**DATA ANALYTICS NSSC’24**

**Objective of the Analysis: Data-Driven Classification of Hazardous Asteroids**

**1. Exploratory Data Analysis (EDA)**

**1.1 Data Inspection**

**Question 1.1.1:** Inspect the dataset and determine the data types of all features (numerical, categorical).

We are use some dataframe methods for inspecting the dataset that is shap, head(), info(), dtypes

**Question 1.1.2:** Calculate and analyse basic statistics for each numerical feature, including range, mean, median, standard deviation, and quartile..

For calculating min, max, range, mean, median, standard deviation, quartile we use inbuilt library called describe() and manually calculate all this things for all the numerical features and store.

**Question 1.1.3:** Identify features that have missing values.

For missing value identification we use isnull() method and calculate null value present in the form of percentage.

**Question 1.1.4:** Identify the numerical and categorical features of the dataset to use for further analysis.

For classification of numerical and categorical features we use select\_dtypes() method and store the features in list format differently.

**Question 1.1.5:** Use imputation to fill the null values in the dataset. How is this process different for numerical and categorical columns?

Select the numerical data types and fill the null value by mean value imputation of this features and in same format for categorical features use mode imputation.

**1.2 Statistical Inference**

**Question 1.2.1:** Plot the distribution of numerical features to assess the skewness of the data. Does this dataset require normalisation? If yes, normalise/scale the dataset. (Hint: Us histograms)

To see the numerical features data distribution, we use two different types of plot that is Histogram and Kernal Density Estimator and using this graph analysis the skewness of the features.

After analysing we fixed the skewness, we use two different type techniques that is Min Max Scaler for normal skewness (skewness value -0.5 to 0.5) and Log Transformation for right or left skew data (skewness value below -0.5 or above 0.5).

**Question 1.2.2:** Identify potential outliers in the numerical columns using any statistical technique (e.g., box plots, z-score, etc.).s)

For outliers detection in numerical features we use box plot method. If any feature contain outlier then we IQR method for removing the outliers.

Many features have huge number of outlier so after the IQR method near about 50% samples are removed.

**Question 1.2.3:** Explore the relationship between different features using scatter plots or correlation matrices. (Hint: Use Seaborn or similar libraries)

I use scatterplot that present on seaborn library and plot for every numerical features and also plot heatmap or correlation matrix for the dataset.

**Question 1.3.1:** Create a pairplot using Seaborn to visualise relationships between multiple numerical features simultaneously.

From seaborn library we use pairplot() and plot a relationship graph for all the numerical features.

**Question 1.3.2:** What do you infer from these plots? How do the diagonal plots and off-diagonal plots in a pairplot differ in the information they provide?

**1.4 Tackling Class Imbalance**

**Question 1.4.1:** Is there a classification bias (class imbalance) in this dataset? If yes, how would you tackle it?

When we analysis the class imbalance then we see the number of not hazardous in 1988 and the number of hazardous is 343.

**Question 1.4.2:**

Discuss the implication of class imbalance on model performance.

**2. Numerical Interpretation and Mathematical Analysis)**

**2.1 Feature Engineering**

**Question 2.1.1:** Combine the approach\_date, month, and year features into a single feature representing the day of the year. Convert it into a ‘datetime’ format.

We combine the three different feature approach\_day, approach\_month and approach\_year and make a new feature called date that is datetime type data.

**Question 2.1.2:** Calculate the ratio of Miss Distance vs. Semi-major axis. Create a 'Time Until Approach' feature based on the difference between the 'Epoch Date Close Approach' and the current date.

Calculate the ratio of miss dist and semi major axis and store the result in the new column name Miss\_Distance\_vs\_Semi\_Major\_Axis. Also create new feature called Time\_Until\_Approach that store the difference value of 'Epoch Date Close Approach' and the current date.

**Question 2.1.3:** Calculate the eccentricity of the orbit, average orbital velocity, and orbital period using Kepler’s Law.

Create new feature called Eccentricity, Average\_Orbital\_Velocity, Orbital\_Period\_seconds and Orbital\_Period\_years.

Formula,

Eccentricity = (Aphelion Dist - Semi Major Axis) / Aphelion Dist

Average\_Orbital\_Velocity = square root of (Gravitational constant \* Mass of the sun / (Semi Major Axis \* astronomical units))

Orbital\_Period\_seconds = 2 \* 3.14 \* square root of( (Semi Major Axis \* astronomical units) ^ 3 / (Gravitational constant \* Mass of the sun))

Orbital\_Period\_years = Orbital\_Period\_seconds / (60\*60\*24\*365.25)

**Question 2.1.4:** Calculate the heliocentric distance, escape velocity, and specific orbital energy.

Formula,

Escape\_Velocity = square root of( 2 \* Mass of the sun / (Semi Major Axis \* astronomical units))

Specific\_Orbital\_Energy = - (Gravitational constant \* Mass of the sun) / (2 \* Semi Major Axis \* astronomical units)

Heliocentric\_Distance = (Semi Major Axis \* astronomical units) \* ( 1- Eccentricity)

**Question 2.1.5:** Calculate the Specific Angular Momentum using the formula: h=sqrt(GMa(1−e2)).

Formula,

Specific\_Angular\_Momentum = sqrt(Gravitational constant \* Mass of the sun \* Semi Major Axis \* astronomical units \* (1 – Eccentricity ^ 2))

**Question 2.1.6:** Calculate the velocity at Perihelion and Aphelion.

Formula,

Velocity\_Perihelion = sqrt( Gravitational constant \* Mass of the sun \* (1 + Eccentricity) / ((Semi Major Axis \* astronomical units \* (1 - Eccentricity))

Velocity\_Aphelion = sqrt( Gravitational constant \* Mass of the sun \* (1 - Eccentricity) / ((Semi Major Axis \* astronomical units \* (1 + Eccentricity))

**Question 2.1.7:** Average the Miss distance of various categories and find the closest approach distance.

Calculate the average of different types of Miss Dist that is Astronomical, lunar, kilometers and miles using mean() method and then calculate the closest approach distance.

