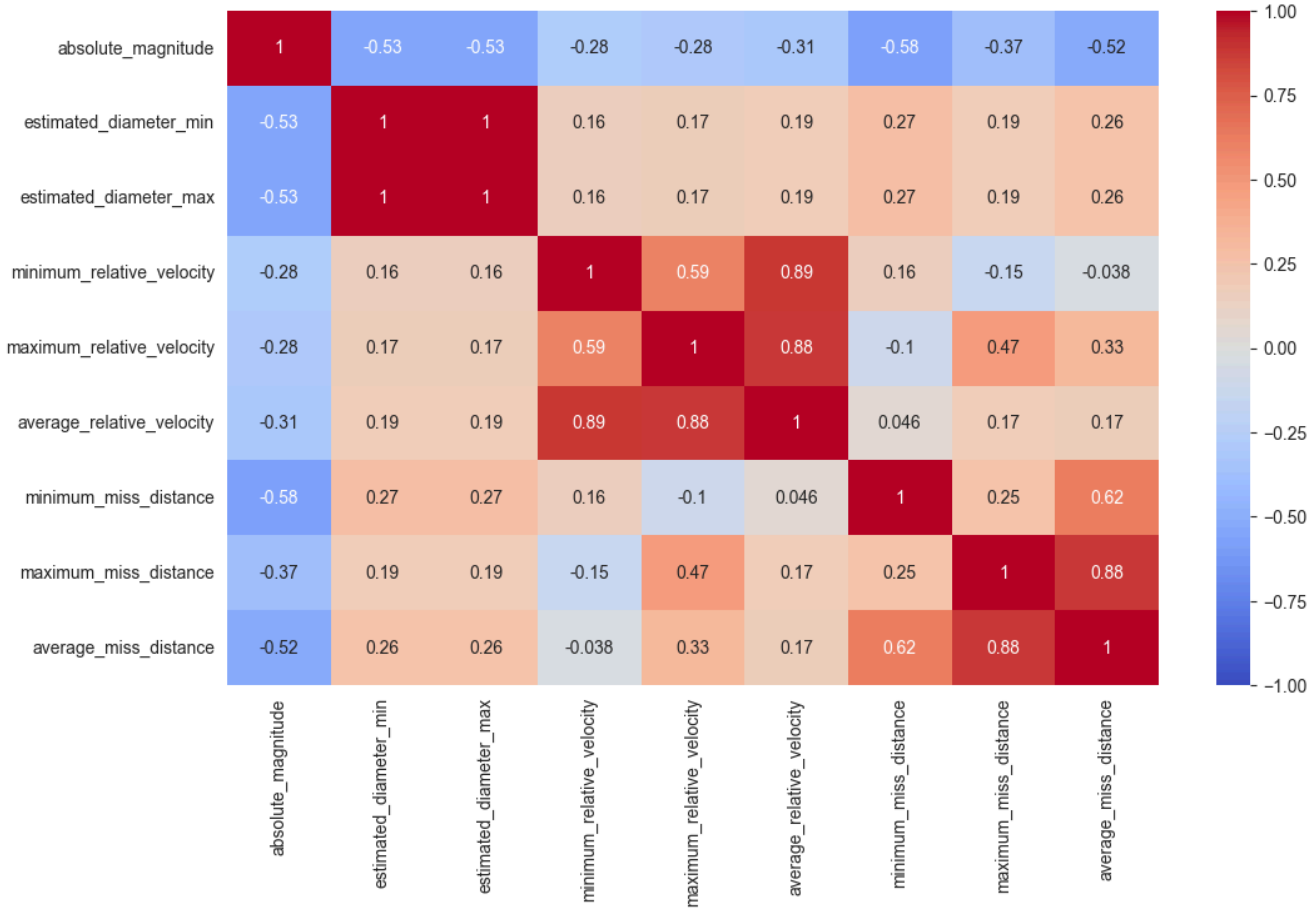
The values of each miss distance classes are as follows:

Class	Count
Near-Midway	16755
Near	8379
Midway	8376
Far	1

The distribution of miss distance classes are as displayed in the following pie graph:



Next we find out the correlation of every numeric feature against each other. Following is the seaborn matrix we get:



Conclusions we derive from the above correlation matrix are:

- There are strong positive correlations between *estimated_diameter_min* and *estimated_diameter_max*, which are perfectly correlated (1.0), as both represent the estimated size of the object. Similarly, *minimum_relative_velocity*, *maximum_relative_velocity*, and *average_relative_velocity* are highly correlated (e.g., *minimum_relative_velocity* and *average_relative_velocity* at 0.89), indicating that as the minimum velocity increases, the average and maximum velocities also tend to increase proportionally. Additionally, *maximum_miss_distance* and *average_miss_distance* have a very strong correlation (0.88), suggesting that the maximum distance at which a NEO could miss Earth is strongly related to the average miss distance.
- Moderate positive correlations are observed between *minimum_miss_distance* and *average_miss_distance* (0.62), indicating that as the minimum miss distance increases, the average miss distance also increases. There is also a moderate positive correlation (0.33) between *maximum_relative_velocity* and *average_miss_distance*, which might suggest that higher maximum relative velocities are somewhat associated with larger miss distances.
- Negative correlations include a strong negative correlation (-0.58) between *absolute_magnitude* and *minimum_miss_distance*, implying that objects with higher absolute magnitude (which means lower brightness) tend to have smaller miss distances, potentially making them more hazardous. Additionally, *absolute_magnitude* shows a

moderate negative correlation (-0.53) with both *estimated_diameter_min* and *estimated_diameter_max*, suggesting that brighter objects (lower absolute magnitude) tend to have larger diameters.

- There are weak or no correlations in some cases, such as between *minimum_miss_distance* and *maximum_relative_velocity* (-0.1), indicating that the relative velocity of the NEO has little to no impact on its minimum miss distance. Similarly, there is a very weak correlation (0.17) between *average_relative_velocity* and *average_miss_distance*, suggesting that the average relative velocity does not significantly influence the average miss distance.

Overall, the strong correlations among velocity measures and among distance measures suggest that these features are closely related, potentially indicating redundant information. The negative correlation between *absolute_magnitude* and miss distance could be useful in assessing the potential hazard level of an NEO, with larger, less bright objects possibly posing a greater risk.

Data Preprocessing Summary

After a thorough exploratory data analysis, we identified the key features that would be instrumental in predicting whether a Near-Earth Object (NEO) is hazardous. To ensure that our model could effectively interpret these features, we performed several preprocessing steps. Categorical data, such as the *min_miss_distance_classification*, was encoded into numerical values, allowing the model to recognize different classifications of NEO distances. We also standardized all features to bring them onto a common scale, ensuring that the model treats each feature equally during training.

By splitting the dataset into training and testing sets, with 80% of the data dedicated to training and 20% reserved for testing, we set up a robust foundation for evaluating our model's performance. With the data now prepped and ready, the next step involves training our classification model to identify hazardous NEOs accurately.

Model Training

Several machine learning models were evaluated to identify the most effective classifier for predicting hazardous NEOs. The models selected include Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Trees, Random Forest, and AdaBoost Classifier. These models were chosen for their diverse approaches to classification, allowing for a thorough comparison of their performance on this dataset. Further analysis was conducted using a confusion matrix, which provided a detailed breakdown of the model's performance. Lastly, the model's performance was quantified using key metrics, including accuracy, precision, recall, and F1 score.

Logistic Regression

Logistic Regression is a widely used linear model for binary classification, predicting the probability of a data point belonging to one of two classes.

The Logistic Regression model was trained on the dataset to learn the relationship between features and the target variable. It then predicted the class labels for the test data.

The model's predictions were compared with the actual values, and the results were visualized using scatter plots and a confusion matrix.

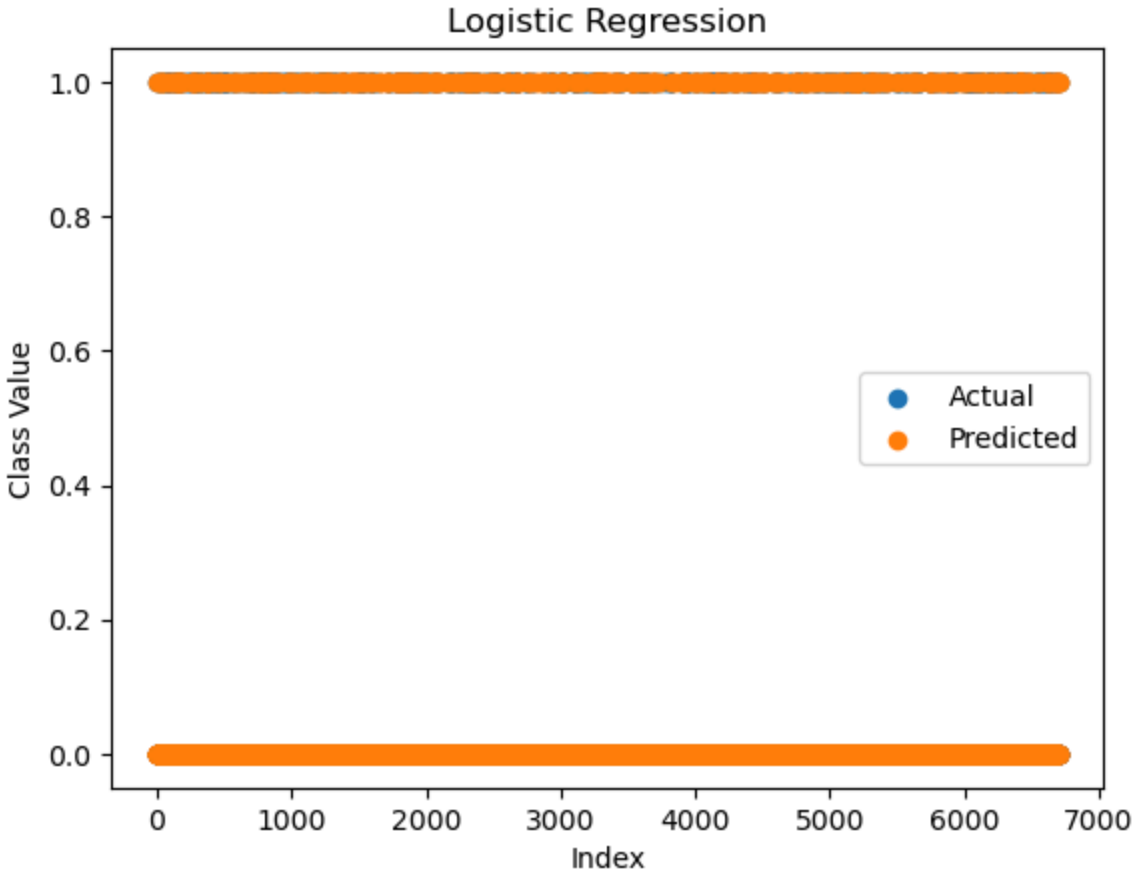


Figure: Scatter plot showing actual vs. predicted values

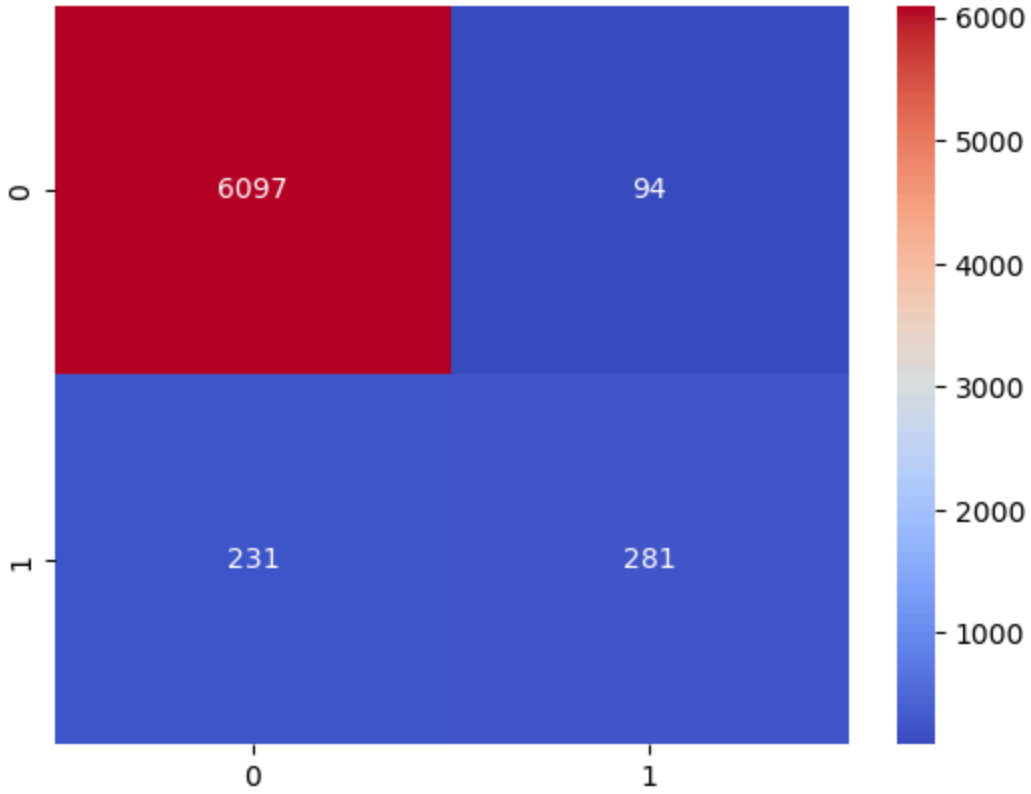


Figure: Confusion Matrix Heatmap

Metric	Value
Accuracy	95.15%
Precision	74.93%
Recall	54.88%
F1 Score	63.36%

The Logistic Regression model performed well, achieving a high accuracy. With the calculated precision, it was effective in making correct positive predictions, though the recall indicates that it missed some true positive cases. The F1 score reflects a moderate balance between precision and recall, suggesting that while the model is accurate, it could benefit from improvements in identifying more true positives.

