Project Title: Iris Dataset Analysis Using Python

Name: Raniya Shareef
Date: 27-05-2025
Course: B.Tech in Artificial Intelligence & Data Science

Table of Contents

1. Introduction

This project focuses on analyzing the Iris dataset, which contains data on three species of iris flowers. Each record includes measurements of sepal length, sepal width, petal length, and petal width. The goal is to understand how these features differ across species and to explore patterns using Python.

2. Objective

The main aim is to perform exploratory data analysis and simple statistical tests to identify differences between the three iris species. The project also includes generating graphs to visualize the relationships between features.

3. Dataset Overview

- The dataset has 150 rows and 5 columns.
- Features:
    - sepal.length (in cm)
    - sepal.width (in cm)

    - petal.length (in cm)
    - petal.width (in cm)
- Target column:
    - variety (species: Setosa, Versicolor, Virginica)

4. Libraries and Tools Used

The analysis was done in Python using the following libraries:

- pandas – to handle and manipulate data
- numpy – for numerical calculations
- seaborn and matplotlib – for plotting graphs
- scipy – for performing ANOVA test

5. Data Analysis Steps

Loading and Exploring Data

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats

df = pd.read_csv('iris.csv')
print(df.info())
print(df.head())
```

Descriptive Stats and Grouping

```
print(df.describe())
print(df.groupby('variety').describe())
print(df.isnull().sum())
```

6. Data Visualizations

We used several plots to understand the dataset better:

- **Histograms**: Showed how each feature is distributed for each species.
- **Scatter Plot**: Compared sepal length with petal length to show species separation.
- **Box Plot**: Compared petal width for each species.
- **Heatmap**: Showed correlation between all numerical features.

(Plots can be added as images here.)

7. ANOVA Test

To check if sepal length differs significantly across the three species, we used the ANOVA test:

```
setosa = df[df['variety'] == 'Setosa']['sepal.length']

versicolor = df[df['variety'] == 'Versicolor']['sepal.length']

virginica = df[df['variety'] == 'Virginica']['sepal.length']

f_stat, p_value = stats.f_oneway(setosa, versicolor, virginica)

print(f"F-statistic: {f_stat:.2f}, p-value: {p_value:.4f}")
```

If the p-value is below 0.05, it means the average sepal length is significantly different between at least two of the species.

8. Key Insights

- Setosa flowers generally have smaller petals compared to the other two species.
- Virginica species usually have larger measurements overall.
- There is a strong correlation between petal length and petal width.
- The ANOVA test confirms significant differences in sepal length across species.

9. Conclusion

This project helped in understanding how simple Python tools can be used to explore and analyze a classic dataset. From basic stats to visualizations and hypothesis testing, we saw how different species of iris flowers can be clearly differentiated using their measurements.

End of Report