**Question 2.1.8:** Calculate Synodic Period and Mean Motion using the orbital period.

Formula,

Synodic\_Period = Orbital\_Period\_years / (Orbital\_Period\_days – 365.25)

Mean\_Motion = 360 / Orbital\_Period\_days

**2.2 Additional Features**

**Question 2.2.1:** Create additional features as per your understanding of the problem for improving accuracy. More marks are awarded for innovative and effective features.

**I have create some new features**

* . **Potential Energy at Closest Approach:**

Calculate the gravitational potential energy of the object at its closest approach to Earth, which can be indicative of the potential impact ene

* **Gravitational Influence Factor:**

Calculate a factor representing the gravitational influence of Earth on the object, which can be useful for assessing potential trajectory change

* **Relative Orbital Energy:**

Compute the relative orbital energy compared to Earth's orbital energy. This can provide insights into the object's dynamic behavior.s.rgy.e.t.

Formula,

Potential\_Energy\_Closest\_Approach = - Gravitational constant \* earth\_mass / Miss Dist.(kilometers)

Gravitational\_Influence\_Factor = Gravitational constant \* earth\_mass / Miss Dist.(kilometers) ^ 2

Relative\_Orbital\_Energy = Specific\_Orbital\_Energy / ( - Gravitational constant \* earth\_mass / earth\_semi\_major\_axis\_km)

**3. Handling Binned Values**

**Question 3.1:** Modify the binned features that have an ordinal relationship in this manner: (very slow = 0, slow = 1, fast = 2, very fast = 3, etc).

Relative Velocity km per sec column has contain ordinal features that is (very slow = 0, slow = 1, fast = 2, very fast = 3) so we use ordinal encoding technique.

Also Orbital Period and Orbit Uncertainity columns have contain ordinal features that is (low=0, medium=1, high=2) so we use ordinal encoding technique.

**Question 3.2:** One-hot encode the binned features whose relationship is not strictly ordinal.

Here only one column has binary feature that is Hazardous (False=0, True=1) so here we apply one hot encoding technique.

**4. Hazardous Classification**

**Question 4.1:** Build a robust and efficient classifier to classify asteroids as Hazardous (1) or Not Hazardous (0).

Here for the classification we choose Random Forest classification algorithm and train the model by given value. After the model train we predict and calculate accuracy score and classification matrix. We achieved 86% accuracy.

**Question 4.2:** Implement K-Fold Cross Validation for training. Train the dataset for all values of K from 2 to 10. Plot the loss and accuracy versus epochs for these K values.

We use kfold validation technique and the value of k is between 2 to 10 and calculate average accuracy score and loss in every fold.

**Question 4.3:** Optimise all the hyperparameters used in the classifier by selecting an appropriate optimisation method.

For model optimization we use GridSearchCV method. After the optimization we found the best parameter for this data that is {'max\_depth': 20, 'max\_features': 'log2', 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100} and the accuracy score is 87%.

**Question 4.4:** Plot the ROC curve and Confusion Matrix to quantify the performance of your classifier.

Using the method roc\_curve, roc\_auc\_score, confusion\_matrix, ConfusionMatrixDisplay we calculate the ROC curve and Confusion matrix value and the display the result in the form of graph.

**Question 4.5:** Use SHAP Values, Permutation Importance, or Partial Dependence Plots to list the most and least useful features.

SHAP is a Explainable AI algorithm and here we use this algorithm for calculating the feature importance value and display in graph format.

Also use Permutation Importance technique and calculate the feature importance value and display in graph format.

**5. Anomaly Detection**

**Question 5.1.1:** Perform anomaly detection using Any inbuilt library of your choice.

We use a library called IsolationForest that is very popular for Anomaly detection. We calculate Anomaly for every features and plot normal and anomaly points in graph, blue point for normal sample and red point for anomaly sample.

**Question 5.1.2:** Perform anomaly detection using Writing your own anomaly detection algorithm.

We calculate z-score for the features and using these value for Anomaly Detection. We calculate Anomaly for every features and plot normal and anomaly points in graph, blue point for normal sample and red point for anomaly sample.

**Question 5.2:** Store the results as a new column in the dataset. Print the number of anomalies detected by each method.

We create a new column called Anomaly and store the Anomaly result calculated by Isolation Forest method and create another method called Anomaly-New for store the Anomaly result calculating by the New method.

**Question 5.3:** Compare the results from both methods by plotting a Confusion Matrix. Print the number of examples flagged by both algorithms.

Compare this two result and calculate confusion matrix and display in graph format.