**The Iris Data Set**

**Introduction:**

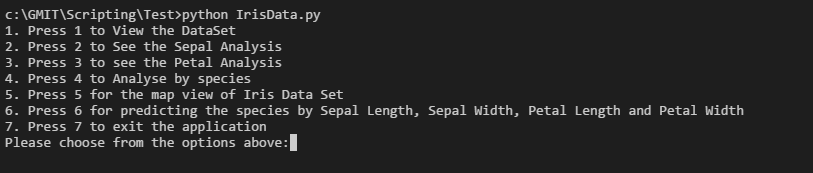
The Iris Data Set consist of 50 samples from each of the three species of Iris flower *(Iris setosa, Iris virginica and Iris versicolor) .* The features or variables present in this data set are Sepal Length, Sepal Width, Petal Length and Petal Width. Each observation in the dataset also contains the value for the class label, the species name. The Four features were measured from each samples (the length and the width) of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. The central goal here is to design a model which makes good classifications for new data, in other words one which exhibits good generalization.

The British Statistician and Biologist Ronald Fisher in his paper ‘*The use of multiple measurements in taxonomic problems’* introduced **Iris Flower Data Set** as an example of Linear Discriminant analysis. This is aliased as **Anderson’s Iris data Set** because Edgar Anderson collected the data to quantify the morphologic variation of Iris Flowers of three related species.

The Fisher’s linear discriminant model data set became a typical test case for many statistical classification techniques in machine learning such as **Support Vector Machines.**

This Data Set is also a good example to explain the difference between supervised and unsupervised techniques in **Data Mining.**

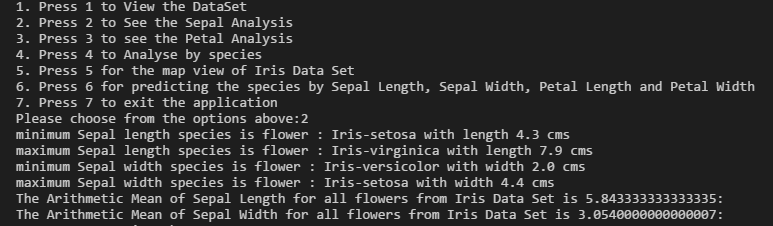
Fisher’s linear discriminant model can be obtained when the object species are known :class labels and clusters are not necessarily the same.

This is a simple script that read data from the Iris Data Set saved in the disk as a comma separated file. The program will present the user with the options appear in the below screenshot, once the application launches:

The user need to choose one of the above options to proceed. The main purpose of the application is to analyse the Iris Data Set. The options 1 to 5 mainly use the default python libraries and option 6 and 7 uses some external libraries like MatPlotLib and Panda.We will go into the functionality covered by each option in the below session.

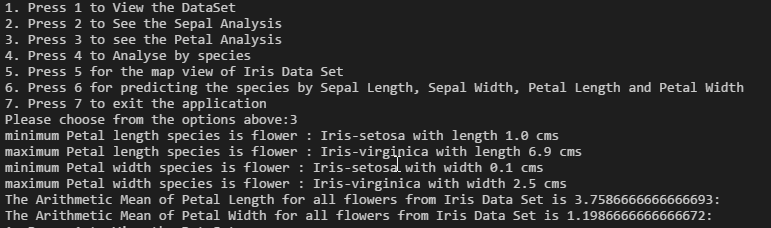
Once the user click the option 1, the application will read the dataset from the csv file and present it to the user.

Option 2 is for the Analysis of the dataset based on the Sepal Length and Sepal width.



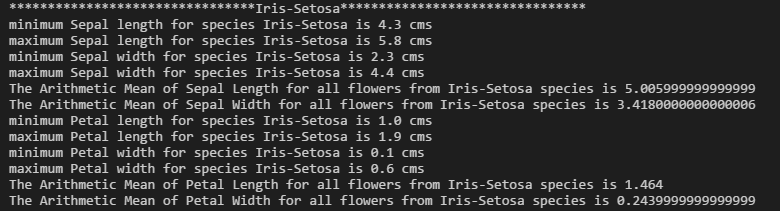
As shown in the screenshot above, once the user choose “2”, the application will display the species with minimum sepal length, species with maximum sepal length, species with minimum sepal width and species with maximum sepal width. It will also display the Arimetic Mean of sepal length and sepal width across all flowers in the data set.

