**Table of contents**

1. Set up and data frame creation
   1. Unzipping
   2. reading data
2. Data Cleaning and Preparation
   1. Column Descriptions
   2. Imputation
   3. Dropping Null values
3. Data Visualization
   1. Feature transformation
   2. Correlation Visualization
      1. Dropping Correlated columns
   3. Histogram Plotting
   4. Box Plots
4. Modeling
   1. train test split of training data
   2. linear regression model
   3. Gradient boosting regression
   4. Hist Gradient Boosting regression
5. model saving
6. Loading actual Test data
   1. Test data preparation
   2. Prediction

# Asteroid Dataset Analysis Report

## Table of Contents

## 1. Set up and Data Frame Creation

### a. Unzipping

* Description: Unzipping the asteroid dataset.
* Status: Complete

### b. Reading Data

* Description: Reading the asteroid dataset.
* Status: Completed

## 2. Data Cleaning and Preprocessing

### Column Description

|  |  |  |  |
| --- | --- | --- | --- |
| Column Description | Null values | Column Status | Drop condition |
| Object Full Name | 654128 | Dropped | Due to high null values |
| a (Semi-Major Axis) | 2 | Used | - |
| e (Eccentricity) | 0 | Used | - |
| i (Inclination) | 0 | Used | - |
| om (Longitude Ascending Node) | 0 | Used | - |
| w (Argument of Perihelion) | 0 | Used | - |
| q (Perihelion Distance) | 0 | Used | - |
| ad (Aphelion Distance) | 6 | Dropped | Correlated with a |
| per\_y (Orbital Period in Years) | 1 | Dropped | Correlated with a |
| data\_arc (Data Arc Span) | 12290 | Used | - |
| Condition Code (Orbit Condition) | 703 | Used | - |
| n\_obs\_used (No of Observations) | 0 | Used | - |
| H (Absolute Magnitude) | 2100 | Used | - |
| NEO (Near Earth Object) | 6 | Used | - |
| PHA (Potentially Hazardous Asteroid) | 13089 | Used | - |
| Diameter | 561590 | Dropped | Due to high null values |
| Extent | 671754 | Dropped | Due to high null values |
| Albedo (Geometric Albedo) | 562541 | Dropped | Due to high null values |
| Rot\_Per (Rotation Period) | 656740 | Dropped | Due to high null values |
| GM (Standard Gravitational Parameter) | 671758 | Dropped | Due to high null values |
| BV (Color Index B-V) | 670951 | Dropped | Due to high null values |
| UB (Color Index U-B) | 670983 | Dropped | Due to high null values |
| IR (Color Index I-R) | 671770 | Dropped | Due to high null values |
| Spec\_B (Spectral Taxonomic Type SMASSII) | 670420 | Dropped | Due to high null values |
| Spec\_T (Spectral Taxonomic Type Tholen) | 670984 | Dropped | Due to high null values |
| G (Magnitude Slope) | 671668 | Dropped | Due to high null values |
| Class (Asteroid Orbit Class) | 0 | Used | - |
| n (Mean Motion) | 2 | Used | - |
| per (Orbital Period in Days) | 6 | Dropped | Correlated with a |
| ma (Mean Anomaly) | 6 | Used | - |
| MOID (Earth Minimum Orbit Intersection Distance) | 13089 | Label | - |

### Imputation

* In the data imputation phase, KNN (K-Nearest Neighbors) imputation was applied to fill missing values in specific features that had a low number of null values
* Using KNN imputation for selected features with low null values
  + data\_arc
  + condition\_code
  + H
  + pha