K-Nearest Neighbors

K-Nearest Neighbors is a simple, non-parametric algorithm used for classification. It classifies a data point based on the majority class among its nearest neighbors in the feature space. KNN is particularly useful when the decision boundary between classes is not linear.

KNN was applied to the dataset, using the training data to identify the nearest neighbors for each data point in the testing set. The model then predicted the class of each test instance by majority vote from its nearest neighbors.

The model's predictions were compared to the actual outcomes, visualized in a scatter plot and further analyzed using a confusion matrix.

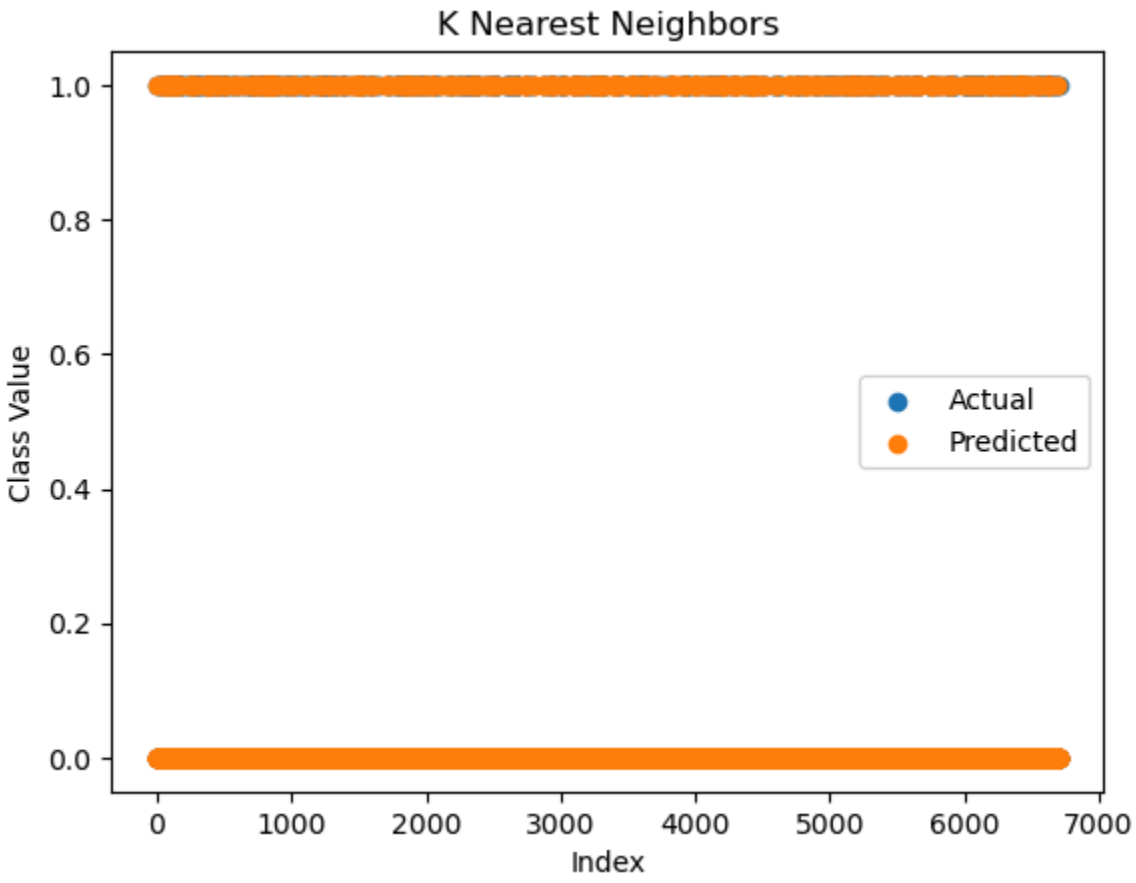


Figure: Scatter plot showing actual vs. predicted values using KNN

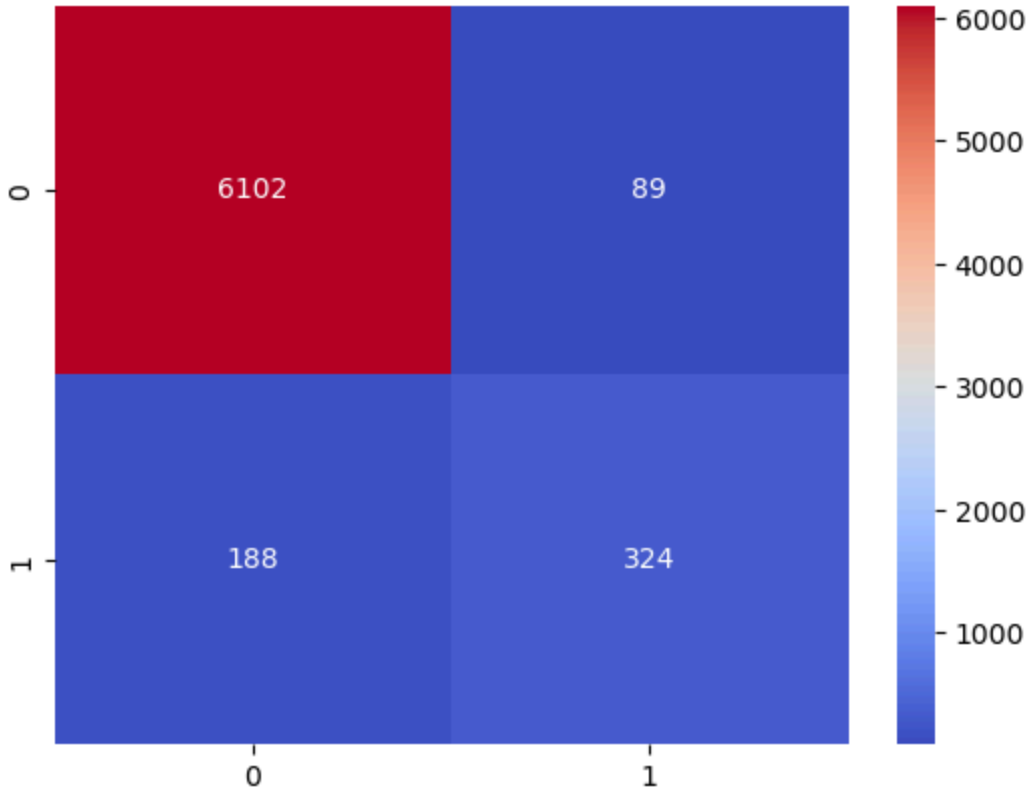


Figure: Confusion Matrix Heatmap for KNN

Metric	Value
Accuracy	95.87%
Precision	78.45%
Recall	63.28%
F1 Score	70.54%

These results suggest that the KNN model performs well, with a good balance between correctly identifying positive cases and minimizing false positives. However, as KNN's performance is sensitive to the choice of 'k' and the distance metric, further tuning could enhance recall, especially if capturing more positive instances is critical.

Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence between features, making it particularly effective for certain types of data, especially when the independence assumption is approximately true.

The Gaussian Naive Bayes algorithm was applied to the dataset, where it calculated the likelihood of each class based on the features and predicted the most probable class for each test instance.

The model's predictions were compared to the actual outcomes, visualized in a scatter plot and further analyzed using a confusion matrix.

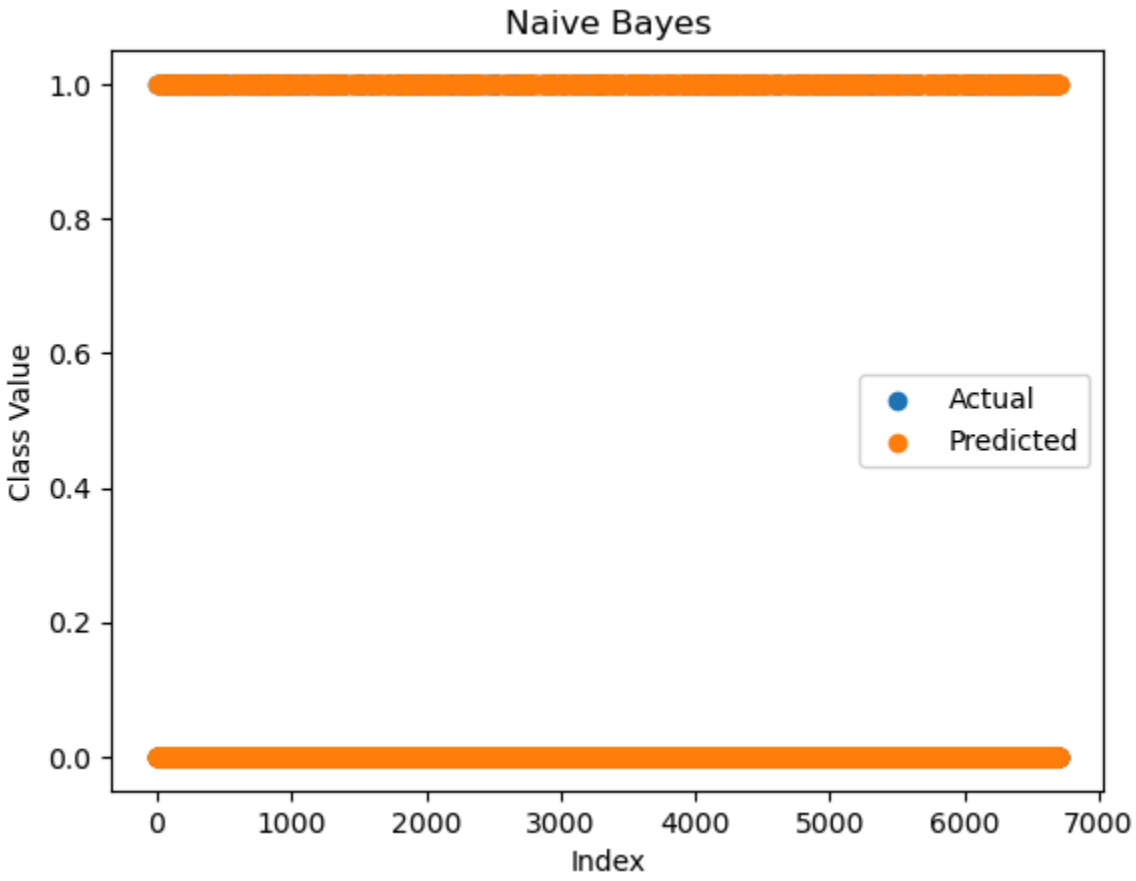


Figure: Scatter plot showing actual vs. predicted values using Naive Bayes

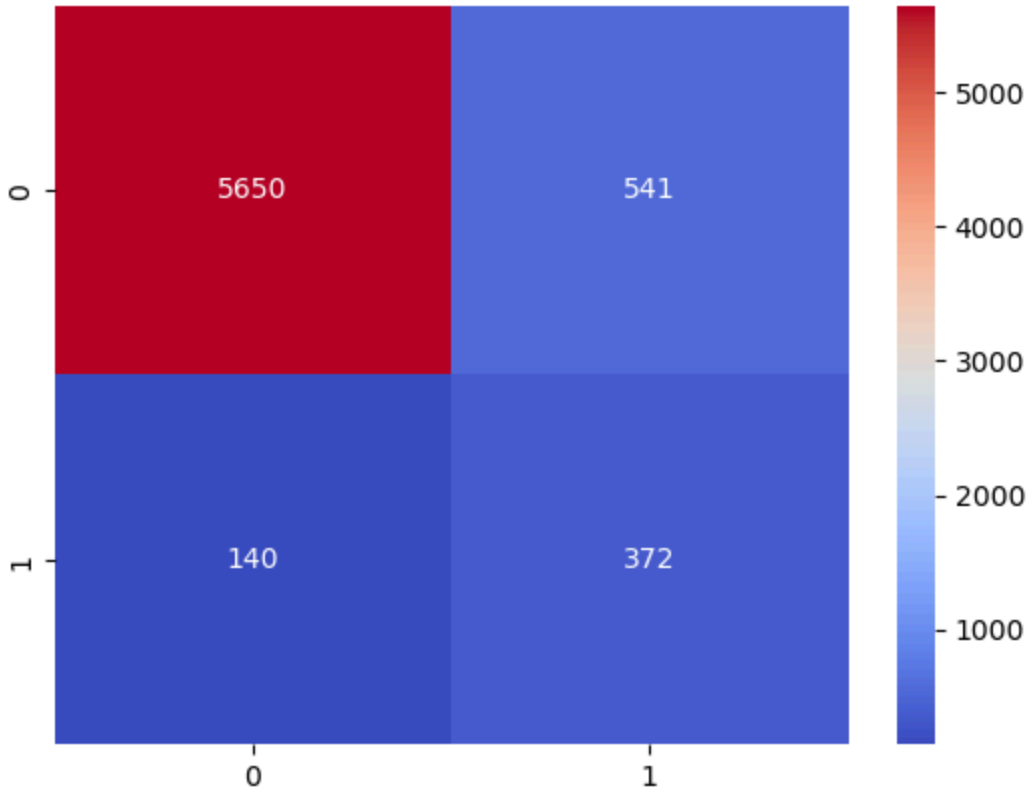


Figure: Confusion Matrix Heatmap for Naive Bayes

Metric	Value
Accuracy	89.84%
Precision	40.74%
Recall	72.66%
F1 Score	52.21%

The Naive Bayes model showed a relatively high recall, indicating it successfully identified the majority of actual positive cases. However, the precision was lower, reflecting a higher rate of false positives. The F1 score suggests a trade-off between precision and recall. These results indicate that while Naive Bayes is effective at capturing positive instances, it tends to predict the positive class more frequently, leading to lower precision. This behavior is typical for Naive Bayes, especially when the independence assumption does not hold strongly in the data.

