NATIONAL UNIVERSITY
of Computer & Emerging Sciences

DATA MINING PROJECT

Group Members

- Syed Muhammad Faizan Abbas Shah Kazmi (21l-5675)
- Muhammad Talal Saleem (21l-5682)
- Muhammad Huzair Amjad (21l-6213)

Instructor's Name

Dr. Esha Tu Razia Babar

Project Report: Asteroid Hazard Classification

Abstract

This report presents the methodology, findings, and insights gained from a data mining project focused on classifying asteroids into hazardous and non-hazardous categories and predicting their hazard potential based on various features. Leveraging a dataset sourced from NeoWs (Near Earth Object Web Service), the project aims to develop a predictive model capable of determining whether an asteroid poses a hazard to Earth, while identifying the key features contributing to its hazardous status. The report details the steps undertaken in data preprocessing, feature engineering, model selection, evaluation metrics, hazard prediction, and visualization. The expected outcome is a robust classification model that accurately predicts asteroid hazard potential, thereby enhancing our ability to assess the risks associated with near-Earth asteroids.

Introduction

Asteroids pose a potential threat to Earth, with the potential for catastrophic consequences if not properly monitored and mitigated. Understanding the hazard potential of near-Earth asteroids is crucial for planetary defense and safeguarding human civilization. This project addresses the need for accurate classification and prediction of asteroid hazard potential by employing data mining techniques on a comprehensive dataset sourced from NeoWs. The development of a robust predictive model holds promise for enhancing our ability to assess and manage the risks posed by near-Earth asteroids.

Problem and Dataset Description

This project aims to leverage machine learning techniques to build a robust predictive model that effectively categorizes asteroids into hazardous and non-hazardous classes. By harnessing a dataset obtained from NeoWs, which encapsulates pertinent details about asteroids like Absolute Magnitude, Estimated Diameter, Close Approach Date, Relative Velocity, and a binary indicator denoting Hazardous status, we endeavor to explore and analyze the intrinsic relationships between these features and the likelihood of an asteroid being hazardous. Through comprehensive data preprocessing, feature engineering, and model selection, our goal is to develop a predictive framework capable of accurately discerning potentially dangerous asteroids from benign ones. Ultimately, this endeavor not only enhances our understanding of asteroid behavior and characteristics but also contributes to the ongoing efforts in planetary defense and space exploration.

Dataset Description

The dataset encompasses a wide range of features relevant to asteroid classification and hazard prediction. Each entry includes details such as Neo Reference ID, Absolute Magnitude, Estimated Diameter, Orbital Parameters, and a binary indicator for Hazardous status. A comprehensive list of features and their descriptions is provided in the project proposal.

Methodology

The methodology involves several key steps:

Data Preprocessing:

Handling missing values, normalizing numerical features, and encoding categorical variables to prepare the dataset for modeling.

- The code starts by importing necessary libraries **(pandas, numpy, matplotlib.pyplot, seaborn)** and loading the dataset using **pd.read_csv().**

	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)	...	Asc Node Longitude	Orbital Period	Perihelion Distance	Perihelion Arg	Aphelic Dis
0	3703080	3703080	21.6	0.127220	0.284472	127.219879	284.472297	0.079051	0.176763	417.388066	...	314.373913	609.599786	0.808259	57.257470	2.00576
1	3723955	3723955	21.3	0.146068	0.326618	146.067964	326.617897	0.090762	0.202951	479.225620	...	136.717242	425.869294	0.718200	313.091975	1.49735
2	2446862	2446862	20.3	0.231502	0.517654	231.502122	517.654482	0.143849	0.321655	759.521423	...	259.475979	643.580228	0.950791	248.415038	1.96685
3	3092506	3092506	27.4	0.008801	0.019681	8.801465	19.680675	0.005469	0.012229	28.876199	...	57.173266	514.082140	0.983902	18.707701	1.52790
4	3514799	3514799	21.6	0.127220	0.284472	127.219879	284.472297	0.079051	0.176763	417.388066	...	84.629307	495.597821	0.967687	158.263596	1.48354

- It checks for missing values and drops columns with significant missing values **(data.isnull().sum()** and **data.drop()).**

```
Missing Values:
 Neo Reference ID                0
Name                            0
Absolute Magnitude              0
Est Dia in KM(min)              0
Est Dia in KM(max)              0
Est Dia in M(min)               0
Est Dia in M(max)               0
Est Dia in Miles(min)           0
Est Dia in Miles(max)           0
Est Dia in Feet(min)            0
Est Dia in Feet(max)            0
Close Approach Date             0
Epoch Date Close Approach       0
Relative Velocity km per sec    0
Relative Velocity km per hr     0
Miles per hour                  0
Miss Dist.(Astronomical)        0
Miss Dist.(lunar)               0
Miss Dist.(kilometers)          0
Miss Dist.(miles)               0
Orbiting Body                   0
Orbit ID                        0
Orbit Determination Date        0
Orbit Uncertainity              0
Minimum Orbit Intersection      0
Jupiter Tisserand Invariant     0
```

- It converts categorical variables to numerical using label encoding **(LabelEncoder)** for non-numeric columns.

```
Equinox
J2000    4687
Name: count, dtype: int64
```

```
Orbiting Body
Earth    4687
Name: count, dtype: int64
```

- It drops additional columns that are not required for analysis.

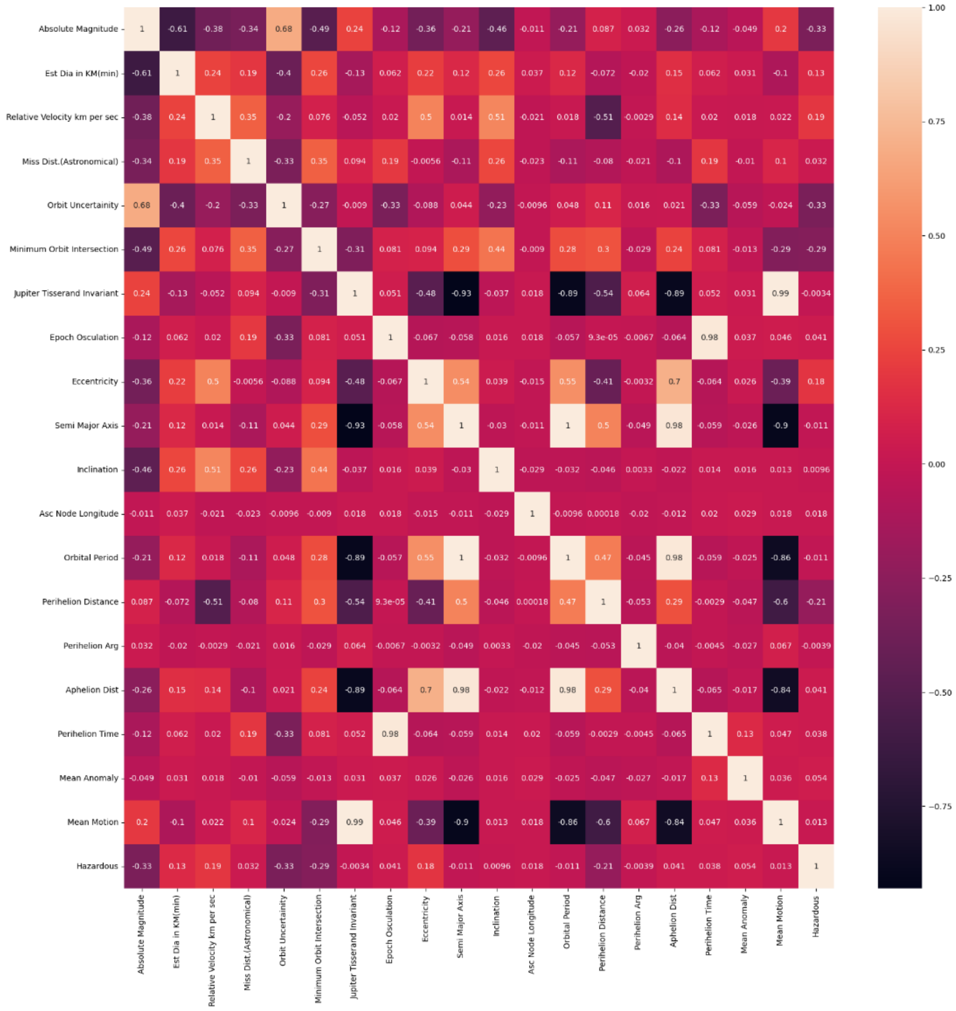
Before:

```
4    3514799    3514799

5 rows × 40 columns
```

After:

```
4          21.6    0.127220    0.28

5 rows × 34 columns
```

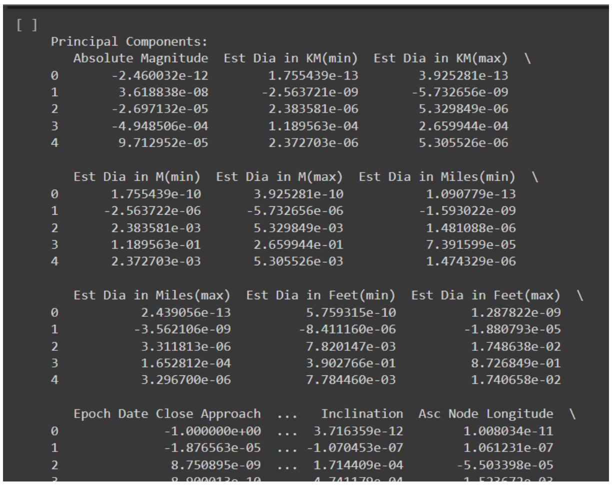
- Visualizes the correlation matrix using a heatmap **(sns.heatmap()).**



Feature Engineering:

Extracting relevant features and performing dimensionality reduction if necessary to enhance model performance.

- Variance thresholding, SelectKBest **(Chi-squared and ANOVA F-value)**, Recursive Feature Elimination **(RFE)**, and Principal Component Analysis **(PCA)** are used for feature selection.



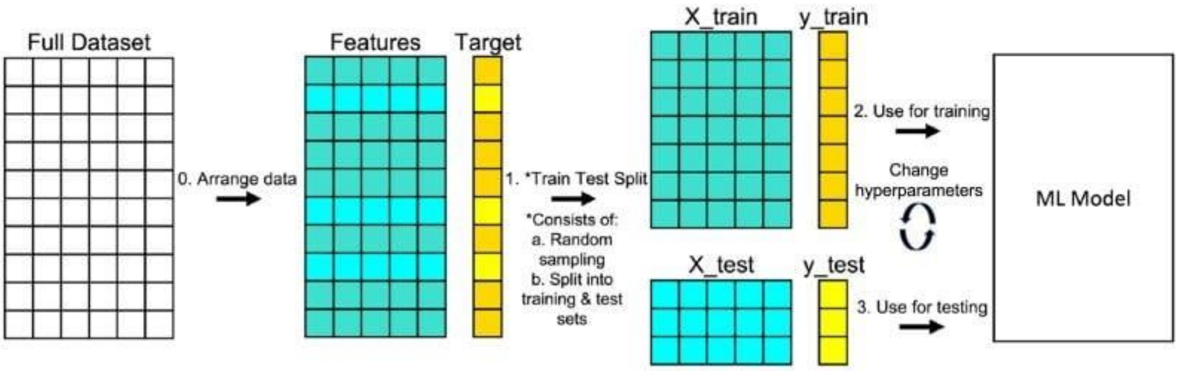
```
[ ]
    Principal Components:
       Absolute Magnitude  Est Dia in KM(min)  Est Dia in KM(max)  \
    0       -2.460032e-12        1.755439e-13        3.925281e-13
    1        3.618838e-08       -2.563721e-09       -5.732656e-09
    2       -2.697132e-05        2.383581e-06        5.329849e-06
    3       -4.948506e-04        1.189563e-04        2.659944e-04
    4        9.712952e-05        2.372703e-06        5.305526e-06

       Est Dia in M(min)  Est Dia in M(max)  Est Dia in Miles(min)  \
    0       1.755439e-10       3.925281e-10           1.090779e-13
    1      -2.563722e-06      -5.732656e-06          -1.593022e-09
    2       2.383581e-03       5.329849e-03           1.481088e-06
    3       1.189563e-01       2.659944e-01           7.391599e-05
    4       2.372703e-03       5.305526e-03           1.474329e-06

       Est Dia in Miles(max)  Est Dia in Feet(min)  Est Dia in Feet(max)  \
    0           2.439056e-13          5.759315e-10          1.287822e-09
    1          -3.562106e-09         -8.411160e-06         -1.880793e-05
    2           3.311813e-06          7.820147e-03          1.748638e-02
    3           1.652812e-04          3.902766e-01          8.726849e-01
    4           3.296700e-06          7.784460e-03          1.740658e-02

       Epoch Date Close Approach  ...   Inclination  Asc Node Longitude  \
    0              -1.000000e+00  ...  3.716359e-12        1.008034e-11
    1              -1.876563e-05  ... -1.070453e-07        1.061231e-07
    2               8.750895e-09  ...  1.714409e-04       -5.503398e-05
```

- These techniques help in selecting the most relevant features for the models.