### Dropping Null Values

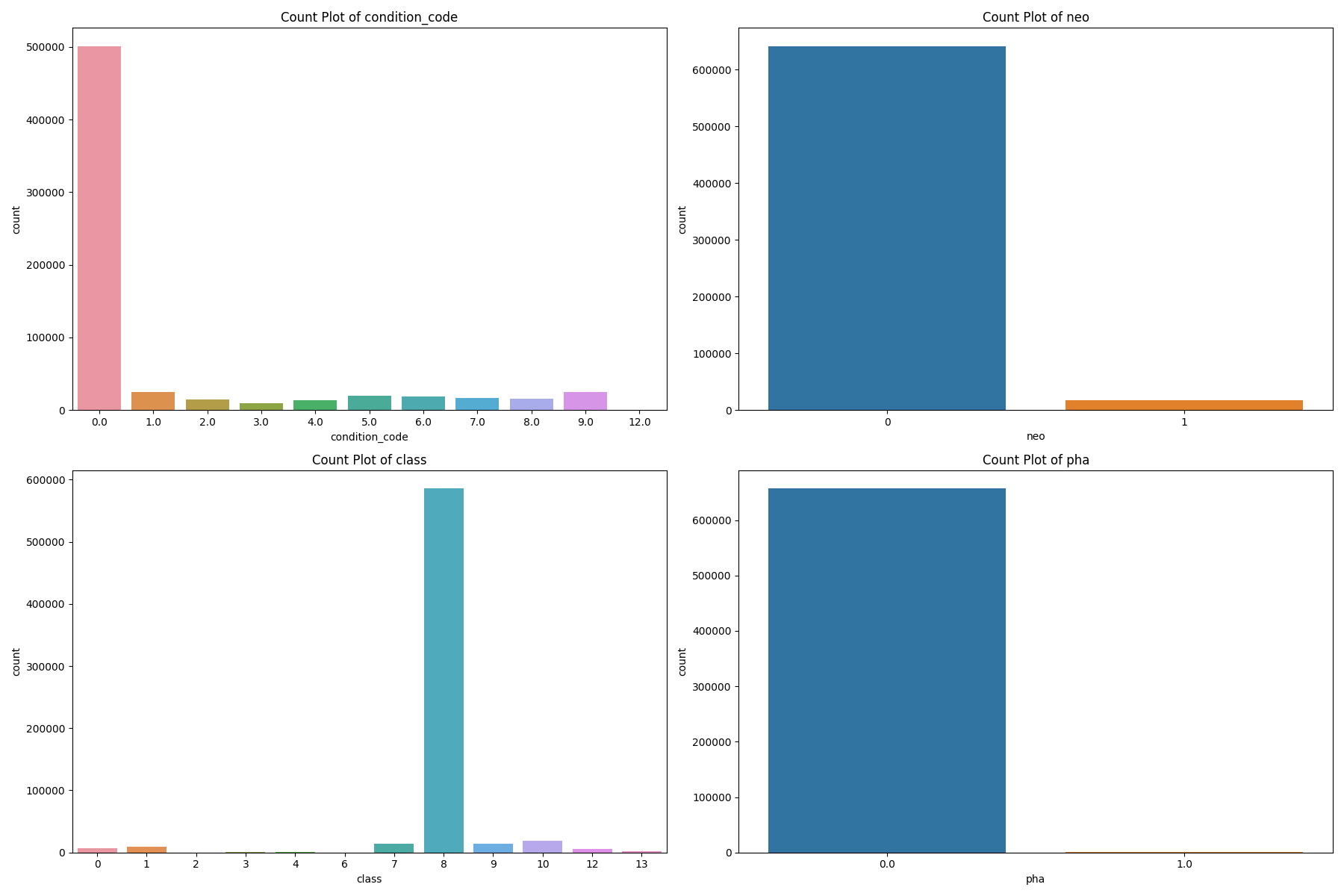
* Dropping columns with high null values: name, diameter, extent, albedo, rot\_per, GM, BV, UB, IR, spec\_B, spec\_T, G

In summary, the data imputation using KNN and the removal of columns with a high number of null values were crucial steps in preparing the dataset for analysis. These actions ensured that the dataset contained relevant and complete information for further exploration and modeling.

