 UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Nouvelles Technologies de l’Information et de la Communication   
 Département d’Informatique et Technologie de l’information

Rapport Exploratory Data Analysis and Visualization Celestial Bodies

**MASTER ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

Students:

•BENSEDDIK ABIR

•BENZAHI WISSAL

2024-2025

**Table of Contents**

1. Introduction ............................................ ........... 3

2. Data Collection and Initial Inspection ...................3

3. Data Cleaning and Transformation ........................5

4. Feature and Sample Analysis ................................6

5. Visualization of Key Patterns .................................6

6. Summary Statistics and Key Insights ..................... 7

7.Model Building and Deployment ............................8

8. Conclusion .......................................................... 9

**1. Introduction**

The Celestial Bodies Dataset was scraped from Wikipedia using requests and BeautifulSoup. The goal was to clean the data and perform exploratory analysis to understand the dataset’s structure, ensuring that it is prepared for future modeling and analysis.

**Dataset Name :** Celestial\_Bodies\_Dataset

**Analysis Date:**

**Analyst:**

**Objective:**

This report presents an Exploratory Data Analysis (EDA) on the Celestial Bodies Dataset, which was scraped from Wikipedia using the requests and BeautifulSoup libraries. The objective is to explore and clean the dataset, identify key features, and derive valuable insights to guide future predictive modeling. The dataset includes textual data related to celestial bodies (planets, stars, moons, etc.), as well as images linked to each entity. The steps of the analysis include cleaning, exploring feature distributions, and identifying relationships between variables that will inform machine learning models and further scientific research.

**2. Data Collection and Initial Inspection**

**Source:**  
 Data was collected using web scraping techniques. Wikipedia pages related to celestial bodies were fetched using the requests library, and the textual content was extracted using BeautifulSoup. Additionally, image URLs for each celestial body were scraped. This method ensured that all collected data came directly from authoritative and structured Wikipedia pages.

**Relevance:**  
 The dataset is valuable for the analysis of celestial bodies because it encompasses various types, including planets, stars, moons, and other celestial entities. By analyzing this data, insights can be gained about celestial body types and their characteristics, which is essential for building classification models.

**Data Overview:**

The dataset consists of two related tables:

**Table 1 (df\_merged):**

Contains 1345 records, each representing a celestial body, with the following features:

* 1. **Title:** The name of the celestial body (e.g., "Apsis").
  2. **URL:** The Wikipedia URL for further details about the celestial body.
  3. **Cleaned\_Content:** The text extracted from the Wikipedia page, cleaned of irrelevant content.
  4. **Type:** The classification of the celestial body (e.g., planet, star, moon).
  5. **Word\_Count:** The number of words in the cleaned content.

**Table 2 (df\_merged\_image):**

Contains information related to the images associated with the celestial bodies, with the following features:

* 1. **Title:** The name of the celestial body (should match the title in Table 1).
  2. **Image\_URL:** The URL of the image related to the celestial body.
  3. **Image\_Saved:** The status of whether the image has been saved.
  4. **Image\_Format:** The format of the image (e.g., JPEG, PNG).
  5. **Image\_Size:** The size of the image.

This data has been collected to provide comprehensive information about celestial bodies, including both textual content and associated images, facilitating integrated analysis and understanding of celestial bodies.

**3. Data Cleaning and Transformation**

**Data Cleaning:**

**Filtering**

Relevant URLs Only relevant Wikipedia pages related to celestial bodies were retained. Irrelevant pages and external links were filtered out, ensuring that we only kept data pertinent to the analysis.

**Data Validation**

The data was validated by checking for missing URLs, invalid links, or incorrect textual descriptions. All entries with invalid URLs or missing data were either removed or filled with placeholder values.

**Handling Missing Values:**

**Text Data**: There were no missing values in the columns that were text-related, which include Title, Cleaned\_Content, and Type.

**Image Data:** URLs and saved paths of missing images were identified and removed in order to retain only complete records for analysis. Total missing images: 481.

**Outlier Detection and Removal:**

Outliers were assessed in terms of word count in Cleaned\_Content. Those with fewer words than the 25th percentile were considered too short to contain useful information and were thus removed.

Threshold: below 778 words.

Rows removed: 250.