Model Training & Selection:

Exploring various classification techniques including Decision Trees, Support Vector Machines, Neural Networks, and ensemble methods to identify the most suitable model for the task.

- It uses libraries **(sklearn,xgboost,lightgbm,catboost,tensorflow,keras)** for machine learning.
- **Split** your data into training and testing sets, then train your chosen model on the training data



- The code trains several machine learning **models** such as Decision Trees, Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machines (SVM), XGBoost, Neural Networks, and Gradient Boosting.

- It uses techniques like **SMOTE** for handling class imbalance.

  Before:

```
Class Distribution:
Hazardous
0    3932
1     755
Name: count, dtype: int64
```

After:

```
Class Distribution after SMOTE:
Hazardous
1    3932
0    3932
Name: count, dtype: int64
```

- **Model evaluation** metrics like accuracy, precision, recall, F1-score, specificity, and confusion matrix are calculated for each model.
- Hyperparameter tuning is performed using **GridSearchCV** to find the best parameters for models like Random Forest, SVM, XGBoost, Gradient Boosting, etc.

Evaluation Metrics:

Assessing model performance using standard classification metrics such as:

- **Accuracy** (Calculate the accuracy of your model on the test set to measure its overall performance)
- **Precision** (Calculate the ratio of true positive **(TP)** predictions to the total number of positive predictions made by the classifier)
- **Recall** (Calculate the ratio of true positive **(TP)** predictions to the total number of actual positive instances in the dataset)
- **F1-score** (It provides a balance between precision and recall)
- **Confusion Matrix** (Analyze the confusion matrix to understand the model's predictions in terms of true positives **(TP)**, true negatives **(TN)**, false positives **(FP)**, and false negatives **(FN)**)

- **ROC curve and AUC (**Plot the Receiver Operating Characteristic **(ROC)** curve and calculate the Area Under the Curve **(AUC)** score to assess the model's ability to discriminate between hazardous and non-hazardous asteroids)

By considering the class imbalance in the dataset.

Evaluation Metrics for Decision Tree Classifier:

```
Accuracy: 0.9974570883661793
Confusion Matrix:
 [[797   4]
 [  0 772]]
Precision: 0.9948453608247423
Recall: 1.0
Sensitivity: 1.0
Specificity: 0.9950062421972534
F1 Score: 0.9974160206718347
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       0.99      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

Evaluation Metrics for Random Forest Classifier:

```
Accuracy: 0.9987285441830897
Confusion Matrix:
[[799   2]
 [  0 772]]
Precision: 0.9974160206718347
Recall: 1.0
Sensitivity: 1.0
Specificity: 0.9975031210986267
F1 Score: 0.9987063389391979
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       1.00      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

Evaluation Metrics for KNN:

```
Accuracy: 0.7889383343928799
Confusion Matrix:
[[538 263]
 [ 69 703]]
Precision: 0.727743271221532
Recall: 0.9106217616580311
Sensitivity: 0.9106217616580311
Specificity: 0.6716604244694132
F1 Score: 0.8089758342922899
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.67      0.76       801
           1       0.73      0.91      0.81       772

    accuracy                           0.79      1573
   macro avg       0.81      0.79      0.79      1573
weighted avg       0.81      0.79      0.79      1573
```

Evaluation Metrics for Naïve Bayes:

```
Accuracy: 0.891290527654164
Confusion Matrix:
[[686 115]
 [ 56 716]]
Precision: 0.8616125150421179
Recall: 0.927461139896373
Sensitivity: 0.927461139896373
Specificity: 0.8564294631710362
F1 Score: 0.8933250155957578
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.86      0.89       801
           1       0.86      0.93      0.89       772

    accuracy                           0.89      1573
   macro avg       0.89      0.89      0.89      1573
weighted avg       0.89      0.89      0.89      1573
```

Evaluation Metrics for SVM:

```
Accuracy: 0.9828353464717101
Confusion Matrix:
[[782  19]
 [  8 764]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       801
           1       0.98      0.99      0.98       772

    accuracy                           0.98      1573
   macro avg       0.98      0.98      0.98      1573
weighted avg       0.98      0.98      0.98      1573

Precision: 0.9757343550446999
Recall: 0.9896373056994818
Sensitivity: 0.9896373056994818
F1 Score: 0.982636655948553
Specificity: 0.9762796504369539
```

Evaluation Metrics for XGBOOST:

```
Accuracy: 0.9980928162746344
Confusion Matrix:
[[798   3]
 [  0 772]]
Precision: 0.9961290322580645
Recall: 1.0
Sensitivity: 1.0
Specificity: 0.9962546816479401
F1 Score: 0.9980607627666451
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       1.00      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

Evaluation Metrics for Neural Network:

```
Accuracy: 0.9847425301970757
Confusion Matrix:
 [[792   9]
 [ 15 757]]
Precision: 0.9882506527415144
Recall: 0.9805699481865285
Sensitivity: 0.9805699481865285
F1 Score: 0.9843953185955786
Specificity: 0.9887640449438202
Classification Report:
               precision    recall  f1-score   support

           0       0.98      0.99      0.99       801
           1       0.99      0.98      0.98       772

    accuracy                           0.98      1573
   macro avg       0.98      0.98      0.98      1573
weighted avg       0.98      0.98      0.98      1573
```

Evaluation Metrics for Gradient Boost:

```
Accuracy: 0.9980928162746344
Confusion Matrix:
 [[798   3]
 [  0 772]]
Precision: 0.9961290322580645
Recall: 1.0
Sensitivity: 1.0
Specificity: 0.9962546816479401
F1 Score: 0.9980607627666451
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       1.00      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

When using GridCV:

Evaluation Metrics for Random Forest:

```
[ ] Accuracy: 0.9972979286813797 Parameters: {'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100}
    Accuracy: 0.9976158937052271 Parameters: {'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}
    Accuracy: 0.9972979286813797 Parameters: {'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 300}

    Best Parameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200}

    Best Accuracy: 0.9980928162746344
    Precision: 0.9974126778783958
    Recall: 0.9987046632124352
    Sensitivity: 0.9987046632124352
    Specificity: 0.9975031210986267
    F1 Score: 0.9980582524271844

    Confusion Matrix:
    [[799   2]
     [  1 771]]

    Classification Report:
                  precision    recall  f1-score   support

               0       1.00      1.00      1.00       801
               1       1.00      1.00      1.00       772

        accuracy                           1.00      1573
       macro avg       1.00      1.00      1.00      1573
    weighted avg       1.00      1.00      1.00      1573
```

Evaluation Metrics for Decision Tree Classifier:

```
Accuracy: 0.9974570883661793
Precision: 0.9961240310077519
Recall: 0.9987046632124352
Sensitivity: 0.9987046632124352
Specificity: 0.9962546816479401
F1 Score: 0.9974126778783958

Confusion Matrix:
[[798   3]
 [  1 771]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       1.00      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

Evaluation Metrics for KNN:

```
Accuracy: 0.87730451366815
Precision: 0.8150163220892275
Recall: 0.9702072538860104
Sensitivity: 0.9702072538860104
Specificity: 0.787765293383271
F1 Score: 0.8858663512714371
Condusion matrix:
[[631 170]
 [ 23 749]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.79      0.87       801
           1       0.82      0.97      0.89       772

    accuracy                           0.88      1573
   macro avg       0.89      0.88      0.88      1573
weighted avg       0.89      0.88      0.88      1573
```

Evaluation Metrics for Naïve Bayes:

```
Accuracy: 0.891290527654164
Precision: 0.8616125150421179
Recall: 0.927461139896373
Sensitivity: 0.927461139896373
Specificity: 0.8564294631710362
F1 Score: 0.8933250155957578

Confusion Matrix:
[[686 115]
 [ 56 716]]

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.86      0.89       801
           1       0.86      0.93      0.89       772

    accuracy                           0.89      1573
   macro avg       0.89      0.89      0.89      1573
weighted avg       0.89      0.89      0.89      1573
```

Evaluation Metrics for SVM:

```
Best Parameters: {'C': 20, 'gamma': 'auto', 'kernel': 'rbf'}

Accuracy: 0.9866497139224412
Precision: 0.982028241335045
Recall: 0.9909326424870466
Sensitivity: 0.9909326424870466
Specificity: 0.982521847690387
F1 Score: 0.9864603481624759

Confusion Matrix:
[[787  14]
 [  7 765]]

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       801
           1       0.98      0.99      0.99       772

    accuracy                           0.99      1573
   macro avg       0.99      0.99      0.99      1573
weighted avg       0.99      0.99      0.99      1573
```

Evaluation Metrics for XGBOOST:

```
Best Accuracy: 0.9974570883661793
Precision: 0.9961240310077519
Recall: 0.9987046632124352
Sensitivity: 0.9987046632124352
Specificity: 0.9962546816479401
F1 Score: 0.9974126778783958

Confusion Matrix:
[[798   3]
 [  1 771]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       1.00      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

Evaluation Metrics for Neural Networks:

```
Best Parameters: {'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': 100, '

Accuracy: 0.9942784488239034
Precision: 0.990990990990991
Recall: 0.9974093264248705
Sensitivity: 0.9974093264248705
Specificity: 0.9912609238451935
F1 Score: 0.9941897998708844

Confusion Matrix:
[[794   7]
 [  2 770]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       801
           1       0.99      1.00      0.99       772

    accuracy                           0.99      1573
   macro avg       0.99      0.99      0.99      1573
weighted avg       0.99      0.99      0.99      1573
```

Evaluation Metrics for Gradient Boost:

```
Best Parameters: {'learning_rate': 0.05, 'max_depth': 6, 'max_features': 'log2',

Accuracy: 0.9993642720915448
Precision: 0.9987063389391979
Recall: 1.0
Sensitivity: 1.0
Specificity: 0.9987515605493134
F1 Score: 0.9993527508090615

Confusion Matrix:
[[800   1]
 [  0 772]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       801
           1       1.00      1.00      1.00       772

    accuracy                           1.00      1573
   macro avg       1.00      1.00      1.00      1573
weighted avg       1.00      1.00      1.00      1573
```

