Week 1.1
Setting Up and Exploring Python Tools for Data Analysis and Machine Learning

Produced By: Abdullah Abdul Majid*(linkedin)*

Completion Date: July 1, 2024

Email: abdullahpisk@gmail.com

Jupyter Notebook Links:

GitHub

Google Drive

Table of Contents:

Problem Statement:

Exploring Machine Learning with Python

Purpose:

The purpose of this assignment to set up a Python environment within Jupyter Notebook and necessary libraries to get started with Machine Learning and Data Analysis. The tasks performed were reading a .csv dataset file, exploring its features and then plotting them.

Process:

Installations and Set Up

- Installed and set up Anaconda for running Jupyter Notebook.
- Downloaded the Iris.csv dataset from Kaggle.
- Installed the pandas, matplotlib and seaborn libraries from pip using *%pip install <package_name>*

Reading and Exploring the Dataset

- Uploaded the Iris.csv file to Jupyter Notebook and read it using pandas' *read_csv* function.
- Used pandas' *head(<no. of rows>)* function to view the first few rows of the dataset.
- Used pandas' *describe()* function to view the datasets statistical properties like minimum and maximum values, mean, percentiles etc.

- Used pandas' *info()* function to view the datasets technical specifications like memory used, datatypes etc.
- Used pandas' *value_counts()* function to view how often the data reoccurs.

Plots

- Used matplotlib's *hist()* function to plot histograms for every iris statistic.
- Used matplotlib's *scatter()* function to plot scatter plot between all possible statistical relationships.
- Used seaborn's *pairplot()* function to plot a pairpplot for the iris.csv dataset.

Findings:

Dataset Exploration

The dataset contains statistics about Petal and Sepal dimensions on 3 different species of Iris flowers, 50 samples for each species. All the outputs are included in the project Jupyter Notebook. The dataset has 6 columns for statistics and 150 rows containing the samples. The datatypes used are float64(4, dimensions), int65(1, sample id) and object(1, species). Every entry in the dataset is unique, i.e there is no repetition. The dataset description is as follows:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Plots

The histograms of every statistic show the variation in dimensions of the species. We conclude that most flowers have a sepal length of around 5cm, sepal width of around 3cm, petal lengths in the 1-2cm range and petal widths of around 0.25cm. Moreover, none of the samples have a petal length in the roughly 2-3cm range and a petal width in the roughly 0.75-1cm range.

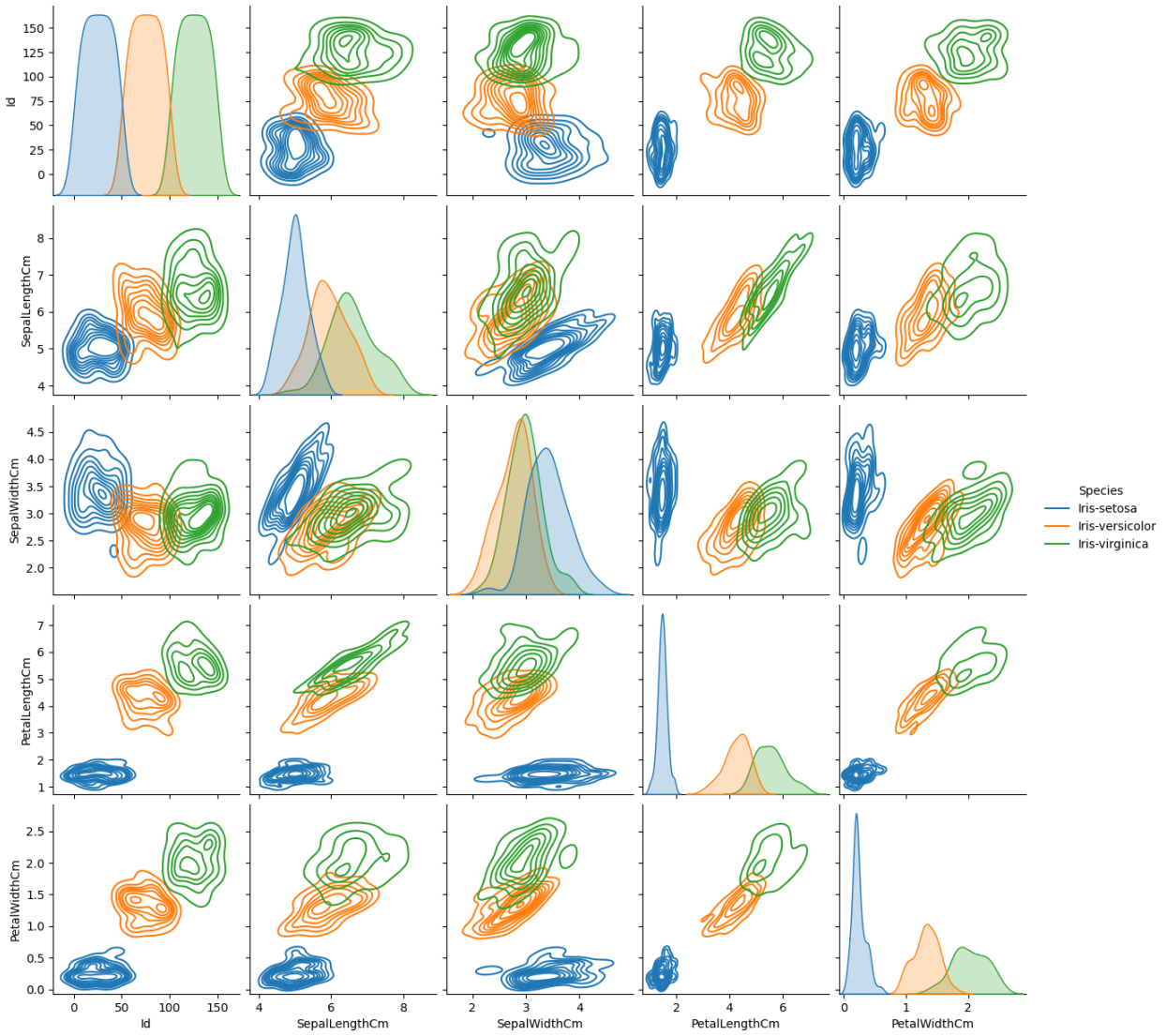
The scatter plots of dimensions show that there is a relationship between petal length and width, both are directly correlated. There seems to be no relation between sepal dimensions. Sepal lengths and petal widths are also somewhat related, and there might be a relation between sepal widths and petal lengths.

Species wise scatter plots show that:

- Iris-Virginica tends to have the longest sepals, while Iris-Setosa tends to have the shortest sepals.
- Iris-Setosa tends to have the widest sepals, while Iris-Versicolor tends to have the thinnest sepals.
- Iris-Virginica tends to have the longest and widest petals, while Iris-Setosa tends to have the shortest and thinnest petals.

The seaborn pairplots and the information from the above scatter plots draw the conclusion that Iris-Virginca tends to be the largest species in terms of petal size, Iris-Versicolor is the mid species and Iris-Setosa tends to have the smallest petals. In addition, Iris-Setosa sepals are the shortest but the widest, Iris-Versiclor sepals are, moderate in length but the thinnest in terms of width and Iris-Virginica sepals are the longest but moderate in width.

All the plots are included in the Jupyter Notebook.



1. Seaborn Pairplot

Conclusion:

I managed to set up a Python environment for Machine Learning and Data Analysis by Installing all the required tools and software. Moreover, I was able to explore the Iris dataset and perform the required plots. Finally, I used the dataset plots and specifications to draw the necessary conclusions and relations about Iris flowers.