Mini Project 3

Mohammadreza Akhtari

Contents

Introduction

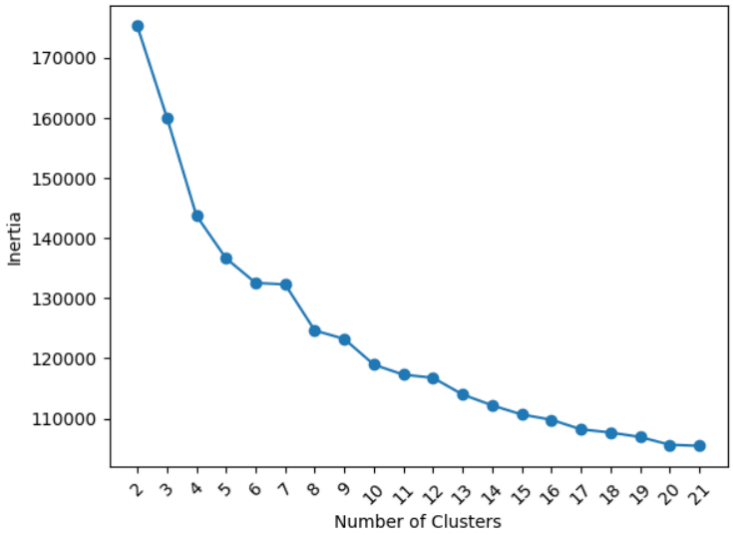
The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. It is intended to identify patterns and clusters within human activity recognition (HAR) datasets. This recognition has various applications, including health monitoring, fitness tracking, behaviour analysis, and so on. To achieve this purpose, data processing and pre-treatment have been done on our data.

Data processing

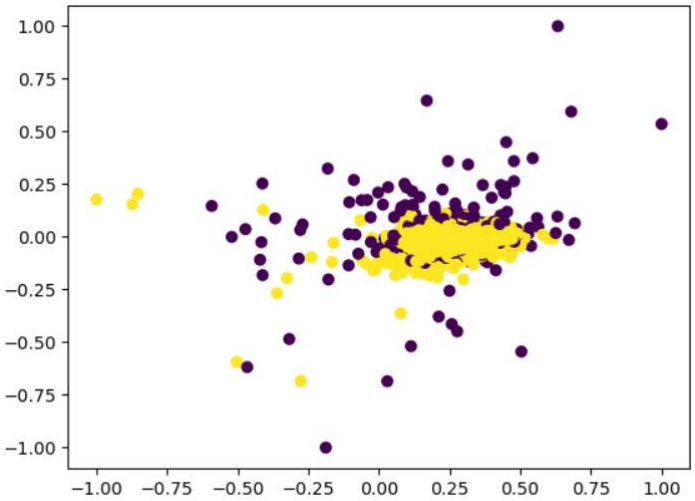
Data is downloaded from the provided link and unzipped to train and test the dataset. Data has been checked for any missing values which was not the case here. The information, summary or description of data have been checked generally for any suspicious input. Then, data is used for clustering by k-means and DBSCAN models before and after dimension reduction using PCA.

K-means without dimension reduction:

First, the dataset is used in a raw mode for our analysis. The k-means algorithm is used for clustering the dataset. Various numbers of clusters are considered at the beginning to see the effect of the number of clusters. As the following image shows, inertia is calculated as the distance of samples to the centre of clusters versus various cluster numbers.

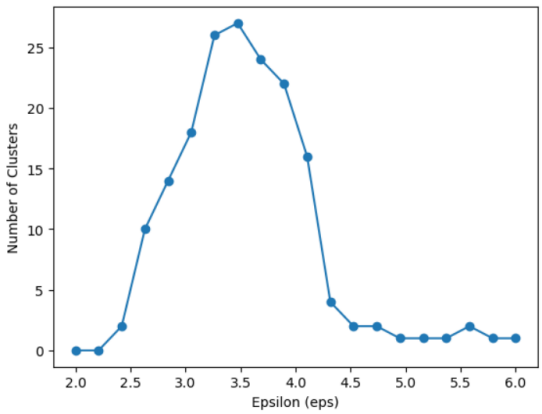


It could be seen that as the number of clusters increases, the centres get closer to data leading to the reduction of inertia values. To find the optimum clusters, I select the one that only changes less than 10 % leading to the 2 clusters. Therefore, doing k-means clustering with 2 clusters leads to the following graph.

