Tarea2: Iris Dataset

1st Samed Rouven Vossberg
Karlsruhe Institue of Technology
Karlsruhe, Germany
Charité – Universitätsmedizin Berlin
Movement Disorder and Neuromodulation Unit, Department of Neurology
Berlin, Germany
Universidad Nacional Autónoma de México
Mexico City, Mexico
samedvossberg@gmail.com

***Abstract*—This paper presents a technical exercise involving the Iris dataset using Python. The study covers data loading, descriptive statistical analysis, and various visualization techniques using pandas, matplotlib, and seaborn. The exercise demonstrates the effective use of Python libraries for basic data analysis. Results are discussed and potential improvements are suggested. The corresponding Jupyter Notebook can be found at https://github.com/SamedVossberg/UNAM_CV/blob/main/iris.ipynb.**

***Index Terms*—Computer Vision, Iris Dataset, UNAM**

I. Introduction

The Iris dataset [1] is a well-known benchmark in data analysis and pattern recognition. It contains 150 samples from three different species of Iris shown in figure 1. Each sample has four features: sepal length, sepal width, petal length, and petal width. Due to its balanced class distribution and clear feature separability, the dataset is commonly used to demonstrate classification algorithms and exploratory data analysis.

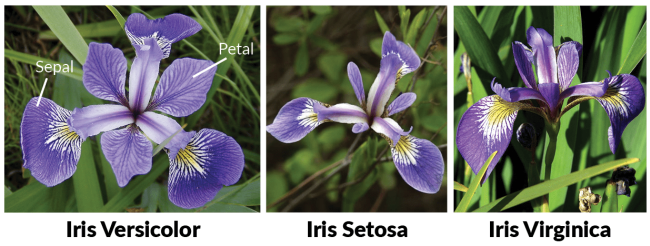

Fig. 1. Pie chart of Iris species frequency. [2]

II. Methodology

The analysis was conducted using a Jupyter Notebook [3]. Initially, the Iris dataset was imported into a pandas DataFrame. The basic properties of the dataset were examined by checking the number of rows, columns, data types, and the first ten rows.

Descriptive statistics, including the mean, standard deviation, minimum, and maximum values for each numerical attribute, were computed. In addition, a 5×5 identity matrix was generated using NumPy and converted to a SciPy compressed sparse row (CSR) format to demonstrate the handling of sparse data.

III. Results and Discussion

The dataset comprises 150 samples with five columns (after including the header). No missing data was detected. The descriptive statistics confirmed the expected distribution of sepal and petal measurements.

Species Distribution

Figure 2 shows the frequency distribution of the three Iris species. The pie chart confirms a balanced distribution among the species.

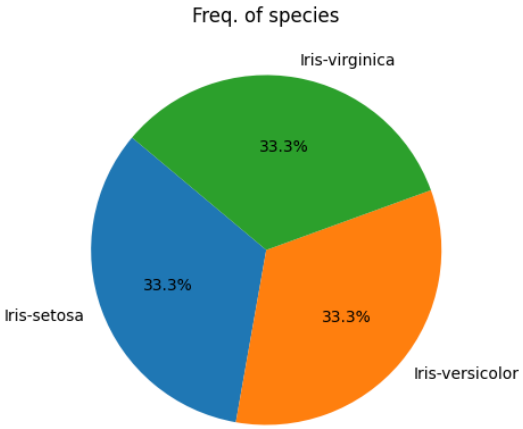

Fig. 2. Pie chart of Iris species frequency.

Relationship Between Features

The scatter plot in Figure 3 demonstrates the relationship between sepal length and sepal width. A positive correlation is observed between these two features, suggesting that as sepal length increases, sepal width tends to increase as well.

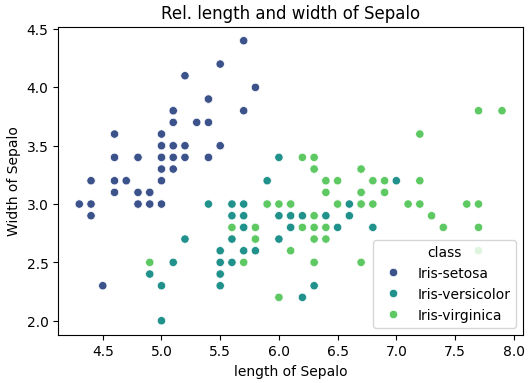

Fig. 3. Scatter plot showing sepal length versus sepal width.

Distribution of Numerical Attributes

Histograms were generated to visualize the distribution of the numerical attributes. Figure 4 illustrates these distributions. Notably, while sepal measurements exhibit a relatively uniform spread, petal measurements show distinct clustering. This differentiation is often crucial in classification tasks.