Decision Tree

A Decision Tree is a versatile machine learning model that splits the data into branches based on feature values, making decisions that lead to predictions. It's particularly useful for its interpretability and ability to handle both categorical and numerical data.

In this analysis, a Decision Tree with a maximum depth of 5 was applied to the dataset. The model was trained to learn decision rules from the training data, which were then used to predict the class of each instance in the test set.

The predicted values were compared with the actual outcomes and visualized through scatter plots and a confusion matrix.

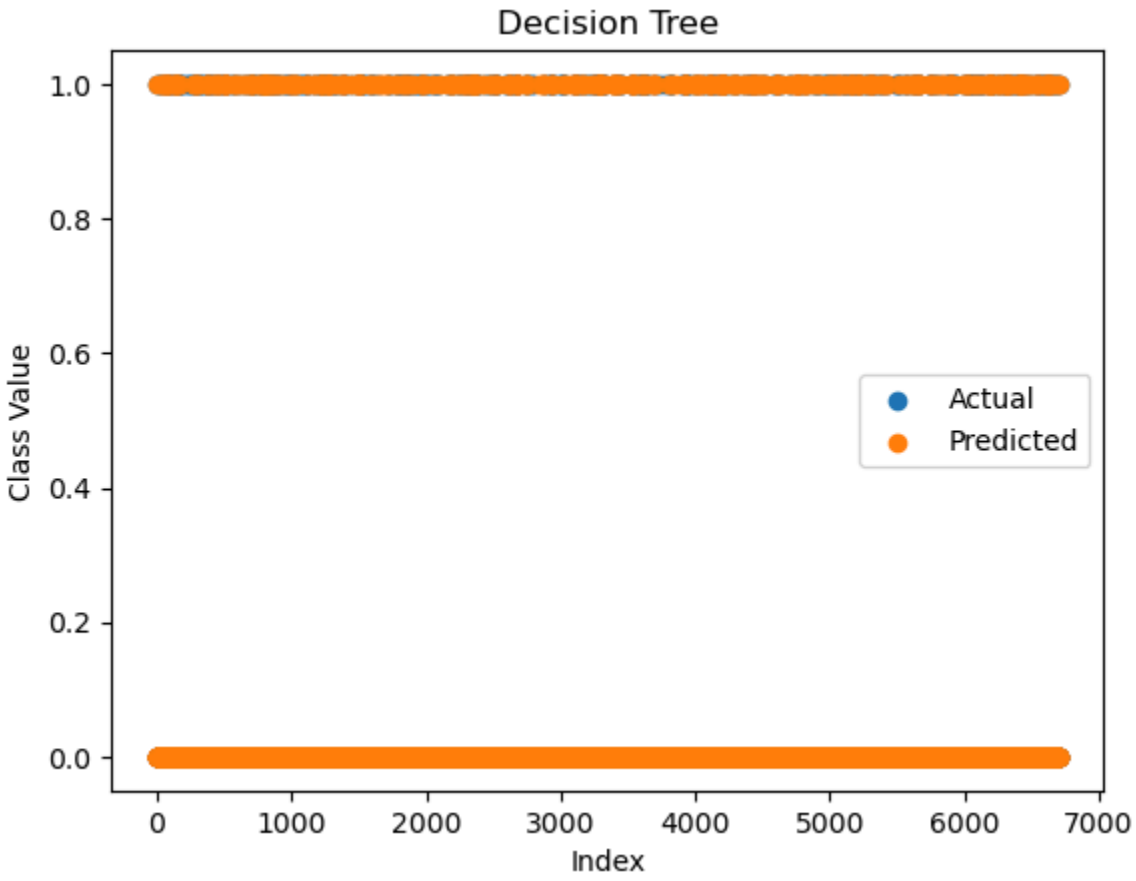


Figure: Scatter plot showing actual vs. predicted values using Decision Tree

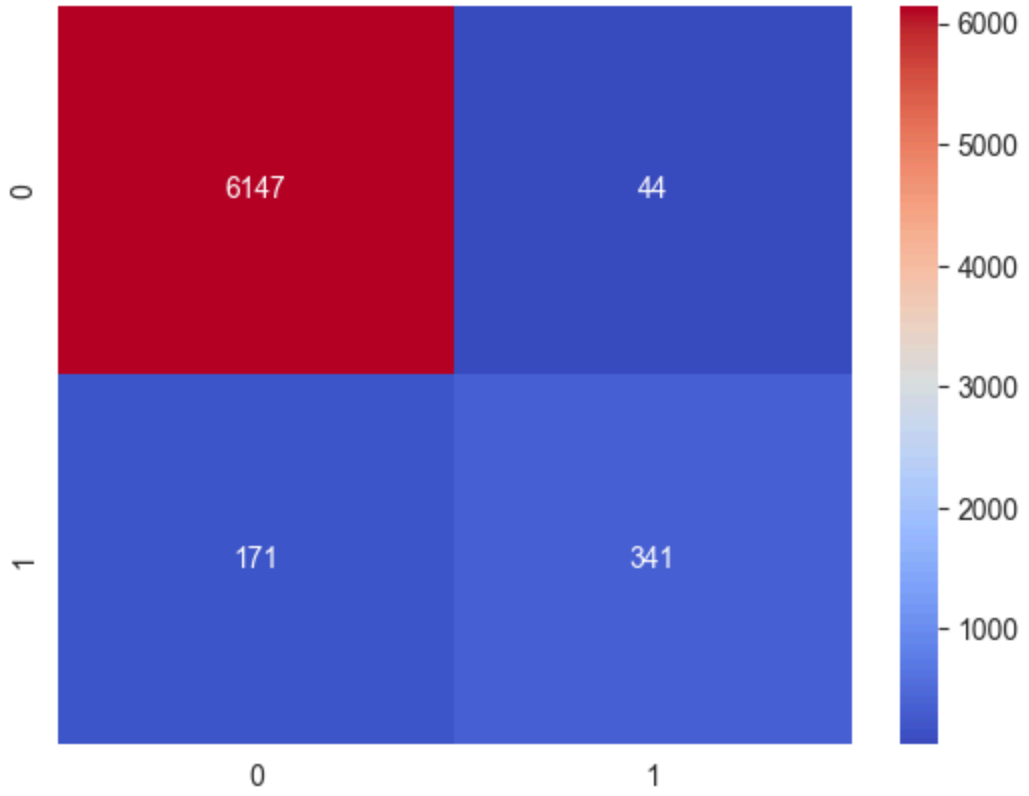


Figure: Confusion Matrix Heatmap for Decision Tree

Metric	Value
Accuracy	96.79%
Precision	88.57%
Recall	66.60%
F1 Score	76.03%

The Decision Tree model delivered strong performance, with a high accuracy and precision, indicating that it was highly effective at making correct positive predictions. The recall shows that it identified most of the actual positive cases, while the F1 score reflects a good balance between precision and recall. The model's performance suggests that it effectively captures the structure of the data, although further tuning or deeper trees might improve recall if more true positives are desired.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their outputs for more accurate and stable predictions. This approach reduces overfitting and improves generalization.

A Random Forest model with a maximum depth of 10 was applied to the dataset. The model was trained on the training data, where it created multiple decision trees and aggregated their predictions to classify the test data.

The model's predictions were compared with the actual values and visualized through scatter plots and a confusion matrix.

