Data Analysis Report – Iris Dataset

Executive Summary

The Iris dataset was analyzed to explore the characteristics of three iris species (Setosa, Versicolor, Virginica) based on sepal and petal measurements. Our analysis shows clear differences in petal length and width across species, making them strong predictors for classification. Outlier handling improved dataset quality, and visualizations revealed patterns useful for species differentiation.

1. Introduction

The Iris dataset is one of the most popular datasets in data science and machine learning. It contains measurements of sepal length, sepal width, petal length, and petal width for three species of iris flowers.

Objective: To clean, visualize, and analyze the dataset to better understand species-level differences.

2. Data Description

- Dataset: Iris dataset (150 rows × 5 columns)
- Features:
  - SepalLengthCm
  - SepalWidthCm
  - PetalLengthCm
  - PetalWidthCm
  - Species (Setosa, Versicolor, Virginica)
- Data Quality Issues:
  - Outliers detected in SepalWidthCm using the IQR method
  - No missing values found

3. Methodology

- Tools Used: Python, Pandas, NumPy, Seaborn, Matplotlib
- Steps:
  1. Data cleaning – checked duplicates, missing values, and outliers
  2. Outlier treatment using the IQR method
  3. Exploratory Data Analysis (EDA) with histograms, boxplots, and count plots
  4. Statistical summary to compare feature distributions across species

4. Analysis & Findings

a. Distribution of Features

- Sepal Length: Normally distributed, overlaps across species
- Sepal Width: Narrower spread, some outliers detected
- Petal Length & Width: Strongly differentiate species
  - Setosa has the smallest petals
  - Virginica has the largest petals
  - Versicolor lies in between

b. Species Counts

Dataset is balanced: 50 samples per species

c. Outliers

Outliers in SepalWidthCm identified and removed to improve dataset quality

d. Relationships

- Positive correlation between petal length and petal width
- Petal features are stronger indicators for classification compared to sepal features

5. Conclusions

- The Iris dataset is clean, balanced, and suitable for classification tasks
- Petal measurements are highly discriminative among species
- Outlier handling improved statistical stability

6. Recommendations

- Use petal length and petal width as primary features in classification models
- Apply machine learning models (e.g., Logistic Regression, SVM, Random Forest) to predict species
- Use visualization dashboards for clearer biological interpretation
- Further validate findings with advanced statistical methods

7. Appendix (Optional)

- Histograms for Sepal/Petal measurements
- Boxplots showing outlier detection
- Correlation heatmap