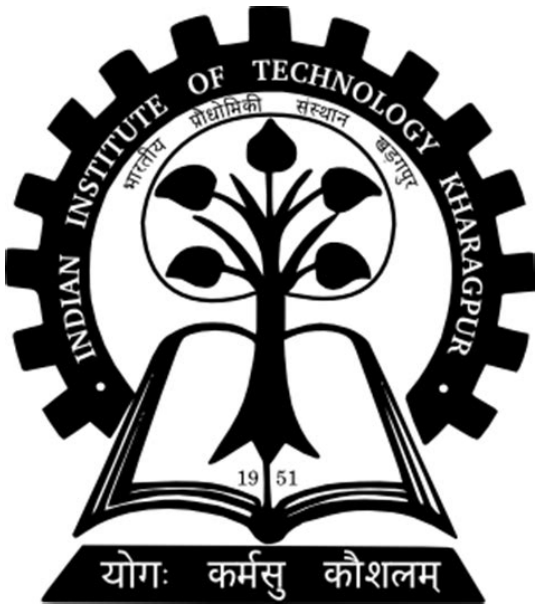
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



ASSIGNMENT 2
(MACHINE LEARNING)

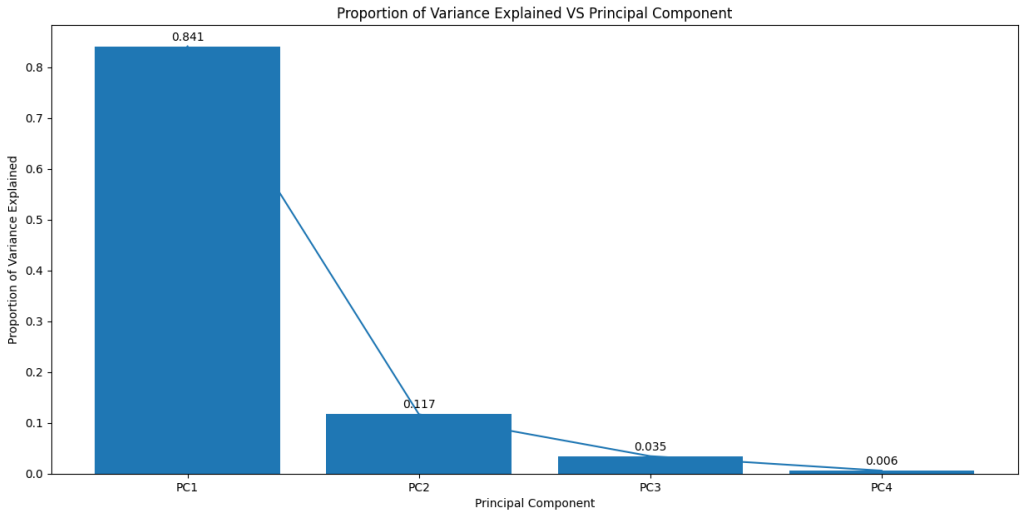
Gandham Heamanth Rao 20CS10027
Talabattula Sai Sahan 20CS30055

UNSUPERVISED LEARNING

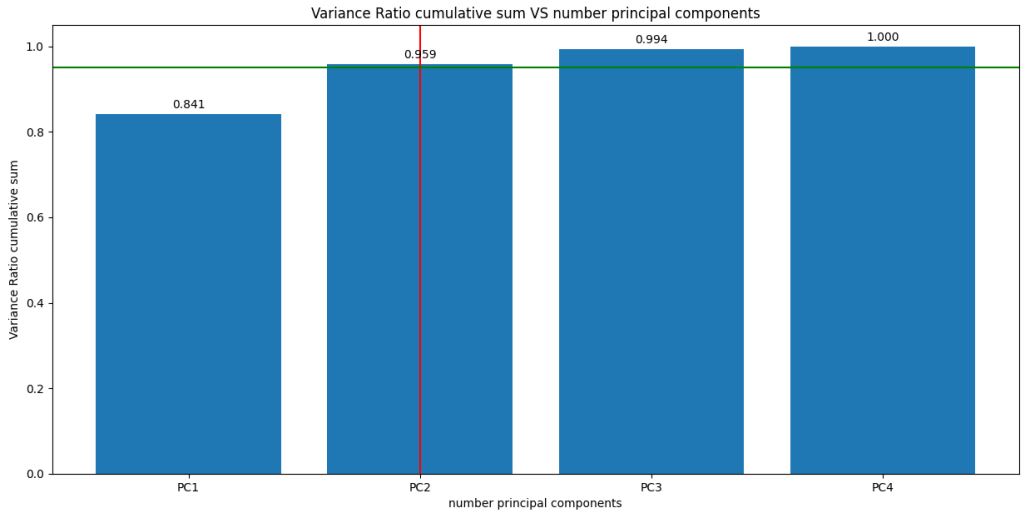
Question:

