Mini Project 3 (**Human Activity Recognition Using Machine Learning**)

Introduction

The focus of this mini project is to recognize human activities using sensor data collected from Smartphone's. The problem being solved is the classification of six different activities performed by individuals, using machine learning techniques on a dataset of 3-axial linear acceleration and 3-axial angular velocity readings.

Data Processing

The dataset was preprocessed before applying the machine learning models. The preprocessing steps included loading the data, checking for missing values, standardizing the features, and reducing the dimensionality of the data using Principal Component Analysis (PCA).

The choice to standardize the features was made to ensure that all features contribute equally to the model performance. PCA was used to reduce the dimensionality of the data, which can improve computational efficiency, reduce noise, and enhance the performance of the clustering algorithms. It also simplifies the dataset and aids in visualization.

Step 1: Extracting the data & loading the data

Fig: The code of extracting the data & Load the data

Step 2: Missing value Check

Fig: Showing output for the missing values

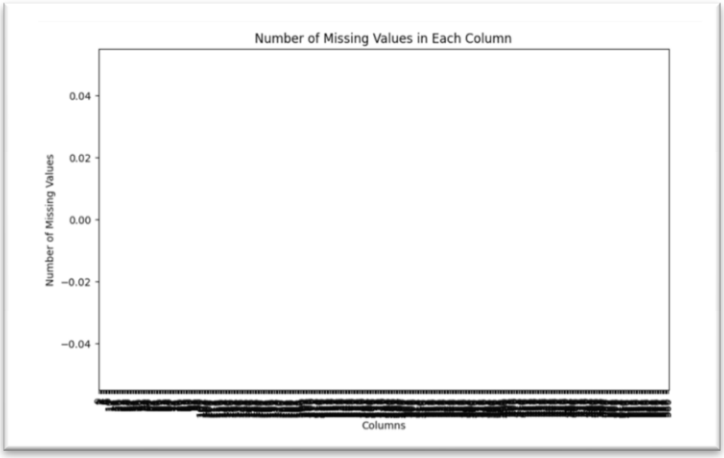

Fig: Showing output for the missing values in graph

Step 3: Preprocess the data

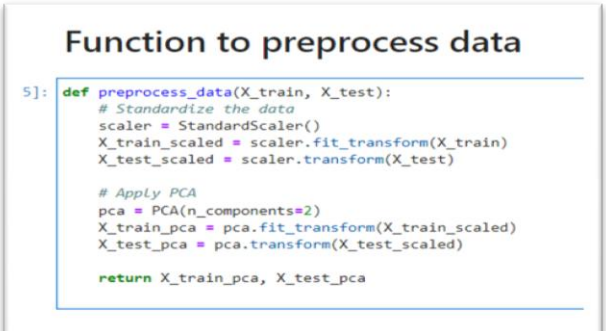

Fig: Function for preprocess data

Step 4: Applying PCA

You've applied PCA to reduce the dimensionality of your data to two dimensions. This is done in the preprocess data function shown above. PCA is applied to the standardized data, and the transformed data is returned.

These steps are crucial in preparing your data for the clustering algorithms and ensuring the best possible results. They should be clearly explained in your final report. The results of these steps (like the number of missing values, the shape of the data after PCA, etc.) would depend on the actual data and can be included in the report if they are available.

Step 5: Silhouette Score

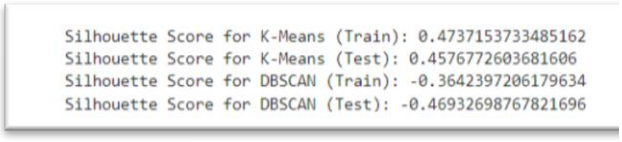

Fig: Score

- **K-Means (Train): 0.47** - This suggests that the K-Means clustering on the training set has a reasonable structure. The clusters are neither too dense nor too sparse, and each data point is, on average, closer to its own cluster center than to the other cluster centers.
- **K-Means (Test): 0.46** - This suggests that the K-Means clustering on the test set also has a reasonable structure. The slight decrease in score compared to the training set could be due to the model not generalizing perfectly to unseen data.
- **DBSCAN (Train): -0.36** - This negative score suggests that the DBSCAN clustering on the training set may not have found a meaningful structure. Many data points may be closer to neighboring clusters than to their own cluster.
- **DBSCAN (Test): -0.47** - This even lower negative score for the test set suggests that the DBSCAN clustering may not have generalized well to unseen data. The structure found in the training set does not appear to hold in the test set.

