Clustering with K-Means and DBSCAN

1. Introduction

We were tasked with clusterization, cluster visualization,finding good parameter values for the models and data processing of the UCI HAR dataset. Additionally we needed to apply dimensional reduction to the dataset and compare results.

The dataset consists of various accelerometer data collected and measured by a smartphone on the person.

2. Data processing

It was slightly difficult to understand how to get the dataset to work but after getting the test and train dataframes to display correctly no preprocessing was needed. Most if not all of the dataframe maneuvering was done with pandas.

An additional column was added which displayed the various states labeled by text, for example 1 = Walking.

Dimensionality reduction was performed with PCA on X_train which reduced the dataframe to a two dimensional dataframe.

3. Modeling

3.1 Choosing parameters by analyzing dataset

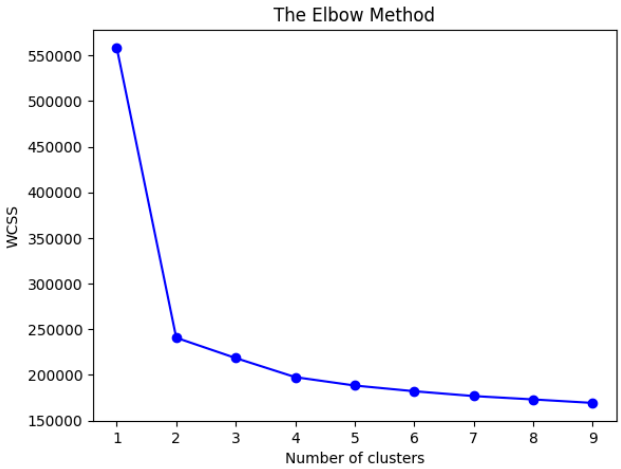

FIG 1

From FIG 1 displaying a graph of the elbow method on the non-dimensionality reduced dataset we can see that the optimal number of clusters is either 2 or 4.

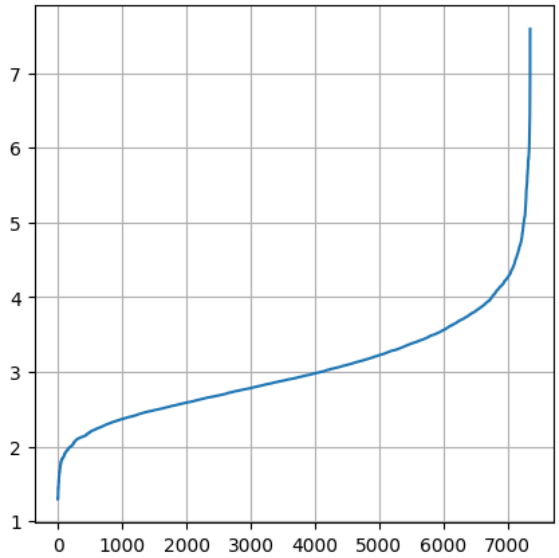

FIG 2

For the optimal number of eps for DBSCAN we can see from FIG 2 that it is around 4,5.

3.2 Choosing parameters by analyzing dimensionality reduced dataset

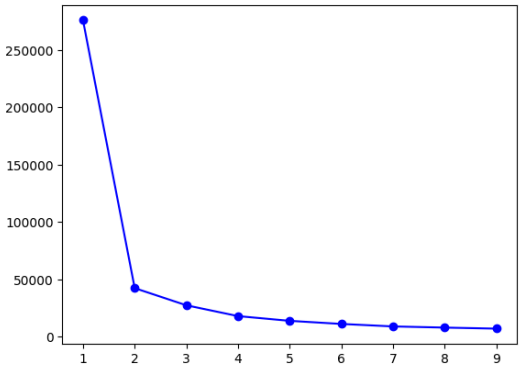

FIG 3

From FIG 3 elbow method graph for the dimensionality reduced dataset we can see that the optimal number of clusters is 2

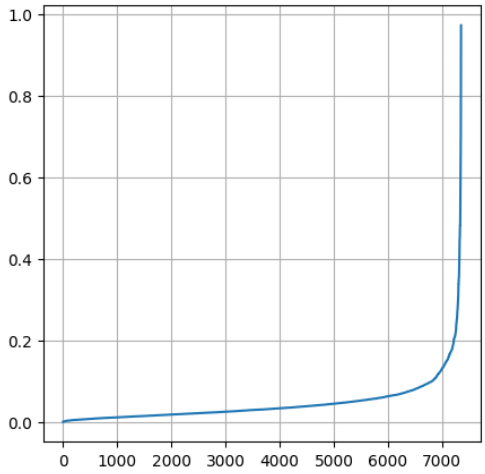

FIG 4

From FIG 4 we can see that the optimal eps is around 0.1, final value chosen for eps was around 0.25.

