Iris Dataset

1. Dataset Overview

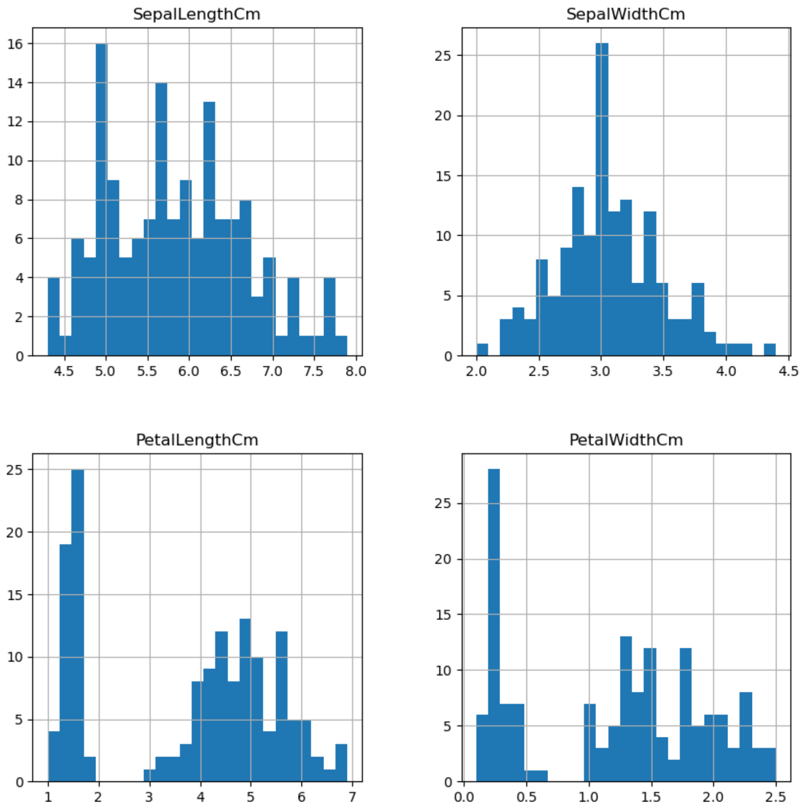
The following dataset that is being used for data analysis and visualization is very well known as the **Iris Dataset**. The Iris refers to the floral morphology of the flower that consists of petals, sepals and their colors. The dataset contains measurements of all the parts that form the iris of the flower. The dataset is collected from University of California Irvine repository[1]. The UCI repository currently maintains around 622 datasets, Iris dataset being one of them. The dataset was initially collected in the year 1988 and is known to be one of the most prominent datasets of all time. Thereby, making this dataset to be used since the advent of **Pattern Classification and Recognition**[2]. After 1988, this dataset was widely being used in the field of **Usage of features measurement in pattern recognition and classification.** The dataset is available in CSV format to be downloaded and used for classification tasks. The dataset consists of 6 columns and 150 rows.
The 6 columns are as follows:
-Id
-Sepal length in cm
-Sepal width in cm
-Petal length in cm
-Petal width in cm
-Iris classes
There are three output classes in the dataset which are:

1. Iris Setosa
2. Iris Versicolor
3. Iris Virginica

There are 6 features present in the dataset thereby making it a multivariate dataset. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other[3].
The best possible accuracy achieved on this dataset in literature as of current state of art is 100% by using decision trees and logistic regression[4].

2. Dataset Analysis and Visualization using pandas and matplotlib

For the current project the dataset is analyzed with the pandas library of python. Python provides an easy and efficient way to analyze data via pandas. The read csv command has been used to study the dataset in the form of a data frame. A data frame is generally a tabular representation of the data with every entry in the data frame is indexed and each column is labeled[5]. Similar commands have been used to analyze data statistically. After observing the dataset it is observed that there are 150 rows in total with 6 columns, wherein the sixth column represents the output class. Out of all the 150 entries there are 50 entries belonging to each output class with no missing values. After data analysis the data is visualized using matplotlib library. The visualization of the data helps in seeing how the data looks like. With the help of matplotlib library we can plot each column with numerical entries to see how the data is varying from one entry to another.



3. References

[1]https://archive.ics.uci.edu/ml/datasets/iris

[2]https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[3]https://archive.ics.uci.edu/ml/datasets/iris

[4]https://www.kaggle.com/code/prakharrathi25/100-accuracy-classification-in-iris-dataset

[5] http://rexa.info/paper/5b3417aa2824988405f9ac934b692af30729b447