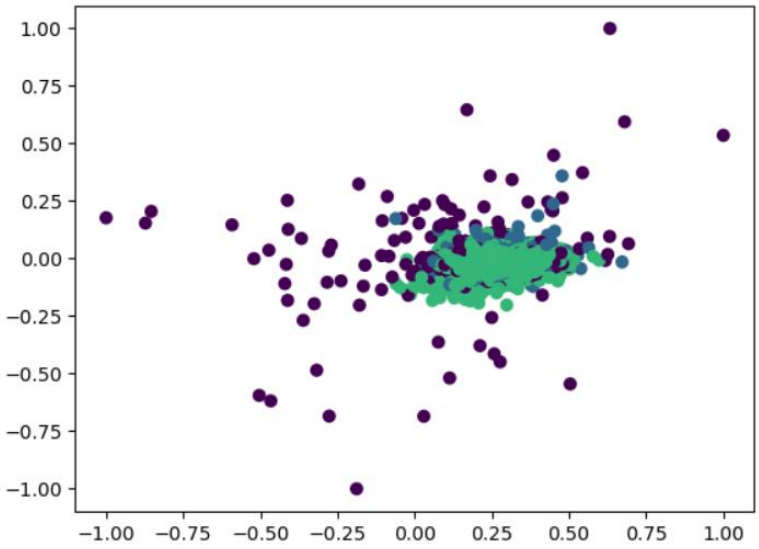


DBSCAN without PCA

DBSCAN is used as another method for clustering. In the beginning, various epsilons are tried to identify the appropriate amount of epsilon, which is 4.5 here. Also, various k numbers are utilized to find the differences in results and obtain more concrete values.



Then, clustering using eps=4.5 and the DBSCAN method delivers the following result. It is worth mentioning that various numbers of sampling are also considered to calculate the best parameters for DBSCAN as the results and graphs can be seen in the Jupyter file as attach.

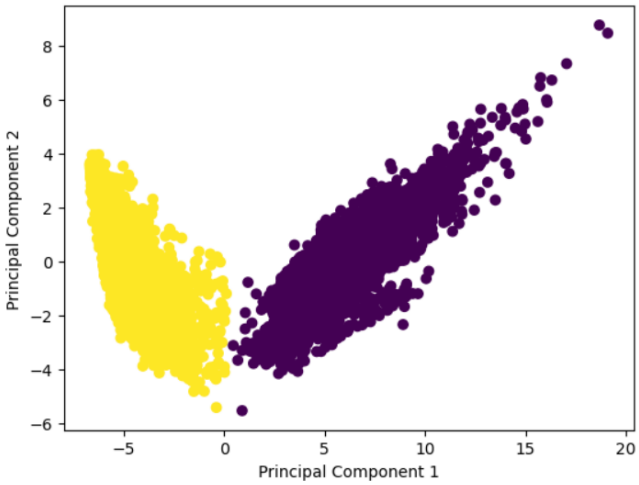


Dimensionality reduction

To get better visualization and processing, PCA is used for the reduction of dimension into 2D space. K-means and DBSCAN models are recalculated to observe the effect of dimensionality reduction on results.

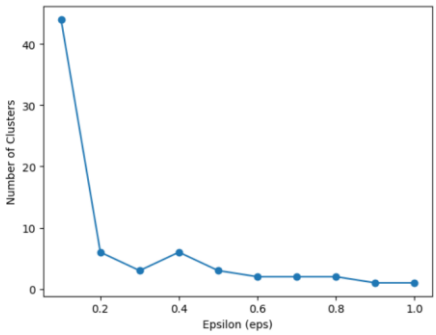
PCA and k-means

After performing PCA for dimensionality reduction, results for the k-means model are visualized below with two obvious clusters.