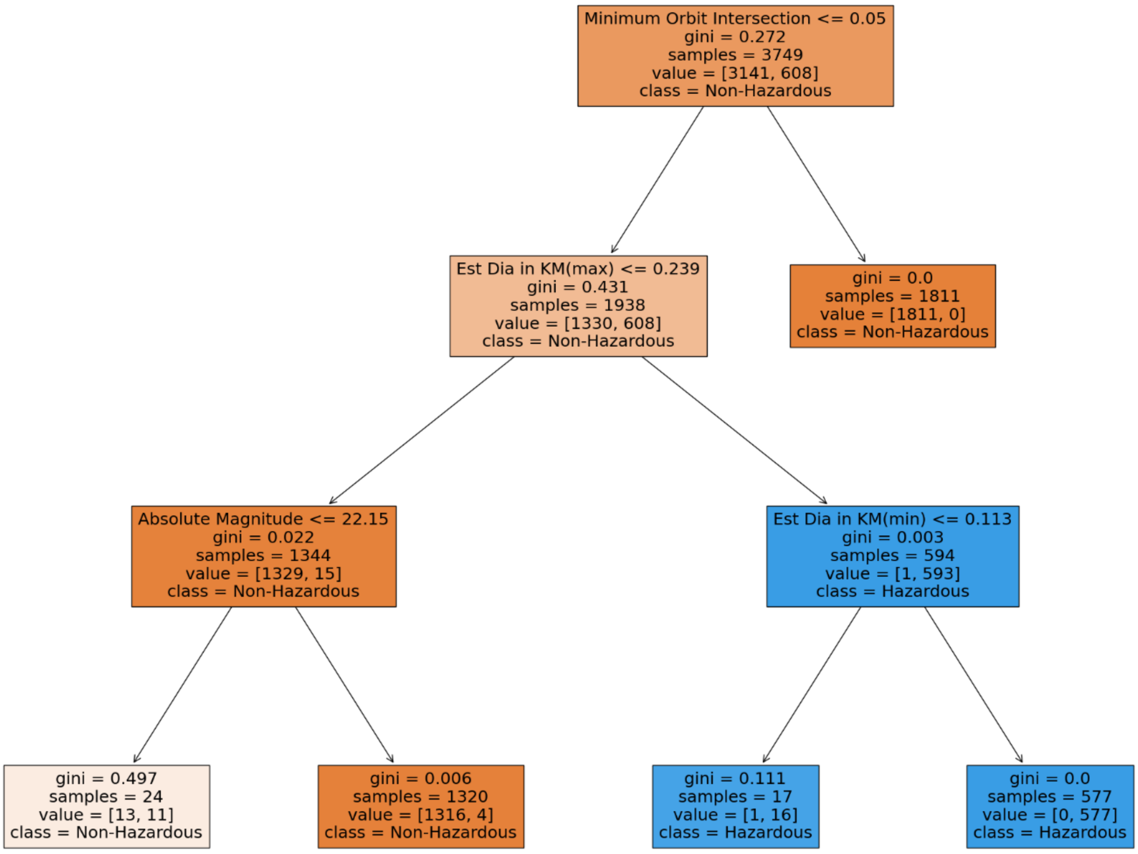
Hazard Prediction:

Utilizing feature importance techniques to identify key factors contributing to asteroid hazard potential.

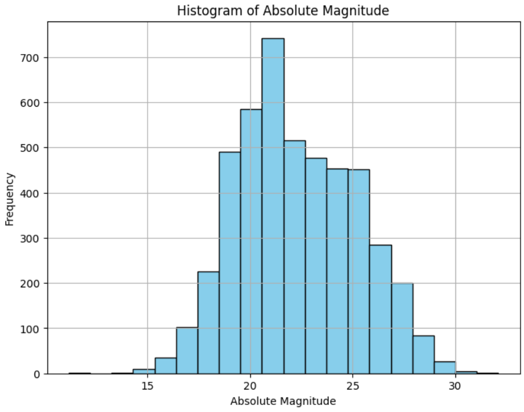
Visualization:

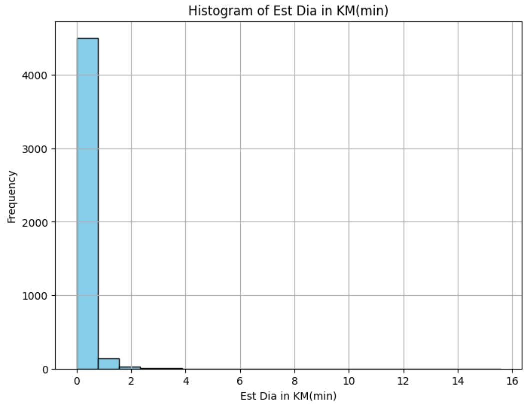
Visualizing the dataset and model predictions using plots and graphs to gain insights into the data distribution and model performance.

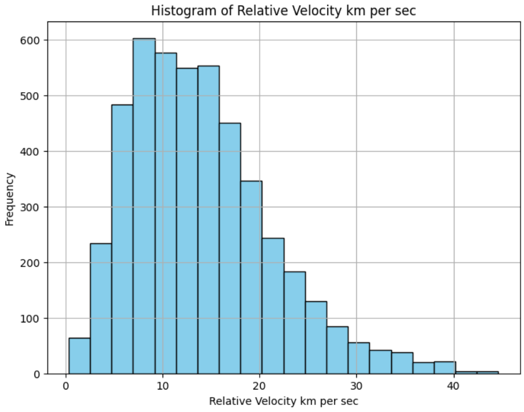
- **Decision Tree** visualization is performed to understand the decision-making process of the model.

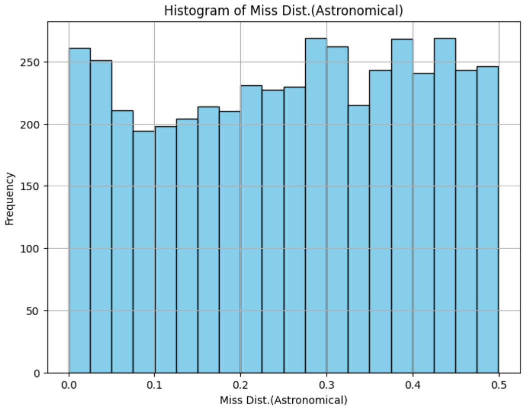


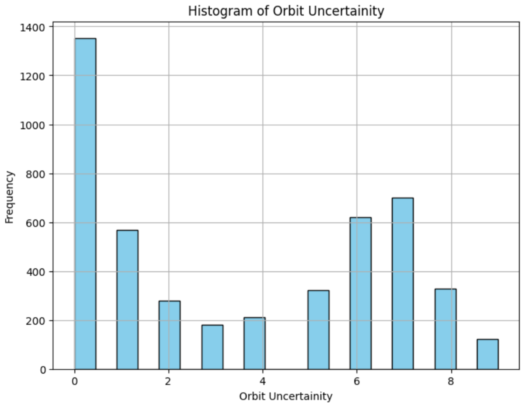
- **Histograms:**

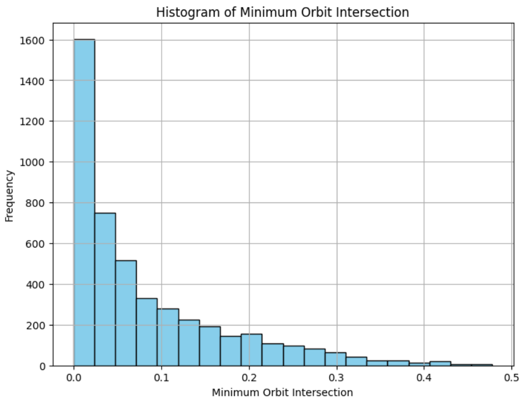

Histogram of Absolute Magnitude
Frequency
Absolute Magnitude

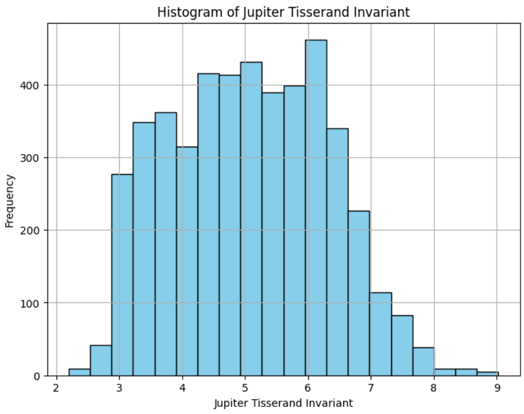

Histogram of Est Dia in KM(min)
Frequency
Est Dia in KM(min)

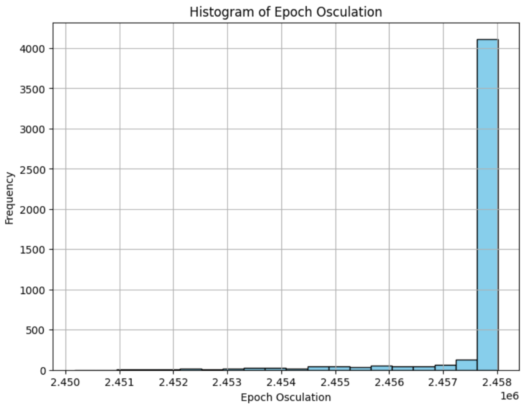

Histogram of Relative Velocity km per sec
Frequency
Relative Velocity km per sec

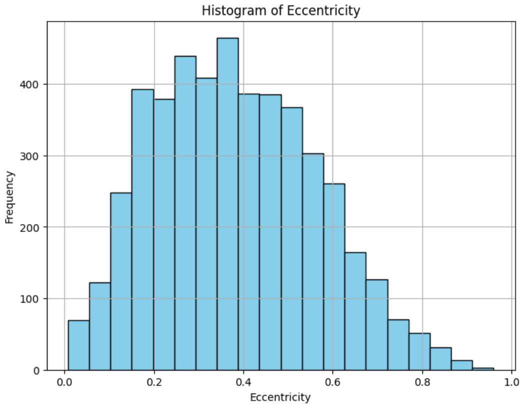

Histogram of Miss Dist.(Astronomical)
Frequency
Miss Dist.(Astronomical)

