Lab 2: Unsupervised Learning

ARI 510

Yola Charara

1. Exploratory Analysis

The dataset consists of six classes representing six different activities performed by 30 volunteers and 561 features vectors. The activities are: “walking, walking upstairs, walking downstairs, sitting, standing, laying.” Figure 1 below shows the distribution of activity labels in the dataset. The histogram shows that the data is quite evenly distributed among the classes. This indicates that class imbalance will not pose a significant concern later in model training.

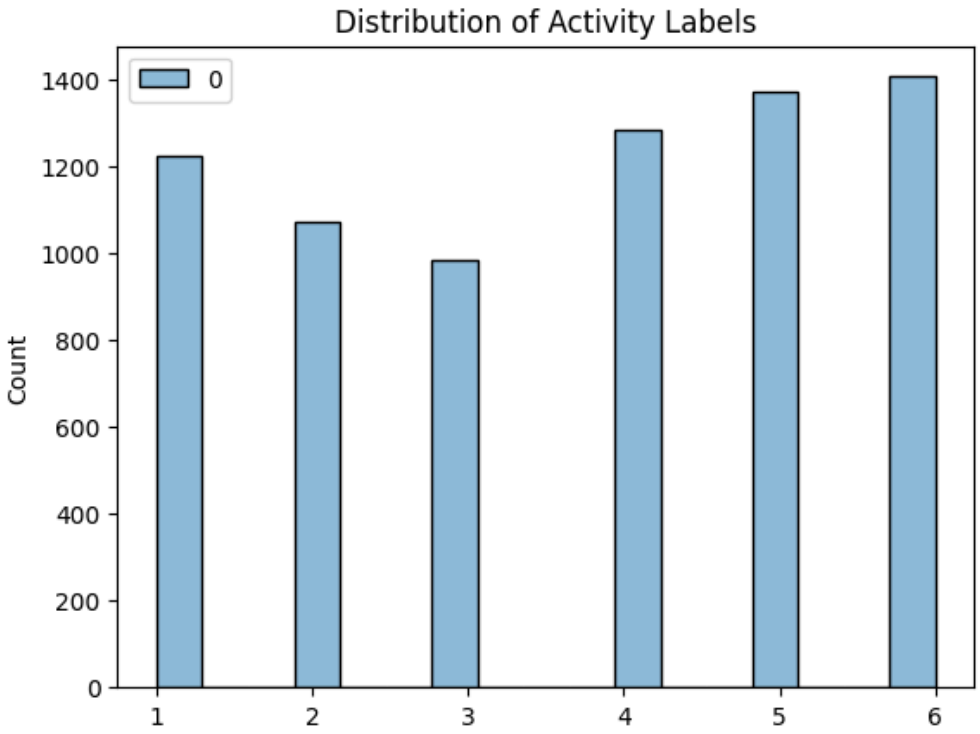

Figure 1: Histogram of the “Distribution of Activity Labels” of the data.

Then, PCA was applied to the feature set to reduce its dimensionality. The reduced features were used to plot the explained variance ratio for each principle component as shown in Figure 2 below. Following the “elbow” approach, we can see that the optimal number of components is k=2.

