PROJECT REPORT

JAIDEEP SINGH KAINTH

PROBLEM

Task of this project is to fulfil the K-means Clustering Algorithm on the iris data.

DATA

Data is taken from the following link:

http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

TOOLS

Language used: Python 3.5

Libraries used: numPy, pandas, deepcopy, matplotlib.pyplot, mpl_toolkits.mplot3d

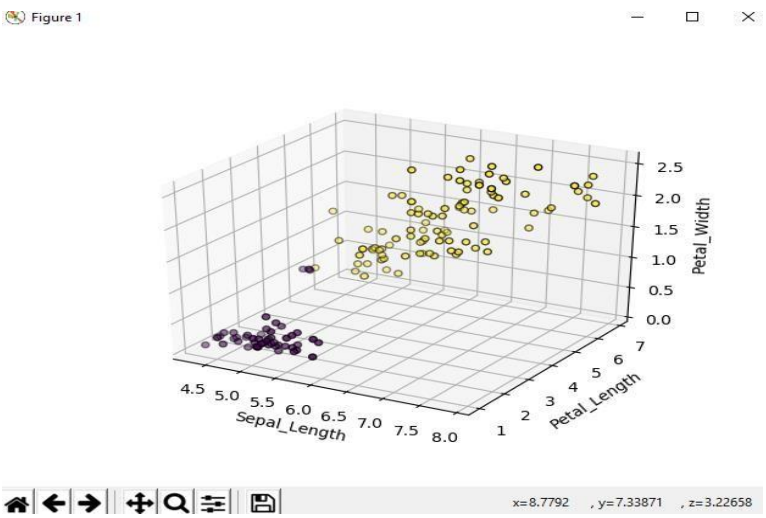
METHOD

In the given data the first four columns represent the dimensions of the flower (Sepal Length, Sepal Width, Petal length, Petal Width) and the fifth column represent the class (Iris-Setosa, Iris-Versicolor, Iris-Virginica). After importing data, I stored it in an array and changed the names of the classes from Iris-Setosa, Iris-Versicolor or Iris-Virginica to 1, 2 and 3 respectively.

First, we will enter the k-fold value that is the number of clusters we want to find. Then based on k we will find k random centroids to begin with. Then we calculate distance of each point from each centroid and assign that point the cluster whose centroids distance is minimum from that point. Then we again compute centroids as the average of all points. We repeat these steps until we find centroids that are stable. After finding clusters I plotted the centroids on graph using only 3 features since it was difficult to plot with 4 features.