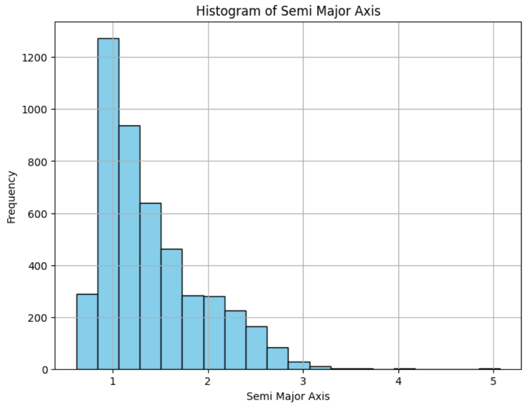

Histogram of Orbit Uncertainity
Frequency
Orbit Uncertainity

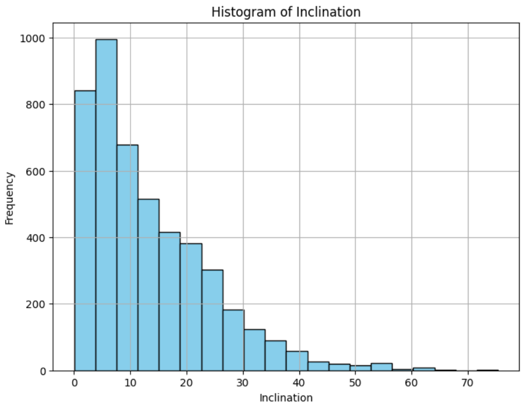

Histogram of Minimum Orbit Intersection
Frequency
Minimum Orbit Intersection

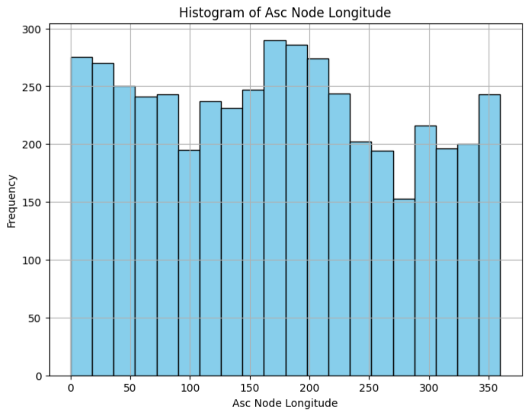

Histogram of Jupiter Tisserand Invariant
Frequency
Jupiter Tisserand Invariant

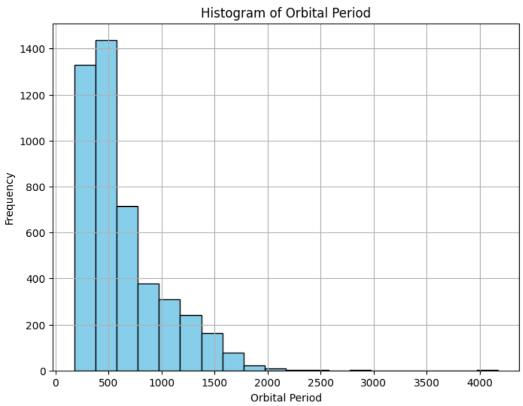

Histogram of Epoch Osculation
Frequency
Epoch Osculation
1e6

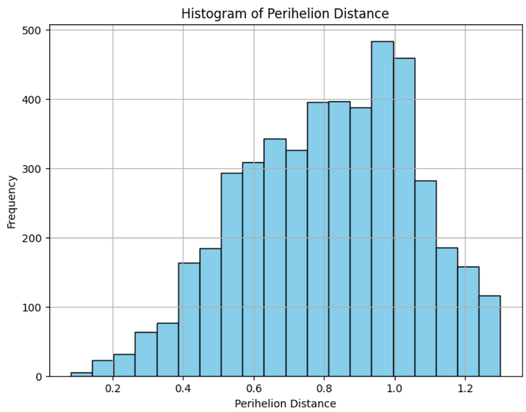

Histogram of Eccentricity
Frequency
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
Eccentricity

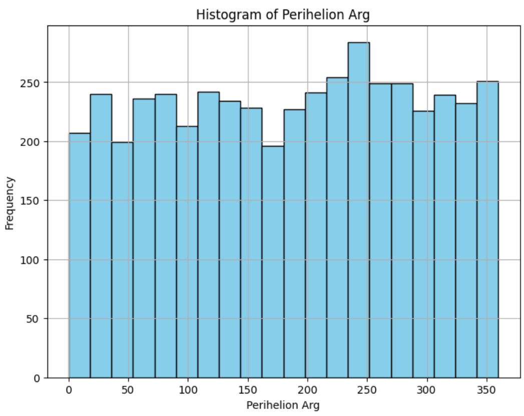

Histogram of Semi Major Axis
Frequency
0
200
400
600
800
1000
1200
1
2
3
4
5
Semi Major Axis

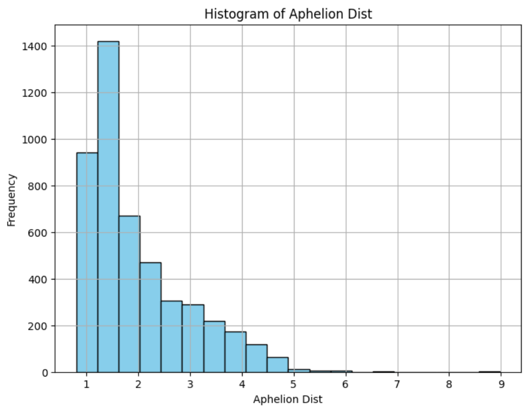

Histogram of Inclination
Frequency
0
200
400
600
800
1000
0
10
20
30
40
50
60
70
Inclination

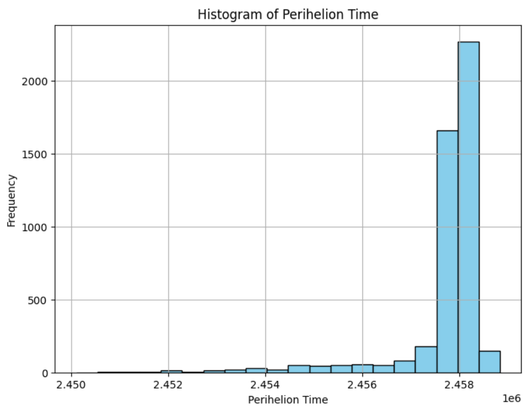
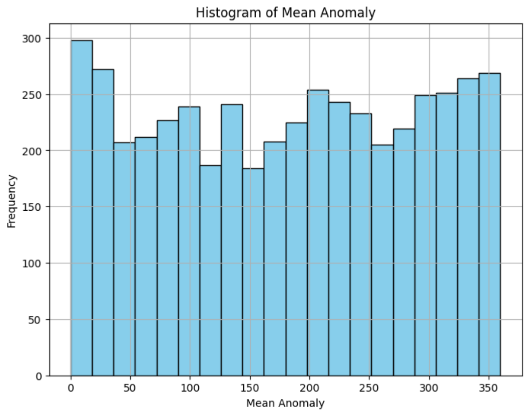
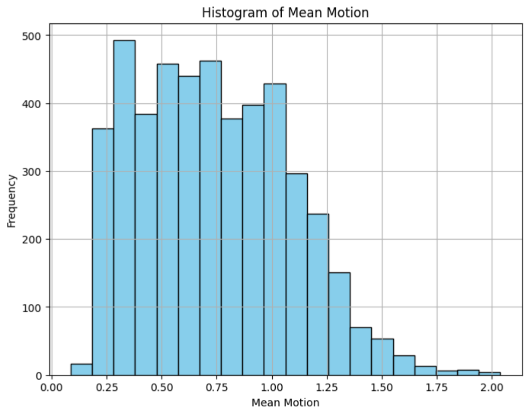
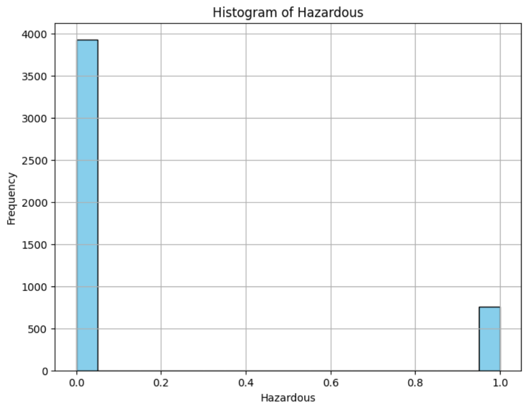

Histogram of Asc Node Longitude
Frequency
0
50
100
150
200
250
300
0
50
100
150
200
250
300
350
Asc Node Longitude


Histogram of Orbital Period
Frequency
0
200
400
600
800
1000
1200
1400
0
500
1000
1500
2000
2500
3000
3500
4000
Orbital Period

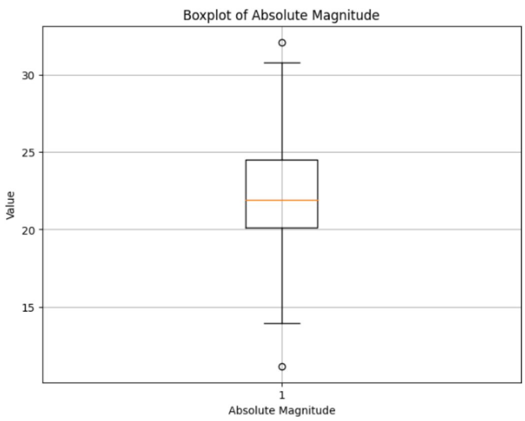
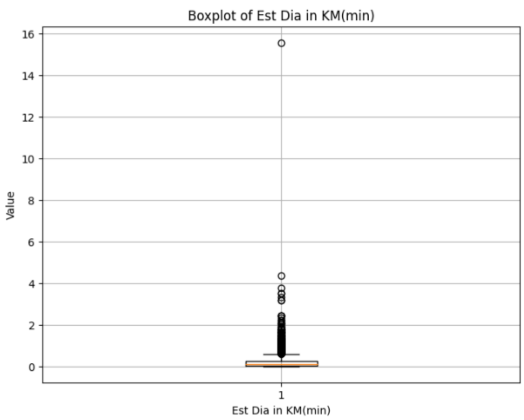

Histogram of Perihelion Distance
Frequency
0
100
200
300
400
500
0.2
0.4
0.6
0.8
1.0
1.2
Perihelion Distance

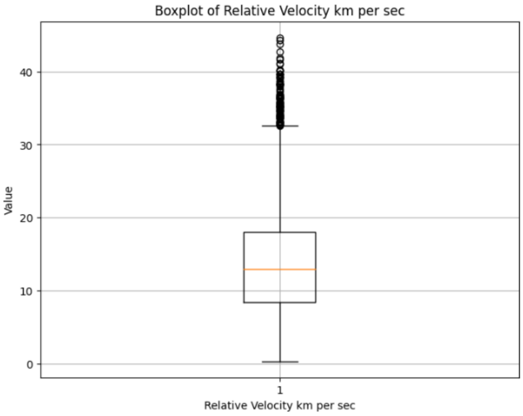

Histogram of Perihelion Arg
Frequency
0
50
100
150
200
250
0
50
100
150
200
250
300
350
Perihelion Arg

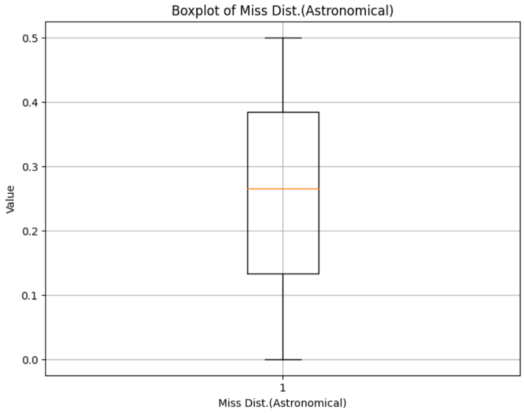

Histogram of Aphelion Dist
Frequency
0
200
400
600
800
1000
1200
1400
1
2
3
4
5
6
7
8
9
Aphelion Dist