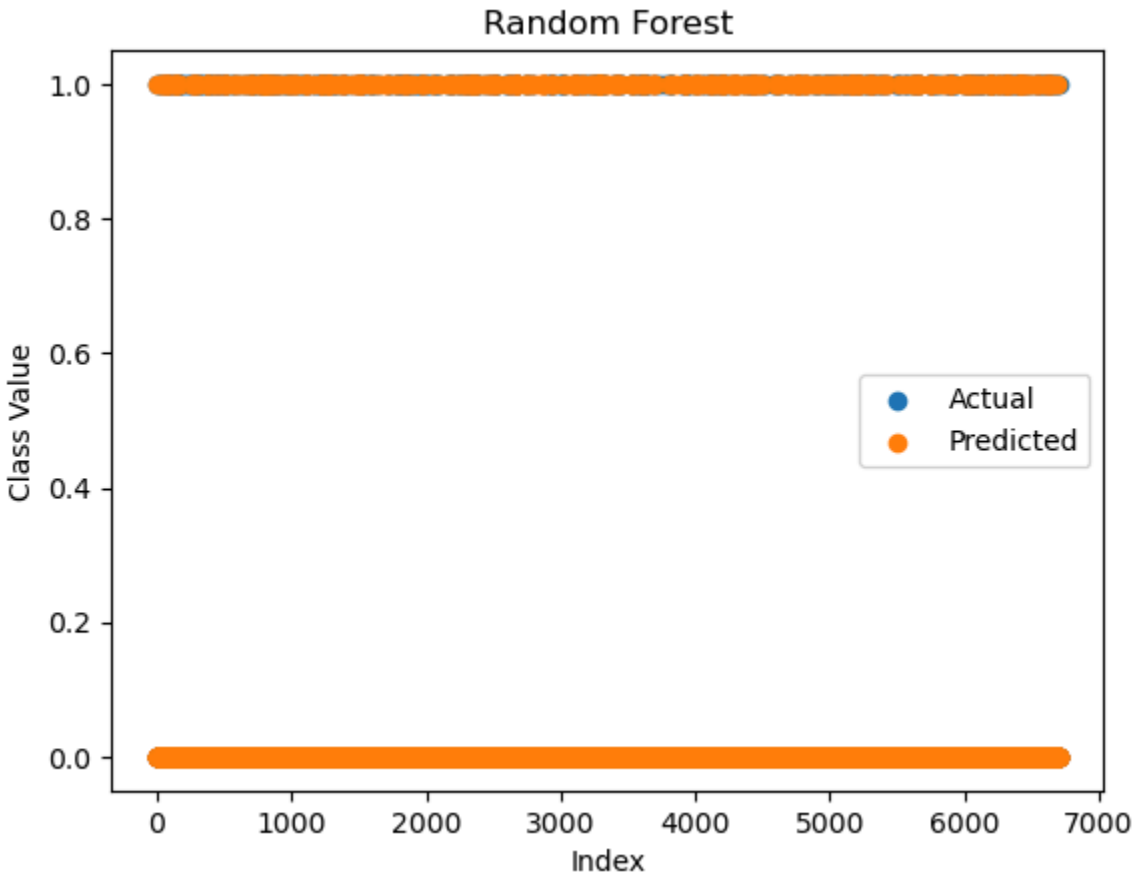


Figure: Scatter plot showing actual vs. predicted values using Random Forest

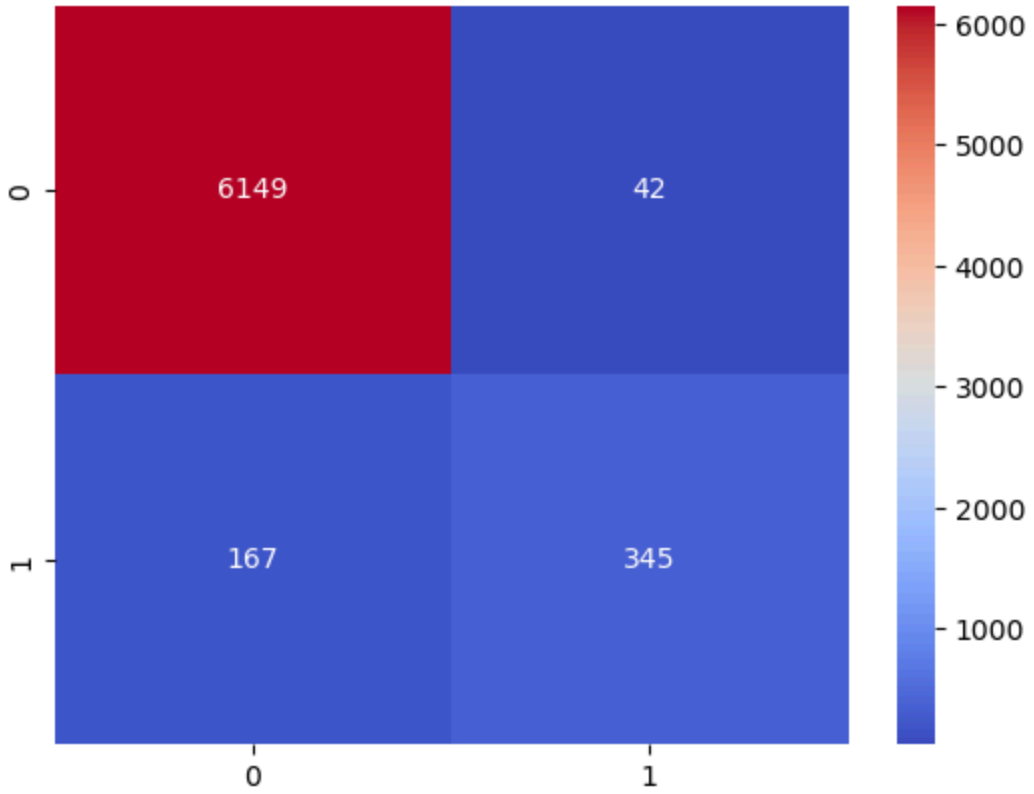


Figure: Confusion Matrix Heatmap for Random Forest

Metric	Value
Accuracy	96.88%
Precision	89.15%
Recall	67.38%
F1 Score	76.75%

The Random Forest model performed exceptionally well, achieving a high accuracy. With the calculated precision, the model was very effective in making correct positive predictions. The recall indicates it captured a substantial portion of true positives, while the F1 score reflects a strong balance between precision and recall. These results suggest that the Random Forest model effectively captures the complexity of the data, providing reliable and robust predictions. Further tuning, such as adjusting the number of trees or depth, could further enhance its recall.

AdaBoost Classifier

AdaBoost is an ensemble learning algorithm that combines multiple weak learners, usually decision trees, to form a strong classifier. It focuses on improving accuracy by giving more weight to incorrectly classified instances during training.

It was applied to the dataset, where it iteratively adjusted the weights of training instances to improve prediction accuracy. The model then made predictions on the test data based on the ensemble of weak learners.

The model's predictions were compared with actual outcomes and visualized using scatter plots and a confusion matrix.

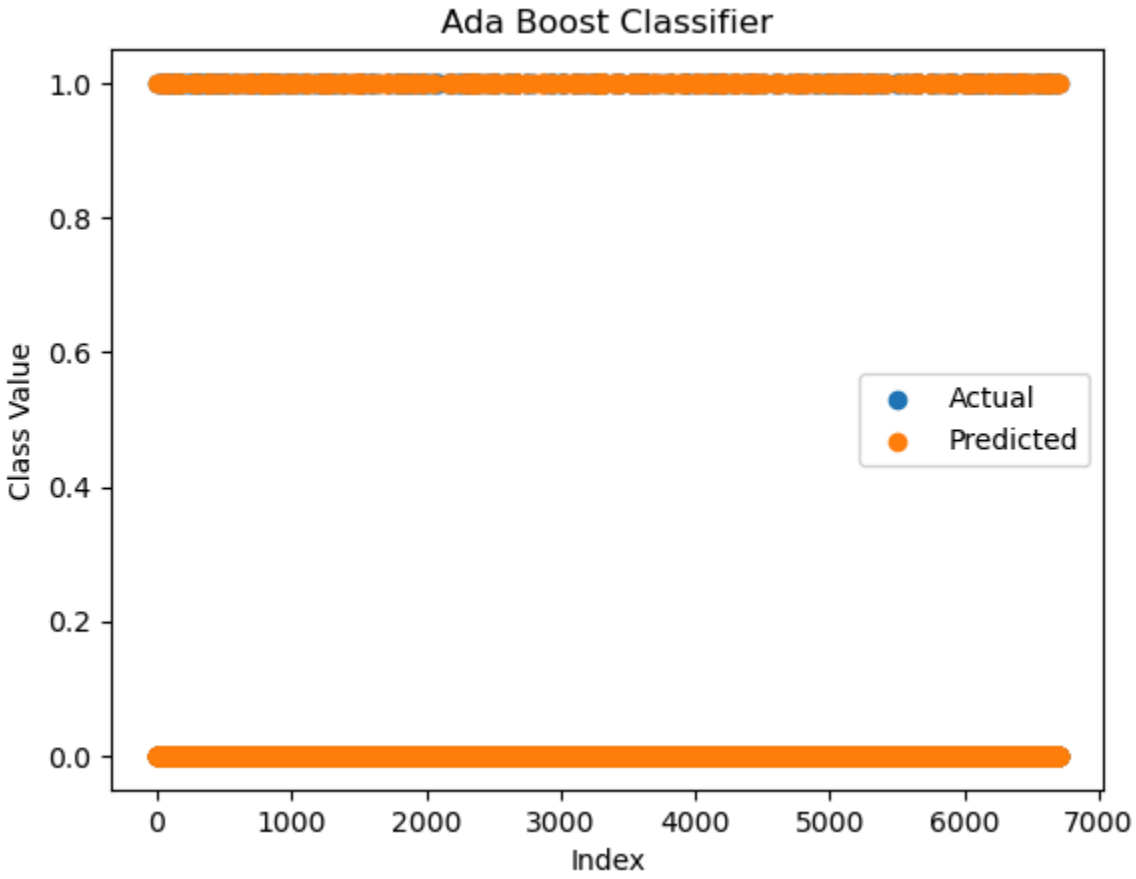


Figure: Scatter plot showing actual vs. predicted values using AdaBoost

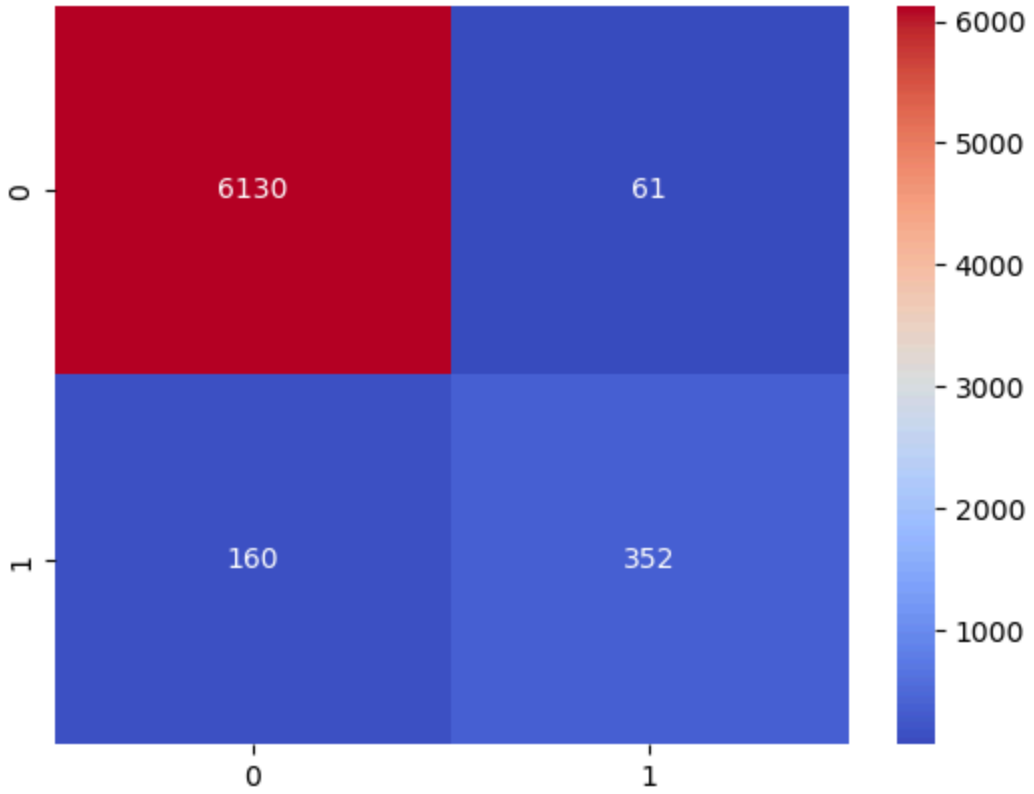


Figure: Confusion Matrix Heatmap for AdaBoost

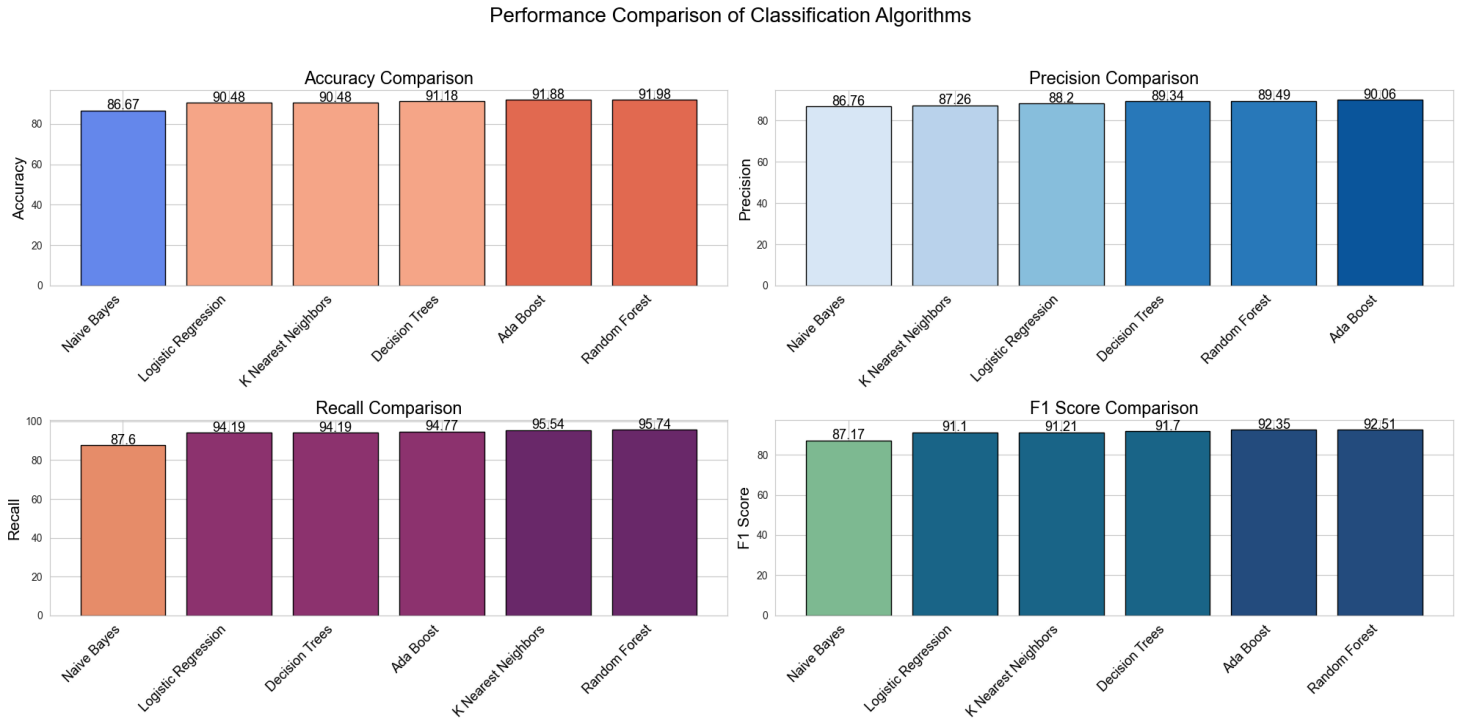
Metric	Value
Accuracy	96.70%
Precision	85.23%
Recall	68.75%
F1 Score	76.11%

The AdaBoost model delivered strong performance, with high accuracy and precision, indicating that it was highly effective at making correct positive predictions. The recall shows it captured a significant portion of true positives, while the F1 score reflects a solid balance between precision and recall. These results suggest that AdaBoost effectively boosts the performance of weak learners, making it a robust choice for classification tasks.

Algorithm Performance Comparison

The following table summarizes the performance of each algorithm, allowing us to identify which models perform best overall.

