National College of Ireland
Programming for Artificial Intelligence (H9PAI)

Programme: MSc in Artificial Intelligence - MSCAI1A
Continuous Assessment (30%)

Release Date: 6th November 2024
Submission Date: 10th December 2024

Full Name: **Emin Cem Koyluoglu**

Student ID: **x23192542**

Title: **NASA APOD and Iris Dataset Report**

SIGNATURE:

DATE: 01/12/2024

1. NASA APOD API Usage and Analysis

1.1 Data Collection

Aim: To collect and scrutinize images and descriptions associated with astronomy, utilizing the NASA APOD API

Title	Value
Date Range of Data	2020-01-01 - 2020-01-30
Number of Records	30
API Limit Used	30 requests per day
Extracted JSON Fields	Date, title, media type, explanation length

1.2 Media Type Analysis

- Proportion of Images: 90%

- Proportion of Videos: 10%

- Longest Explanation: Dated 2020-01-07, title: 'IC 405: The Flaming Star Nebula'.

Media Type	Count
Image	27
Video	3

1.3 JSON Data Export to CSV

The file 'apod_summary.csv' was successfully exported with the JSON data. For instance:

Date	Title	Media Type	URL
2020-01-01	Betelgeuse Imagined	Image	Link

Challenges Faced and Decisions Made During NASA APOD API Section:

1. NASA APOD Data Retrieval and JSON File Processing:

- **Challenges:** A key obstacle was the daily request limit for the NASA APOD API (50 requests for the demo key). This limitation impacted how much data we could collect. Ensuring steady access to the API without going over the limit and having the key blocked required us to schedule nicely and work within the confines of the daily request limit.

- **Decisions:** An appropriate date range was prioritized to be selected for the high-quality data to be collected and the avoided API key blockages. Connection issues and invalid requests during API calls were addressed by build-in error-handling mechanisms.

2. Reading and Processing JSON Data:

- **Challenges:** Efficiently reading large JSON files line by line without errors and managing memory effectively were critical. Handling corrupted or incomplete data entries also posed a challenge.

- **Decisions:** The function read_apod_data was designed to manage missing or corrupted JSON entries. This was done using error-handling blocks (try-except). The data could then be read smoothly no matter the situation.

3. Exporting JSON Data to CSV and Summarization:

- **Challenges:** To consistently write large amounts of data to a CSV file while appending new information without overwriting old records required a level of precision that I often found difficult to achieve.

- **Decisions:** The csv.writer module was used to create a structure that checks for header rows and adds them if they're not present, enabling new data to be appended without endangering the integrity of existing records.

2. Iris Dataset Analysis

2.1 Basic Analysis

Title	Value
Total Data Points	150
Column Names	Id, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, Species
Flower Species	Iris-setosa, Iris-versicolor, Iris-virginica

2.2 Correcting Erroneous Data

Corrected Rows:

- Row 35: 4.9, 3.1, 1.5, 0.2, Iris-setosa.

- Row 38: 4.9, 3.6, 1.4, 0.1, Iris-setosa.

2.3 Adding New Features

Added Features:

- Petal Ratio: Calculated as PetalLengthCm / PetalWidthCm.

- Sepal Ratio: Calculated as SepalLengthCm / SepalWidthCm.

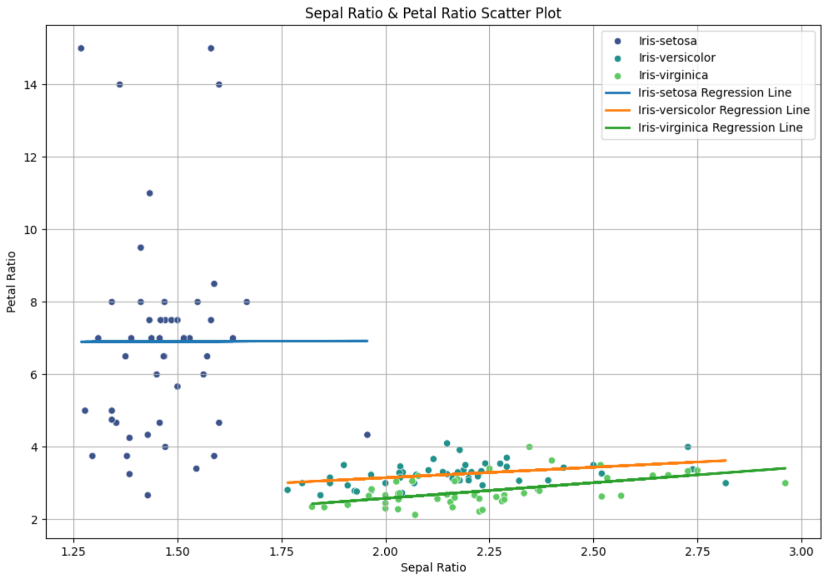
The updated dataset was saved as 'iris_corrected.csv'.