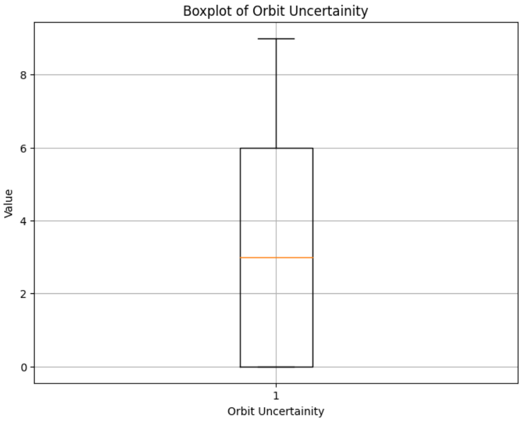

Histogram of Perihelion Time
Frequency
Perihelion Time
1e6

Histogram of Mean Anomaly
Frequency
Mean Anomaly

Histogram of Mean Motion
Frequency
Mean Motion

Histogram of Hazardous
Frequency
Hazardous

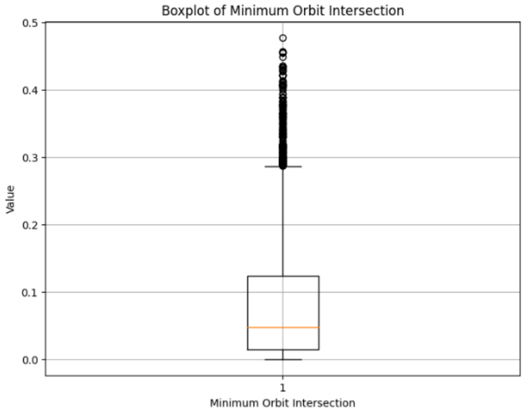
- **Box Plots:**

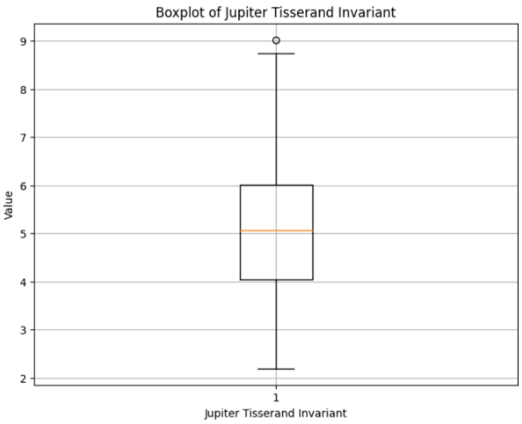

Boxplot of Absolute Magnitude
Value
Absolute Magnitude

Boxplot of Est Dia in KM(min)
Value
Est Dia in KM(min)

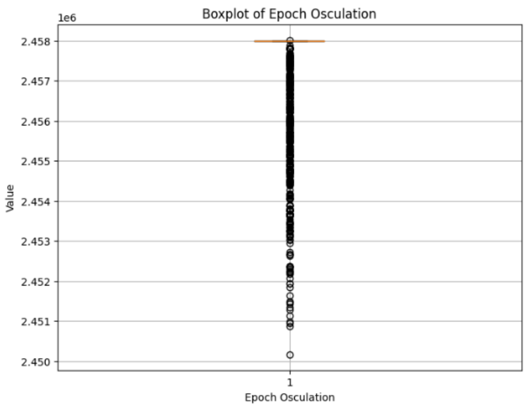

Boxplot of Relative Velocity km per sec
Value
0
10
20
30
40
1
Relative Velocity km per sec

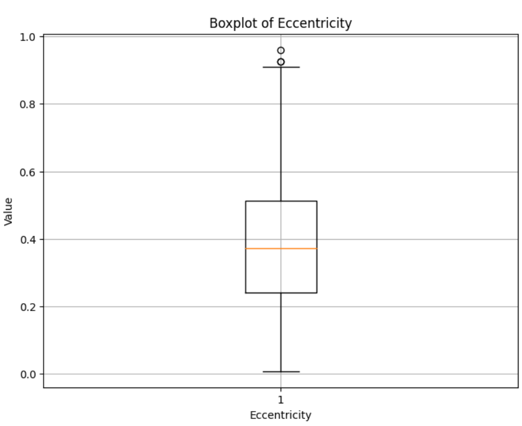

Boxplot of Miss Dist.(Astronomical)
Value
0.0
0.1
0.2
0.3
0.4
0.5
1
Miss Dist.(Astronomical)

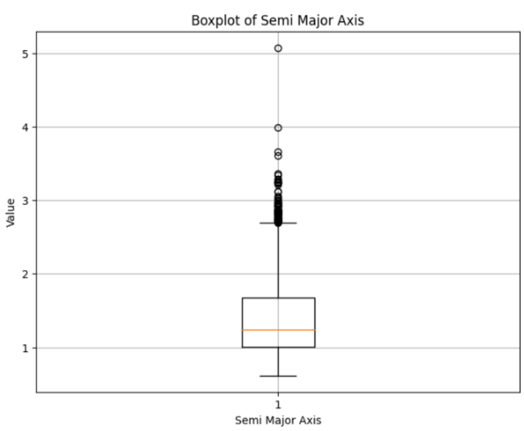

Boxplot of Orbit Uncertainity
Value
0
2
4
6
8
1
Orbit Uncertainity

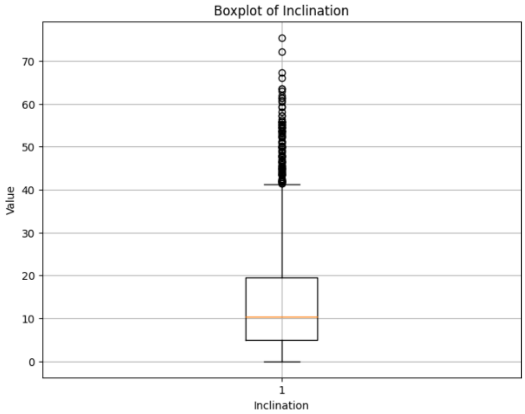

Boxplot of Minimum Orbit Intersection
Value
0.0
0.1
0.2
0.3
0.4
0.5
1
Minimum Orbit Intersection

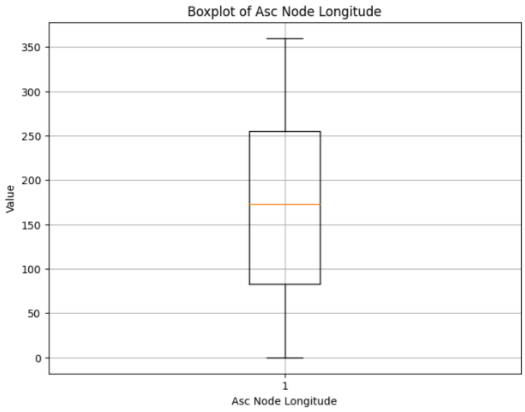

Boxplot of Jupiter Tisserand Invariant
Value
2
3
4
5
6
7
8
9
1
Jupiter Tisserand Invariant

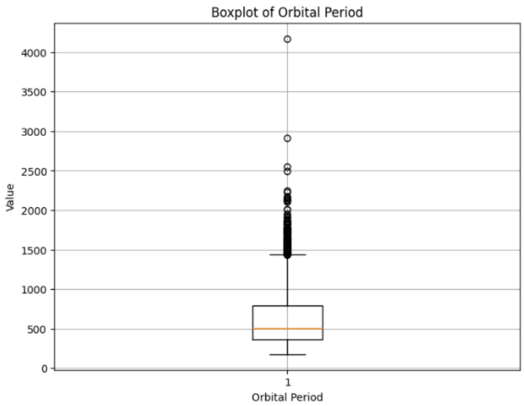

Boxplot of Epoch Osculation
1e6
Value
2.450
2.451
2.452
2.453
2.454
2.455
2.456
2.457
2.458
1
Epoch Osculation

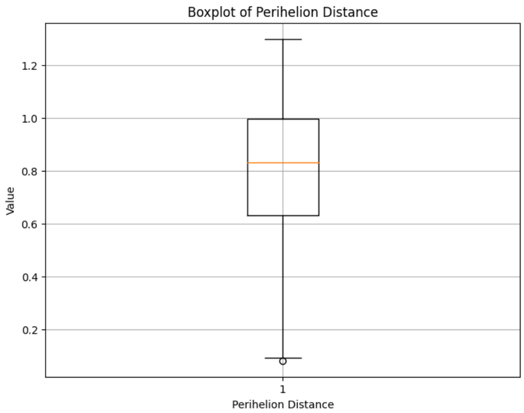

Boxplot of Eccentricity
Value
0.0
0.2
0.4
0.6
0.8
1.0
1
Eccentricity

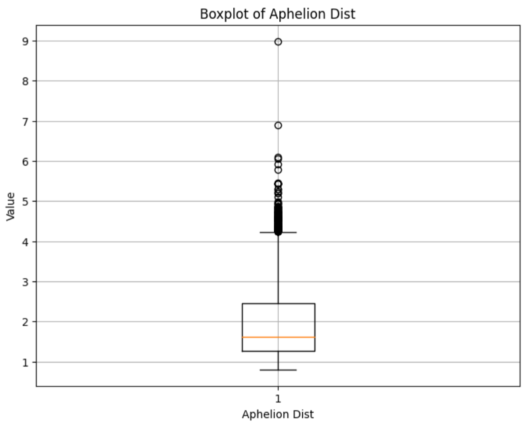

Boxplot of Semi Major Axis
Value
1
2
3
4
5
1
Semi Major Axis

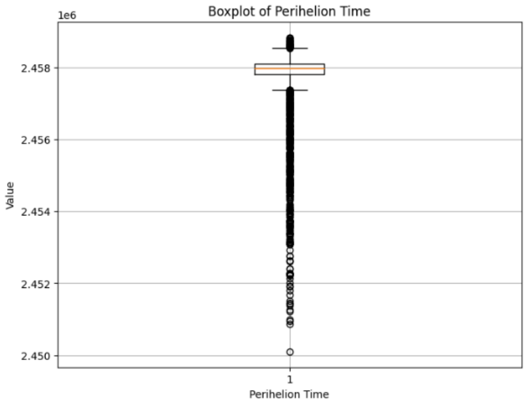

Boxplot of Inclination
Value
0
10
20
30
40
50
60
70
1
Inclination

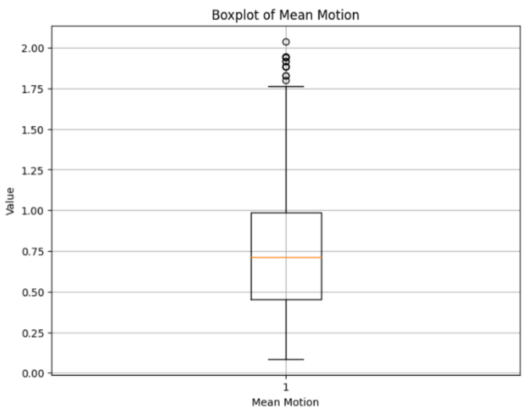

Boxplot of Asc Node Longitude
Value
0
50
100
150
200
250
300
350
1
Asc Node Longitude

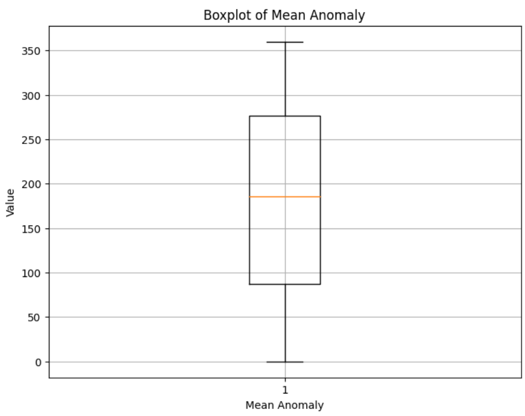

Boxplot of Orbital Period
Value
0
500
1000
1500
2000
2500
3000
3500
4000
1
Orbital Period


Boxplot of Perihelion Distance
Value
0.2
0.4
0.6
0.8
1.0
1.2
1
Perihelion Distance