Classification Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	95.151425	74.933333	54.882812	63.359639
K Nearest Neighbors	95.867522	78.450363	63.281250	70.054054
Naive Bayes	89.840370	40.744797	72.656250	52.210526
Decision Trees	96.777562	88.341969	66.601562	75.946548
Random Forest	96.881993	89.147287	67.382812	76.751947
Ada Boost	96.702969	85.230024	68.750000	76.108108



Random Forest and Decision Tree classifiers lead in accuracy, both achieving over 96.7%, with Random Forest slightly edging out with the highest precision (89.15%) and a strong F1 score (76.75%). However, while these models perform well overall, it's important to consider the imbalanced nature of the dataset.

Given that non-hazardous asteroids are significantly more common, there's a risk that these models might favor the majority class, leading to a higher accuracy at the cost of misclassifying hazardous asteroids. This concern is particularly evident in the recall scores, where even though

the Random Forest and Decision Tree perform relatively well, they might still miss a critical number of hazardous asteroids.

Naive Bayes, while having the lowest accuracy (89.84%), achieved the highest recall (72.66%), indicating that it is more effective at identifying hazardous asteroids, even at the cost of making more false positives. This suggests that while accuracy is a critical metric, recall is also vital in contexts where identifying true positives (hazardous asteroids) is crucial.

In summary, while algorithms like Random Forest and Decision Tree offer high overall performance, their reliance on the majority class could lead to a potential underestimation of hazardous asteroids. This underlines the importance of considering class imbalance and the need for strategies like re-sampling or adjusting decision thresholds to ensure that hazardous asteroids are not overlooked by the model.

Addressing Data Imbalance

To ensure that the classification models do not favor the majority class (non-hazardous asteroids) due to the inherent data imbalance, undersampling is applied using the *RandomUnderSampler* method. This approach balances the dataset by reducing the number of non-hazardous asteroids, creating an equal representation of both classes. After this adjustment, the dataset was split into training and testing sets, enabling a more accurate evaluation of model performance without the bias introduced by the original imbalance.

After addressing the class imbalance using the undersampling technique, the same classification algorithms were re-applied to the balanced dataset. Below are the performance metrics and visualizations for each algorithm.

Logistic Regression

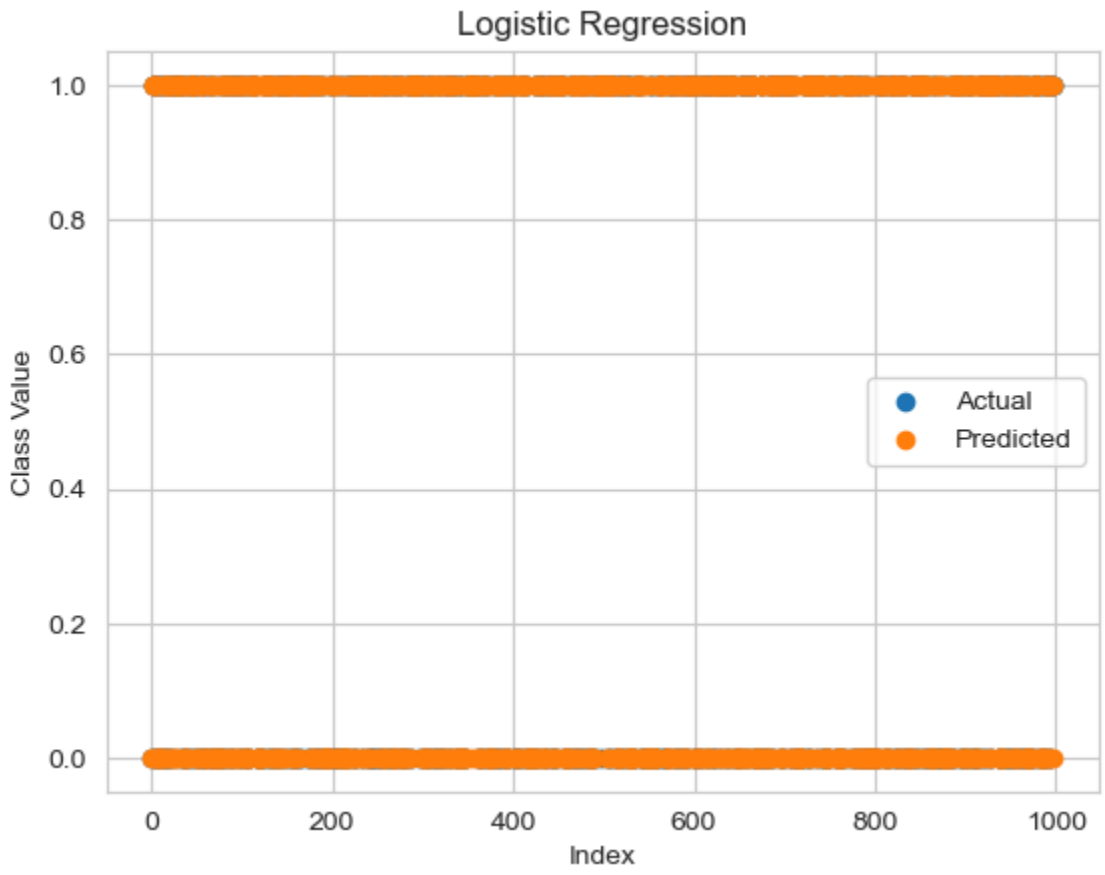


Figure: Scatter plot showing actual vs. predicted values using Logistic Regression

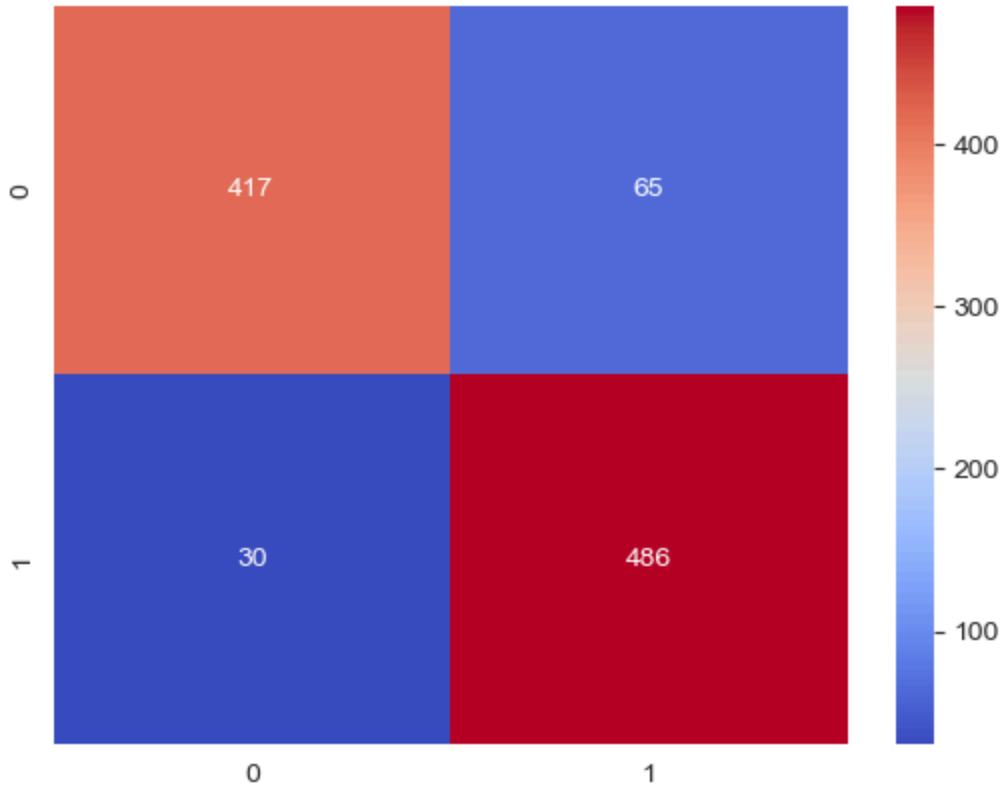


Figure: Confusion Matrix Heatmap for Logistic Regression

Metric	Value
Accuracy	90.48%
Precision	88.20%
Recall	94.19%
F1 Score	91.10%

While the accuracy decreased slightly from 95.15% to 90.48%, this was accompanied by a significant improvement in precision (from 74.93% to 88.20%) and a substantial increase in recall (from 54.88% to 94.19%). The F1 score also rose dramatically, from 63.36% to 91.10%, indicating a much better balance between precision and recall. These results suggest that undersampling effectively addressed the data imbalance, leading to a model that is far more reliable in identifying hazardous asteroids, even if overall accuracy is slightly reduced.

K-Nearest Neighbors

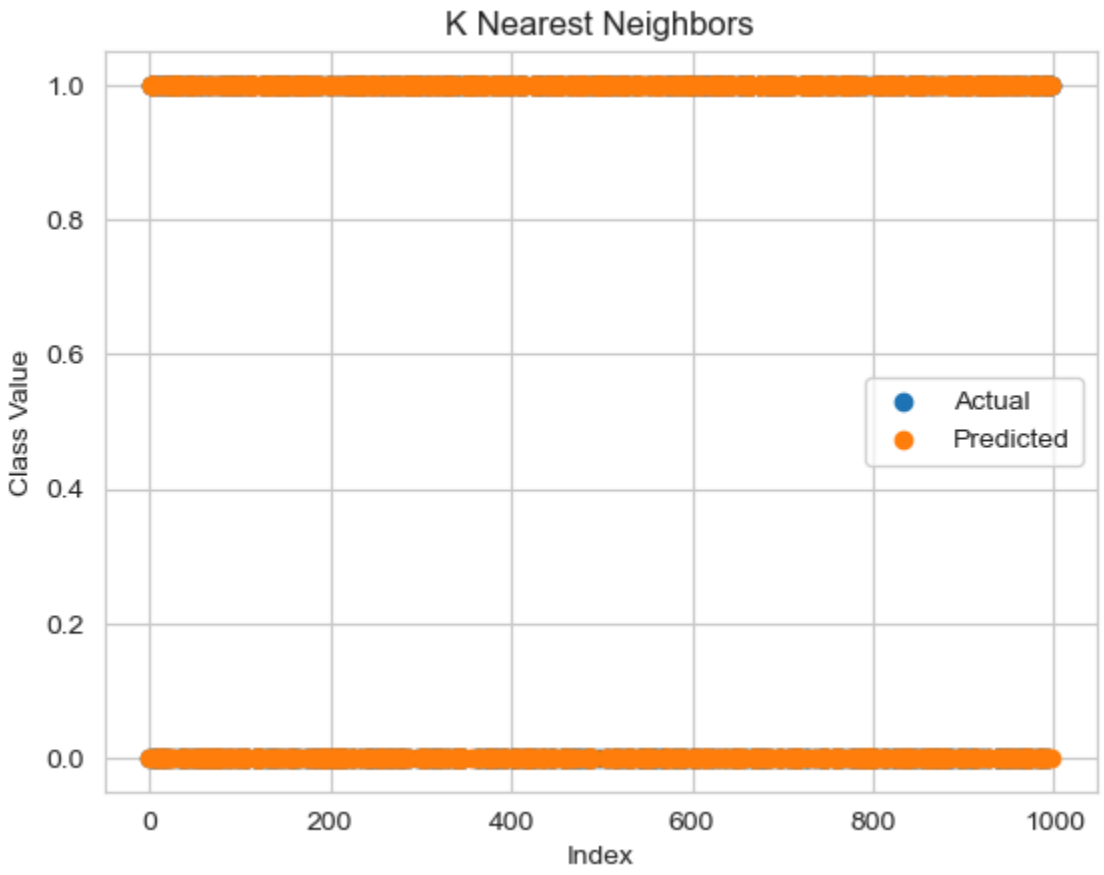


Figure: Scatter plot showing actual vs. predicted values using KNN

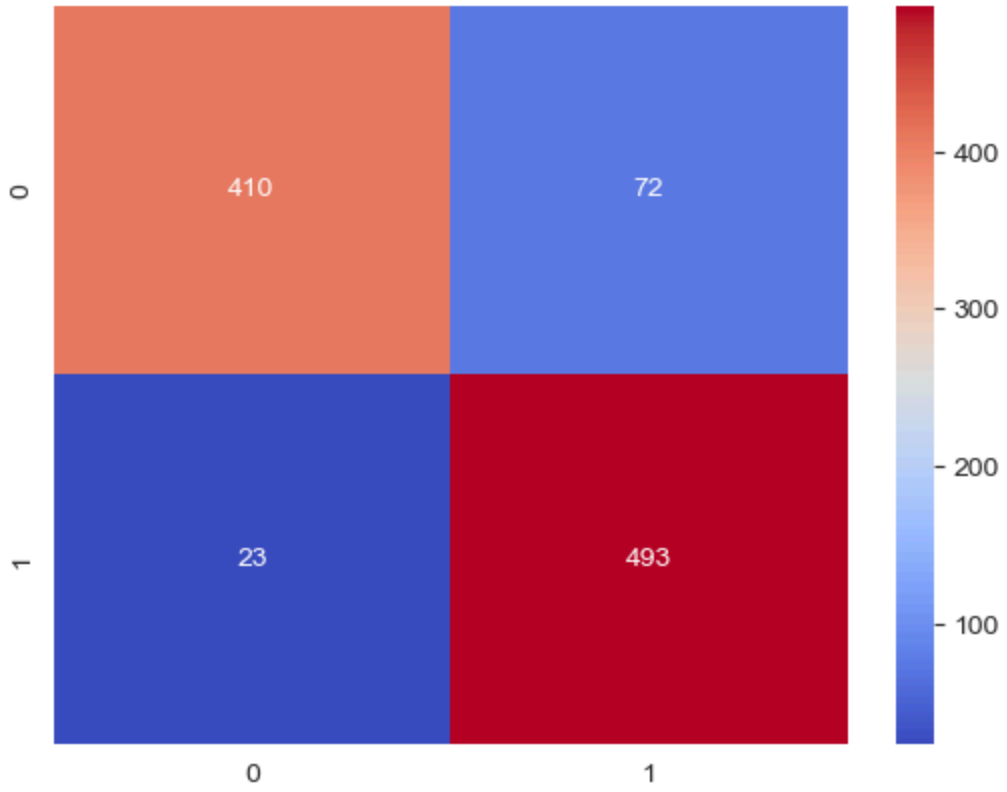


Figure: Confusion Matrix Heatmap for KNN

Metric	Value
Accuracy	90.48%
Precision	87.26%
Recall	95.54%
F1 Score	91.21%

Although the accuracy declined from 95.87% to 90.48%, the model's precision improved from 78.45% to 87.26%. More importantly, recall jumped from 63.28% to 95.54%, indicating a substantial enhancement in detecting hazardous asteroids. The F1 score, reflecting this improved balance, rose from 70.05% to 91.21%. These shifts demonstrate that undersampling effectively improved the model's sensitivity to hazardous asteroids, even though it slightly reduced the overall accuracy.