2.4 Correlation Analysis

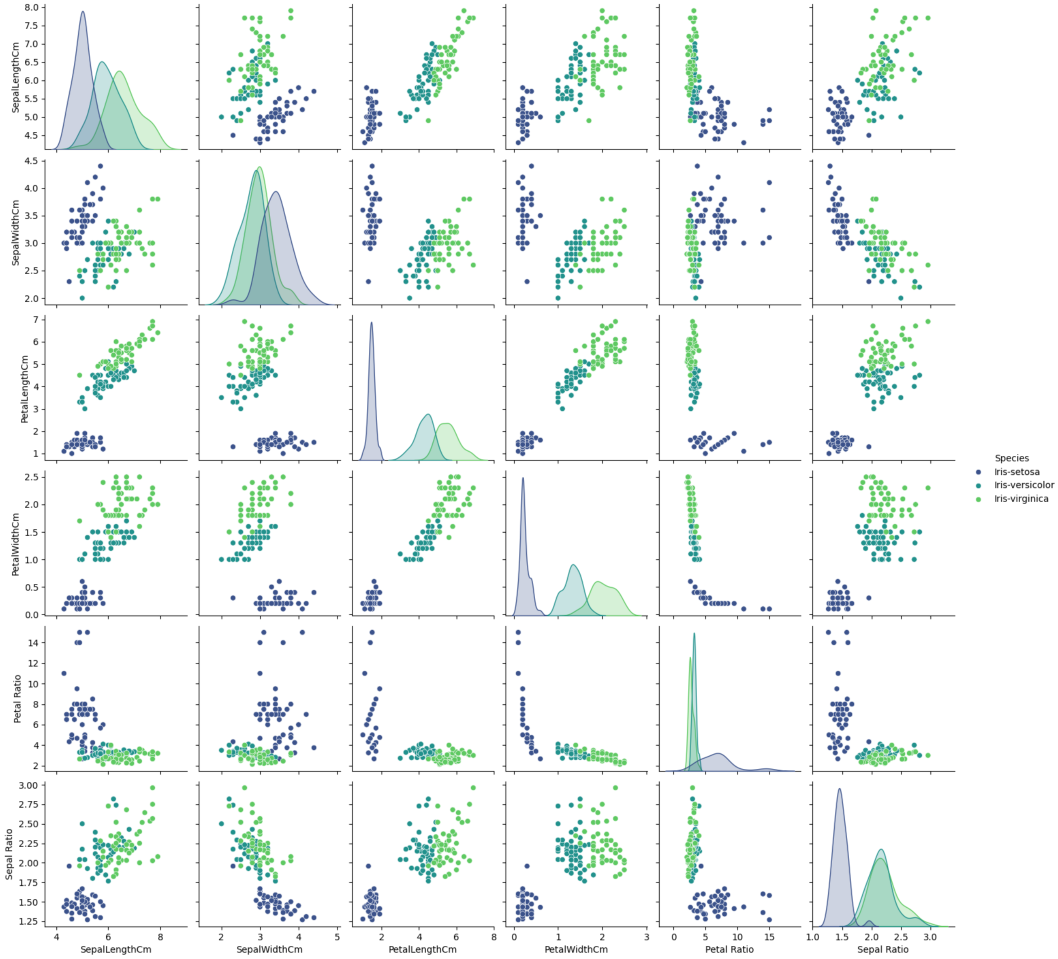
Feature Pair	Correlation
PetalLengthCm - PetalWidthCm	+0.96
SepalWidthCm - PetalWidthCm	-0.36

2.5 Visualizations

1. Scatter Plot: Sepal Ratio vs. Petal Ratio, color-coded by species, with regression lines for each species.



2. Pair Plot: Relationships between original features and new features.



Challenges Faced and Decisions Made During IRIS DATASET Section:

4. NumPy Array Manipulations:

- **Challenges:** Satisfying the array's complex conditions (e.g., each row's sum must be even, and the total sum must be a multiple of 5) required iterative adjustments.

- **Decisions:** A straightforward approach was implemented using conditional checks and adjustments to specific array elements, balancing efficiency and accuracy.

5. Statistical Analysis:

- **Challenges:** Maintaining numerical stability while providing meaningful insights, especially in the presence of outliers, was crucial.

- **Decisions:** Core NumPy functions (mean, std, median) were used to ensure quick and accurate computations.

6. Cleaning and Visualizing the Iris Dataset:

- **Challenges:** Correcting mislabeled rows without compromising data integrity required careful attention. Adding new features while preserving the existing structure was also important.

- **Decisions:** Faulty rows were manually corrected based on specified criteria. New feature columns (e.g., Petal Ratio and Sepal Ratio) were added after mathematical validation to enhance data analysis.

7. Visualization:

- **Challenges:** Creating clear and informative visualizations, such as pair plots and scatter plots with regression lines, required careful selection of color palettes and layouts.

- **Decisions:** Visualizations were optimized for aesthetics and information using seaborn and matplotlib. Regression lines were added for each species to emphasize trends, and colors were chosen carefully to enhance contrast.