4. Cluster visualization

4.1 Normal dataset

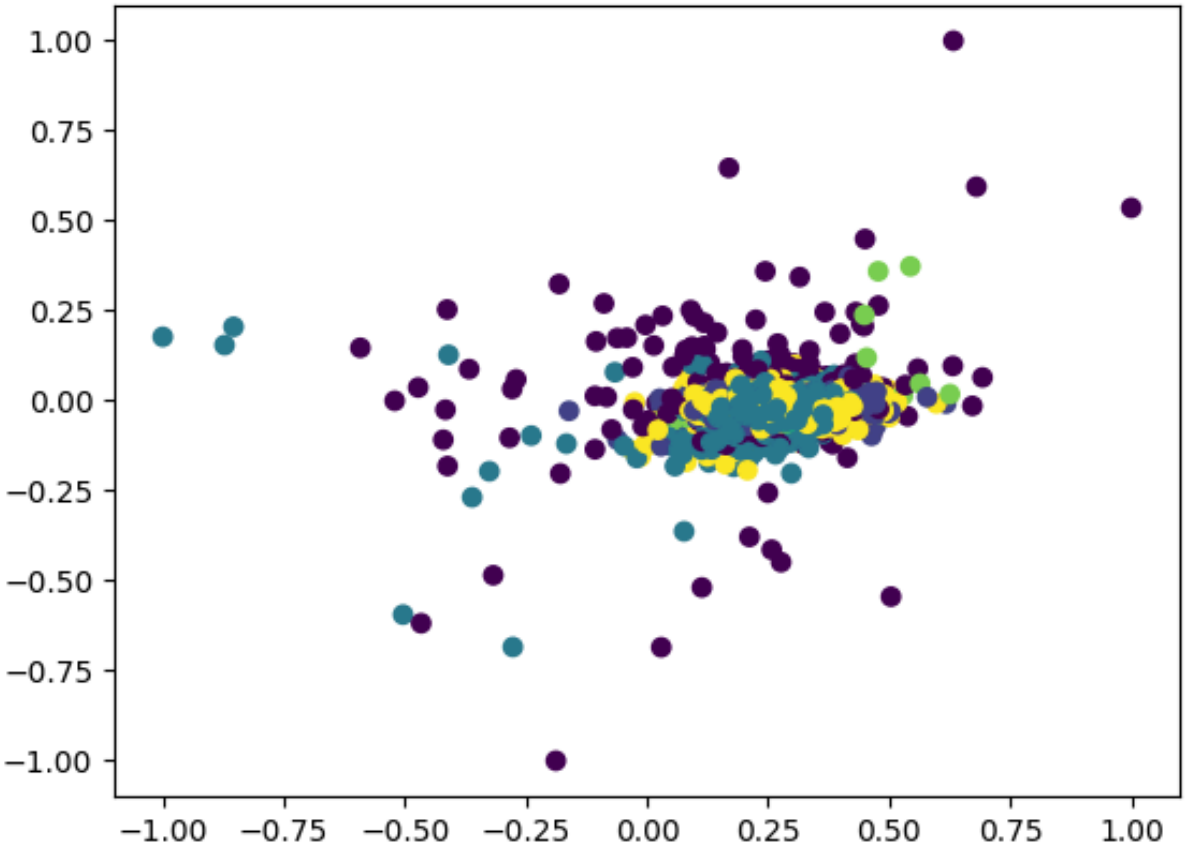

FIG 12 K-means

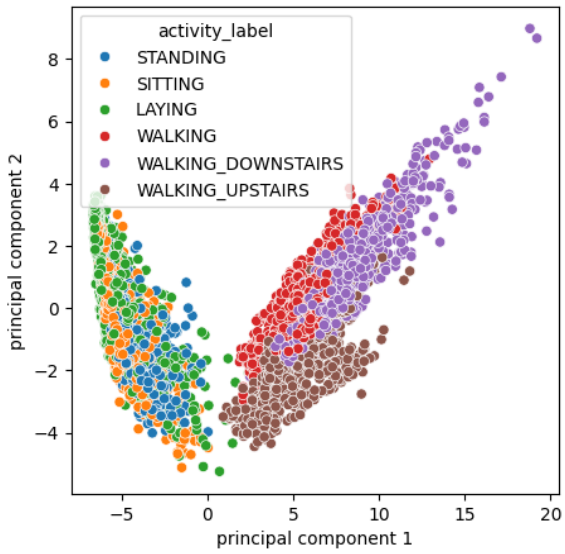

FIG 5 DBSCAN

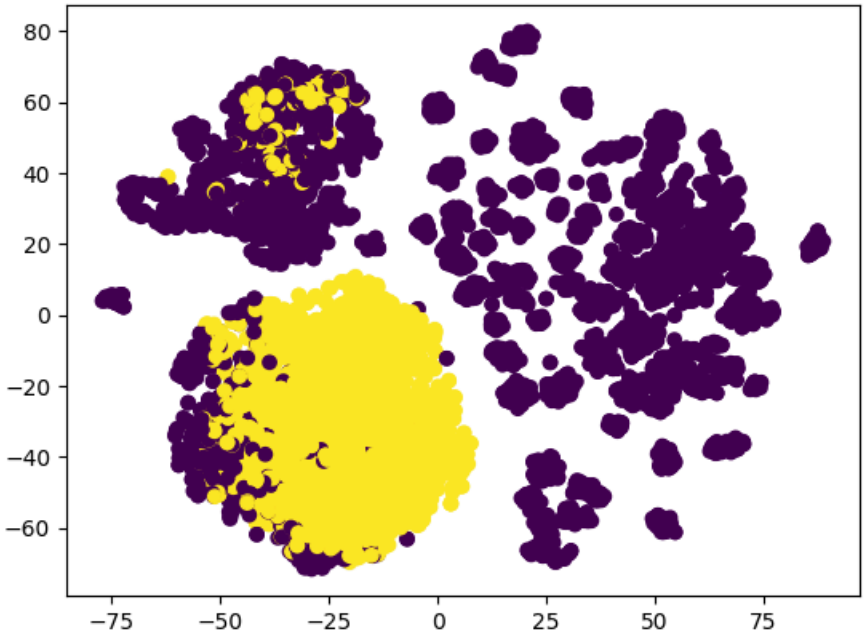

FIG 6 TSNE

