Report

Machine Learning(CS6000)

Indian Institute of Technology, Kharagpur

Assignment-2

-Atishay Jain(20CS30008)
-Gaurav Malakar(20CS10029)

Ques-1:

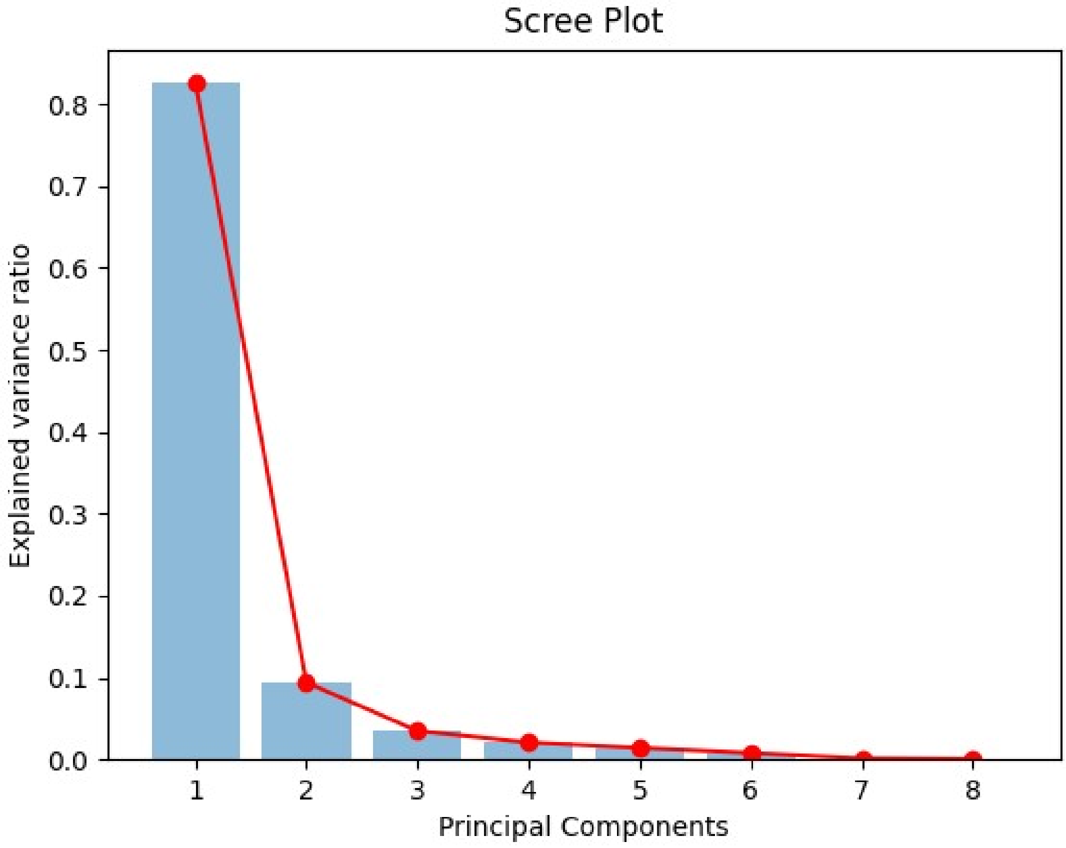
Dataset-D
Abalone Data
No. of entries in dataset: 4177
No. of attributes: 8

Required variance after applying Principal Component Analysis(PCA) ≥ 0.95

We have used the inbuilt function pca.fit_transform() to find the most important principal components.

We need to pass the normalised dataset to this function.