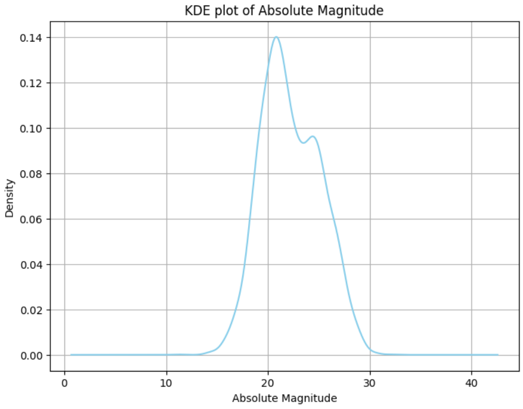

Boxplot of Aphelion Dist
Value
1
2
3
4
5
6
7
8
9
1
Aphelion Dist

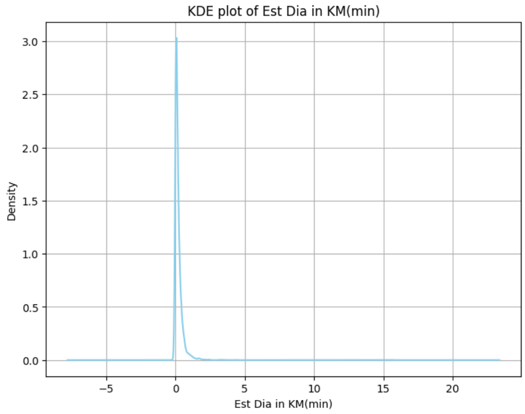

Boxplot of Perihelion Time
1e6
Value
2.450
2.452
2.454
2.456
2.458
1
Perihelion Time

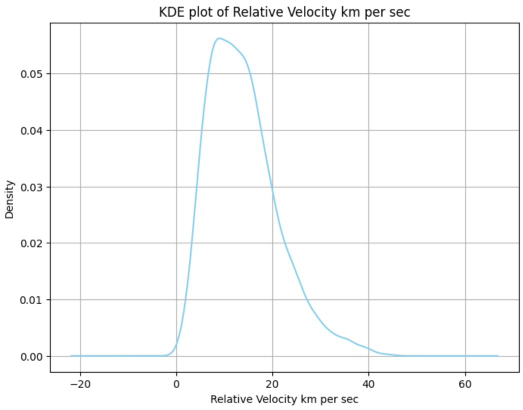

Boxplot of Mean Motion
Value
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1
Mean Motion

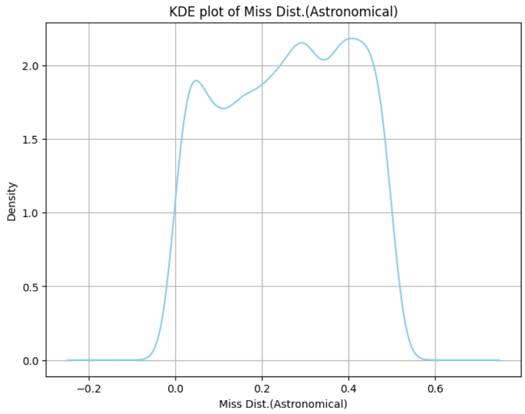

Boxplot of Mean Anomaly
Value
0
50
100
150
200
250
300
350
1
Mean Anomaly

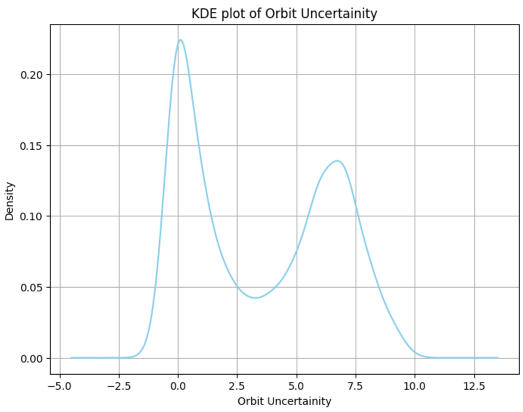
- ## Kernel Density Estimate (KDE):

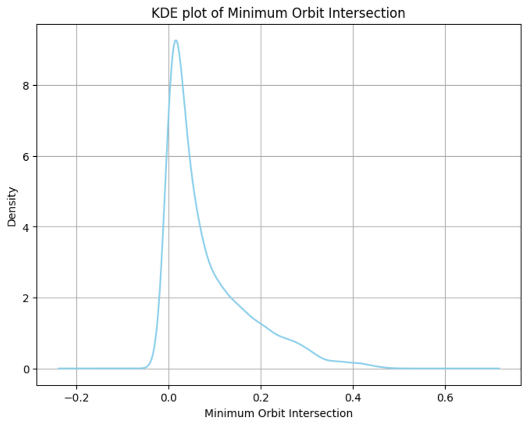


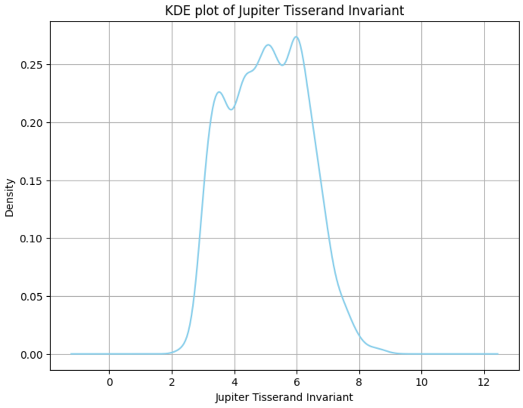


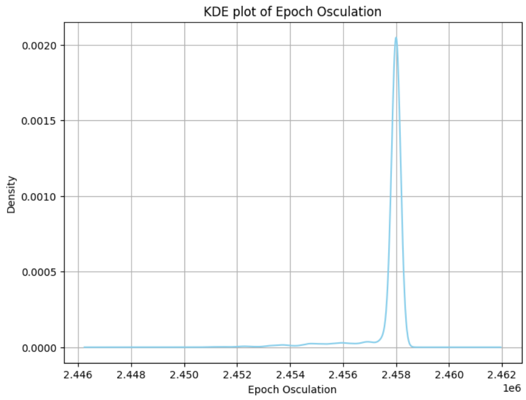


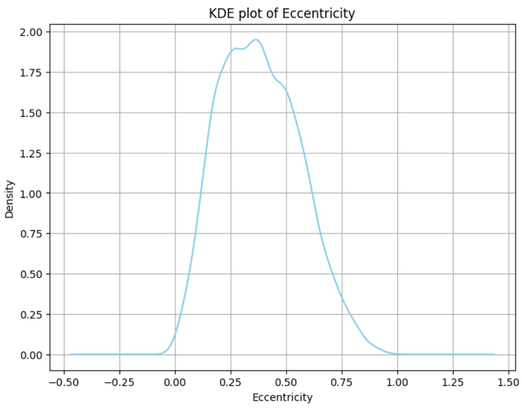


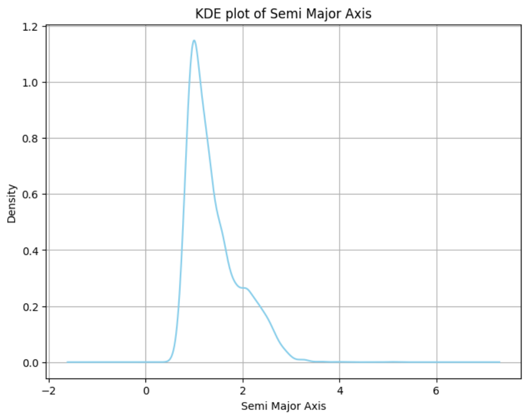




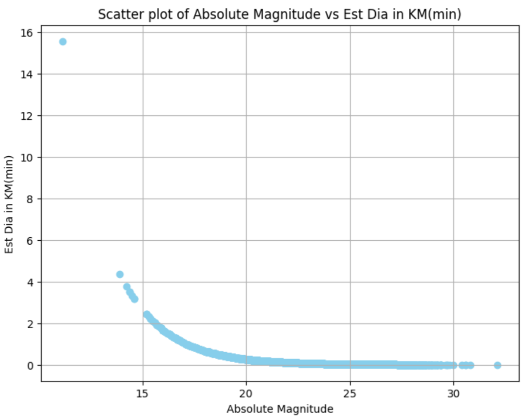


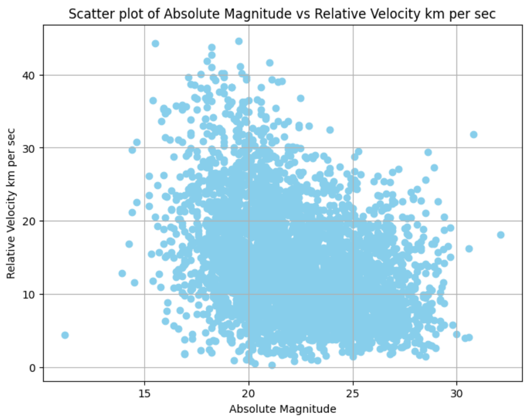


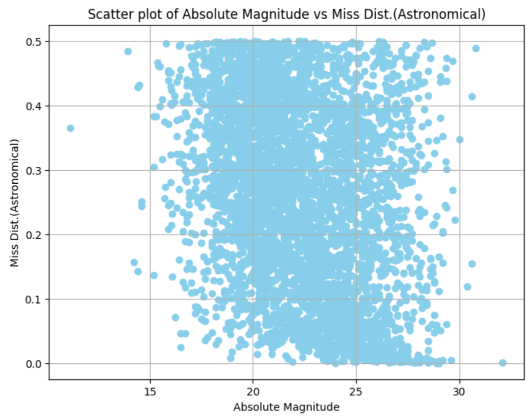


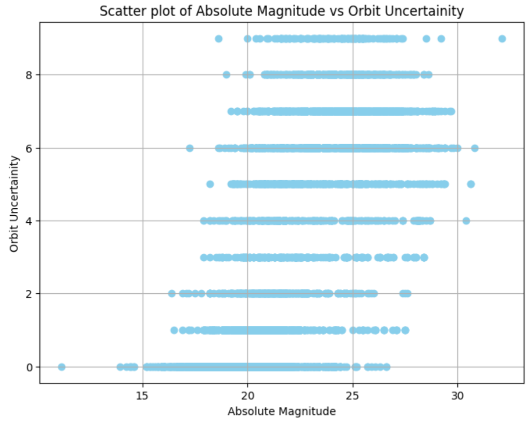


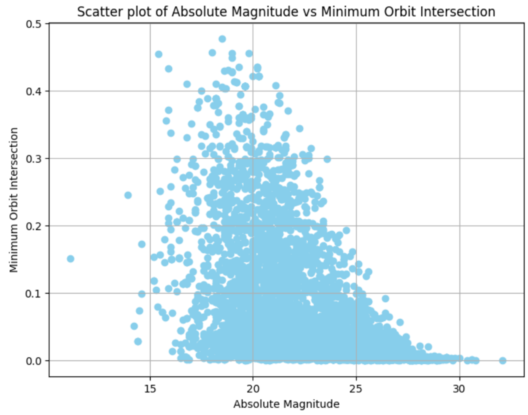
- **Scatter Plot:**

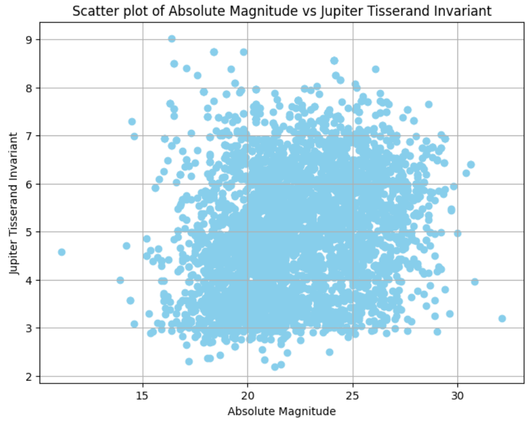


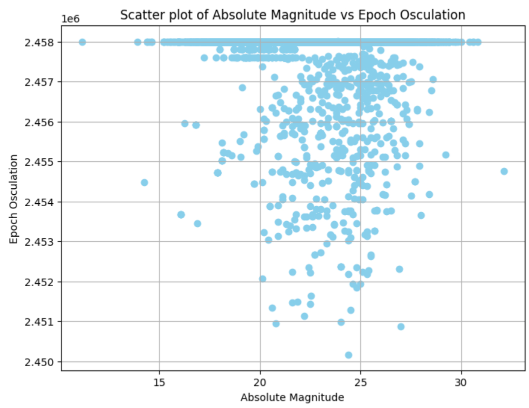


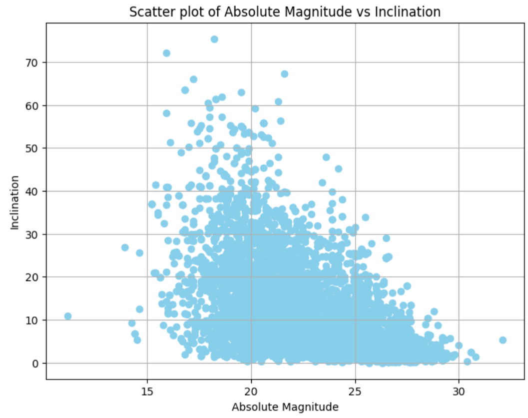

Scatter plot of Absolute Magnitude vs Miss Dist.(Astronomical)
Miss Dist.(Astronomical)
Absolute Magnitude