Naive Bayes

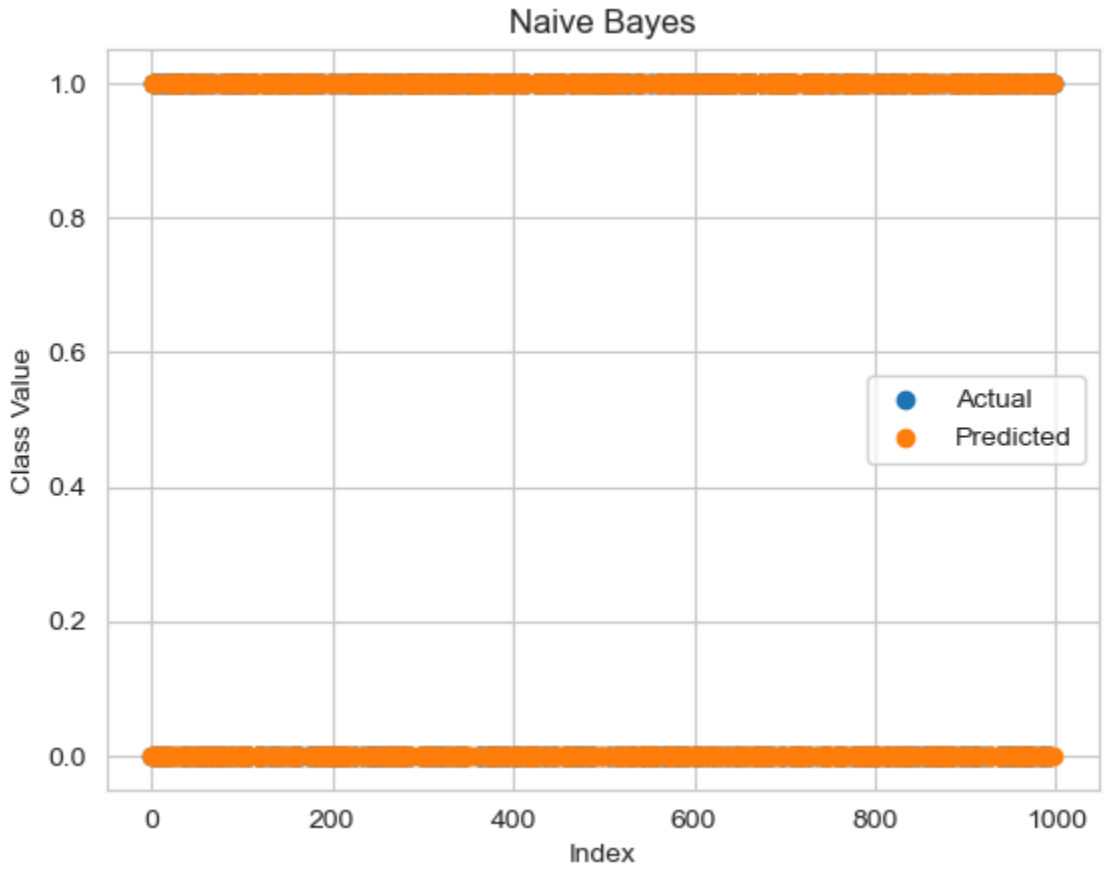


Figure: Scatter plot showing actual vs. predicted values using Naive Bayes

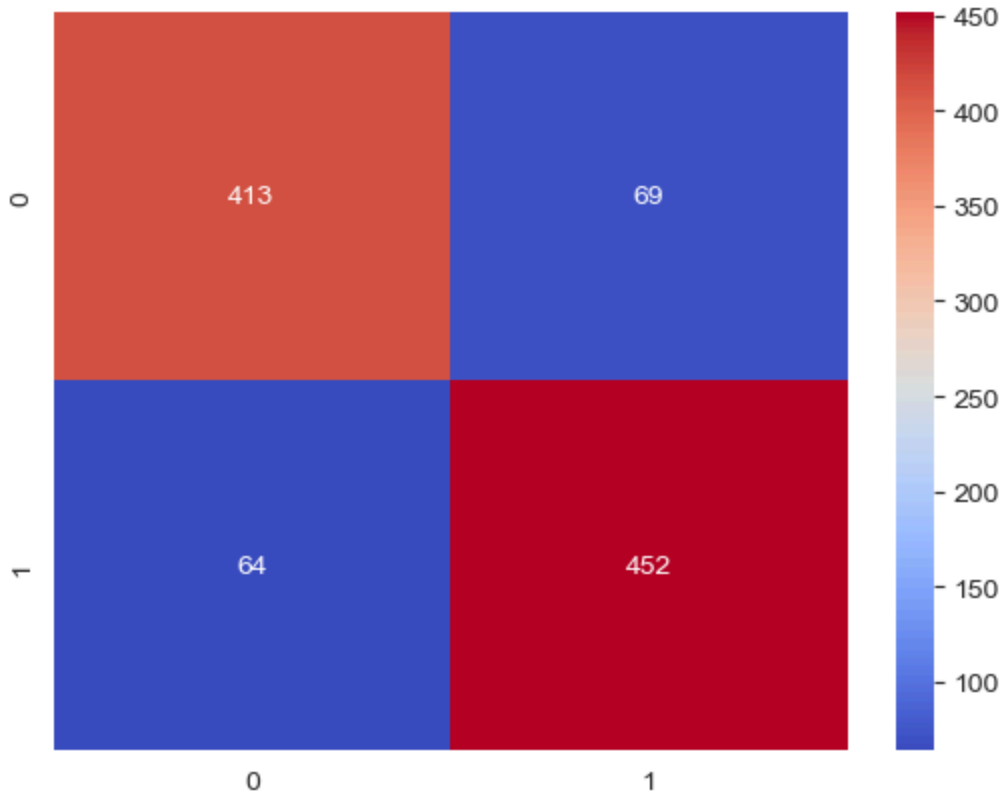


Figure: Confusion Matrix Heatmap for KNN

Metric	Value
Accuracy	86.67%
Precision	86.76%
Recall	87.60%
F1 Score	87.17%

The model's accuracy dropped slightly from 89.84% to 86.67%, yet the precision surged from 40.74% to 86.76%. Most notably, recall improved from 72.66% to 87.60%, which significantly boosted the F1 score from 52.21% to 87.17%. This drastic improvement in precision and recall highlights the model's enhanced ability to correctly classify hazardous asteroids, reflecting a substantial gain in both confidence and reliability following the adjustment for class imbalance.

Decision Tree

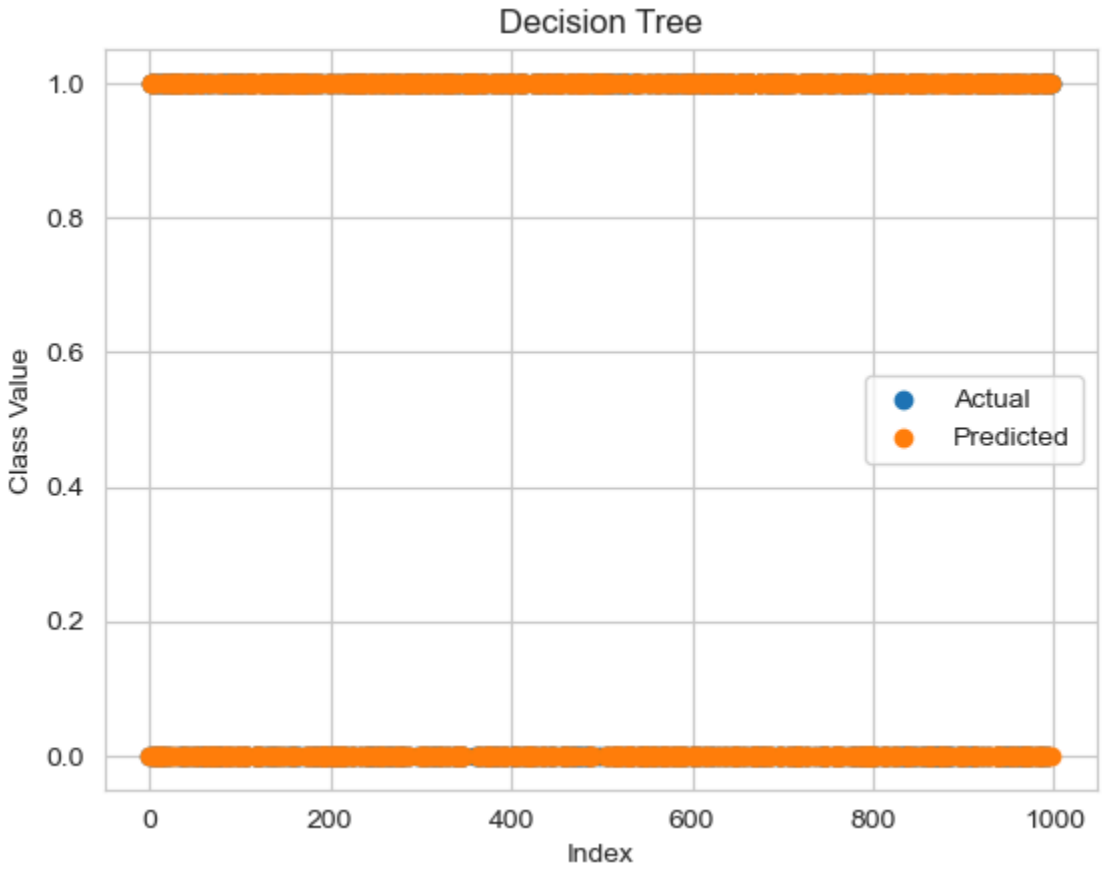


Figure: Scatter plot showing actual vs. predicted values using Decision Tree

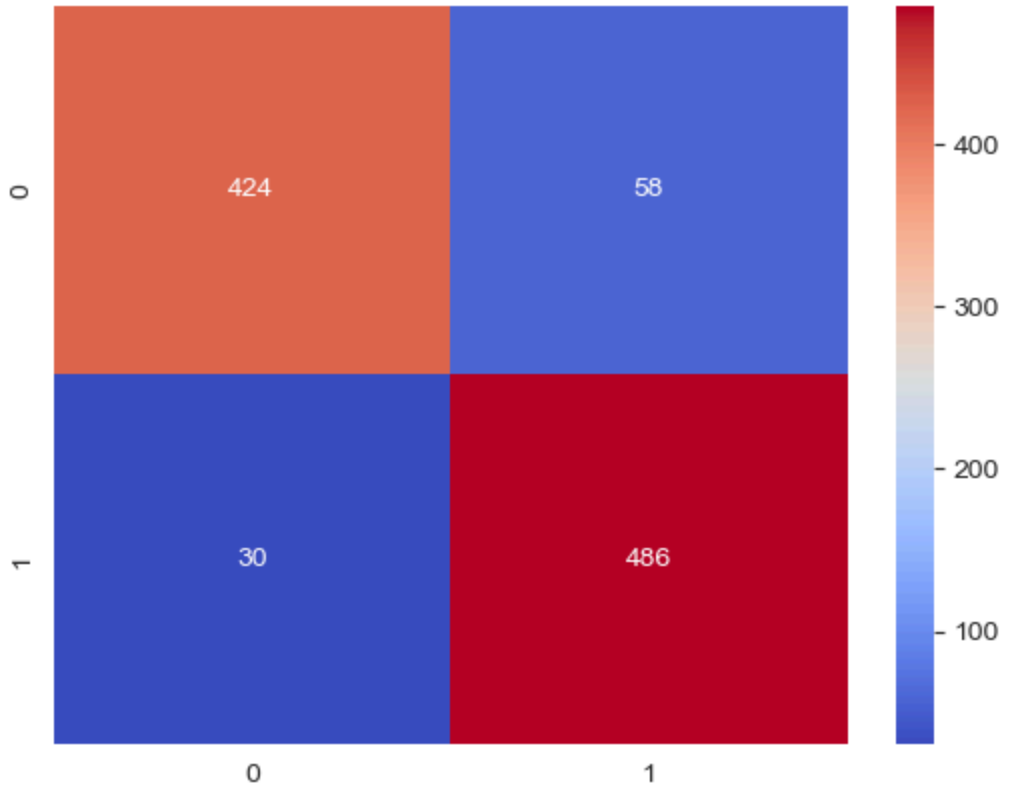


Figure: Confusion Matrix Heatmap for Decision Tree

Metric	Value
Accuracy	91.18%
Precision	89.34%
Recall	94.19%
F1 Score	91.70%

While accuracy decreased from 96.79% to 91.18%, both precision and recall improved, with precision moving from 88.57% to 89.34%, and recall from 66.60% to 94.19%. The F1 score also increased significantly, from 76.03% to 91.70%. These enhancements, particularly in recall and F1 score, indicate that the model is now more adept at identifying hazardous asteroids, achieving a better overall balance in prediction despite the slight drop in accuracy.