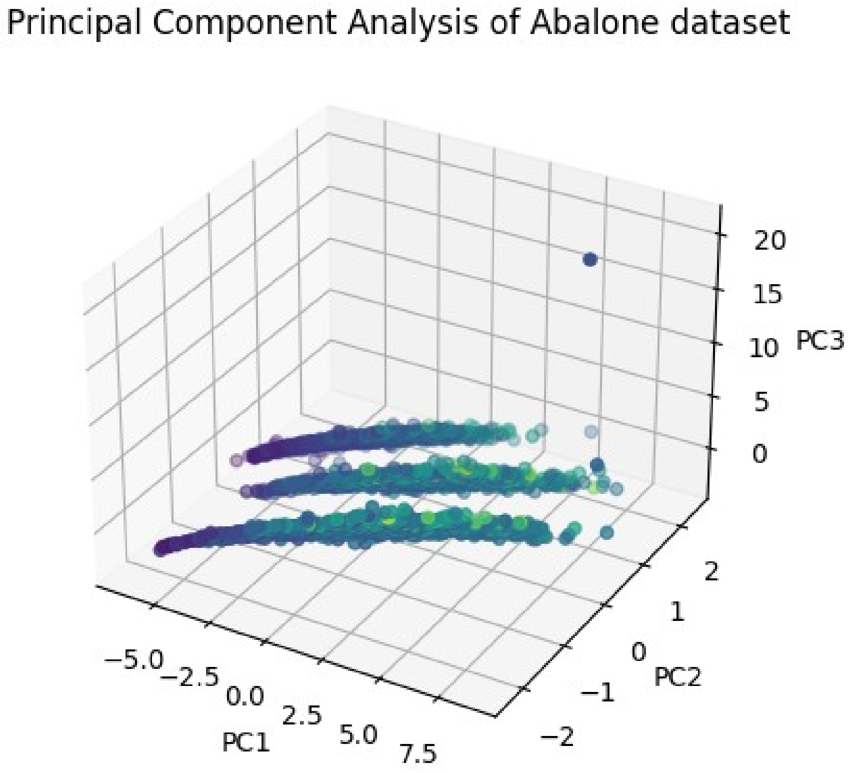
We can see the variance ratio of each principal component axes in the scree plot below



More specifically, the values are [0.8255748, 0.09384237, 0.03491563, 0.02091799, 0.01425899, 0.00807432, 0.0015841, 0.00083172]

So, we got 3 principal axes which will preserve a total variance of atleast 0.95.
as 0.8255748 + 0.09384237 + 0.03491563 > 0.95

The graph for PCA with the 3 principal components derived from the pca.fit_transform() function is shown below.



The color shade of point represent its label.

Now, we need to cluster these data points using k-Means clustering algorithm.

We will find the normalised mutual information (NMI) score of the labels generated by clustering algorithms and the orginal labels.

K-Means Clustering Algorithm:

First we randomly select k centroids, find the distance of every point with every centroid and assign the nearest centroid index as its label.

Then we find the mean of the points assigned the same label and use these as the new centroids.

We will keep reating this unitl the new list of centrods is same as the last one or we have done maxIter no. of iterations.

Normalised Mutual Information Score Algorithm:

We will find the unique the elements of both X and Y first.

Then find the probability of each of these elements.

Then we will find joint probalility distribution $p_{XY}(x, y)$.

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)}\right)$$

$$H(X) = -\sum_i p_i \log p_i \,,$$

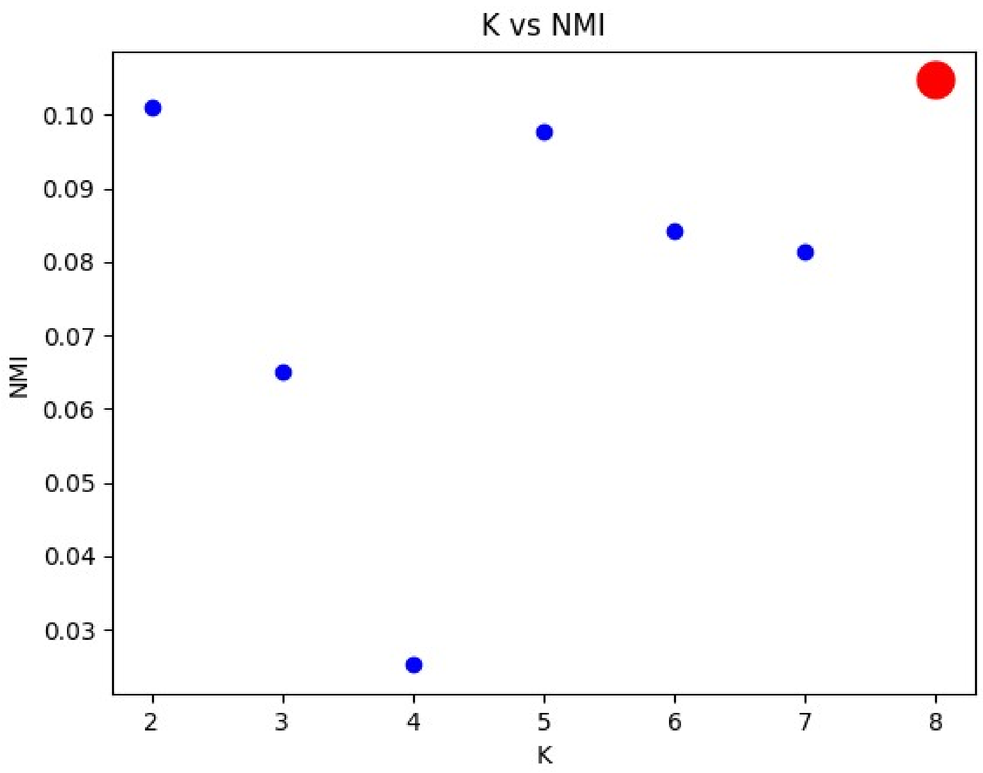
Similarly H(Y)