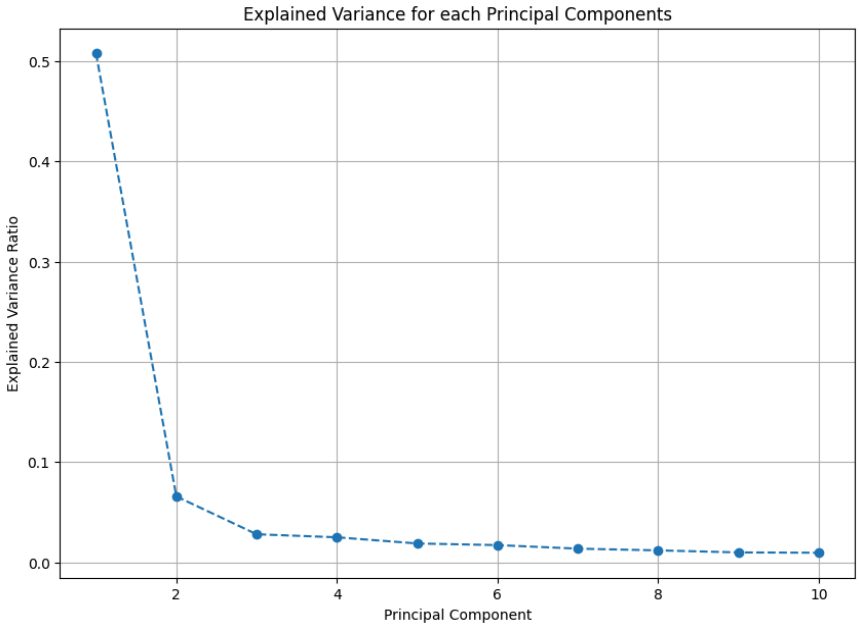

Figure 2: Graph showing the "Explained Variance for each Principle Components".

The first two principal components were used to plot the data into a reduced 2D space, as shown in Figure 3 below. A scatter plot was made showing the six different activity labels. Each class is represented in a different color and seen in a cluster. An overlap between clusters can be seen. Some classes are more overlapping than others, which can be because some activities may have a lot of similarities in their data. Most clusters are distinguishable, suggesting that PCA could partially divide the data by activity.

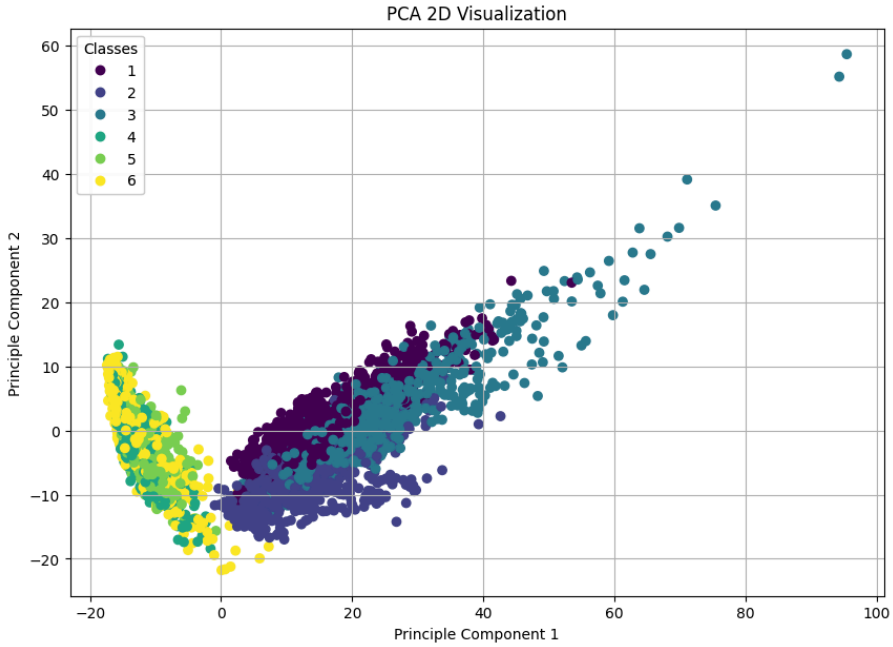

Figure 3: Scatter plot for "PCA 2D Visualization" based on activity labels.

2. Clustering

K-Means Clustering and Silhouette Score were used on the full features dataset to identify the optimal number of clusters as shown in Figure 4 below. When the number of clusters is equal to 2, a peak in the Silhouette Score is seen meaning that 2 is the optimal number of clusters for this dataset.