Random Forest

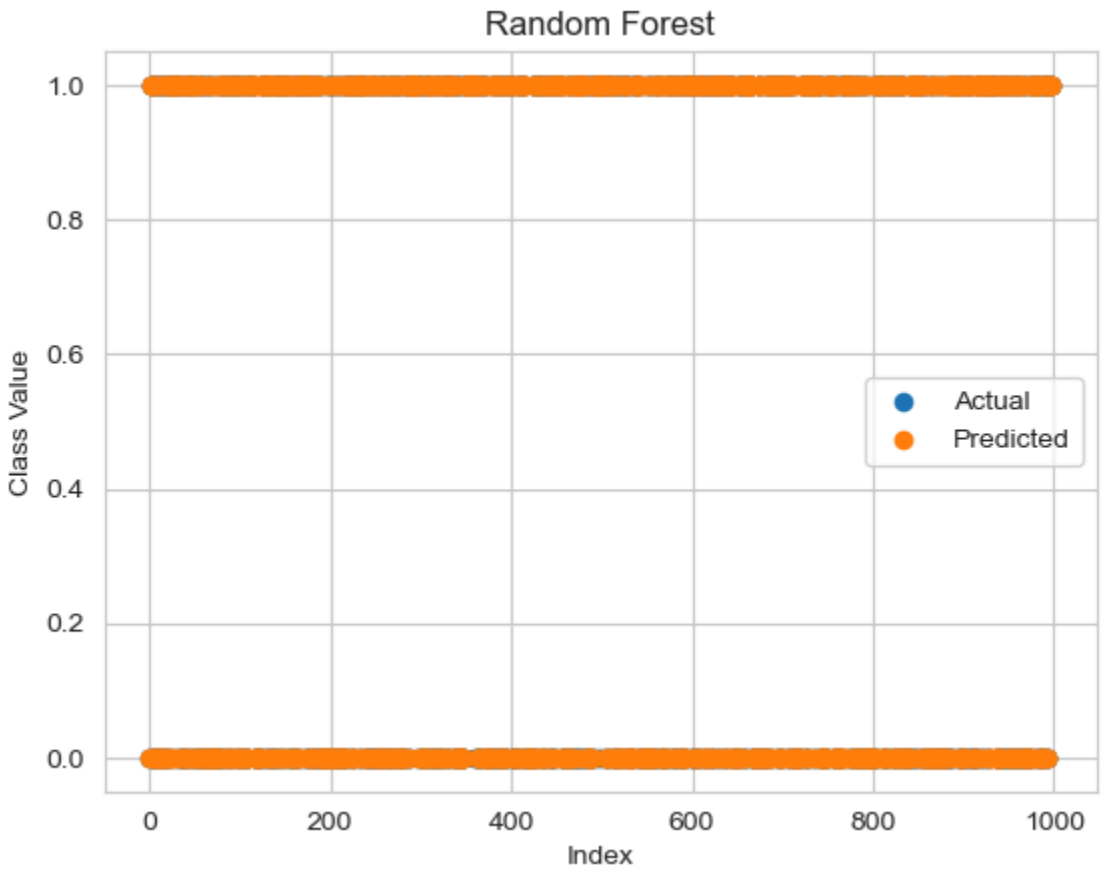


Figure: Scatter plot showing actual vs. predicted values using Random Forest

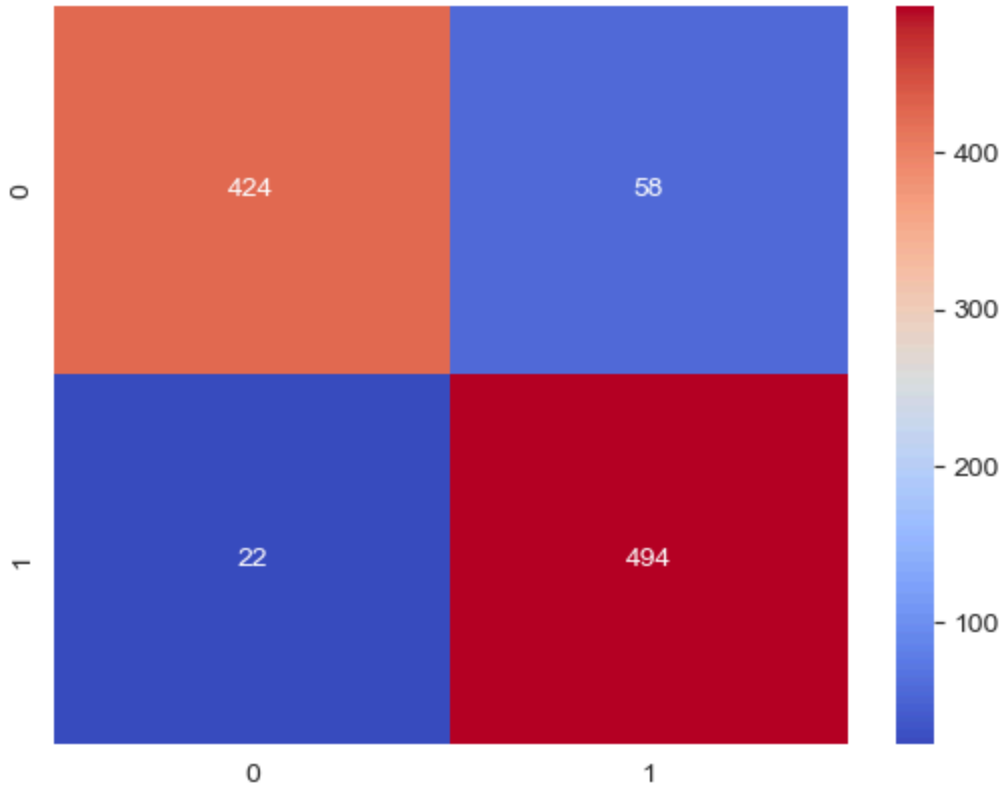


Figure: Confusion Matrix Heatmap for Random Forest

Metric	Value
Accuracy	91.98%
Precision	89.49%
Recall	95.74%
F1 Score	92.51%

The Random Forest model, after undersampling, has demonstrated its strongest performance yet. Although the accuracy saw a minor decrease, the key metrics of precision, recall, and F1 score all showed substantial improvements, reflecting the model's heightened ability to distinguish between hazardous and non-hazardous asteroids. With a near-perfect balance, as evidenced by the F1 score of 92.51%, Random Forest stands out as the most robust classifier in this analysis. The increase in recall, in particular, signifies a major advancement, ensuring that nearly all hazardous asteroids are correctly identified, making this model exceptionally reliable for critical predictions.

AdaBoost Classifier

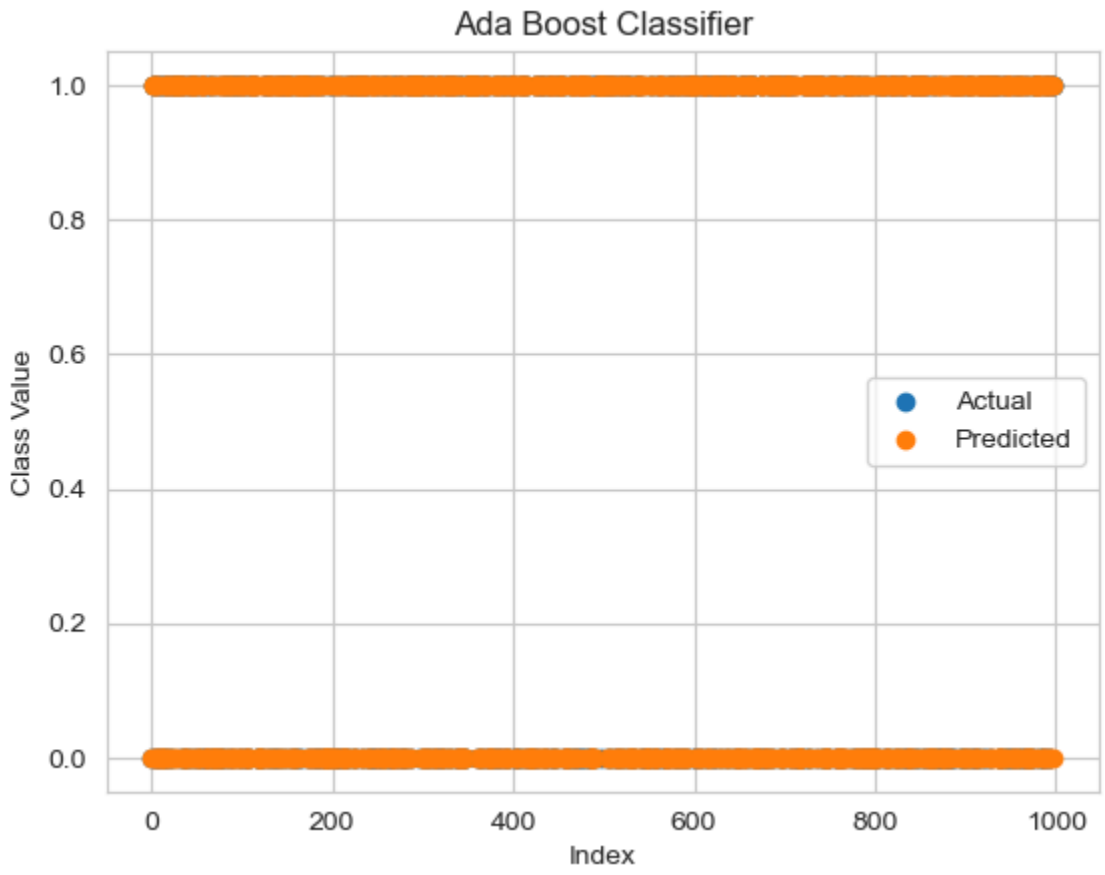


Figure: Scatter plot showing actual vs. predicted values using AdaBoost Classifier