1) Apply PCA (select number of components by preserving 95% of total variance). (in-built function allowed for PCA).
2) Plot the graph for PCA.
3) Using the features extracted from PCA, apply K-Means Clustering. Vary the value of K from 2 to 8. Plot the graph of K vs normalised mutual information (NMI). Report the value of K for which the NMI is maximum. (in-built function not allowed for K-Means).

GRAPH OF PROPORTION VAR vs PCA

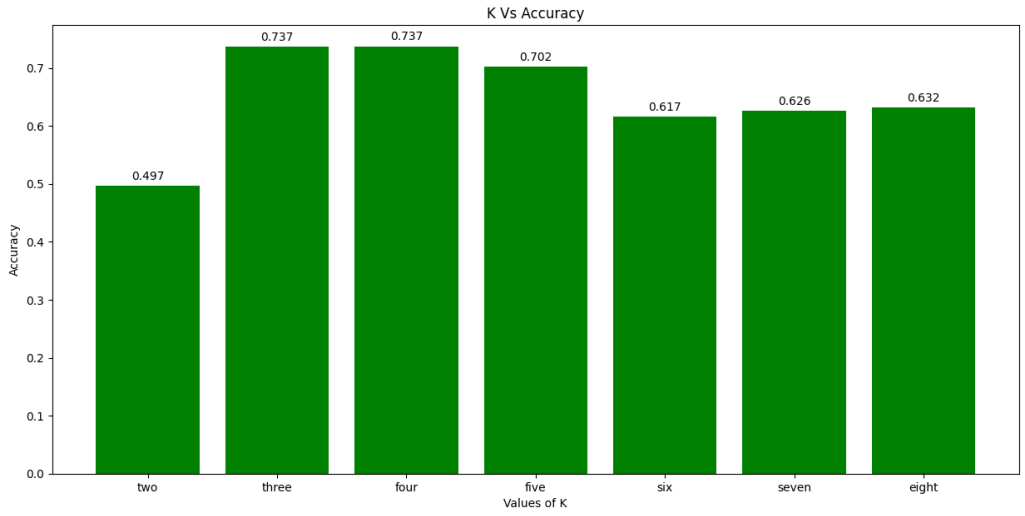


GRAPH OF VARIANCE OF CUMULATIVE SUM vsVARIANCE



So from the above two observations the variance ratio cumulative sum reaches greater than or equal to 0.95 at 2 principal components.

K vs NMI



ALGORITHM DONE:

1. First calculated the pca with the number of components 4 and had the cumulative variance crossing 0.95 at 2 components.
2. So after that the number of components is set to 2 and data is fit according to that.

3. Now apply k means clustering for k from 2-8.
4. K-means clustering is done by taking k random points.
5. After that we cluster it based on euclidean distance and again find the center of each distribution.
6. If the new centers match with old points then we stop or else we continue with the above two steps.
7. And using NMI I have compared the clusters and it was found that for k = 4 it has highest accuracy.

OBJECTIVES DONE:

1. Has normalized the given data.
2. Applied PCA on the given data.
3. Plot of the pca and apply the pca to the best number of components, that is 2.
4. Applied the k-means clustering and plotted k vs NMI.
5. Found that for k = 4 NMI is highest.