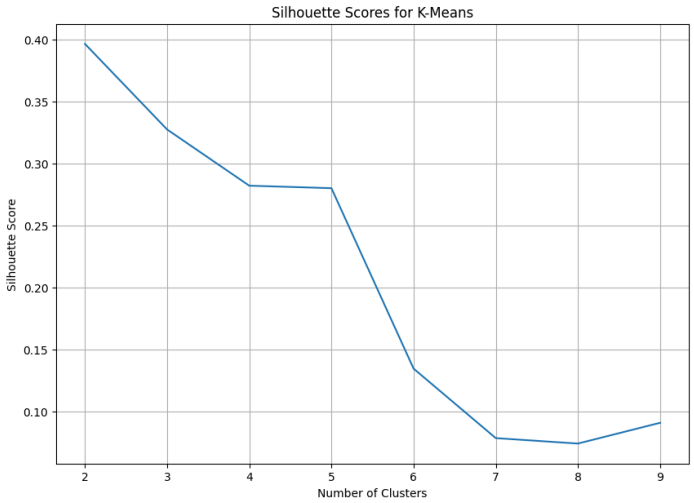

Figure 4: Graph representing the "Silhouette Score for K-Means" on the full features dataset.

Figure 5 below represents the K-Means Clustering for PCA in a 2D scatter plot using the cluster labels. This clustering shows a pretty similar shape as Figure 3, which was the scatter plot for the activity labels, but the clusters are better grouped in Figure 5. K-Means clustering enforces the data into two clear clusters. The number of clusters was chosen beforehand based on the optimal silhouette score and applied in the algorithm, which in this case is two.

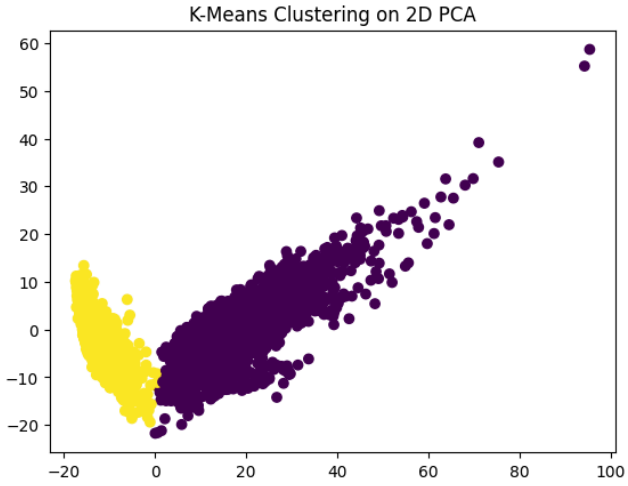

Figure 5: Scatter Plot representing "K-Means Clustering on 2D PCA" based on cluster labels.

Figure 6 below shows the DBSCAN Clustering on similar 2D PCA transformed data. This method determines clusters based on their density. Unlike K-Means Clustering, DBSCAN doesn't enforce data into a pre-defined number of clusters. DBSCAN Clustering couldn't identify the clusters well, suggesting that the clusters have a low density. Changing the "eps" and "min_samples" parameters in the algorithm determines how well we see the clusters. Figure 6 parameters were set as follows: eps=8, min_samples=5, which mainly shows one cluster. When changing the parameters to: eps=15, min_samples=1, as seen in Figure 7, we can see more clusters that are also found in Figure 3, suggesting that it could capture the density of the six different clusters, but the grouping of the clusters in Figure 7 is not quite the same as in Figure 3.

We can determine that K-Means Clustering performs better at grouping clusters and avoiding overlaps in this dataset.