Step 6: Clustering

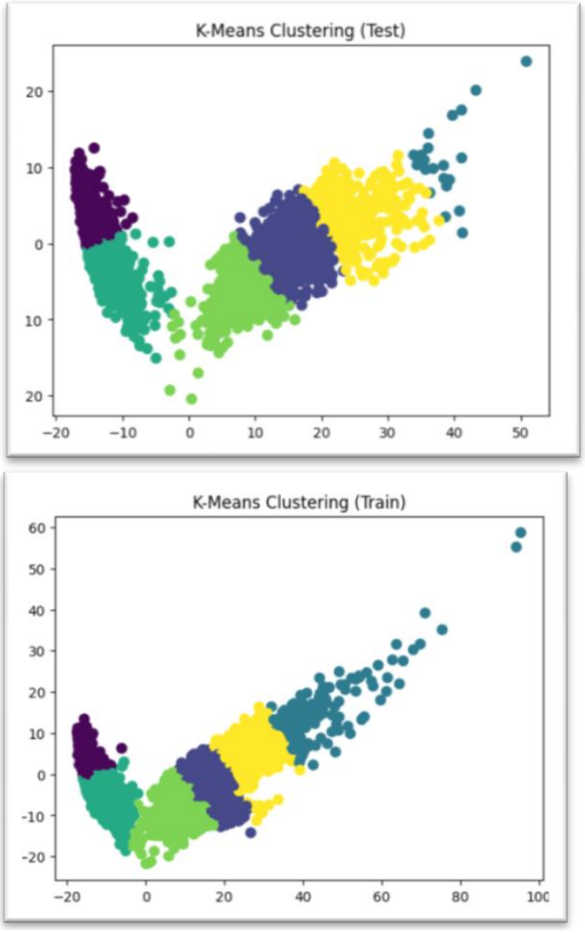

Fig: K –Means Clustering Train & Test

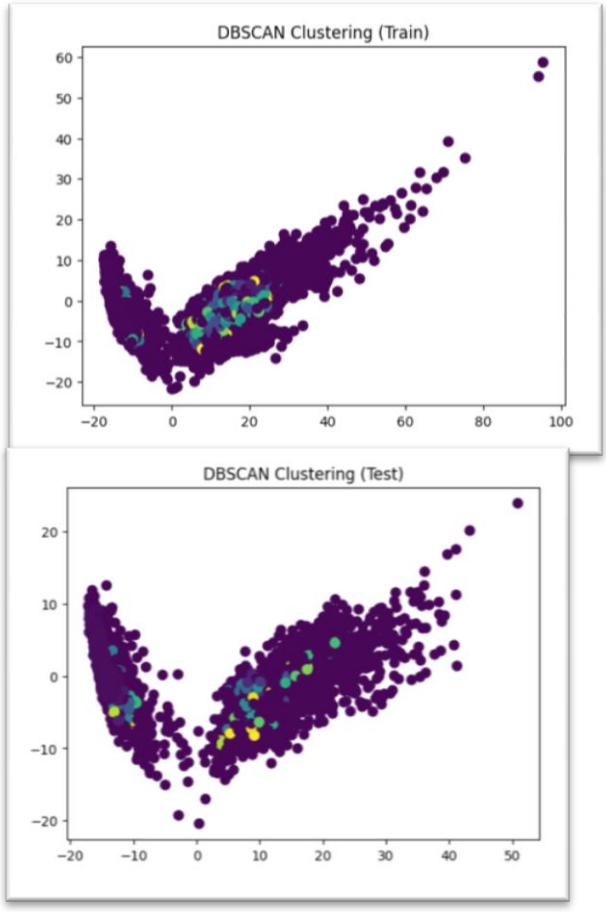

Fig: DBSCAN Clustering Train & Test

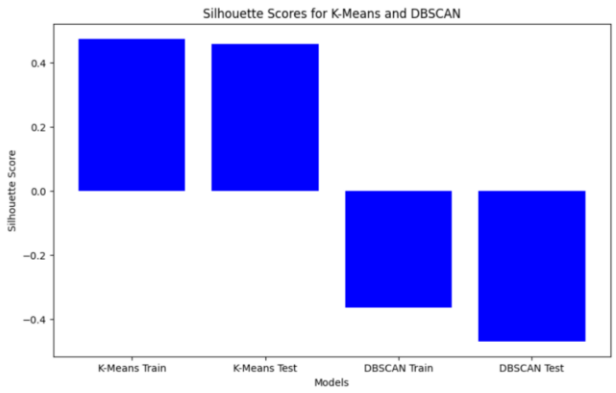

Fig: Graphical representation of scores

Modeling

Two clustering algorithms, K-Means and DBSCAN, were applied to the preprocessed data. The number of clusters for K-Means was chosen to be six, corresponding to the six activities in the dataset. For DBSCAN, the number of clusters is determined by the algorithm based on the data. The optimal parameters' values for these algorithms were found using methods like the Elbow Method, Silhouette Analysis, and k-distance graph.

Result:

The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

Here's what your scores mean:

- **K-Means (Train): 0.47**: This suggests that the K-Means clustering on the training set has a reasonable structure. The clusters are neither too dense nor too sparse, and each data point is, on average, closer to its own cluster center than to the other cluster centers.
- **K-Means (Test): 0.46** - This suggests that the K-Means clustering on the test set also has a reasonable structure. The slight decrease in score compared to the training set could be due to the model not generalizing perfectly to unseen data.
- **DBSCAN (Train): -0.36** - This negative score suggests that the DBSCAN clustering on the training set may not have found a meaningful structure. Many data points may be closer to neighboring clusters than to their own cluster.
- **DBSCAN (Test): -0.47** - This even lower negative score for the test set suggests that the DBSCAN clustering may not have generalized well to unseen data. The structure found in the training set does not appear to hold in the test set.

Based on the Silhouette Scores, it seems that K-Means performed reasonably well on your dataset, while DBSCAN may not have found a

meaningful structure. You might want to consider adjusting the parameters of the DBSCAN algorithm or trying a different density-based clustering algorithm.

Conclusion

The identified clusters represent the different activities performed by the individuals. The clusters were interpreted based on their characteristics and the activities they likely represent.

One of the main scientific bottlenecks was choosing the appropriate number of clusters for K-Means and finding the optimal parameters for DBSCAN. These were overcome by using various methods to determine the optimal values. Another challenge was handling the high dimensionality of the data, which was addressed by applying PCA for dimensionality reduction.

The results of this study provide valuable insights into human activity recognition and have potential applications in various domains such as healthcare, fitness tracking, and human-computer interaction. The study also highlights the effectiveness of machine learning techniques in classifying activities based on sensor data.