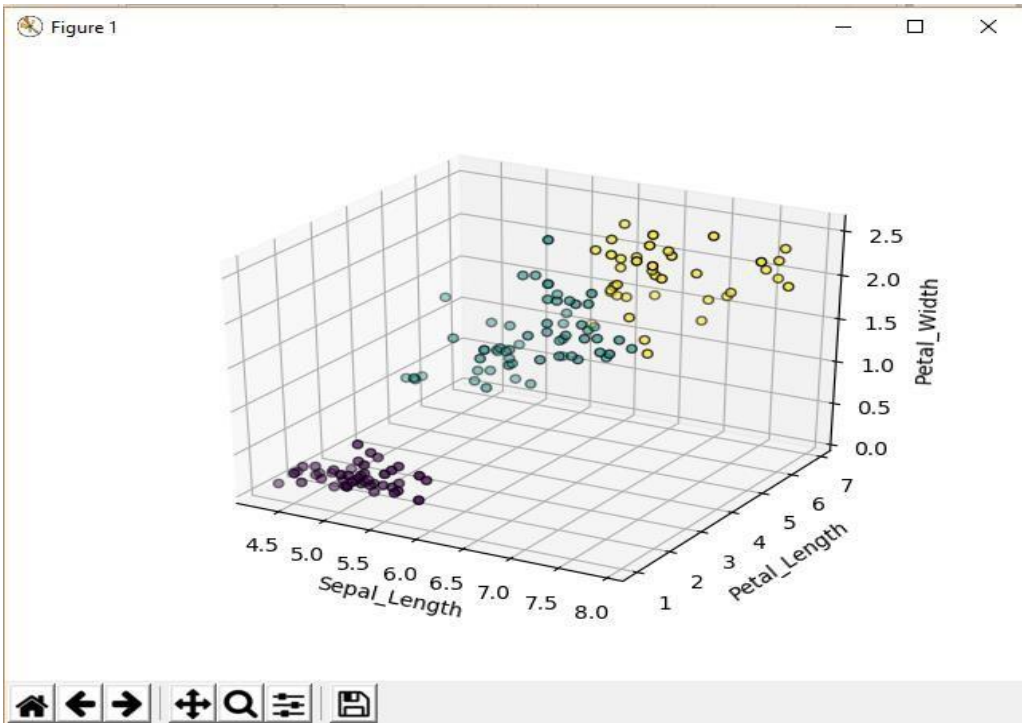
RESULT

K=2:



Centroids are: [[5.00566038 3.36037736 1.56226415 0.28867925]
[6.30103093 2.88659794 4.95876289 1.69587629]]

K=3:

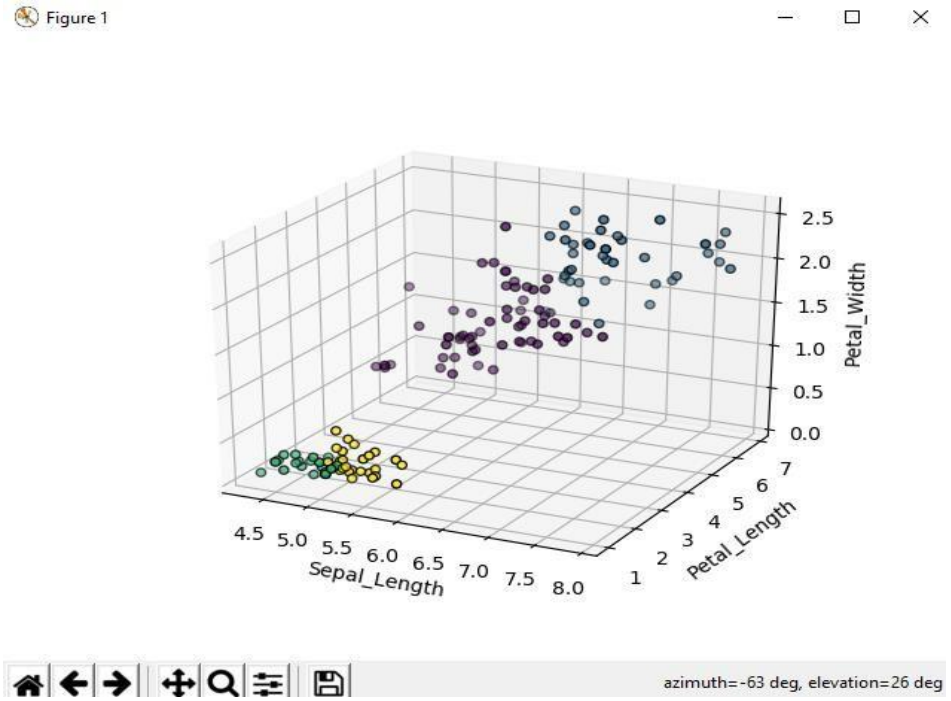


Centroids are: [[5.006 3.418 1.464 0.244]

[5.88360656 2.74098361 4.38852459 1.43442623]

[6.85384615 3.07692308 5.71538462 2.05384615]]

K=4:



Centroids are: [[5.9016129 2.7483871 4.39354839 1.43387097]

[6.85   3.07368421 5.74210526 2.07105263]

[4.725   3.13333333 1.42083333 0.19166667]

[5.26538462 3.68076923 1.50384615 0.29230769]]

CONCLUSION:

Using the elbow method, we can see that the optimal value for k is 3. Also, since we have 3 classes for the flowers so there will be three clusters one for each class.