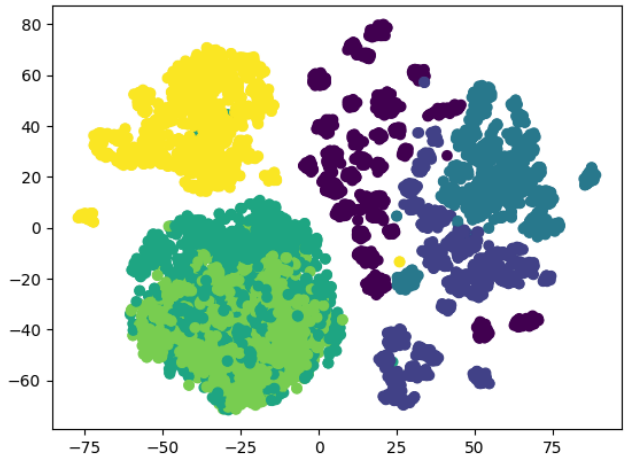

FIG 7 Optimal TSNE

4.2 With dimensionality reduction

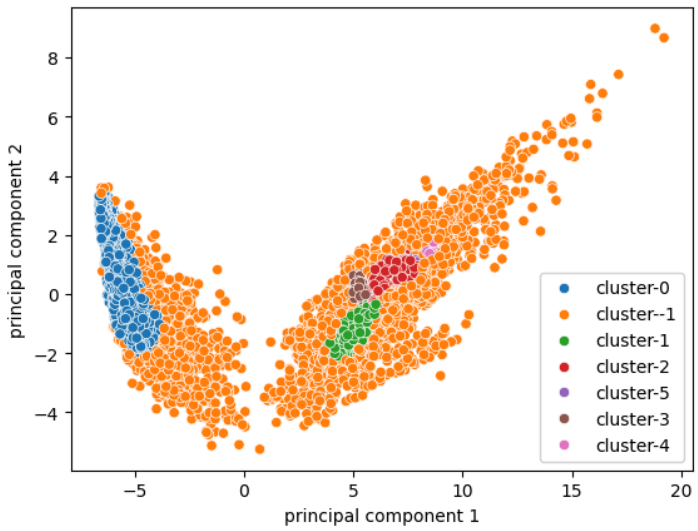

FIG 8 dimensionality reduced DBSCAN

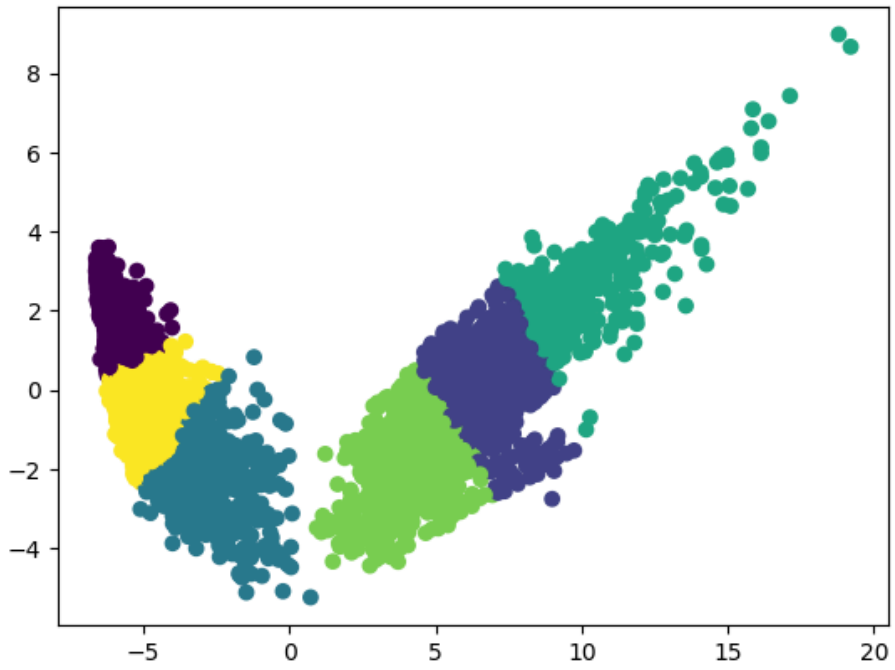

FIG 11 Dimensionality reduced K-means