NMI = 2 * I(X;Y) / (H(X) + H(Y))

We will find the k clustering and its normalised mutual inforamtion score with the origianl labels for k = 2 to 8.

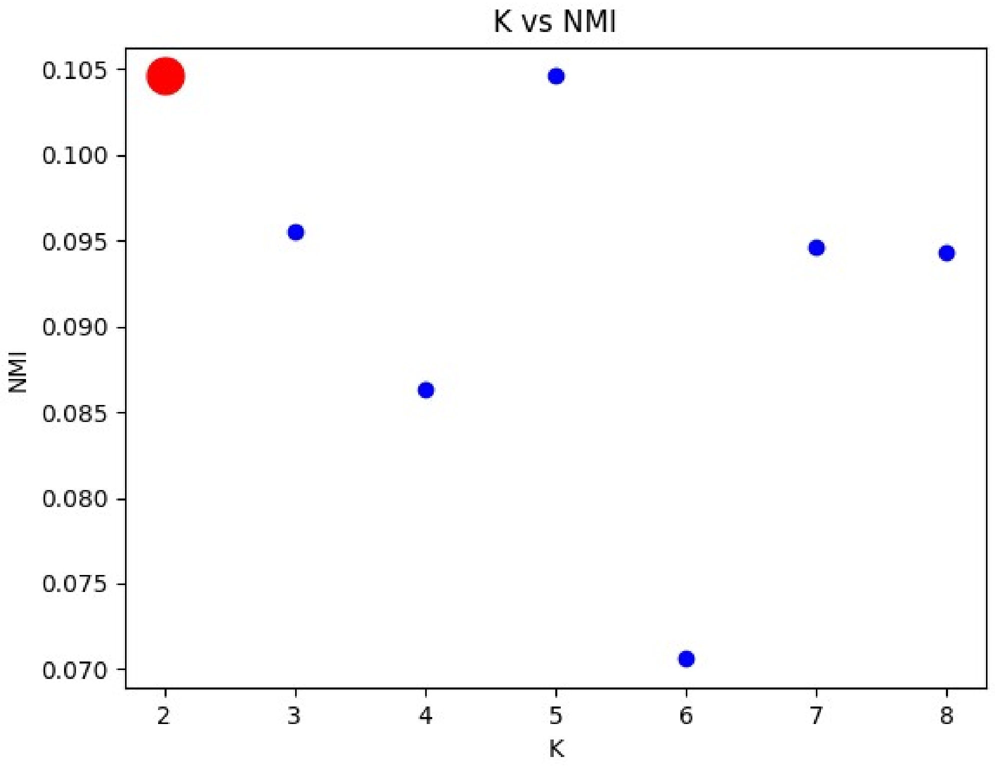
We need to select that value of k for which NMI score is maximum.

The graph of K vs NMI is shown below



As, we can see here, for k=7, the NMI score is maximum. Although we can get different values in different runs of the algorithm as initially we chose the centroids randomly.

As we can we see below,


K vs NMI
NMI
K