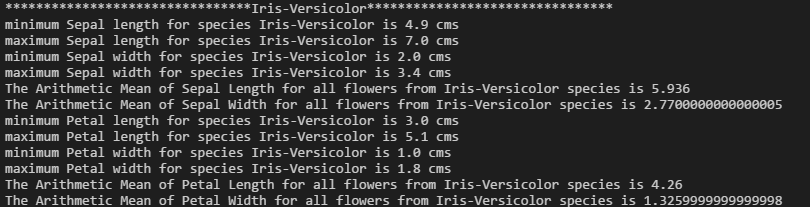
Option 3 is for Anlaysing the flowers based on their Petal attributes. Once the user choose this option, he/she can view the species with minimum petal length, species with maximum petal length, species with minimum petal width and species with maximum petal width. The application will also display the Arimetic Mean of petal length and petal width across all species in the data set, as shown below.

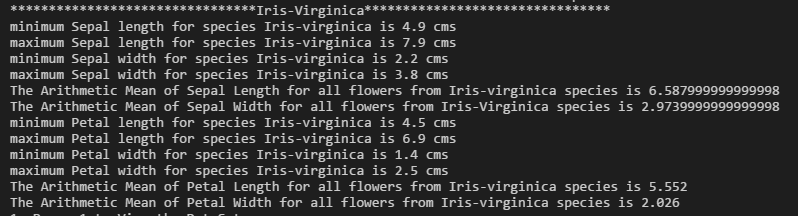


All the above analysis were done across all species. Option 4 will help the users to analyse the flowers by species. That is the results displayed will be grouped by the species Iris-Setosa, Iris-Virginica and Iris-Versicolor.

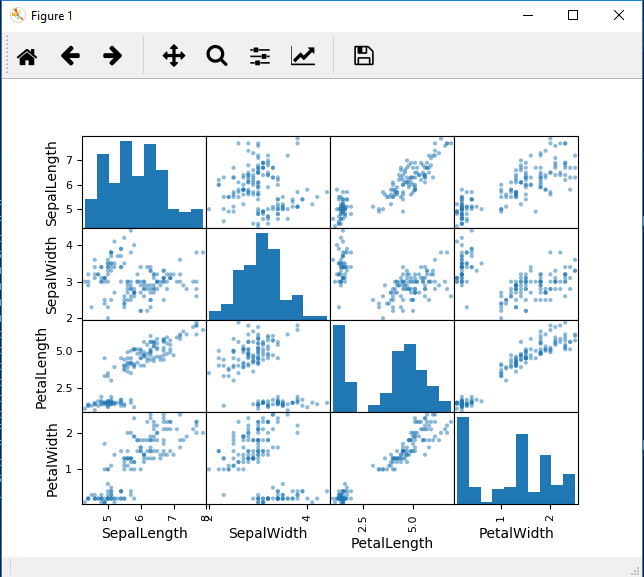
Option 4 will display the minimum petal length, maximum petal length, minimum petal width, maximum petal width and minimum sepal length, maximum sepal length, minimum sepal width, maximum sepal width for each species. The application will also display the Arimetic Mean of lengths and widths of sepal and petals for each species.







Option 5 can be used by the user to analyse the dataset using the scatter plot representation and histogram representation. The application uses the Panda and MatPlotLib libraries to plot the graphical representation of the graphs. The graphical representation of the Iris Data set is as follows.



References :

<https://medium.com/codebagng/basic-analysis-of-the-iris-data-set-using-python-2995618a6342>

<https://www.kaggle.com/saidaantonyan/iris-species-prediction-with-knn-algorithm>

<https://www.dataquest.io/blog/k-nearest-neighbors-in-python/>

<https://stackoverflow.com/questions/37812325/pandas-scatter-plot-with-different-color-legend-for-each-point>

<https://stackoverflow.com/questions/1025379/decimal-alignment-formatting-in-python>

<https://stackoverflow.com/questions/275018/how-can-i-remove-chomp-a-trailing-newline-in-python>

<https://stackoverflow.com/questions/15109165/loading-a-dataset-from-file-to-use-with-sklearn>

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>

<http://www.learn4master.com/machine-learning/visualize-iris-dataset-using-python>

<https://stackoverflow.com/questions/11106823/adding-two-pandas-dataframes>

<http://scikit-learn.org/stable/modules/neighbors.html>

<https://www.kaggle.com/xuhewen/iris-dataset-visualization-and-machine-learning?scriptVersionId=684733>