## 3. Data Preparation

### Feature Transformation

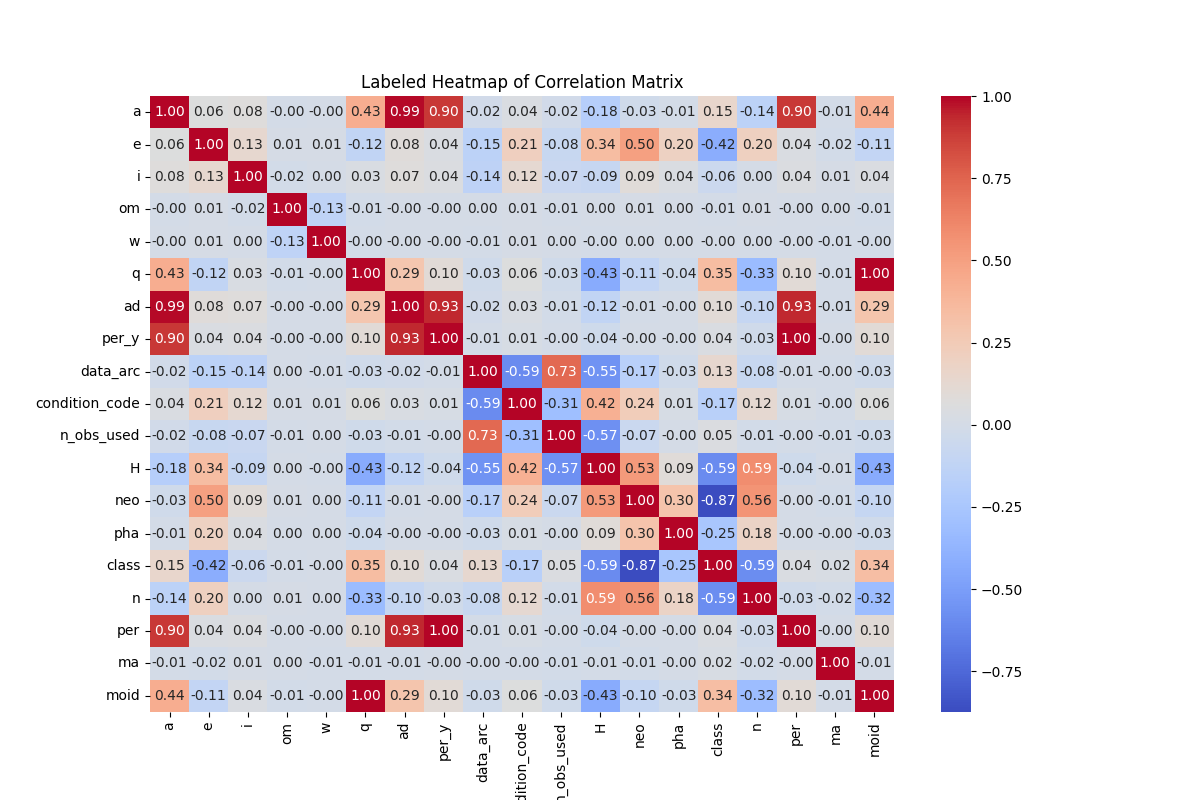
* Transformed categorical features to numerical counts with LabelEncoder
  + Class
  + Neo
  + Pha
  + Conditional\_code



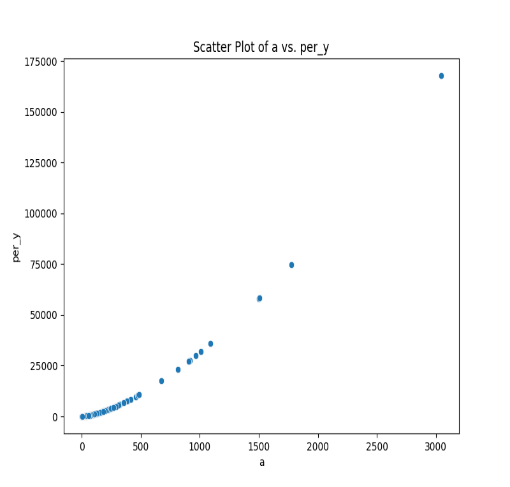
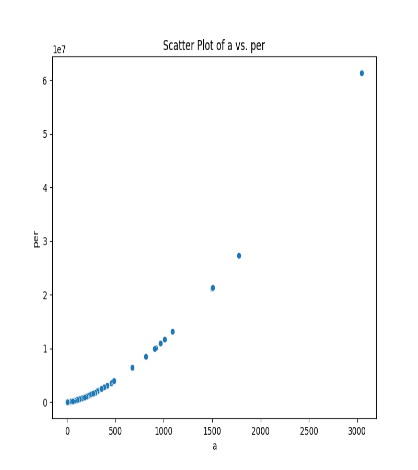
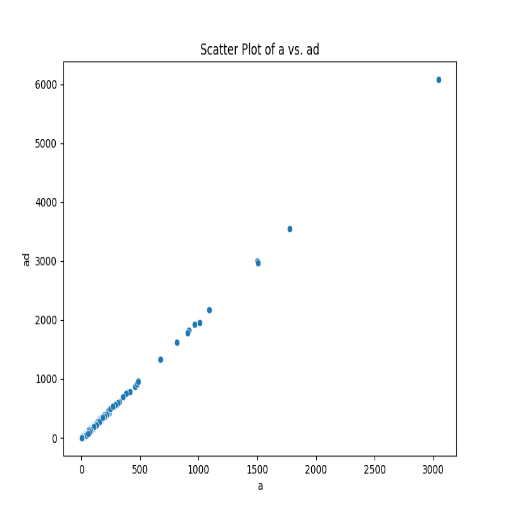
### Correlation Visualization