**Save the Data** After cleaning, the final dataset was saved as a new CSV file, ready for exploration and modeling.

**Data Transformation:**

**Scaling:**  
 The Word\_Count feature was scaled using Min-Max normalization to ensure the feature values were within the range [0, 1]. This scaling ensures that all features are on the same scale and suitable for use in machine learning algorithms.

**Encoding**:  
 The Type column was encoded using Label Encoding, converting categorical values (e.g., "planet", "star") into numerical values (e.g., 6 for planet, 8 for star).

**Additional Transformations**:  
 The Word\_Count feature showed a skewed distribution, so a log transformation was applied to normalize the distribution and stabilize the variance.

**4. Feature and Sample Analysis**

• **Univariate and Bivariate Analysis:**

* **Central Tendency and Dispersion:** The Word\_Count feature showed significant variability, with a mean of 2,925 words and a standard deviation of 2,988. This wide spread indicates that some celestial bodies have very detailed descriptions, while others have shorter ones.
* **Class Imbalance:** The dataset showed class imbalance, with planets being the most common type (763 records), followed by stars (113 records). This imbalance may affect the performance of machine learning models and needs to be addressed during the modeling phase.

• **Correlation Analysis:** A correlation heatmap was created to understand relationships between variables. The correlation between Word\_Count and Type\_Encoded was very weak (0.07689), indicating that word count alone does not provide much insight into the celestial body type.

• **Dimensionality Reduction:** Since the dataset was not highly dimensional, no dimensionality reduction techniques like PCA (Principal Component Analysis) were applied.

**5. Visualization of Key Patterns**

**Visualization Techniques:**  
 Several visualizations were used to gain insights:

* + **Histograms** were created to show the distribution of Word\_Count across all records.

The histogram was right-skewed, with most of the celestial body descriptions having few words and a few quite long ones.

**Box plots** illustrated the distribution of word counts across different celestial body types.

**Pie Chart**: The proportion contributed by each type of celestial body; the planets make up the majority of this dataset.

* + **Bar Chart**: The frequency of each type; additional insights are provided in the infrequent categories of spacecraft and moons.
  + **Correlation Heatmap**:

A heatmap visualized the weak relationship between word count and type encoding.

**Observed Patterns:**

* The planet type had the highest frequency, while spacecraft and black hole types were less common.
* There was significant variability in word counts, suggesting that some celestial bodies had more detailed descriptions than others.

**6. Summary Statistics and Key Insights**

**Central Tendency and Variability:**

Word Count:

Mean: 2,925 words

Median: 1,841 words

Range: 5 to 24,367 words

**Interpretation of Statistics:**

Planets are the most dominant in both text and image data.

The descriptions of stars and galaxies were longer, whereas moons had shorter descriptions.

The images were not available consistently, and missing data in this regard erased 35% of the entries.

**Hypothesis Generation:**

Longer text descriptions relate to higher-resolution images.

Some celestial types, such as spacecraft, are less represented in both text and images.

**Model Building and Deployment**

**Data Preparation:**

Scaled the features (Word\_Count, Type\_Encoded) for training.

**Model Selection:**

Three models were selected for classification:

Random Forest: Gave the maximum accuracy of 99.45%.

Logistic Regression: Accuracy of 96.70%.

**Model Training and Evaluation:**

Trained the Random Forest model on the entire dataset and evaluated it.

Classification Report:

Precision: 99.7%

Recall: 99.4%

F1-Score: 99.5%

**Deployment:**

The trained model was then saved using joblib and deployed for prediction. A function was implemented to predict the celestial type and return the corresponding URL for user queries.

**. Conclusion**

The Celestial Bodies Dataset is a very informative dataset that allows for a deep analysis and classification of celestial objects.

**Key Findings:** Planets and stars are leading in this dataset, considering textual and visual representations. Random Forest was the best model that achieved 99.45% accuracy in classification.

**Further Steps**: Extend the dataset to other celestial types as well. Improve the quality and quantity of images for the underrepresented categories. Develop more advanced models, integrating image features for multimodal classification.

**Importance of Insights:**

Analysis shows how publicly available data is transformed into actionable insights. These findings will help in understanding the different celestial phenomena and improving the classification models for future research.