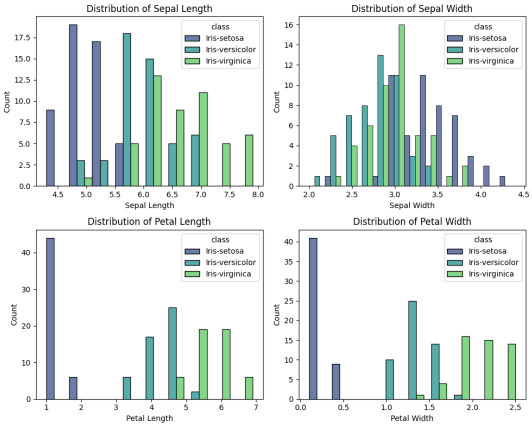

Fig. 4. Histograms for the numerical attributes.

Pairwise Relationships Among Features

A pairplot was created to display the relationships between each pair of features. As shown in Figure 5, the pairplot confirms that petal measurements are more discriminative than sepal measurements, as they form more distinct clusters.

Detailed Correlation Analysis

Figure 6 provides a jointplot with hexagonal binning for sepal length and sepal width. This visualization offers a detailed view of data density and dispersion, emphasizing regions with a high concentration of samples.

Additional Insights on the Iris Dataset

The Iris dataset has long been a cornerstone in data analysis due to its simplicity and rich informational content. Notably, the petal dimensions provide better separation among the

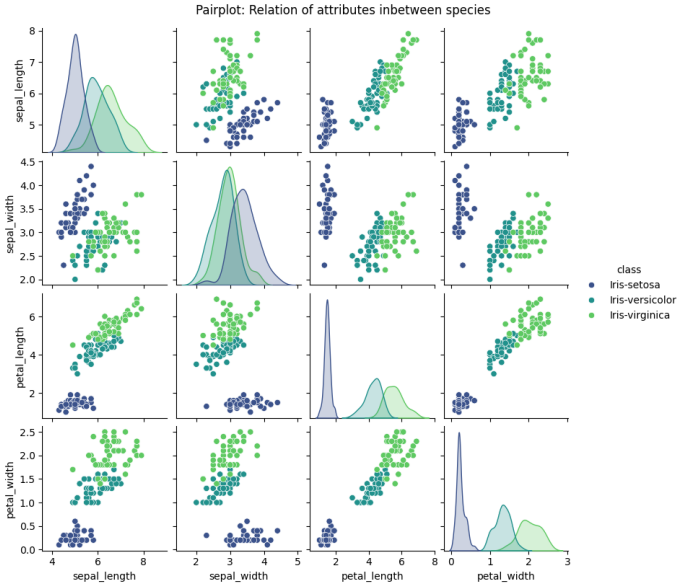

Fig. 5. Pairplot displaying pairwise relationships among features.

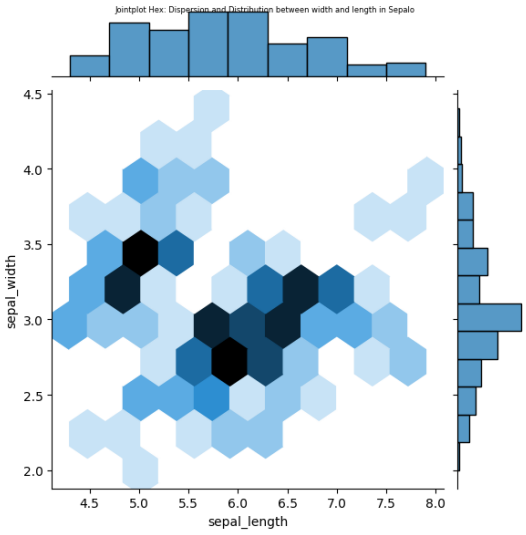

Fig. 6. Jointplot with hexagonal binning of sepal dimensions.

species compared to the sepal dimensions. This distinction has led to its widespread use in benchmarking classification algorithms. Furthermore, the clear clustering observed in the petal features underlines the importance of feature selection in machine learning tasks. These insights demonstrate how visualizations can uncover underlying patterns, guiding the choice of appropriate analytical methods.

IV. Conclusion

This report detailed the process of loading, preprocessing, and analyzing the Iris dataset. The methodology included both statistical analysis and a variety of visualizations, which were integrated within the text to enhance clarity. Python libraries such as pandas, NumPy, matplotlib, seaborn, and SciPy proved

effective for these tasks. The study not only reaffirms the utility of the Iris dataset as a teaching tool but also highlights its relevance in demonstrating key concepts in data analysis and pattern recognition. Future work may involve applying advanced statistical or machine learning techniques to further explore the dataset.

REFERENCES

[1] R. A. Fisher, "Iris," 1936. [Online]. Available: https://archive.ics.uci.edu/dataset/53

[2] "Data Science Example - Iris dataset." [Online]. Available: http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html

[3] "UNAM_cv/iris.ipynb at main · SamedVossberg/UNAM_cv." [Online]. Available: https://github.com/SamedVossberg/UNAM_CV/blob/main/iris.ipynb