UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Nouvelles Technologies de l'Information et de la Communication

Département d'Informatique et Technologie de l'information

Exploratory Data Analysis (EDA) Report

Exploratory Data Analysis and Visualization of Celestial Bodies

- **Prepared by:**
  - "Bensddik Abir"
  - "Benzahi Wissal"

- **Analysis Date:**
  - "05 Nov 2024"

2024-2025

Table of Contents

1. Introduction:

The Celestial Bodies Dataset was scraped from Wikipedia using requests and BeautifulSoup for scraping data on a wide array of celestial objects: planets, stars, moons, and galaxies. This project cleaned the dataset and performed EDA on it to understand the structure of the data, find any possible insights, and prepare the data for modeling and prediction in the future.

- **Dataset Name:** Celestial Bodies Dataset
- **Source:** Wikipedia

Objective:

This analysis focuses on collecting and analyzing data from celestial bodies like planets, stars, moons, galaxies, and other relevant objects in astronomy. Using the publicly available dataset from Wikipedia, we want to find insights using higher-order data analysis methods that can reveal information about text and image data. Cleaning of data, exploratory analysis, and visualization will be performed, including the application of machine learning models for further investigation.

2. Objectives:

- **Data Collection from Wikipedia:** Data were collected from publicly available sources, such as Wikipedia, to gather texts and images about celestial bodies.
- **Data Structuring (Text and Images) for Analysis:** The collected data had to be organized into an analyzable format; both textual descriptions and images showing were used.
- **Data Loading and Cleaning:** Data were cleaned and prepared in a manner that irrelevant data are either ignored or represented as missing data.
- **Perform Exploratory Data Analysis (EDA):** In this step, we examined the data to find patterns and trends and the relationship among variables, which is imperative for any modeling.
- **Model Building and Deployment:** Based on the insights derived from the EDA, we have then built a machine learning model and then deployed it for making the predictions.

3. Data Collection and Initial Inspection:

Source:

The dataset was collected using a Wikipedia scraping method via the Requests and BeautifulSoup libraries, allowing us to collect both Wikipedia URLs and image links related to celestial bodies. In this way, all data collected would be directly from authoritative and structured Wikipedia pages.

Relevance:

The dataset is helpful in analyzing celestial bodies as it encompasses different types like planets, stars, moons, and other celestial entities. From this analysis, the insight about celestial body type and their characteristic features would be gained, which helps in building the classification models.

Data Overview:

The dataset consists of two related tables:

Table 1: Celestial Bodies Textual Data (df_merged)

This table contains the primary textual data about celestial bodies, including their descriptions and classifications.

Key Features:

a. **Title:** The name of the celestial body (e.g., "Moon").
b. **URL:** The Wikipedia URL for further details about the celestial body.
c. **Cleaned_Content:** The text extracted from the Wikipedia page, cleaned of irrelevant content.
d. **Type:** The classification of the celestial body (e.g., planet, star, moon).
e. **Word_Count:** The number of words in the cleaned content.
f. **Type_Encoded**: Numerical encoding of the Type column for machine learning purposes (e.g., Planet: 6, Star: 8).
g. **Is_Valid_URL**: Boolean indicating whether the URL is valid (True) or broken (False).

Dataset Overview:

- **Total Records:** 1345

Table 2: Celestial Bodies Image Data (df_merged_image)

This table contains image-related information associated with the celestial bodies, including details on availability, format, and size.

Key Features:

a. **Title:** The name of the celestial body .
b. **Image_URL:** The URL of the image related to the celestial body.

c. **Image_Saved:** The status of whether the image has been saved.

d. **Image_Format:** The format of the image (e.g., JPEG, PNG).

e. **Image_Size:** The size of the image.

Dataset Overview:

- **Total Records:** 6524

4. Data Cleaning and Transformation:

Data Cleaning:

1. Filtering Relevant URLs

Objective: To retain only entries related to celestial bodies by removing irrelevant data and external links.

Process:

- Checked URLs against keywords and filtered out irrelevant pages.
- Pages like unrelated topics or advertisements were excluded.

Outcome: The dataset focused on celestial bodies, enhancing the relevance of the analysis.

2. Data Validation

a) Text Data:

Verified columns such as Title, Cleaned_Content, and Type for completeness.

Result: No missing values were found in these text-related columns.

b) Image Data:

Image_URL and Image_Saved columns checked for missing values or invalid paths.

Result: 481 missing images were found, and their corresponding rows were removed to keep the dataset complete for image analysis.

3. Handling Missing Values

a) Text Data:

No missing values in the data; thus, nothing to be done.

b) Image Data:

Records with missing image URLs or saved paths were removed.

Action Taken: The remaining dataset had only complete records.

4. Outlier Detection and Removal

Objective: Outlier removal for Word_Count column; extremely short content might not make much sense.
Process: The word count for the 25th percentile of cleaned content was calculated:

Threshold = 778 words.

These were considered too short to have any significant information, so they were removed.

Outcome: 250 rows of data for too little content. Dataset resulted in high-quality, meaningful textual data.
Saving Cleaned Data

Saved the cleaned datasets into new CSV files, “merged_celestial_bodies.csv” and “merged_images_celestial_bodies.csv”, in a prepared-for-analysis manner.