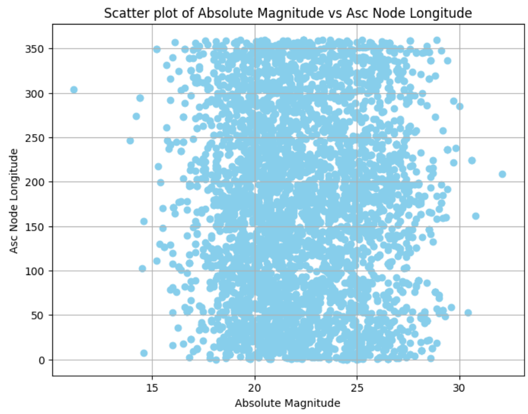

Scatter plot of Absolute Magnitude vs Orbit Uncertainity
Orbit Uncertainity
Absolute Magnitude

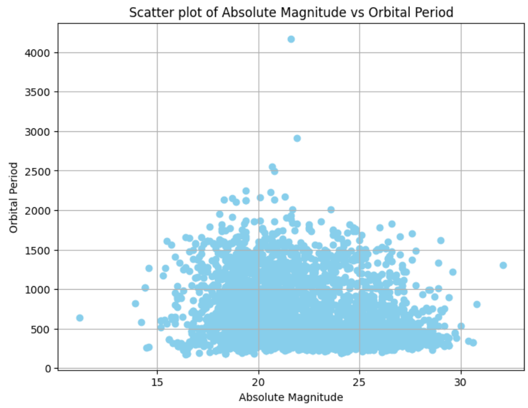

Scatter plot of Absolute Magnitude vs Minimum Orbit Intersection
Minimum Orbit Intersection
Absolute Magnitude

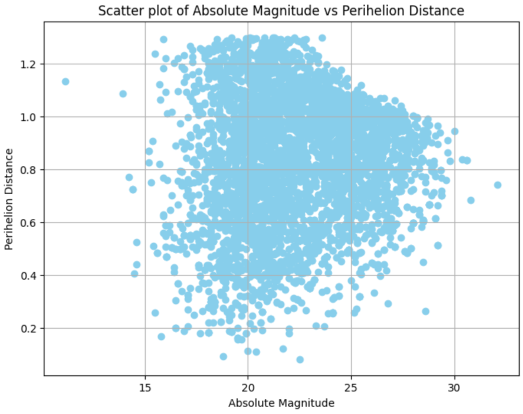

Scatter plot of Absolute Magnitude vs Jupiter Tisserand Invariant
Jupiter Tisserand Invariant
Absolute Magnitude

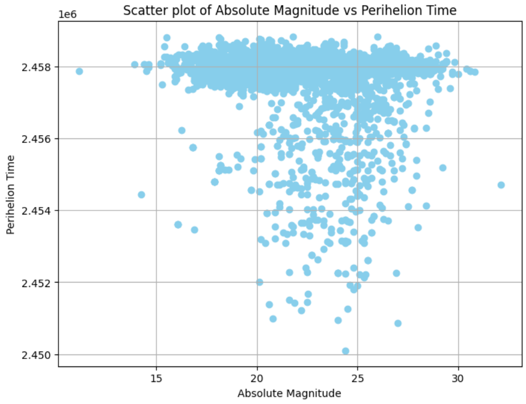

Scatter plot of Absolute Magnitude vs Epoch Osculation
1e6
Epoch Osculation
Absolute Magnitude


Scatter plot of Absolute Magnitude vs Inclination
Inclination
Absolute Magnitude

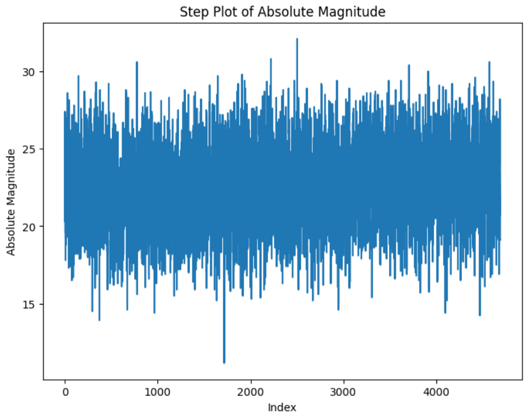

Scatter plot of Absolute Magnitude vs Asc Node Longitude
Asc Node Longitude
Absolute Magnitude

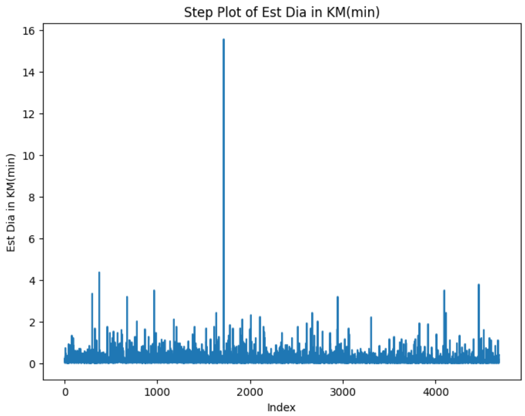

Scatter plot of Absolute Magnitude vs Orbital Period
Orbital Period
Absolute Magnitude

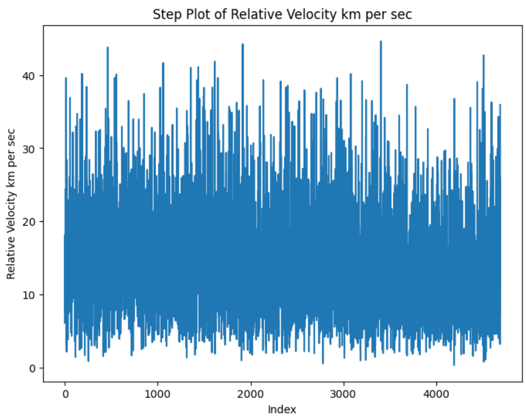

Scatter plot of Absolute Magnitude vs Perihelion Distance
Perihelion Distance
Absolute Magnitude

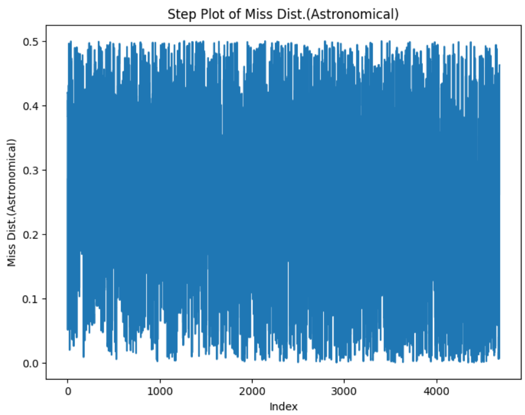

Scatter plot of Absolute Magnitude vs Perihelion Time
1e6
Perihelion Time
Absolute Magnitude