Visualized data with Heatmap and plots b/w

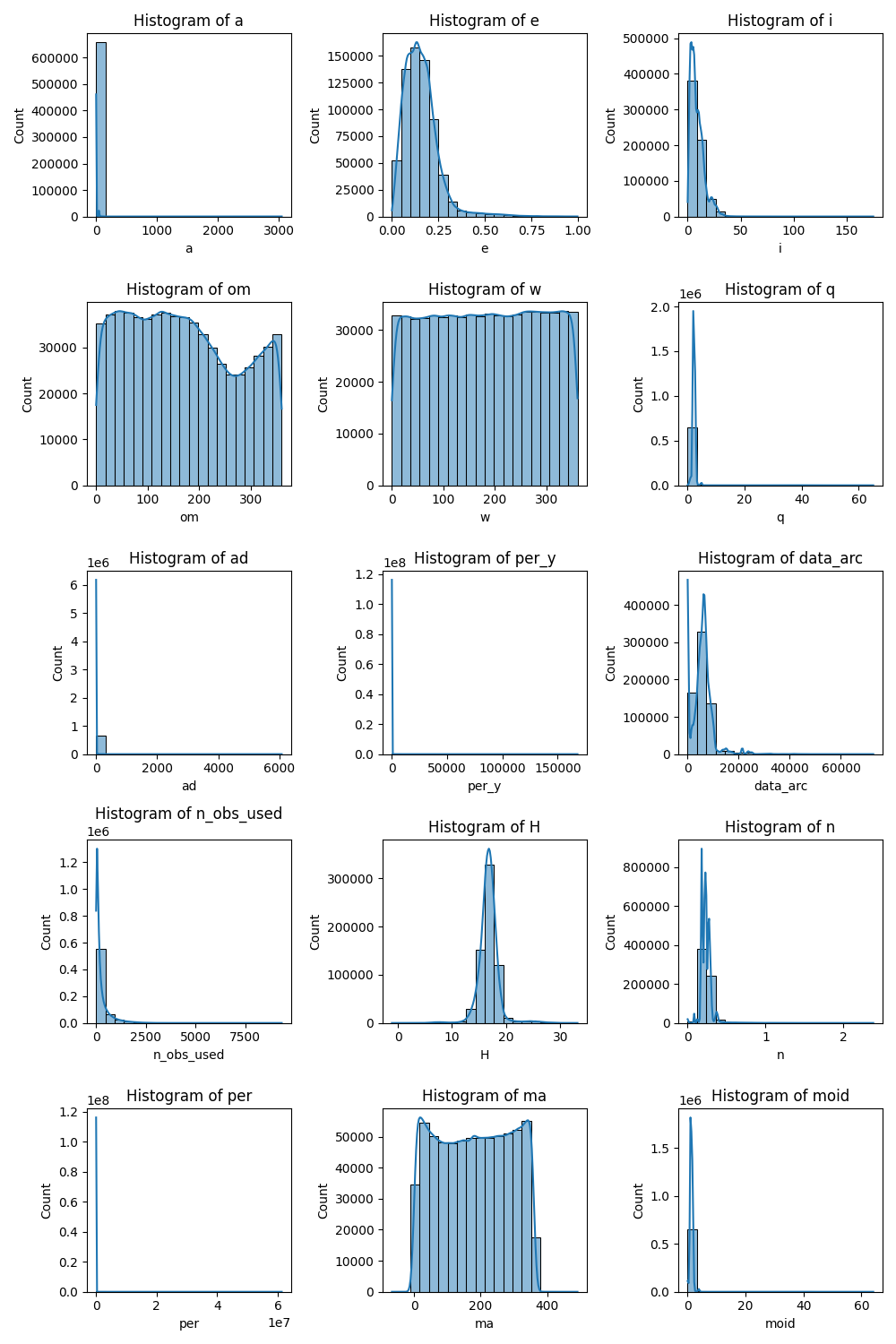
* a vs per
* a vs per\_y
* a vs ad



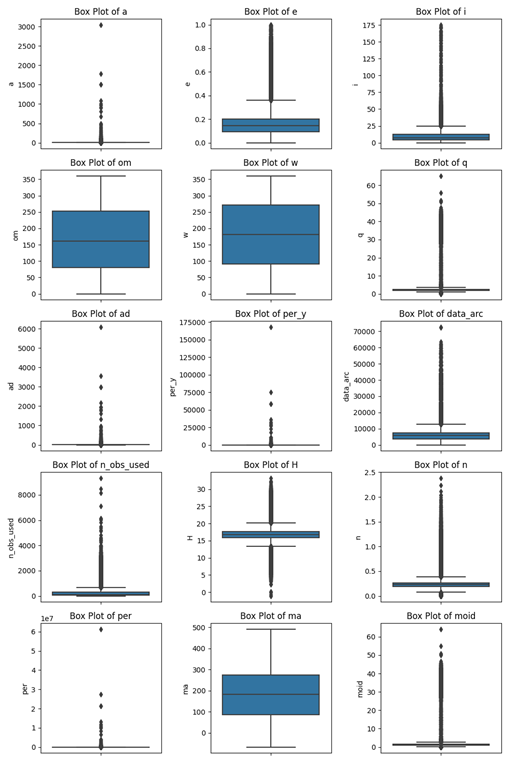
* + Dropping Correlated Columns
    - per
    - per\_y
    - ad



### Histogram Plotting



### Box plots



## 4. Modeling

### Train-Test Split of Training Data

* 20% test split
* Training data shape: (526937, 15)
* Test data shape: (131735, 15)

### Linear Regression Model

* Linear Regression, a simple yet powerful regression model, was applied to predict a continuous target variable using the training data features.
* Performance Metrics:
  + - Mean Squared Error: 0.0013448835921428656
    - Root Mean Squared Error: 0.036672654555443156
    - Mean Absolute Error: 0.020814241770678144
    - R-squared (Coefficient of Determination): 0.9997452537976232

### Gradient Boosting Regression

* Gradient Boosting Regression, an ensemble learning technique, was used to predict the target variable, leveraging multiple decision trees for complex relationship modeling.
* Performance Metrics:
  + - Mean Squared Error: 0.0006807442439919126
    - Root Mean Squared Error: 0.026091075945462897
    - Mean Absolute Error: 0.014509631039392745
    - R-squared: 0.999871054259298

### Hist Gradient Boosting Regression

* Hist Gradient Boosting Regression, optimized for large datasets and missing data handling, was employed to predict the target variable.
* Performance Metrics:
  + - Mean Squared Error: 0.05544783520303665
    - Root Mean Squared Error: 0.23547364014478703
    - Mean Absolute Error: 0.030612144613337926
    - R-squared: 0.9894971389862222

## 5. Model Saving via Pickle

Pickle is a convenient way to save and load machine learning models in Python, it's essential for reducing repeated training time, especially when dealing with untrusted large data sources.

## 6. Actual Test Data

Upon loading the test dataset, it became evident that, like the training data, it also contained null values. To ensure that the data was in a suitable state for analysis, the following preprocessing steps were taken:

### Test Data Preparation

* **Dropping Columns with Null Values**: Columns in the test dataset that contained a high number of null values were removed. This helped in cleaning the dataset and improved the quality of the data used for predictions.
* **Dropping Correlated Columns**: As in the training data, columns that were highly correlated with each other were dropped. This step was necessary to mitigate multicollinearity, which can lead to unstable predictions in certain models.

### Prediction

In the prediction phase, two machine learning models were utilized to make predictions on the test data. These models were selected based on their suitability for handling different scenarios within the test dataset:

* **Gradient Boosting Regressor**: This model was applied to the test data for rows that did not contain null values. It's a powerful model for regression tasks and is well-suited for making predictions on data with complete information.
* **Hist Gradient Boosting Regressor**: For rows with null values, the Hist Gradient Boosting Regressor was employed. This model has the capability to handle missing data effectively and produce reliable predictions even when there are gaps in the input data.

The use of these two models allowed for a comprehensive and robust approach to prediction, ensuring that both complete and incomplete information in the test data could be processed effectively, providing valuable insights and predictions for your analysis.