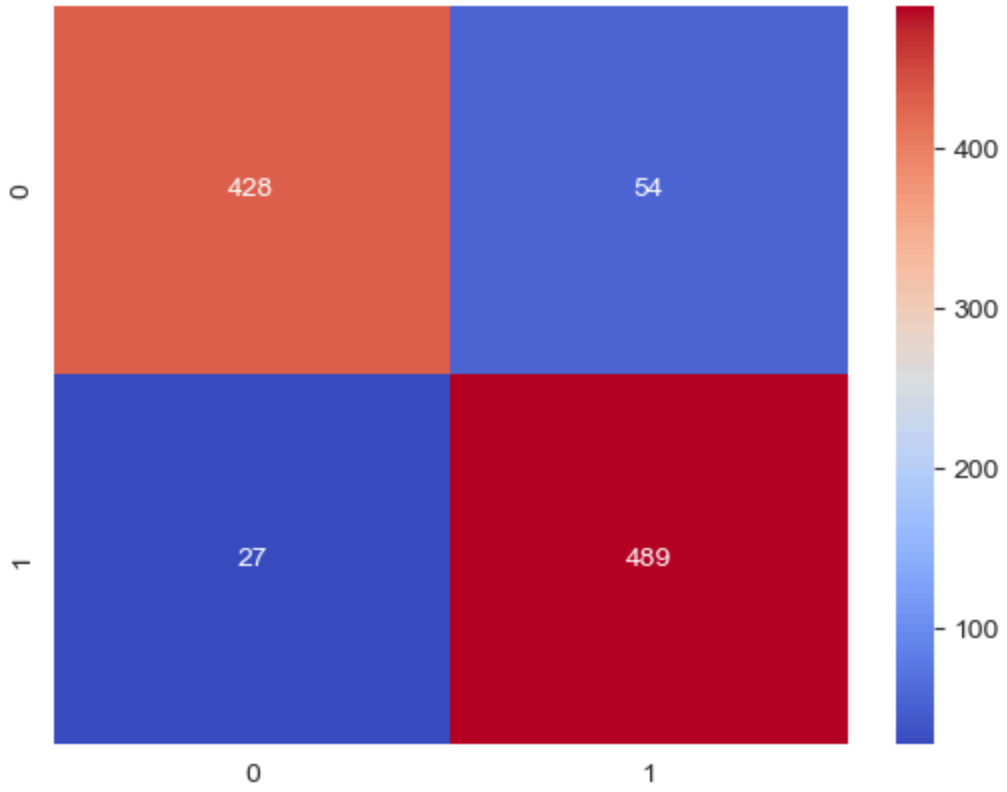


Figure: Confusion Matrix Heatmap for AdaBoost Classifier

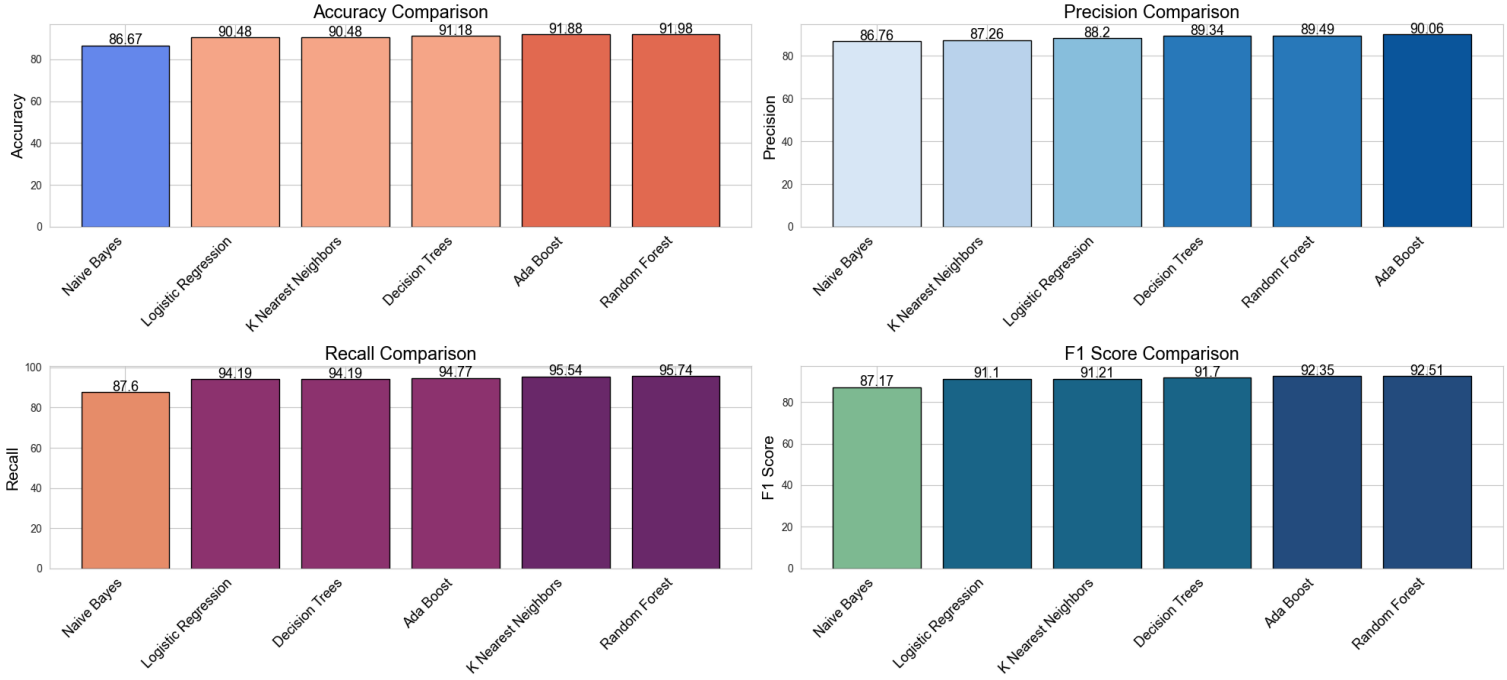
Metric	Value
Accuracy	91.88%
Precision	90.06%
Recall	94.77%
F1 Score	92.35%

While the accuracy dipped slightly, the key takeaway is the significant enhancement in both precision and recall, leading to a much higher F1 score of 92.35%. This suggests that AdaBoost has become more adept at correctly identifying hazardous asteroids, reducing the risk of misclassification. The refined precision and recall indicate that the model is now more finely tuned, balancing the trade-offs between identifying true positives and minimizing false alarms, thus offering a more dependable tool for making critical decisions.

Algorithm Performance Comparison

Classification Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	90.480962	88.203267	94.186047	91.096532
K Nearest Neighbors	90.480962	87.256637	95.542636	91.211841
Naive Bayes	86.673347	86.756238	87.596899	87.174542
Decision Trees	91.182365	89.338235	94.186047	91.698113
Random Forest	91.983968	89.492754	95.736434	92.509363
Ada Boost	91.883768	90.055249	94.767442	92.351275

Performance Comparison of Classification Algorithms With Undersampling



Balanced Bagging on Random Forest

Amongst other resampling methods we also tested imbalance learn's *BalancedBaggingClassifier* on the original Random Forest model, as that was the original model with the highest accuracy, to get the best possible performance. Upon getting the classification report, we observe that all the metrics have values slightly greater to the ones we got from undersampled Random Forest. Following are the depictions:

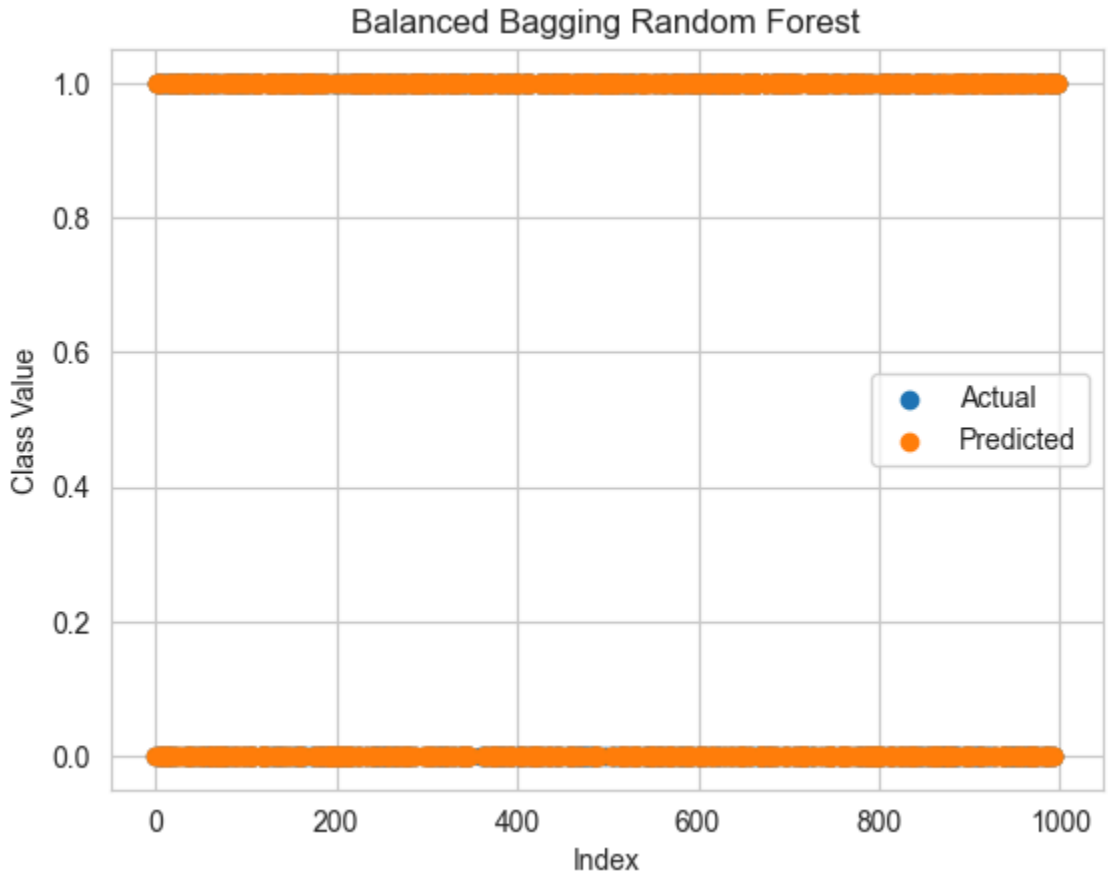


Figure: Scatter plot showing actual vs. predicted values using Balanced Bagging Random Forest

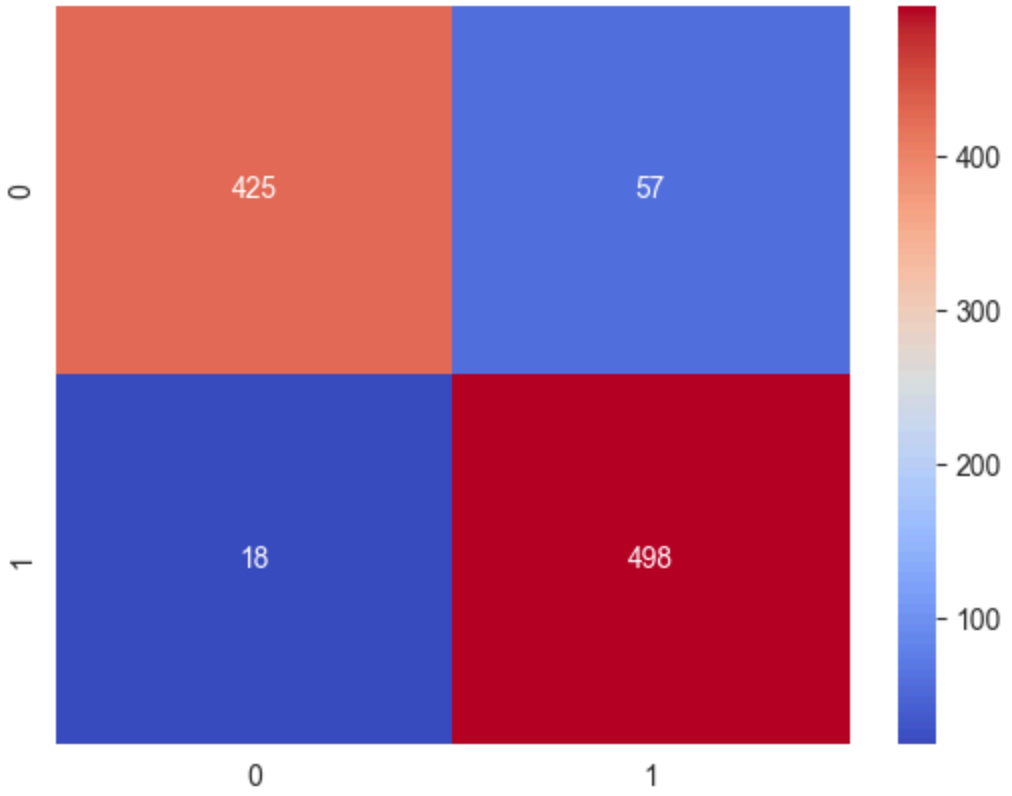


Figure: Confusion Matrix Heatmap for Balanced Bagging Random Forest Classifier

Metric	Value
Accuracy	92.48%
Precision	89.72%
Recall	96.51%
F1 Score	93.00%

Conclusion

A Trade-Off Between Accuracy and Recall:

After creation and careful observation of the performance of all the aforementioned models, we can conclude that the original Random Forest model gives the best accuracy of 96% amongst all the other original models. However the recall score, being quite low for this model, might lead to a significant number of false negatives.

We observe that all the metrics of the classification report have similar close values for the undersampled models and also the balanced bagging random forest model, which ultimately results in a high F1 score. This leads to a decrease in the number of false positives and false negatives from the resampled data. However accuracy also decreases for all the models with the increase in other metric values.

We can tackle the problem of false negatives by using the undersampled or the balanced bagging Random Forest model, but compromising the accuracy by approximately 4%.

Potential Applications:

1. **Early Warning Systems:** The methodologies and findings from this project could be integrated into early warning systems to predict and mitigate risks posed by hazardous NEOs.
2. **Space Mission Planning:** Understanding the characteristics of hazardous NEOs can assist in planning space missions aimed at deflecting, destroying or studying these objects.

Summarizing The Project:

Thus we conclude our project by creating several models which are highly accurate in successfully classifying Near Earth Objects into hazardous and non-hazardous categories, providing key insights into the factors that determine an NEO's potential threat to Earth. The findings underscore the importance of continued monitoring and analysis of NEOs to enhance planetary defense strategies. While the current models are robust, future enhancements and a broader dataset could further improve the accuracy and reliability of hazard predictions.