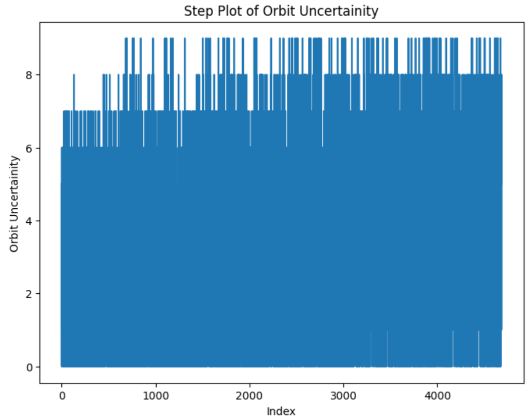
- **Bar Graphs:**

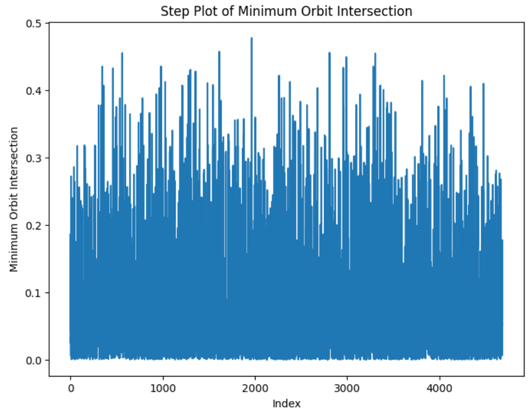

Step Plot of Absolute Magnitude
Absolute Magnitude
Index

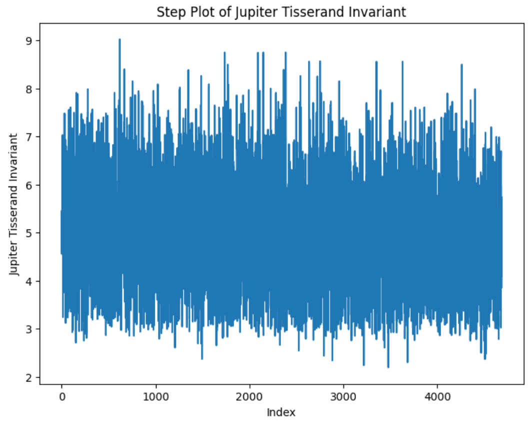

Step Plot of Est Dia in KM(min)
Est Dia in KM(min)
Index

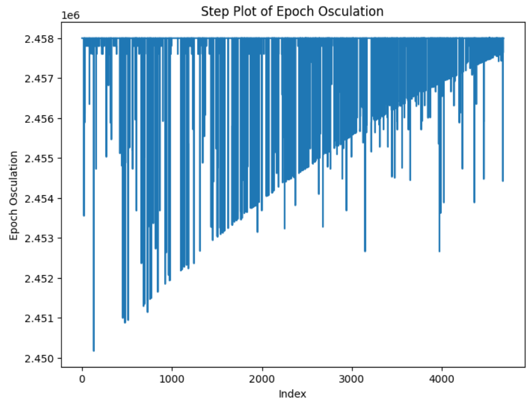

Step Plot of Relative Velocity km per sec
Relative Velocity km per sec
Index

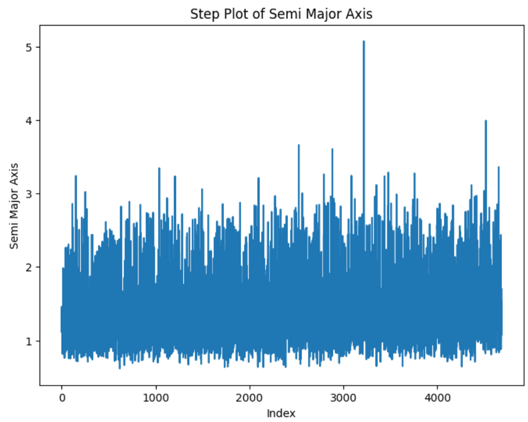

Step Plot of Miss Dist.(Astronomical)
Miss Dist.(Astronomical)
Index

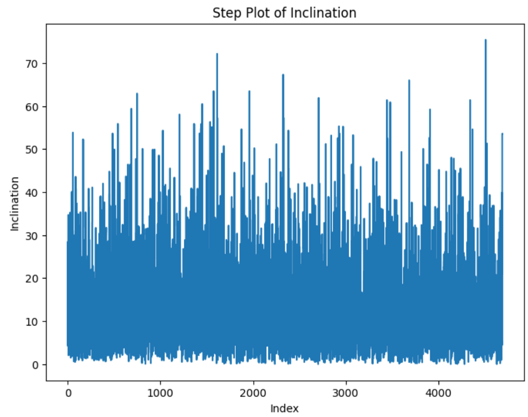

Step Plot of Orbit Uncertainity
Orbit Uncertainity
Index

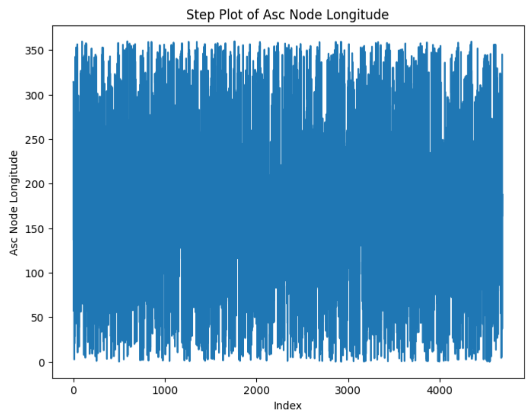

Step Plot of Minimum Orbit Intersection
Minimum Orbit Intersection
Index

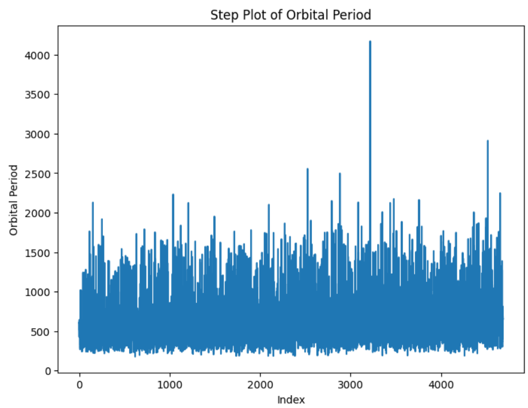

Step Plot of Jupiter Tisserand Invariant
Jupiter Tisserand Invariant
Index

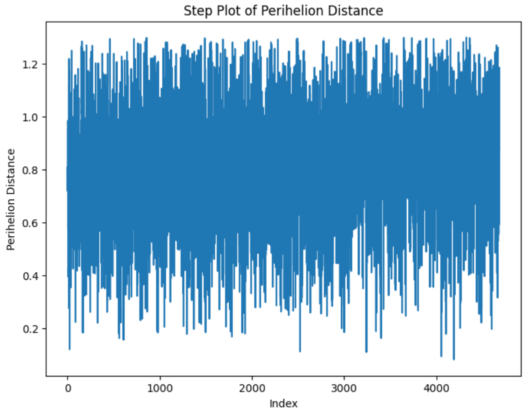

Step Plot of Epoch Osculation
1e6
Epoch Osculation
Index

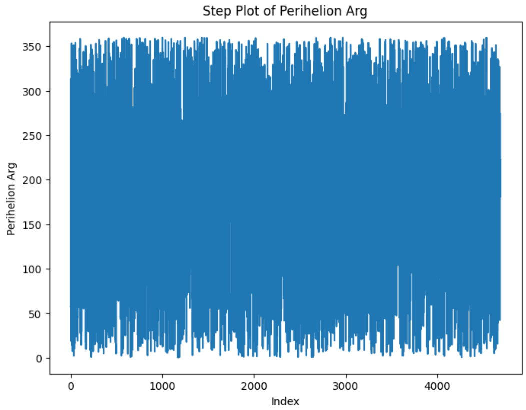

Step Plot of Semi Major Axis
Semi Major Axis
Index

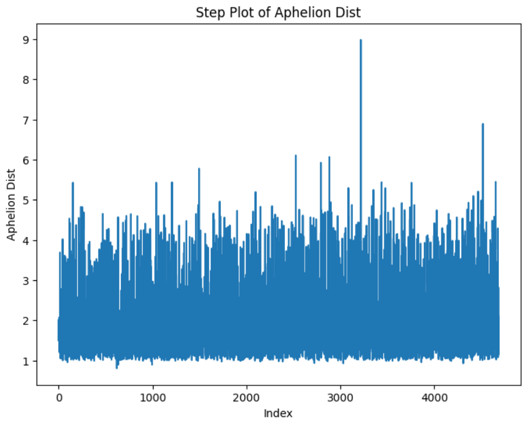

Step Plot of Inclination
Inclination
Index

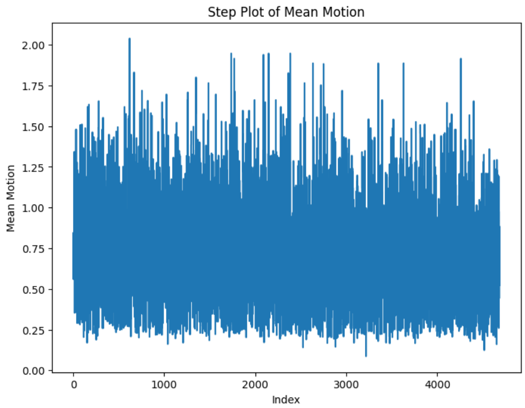

Step Plot of Asc Node Longitude
Asc Node Longitude
Index


Step Plot of Orbital Period
Orbital Period
Index



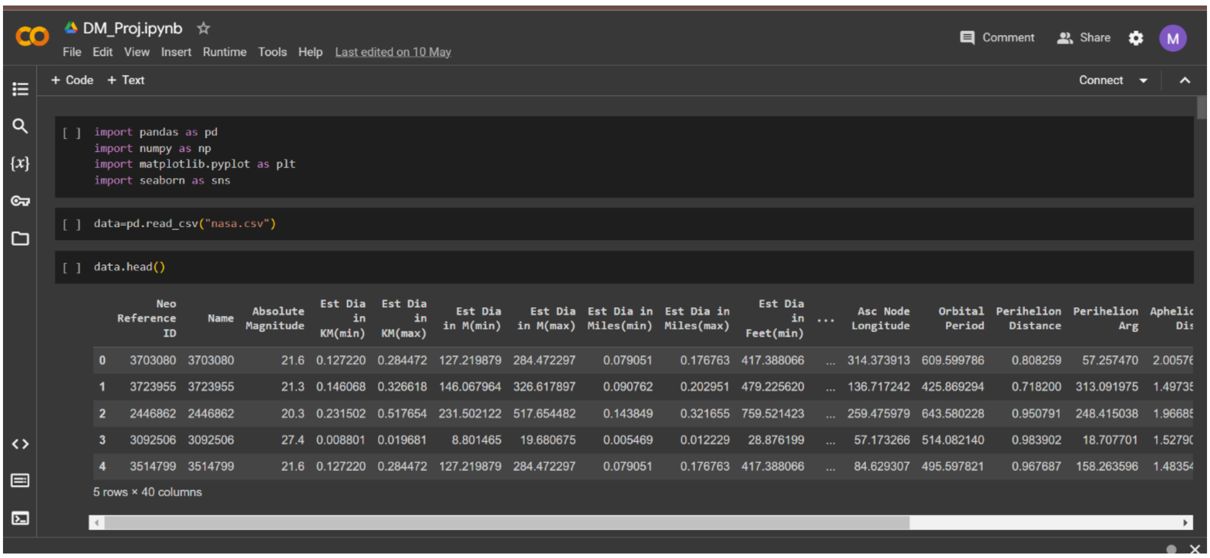






Software Tools:

Jupyter Notebook or Google Colab will be utilized for data exploration, preprocessing, model development, and evaluation. Libraries such as **Pandas, NumPy, Scikit-learn,** and **Matplotlib** will facilitate data manipulation, modeling, and visualization.



Expected Results and Evaluation Techniques:

The expected outcome is a classification model that accurately predicts whether an asteroid is hazardous or not, based on its features. Model evaluation will be conducted using standard classification metrics such as accuracy, precision, recall, and F1-score. Additionally, ROC-AUC curve analysis will assess the model's ability to discriminate between hazardous and non-hazardous asteroids.

Preliminary Results and Dataset Exploration:

The dataset will be explored through Exploratory Data Analysis (EDA) to understand feature distribution, detect missing values or outliers, and identify potential challenges. Visualization techniques will be employed to gain insights into feature-target relationships. Preprocessing steps such as handling missing values, encoding categorical variables, and feature scaling will be performed as needed before model training.

Outline of the Work-to-Do:

Data preprocessing: Handling missing values, normalizing features, and encoding categorical variables.

Feature engineering: Extracting relevant features and performing dimensionality reduction if necessary.

Model training: Experimenting with different classification algorithms and hyperparameters.

Model evaluation: Assessing model performance using appropriate evaluation metrics.

Feature importance analysis: Identifying key features influencing asteroid hazard potential.

Visualization: Visualizing data distributions, model predictions, and feature importance to facilitate interpretation.

<u>Conclusion</u>

This project stands as a pivotal response to the pressing necessity for precise classification and predictive modeling concerning the potential hazards posed by **asteroids**, employing sophisticated data mining methodologies. Through the development of a resilient **classification model** and the discernment of pivotal features, the overarching goal is to significantly elevate our capacity to evaluate and effectively mitigate the risks entwined with near-Earth asteroids. By delineating and comprehensively analyzing various asteroid properties and orbital characteristics, this endeavor strives to furnish stakeholders with invaluable insights into the nature and severity of **potential threats**, thereby empowering informed decision-making and proactive risk management strategies. Moreover, the iterative nature of this project ensures that further refinement and validation of the model will be rigorously pursued, leveraging the insights gleaned throughout its course to continually enhance the **accuracy, reliability, and applicability** of the developed predictive framework. In doing so, this initiative not only contributes to the advancement of scientific understanding in the field of asteroid hazard assessment but also holds the potential to safeguard lives, infrastructure, and planetary well-being through proactive and informed risk mitigation measures.

References

[1] NeoWs (Near Earth Object Web Service). Follow: https://catalog.data.gov/dataset/

[2] Pandas documentation. Follow: https://pandas.pydata.org/docs/

[3] Scikit-learn documentation. Follow: https://scikit-learn.org/stable/

[4] Matplotlib documentation. Follow: https://matplotlib.org/stable/index.html

[5] Jupyter Notebook documentation. Follow: https://docs.jupyter.org/en/latest/

[6] SMOTE for Imbalanced Classification with Python. Follow: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[7] Decision tree – Scikit Learn. Follow: https://scikit-learn.org/stable/modules/tree.html

[8] Random Forest Classifier. Follow: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[9] XGBOOST documentation: Follow: https://xgboost.readthedocs.io/en/stable/

[10] LightGBM documentation: Follow: https://lightgbm.readthedocs.io/en/stable/

[11] catBoost documentation. Follow: https://catboost.ai/