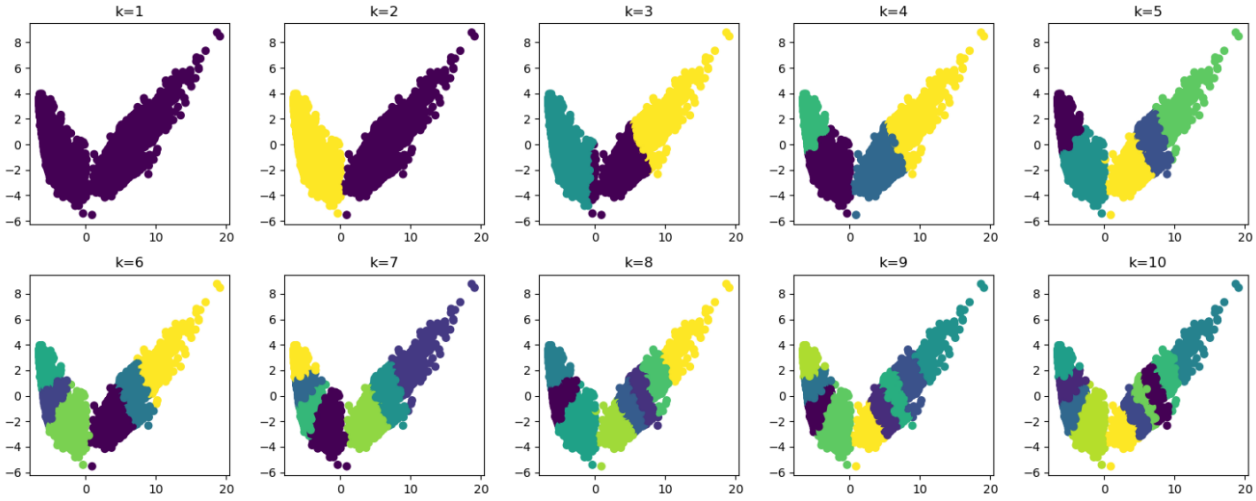


DBSCAN with PCA

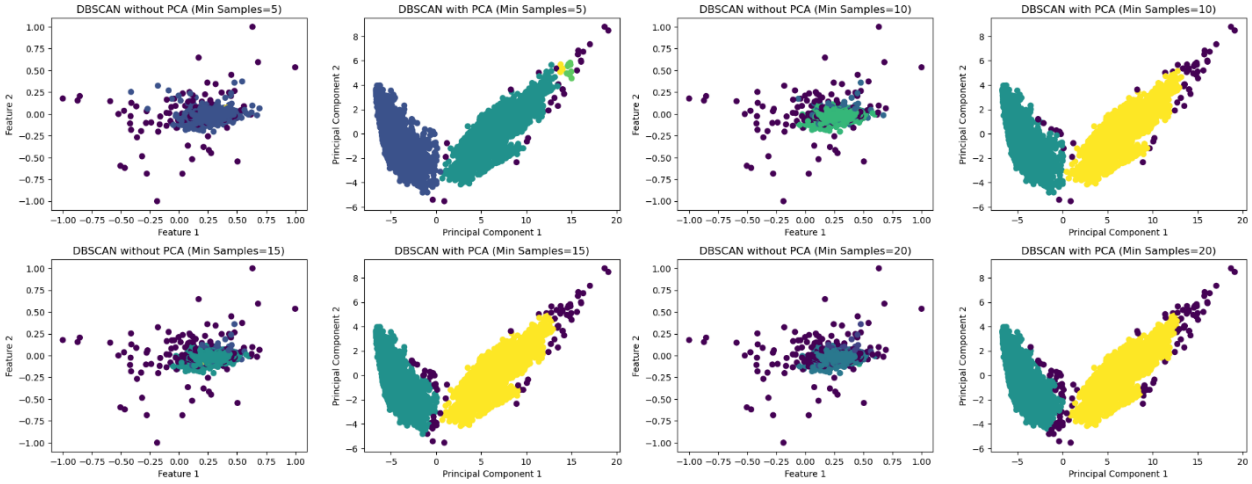
DBSCAN is again used but based on the data with reduced dimensionality. The model is recalculated again for various parameters as the following graph. The optimal epsilon is observed to be 0.6 in this case for DBSCAN with PCA method for dimensionality reduction.



Although 2 clusters are the best, various numbers of clusters are explored to see the effect of this parameter as below:



To see the effect of dimensionality reduction together without PCA, the following graphs are drawn near each other for better investigation.'



Conclusion

As can be seen, the visualization of data without dimensionality reduction is not that good. However, two obvious cluster are observable after PCA or dimensionality reduction which is basically used for better visualization. Therefore, it is important to use dimensionality reduction to see clusters and differences. Otherwise, it would be hard to see the differences or to visualize clusters. However, to select the best parameters such as epsilon, there could be more investigation and dividing epsilon into smaller intervals. However, a bottleneck here is the computational problem which makes it impossible to investigate all the possibilities.