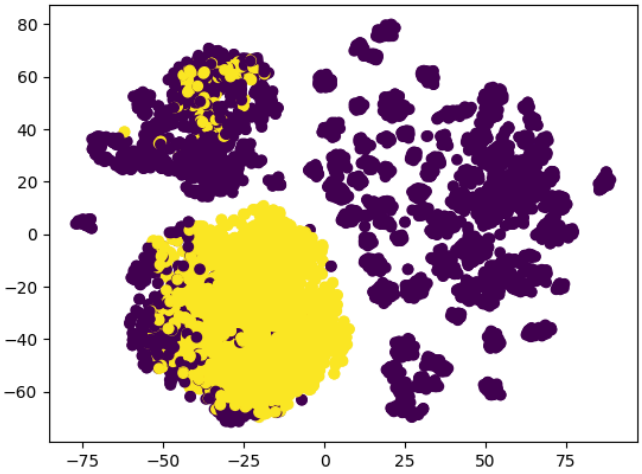

FIG 9 TSNE

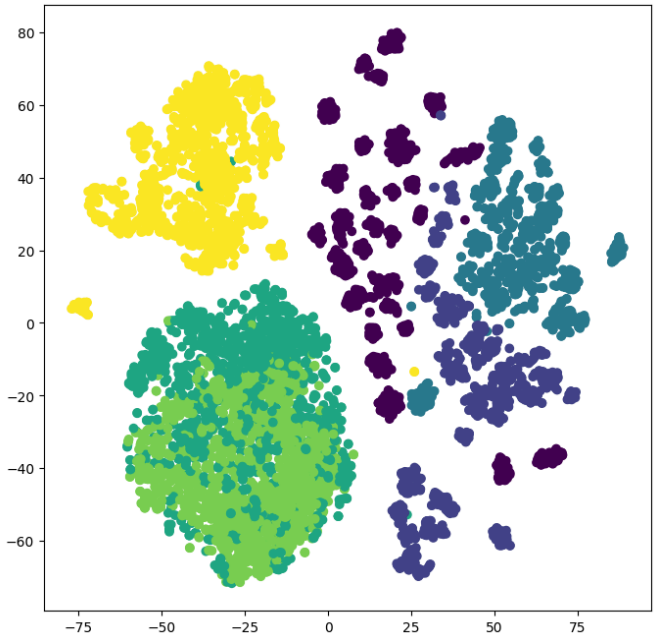

FIG 10 Optimal TSNE

5. Conclusion

We can infer from cluster visualizations that PCA dimensionality reduction improved the model greatly.