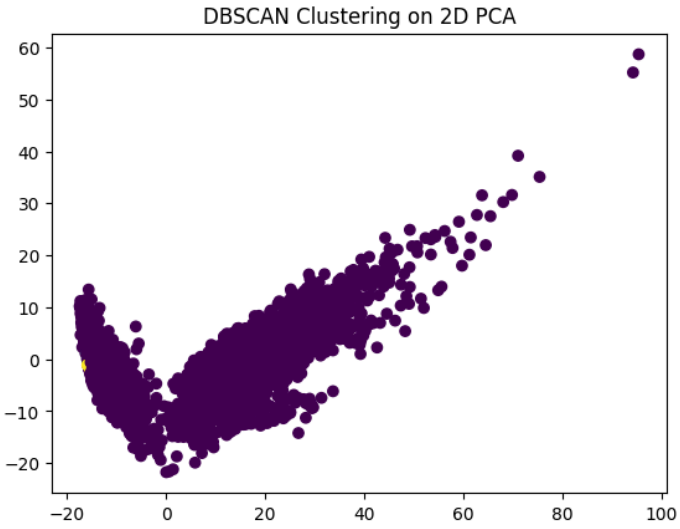

Figure 6: Scatter Plot showing the "DBSCAN Clustering on 2D PCA" (esp=8, min_samples=5).

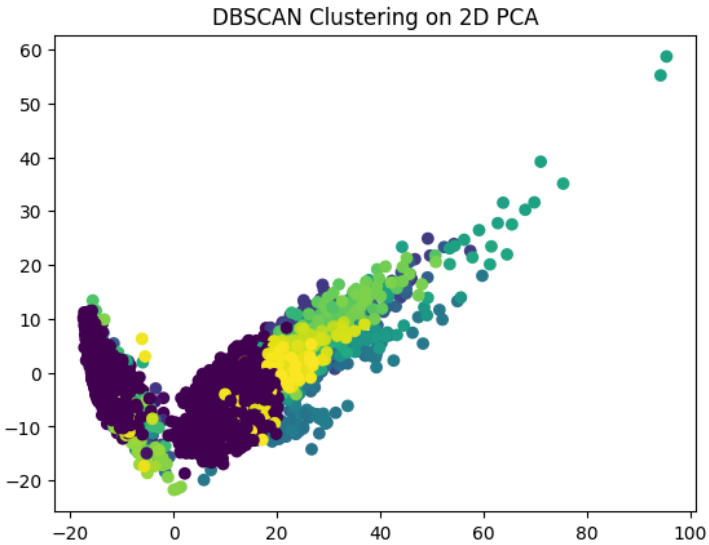

Figure 7: Scatter Plot showing the "DBSCAN Clustering on 2D PCA" (esp=15, min_samples=1).

3. Classification

For k = 2

Model	Precision	Recall	F1-Score
Full Feature Set (Logistic Regression)	**0.96**	**0.96**	**0.96**
PCA Reduced Set (Logistic Regression)	**0.55**	**0.57**	**0.53**
Random Forest Reduced Set	**0.34**	**0.37**	**0.30**

Table 1: Variations of Classification Models for k=2.

For k = 50

Model	Precision	Recall	F1-Score
Full Feature Set (Logistic Regression)	**0.96**	**0.96**	**0.96**
PCA Reduced Set (Logistic Regression)	**0.91**	**0.91**	**0.91**
Random Forest Reduced Set	**0.89**	**0.89**	**0.89**

Table 2: Variations of Classification Models for k=50.

A Logistic Regression classifier was performed on the full set of 561 features vectors. Tables 1 and 2 above show similar and good scores on the performance metrics ("precision," "recall," and "f1-score") which are 0.96.

When setting k = 2, like seen in Table 1, we can observe poor performance for the "PCA Reduced Set," and the "Random Forest Reduced Set" having very low scores on the performance metrics. But when changing the setting to k = 50 as seen in Table 2, the performance metrics for these two models increased and showed pretty good scores.

A higher value of k allows the model to keep more principal components and features in the dataset. So, reducing the value of k may result in a loss of important information for the dataset. In this situation, selecting k=50 means that we now have 50 principal components, and that accounts for more percentage of the variation in the initial data set, compared to k = 2. Despite the fact that this dataset only has six classes, the data's complexity may need the use of additional features to correctly differentiate among the six classes. Thus, adding additional principal components allows the model to collect more information, resulting in better prediction performance.

4. Reflection

Completing this assignment was a good way to study and analyze different types of models and how well they describe the data. The most surprising part was getting lower scores on the reduced data because normally a reduced model gets rid of unnecessary information to achieve higher scores on the performance metrics. Overall, this was an interesting assignment that helped me better comprehend the way each model works.