Data Transformation:

1. Scaling

Objective: To scale the numerical features, Word_Count to bring their magnitudes consistent for machine learning algorithms compatibility.

Process: Min-Max normalization was applied to scale the Word_Count in the range of [0,1].

Result: Numerical data that is consistently scaled for machine learning.

2. Encoding

Objective: Convert categorical data (Type) into numerical values that will be understandable by the modeling.

Process: Used Label Encoding to map celestial body types to numerical values.

Example Mapping:

Planet → 6

Star → 8

Moon → 5

Outcome: Encoded Type into a numerical format as Type_Encoded.

3. Additional Transformations

Objective: Address skewness in the Word_Count feature to improve the analysis and modeling results.

Process:

Applied log transformation to normalize the distribution and stabilize variance.

Verified that the transformed data followed a more symmetric distribution.

Outcome: Reduced skewness, making the feature more suitable for modeling.

Key Statistics After Cleaning

5. Feature and Sample Analysis:

● Univariate and Bivariate Analysis:

1. Central Tendency and Dispersion

The following are key statistics for the variable **Word_Count,** which highlighted a high dispersion:

- **The mean:** 2,925 words
- **Standard deviation** of 2,988 words
- **Minimum** 5 - **Maximum** 24,367 words.

Interpretation

Some celestial bodies, especially planets and stars, contain highly detailed descriptions, but others are smaller or even less documented, hence shorter.

2. Class Imbalance

Observation: Significant imbalance between the different types of celestial bodies.

- **Planets :** 763 records (Most frequent).
- **Stars:** 113 records.

Other classes, such as moons or spacecraft, have considerably fewer entries.

Implication:

- This may lead to a bias in machine learning models toward the majority class.
- Oversampling, undersampling, or weighted models might be necessary at the time of modeling.

3. Correlation Analysis

a) Correlation Matrix:

A heatmap was created to show the relationship between numerical features.

Key Observation:

The correlation between Word_Count and Type_Encoded is 0.07689, which is very weak.

b) **Interpretation:**

- This weak correlation indicates that the type of celestial body is not well predicted by text length alone.
- Other features, like textual content or semantic analysis, may be more predictive.

6. Visualization of Key Patterns:

- **Visualization Techniques:**

a) Histograms:

The distribution of Word_Count for all the records.

Observation: The distribution was right-skewed. Most of the celestial body descriptions were short, while a few had very long descriptions.

b) **Box Plots:**

The distribution of Word_Count across the different types of celestial body.

Observation: Some types, like planets, had a higher range of word counts than other categories, such as spacecraft or moons.

c) Pie Chart:

The percentage of each type of celestial body.

Observation: The planets were dominant in the dataset, being the majority, while the less frequent ones included spacecraft and black holes.

d) Bar Chart:

The frequency of each type of celestial body.

- The observations were planets and stars most often, while the other categories were moons and space craft significantly lower; these observations reinforce the concept of class imbalance.

e) Heatmap Correlation

The relationships among the numeric variables; among others, Word_Count versus Type_Encoded

Observation: The relationship between the Word_Count and Type_Encoded had a very poor correlation - meaning that size is a bad predictor for the text classification.

- **<u>Observed Patterns:</u>**

a) Frequency of Types:

Most of the dataset was occupied by planet types, of course. The other classes include Space Craft and Black Hole Types and are fairly less common compared to each other, adding up to cause the observed imbalance in this class.

b) Dispersion in the Word Count :

The distribution of Word Count was vast, indicating huge discrepancies in description details of most celestial body types. There's a notable difference; planet types included more verbose explanations, where spacecraft have notably concise summaries.

7. Summary Statistics and Key Insights:

a) Central Tendency and Variability

Word Count:

Mean: 2,925 words

Median: 1,841 words

Range: 5 to 24,367 words

b) 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.

c) Hypothesis Generation

Text and Image Relationship:

Celestial bodies with longer textual descriptions may correlate with higher-resolution images.

Representation Gap:

Certain celestial types, such as spacecraft, are underrepresented in both text and images, which may limit the generalizability of insights for these categories.

8. Model Building and Deployment:

a) **Data Preparation:** Features (Word_Count, Type_Encoded) were scaled for consistent input.

b) Model Selection:

- Random Forest had the best accuracy of 99.45%.
- Logistic Regression gave a good baseline of 96.70%.

c) Model Training and Evaluation:

Trained Random Forest model with excellent performance:

Precision: 99.7%

Recall: 99.4%

F1-Score: 99.5%

d) Deployment:

The model was saved using joblib.

A prediction function was implemented to classify celestial bodies and retrieve their corresponding Wikipedia URLs.

9. Conclusion:

The dataset of the celestial bodies propounds the significance of publicly available data in carrying out meaningful analysis and classification of astronomical objects.

Findings: The most represented classes were planets and stars, and Random Forest was able to achieve a classification accuracy of 99.45%.

Future Steps: Expand the celestial type coverage, improve characteristics in image quality and representation as well as process advanced models that consider both text and image data.

Insight Significance: the idea that to many, well-structured data transforms into actionable knowledge, thus, improving the understanding of celestial phenomena while developing methodologies for future research in classification improvement.