Machine Learning Mini Project 3: Clustering of Human Activity Recognition Using Smartphones Dataset

Sarosh Krishan, *EDISS Programme*, *Abo Akademi*, Turku, Finland, sarosh.krishan@abo.fi

I. Introduction

Human Activity Recognition (HAR) using smartphones is a crucial area in the realm of ubiquitous computing, enabling applications in various fields like healthcare, fitness tracking, and security. The core problem tackled in this study is the categorization of different human activities into clusters based on data from smartphone sensors, without the aid of explicit labels. The dataset comprises accelerometer and gyroscope readings from a Samsung Galaxy S II worn on the waist by thirty volunteers performing six types of activities. This report delves into unsupervised learning techniques, specifically K-Means and DBSCAN clustering algorithms, to discern inherent groupings within the data.

II. Data Processing

The raw dataset was first structured into a comprehensive dataframe by labeling columns using extracted feature names and merging training and test sets, acknowledging that unsupervised learning does not require split datasets. Key observations during exploratory data analysis (EDA) highlighted a natural division between active and passive activities, guiding the choice to retain most sensor-derived features while excluding subject identifiers and activity labels to prevent bias.

A. Exploratory Data Analysis (EDA) and Feature Selection

Firstly, the count of each type of activity (Standing, Sitting, Laying, Walking, Downstairs, Upstairs) is plotted for each user. This gives us an understanding of how the activity data is distributed across all users.

There are a few notable observations. Firstly for most users, walking upstairs and donwstairs data is less than their other activities. This also makes intuitive sense as most people spend less time walking on stairs than doing other activities. Subject 30 was an anomaly in this case since they spent most of their time either laying or walking.

Most participants spent a most of their time performing passive activities (Standing, sitting, laying) rather than active ones. With a notable exception of subject 1 who spent significantly more time walking than doing anything else.

These observations suggest that there is a significant division between active and passive activities, and that the clustering algorithms might be able to detect that division prominently. Additionally, besides subjects 1 and 30, no other anomalies were detected. Showing that the subject number isn't very relevant and should be removed in the clustering process.

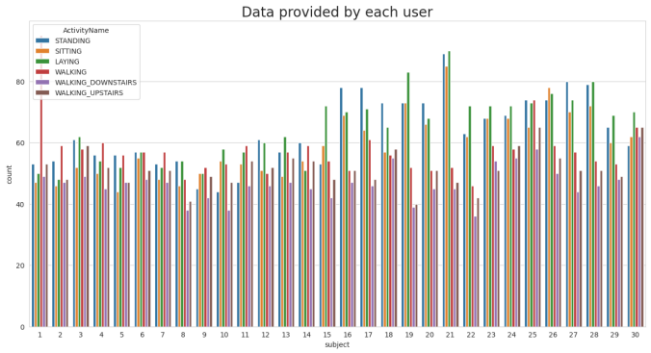

Fig. 1. Data Division Based on Subject.

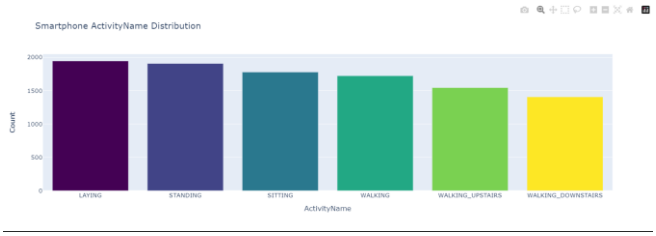

Fig. 2. Histogram Distribution of Activities.

The densities of the different activities were plotted against the mean magnitude of the body acceleration of the subjects. This value indicates the degree of movement in the subject's body. This value alone was able to distinguish clearly between the stationary and moving activities.

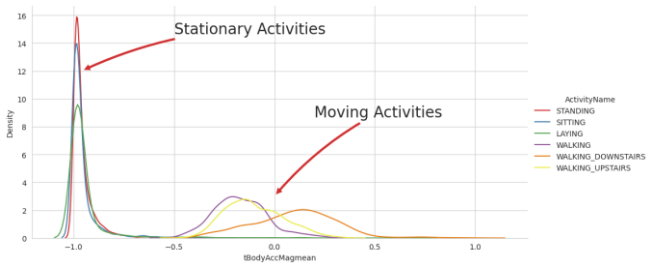

Fig. 3. Density Distribution of Activities based on Mean Acceleration.

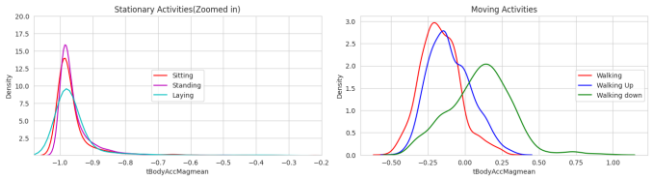

Fig. 4. Density Distribution of Activities based on Mean Acceleration separated into stationary and moving activities.

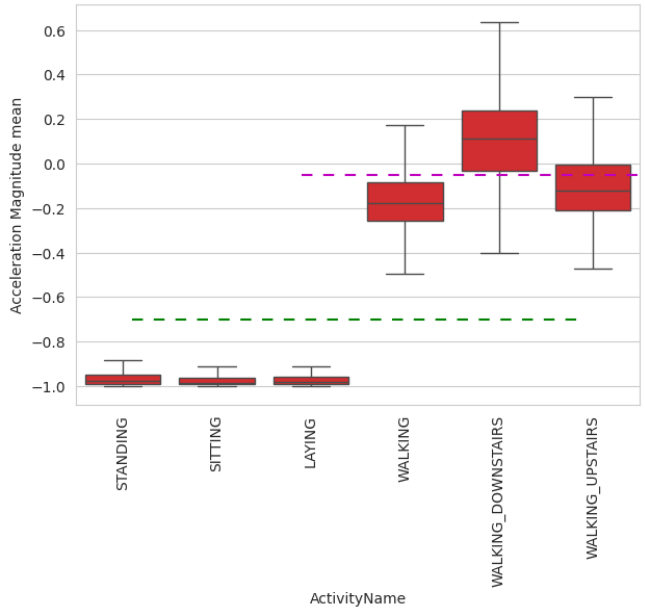

Fig. 5. Boxplot Distribution of Activities based on Mean Acceleration.

After the degree of movement, the variation in gravity and it's correlation with the different activities was tested. This was done by plotting the mean gravity angle in the x and y axis in a box plot. This shows that laying can be easily distinguished from the rest of the activities using these features.

All in all, the features measured by the sensors provide significant details about the activities being performed. Thus, only the subject name, activity type and activity name were removed from the dataset being used by the clustering algorithm. This was done to not perform any dimensionality reduction before task 1.

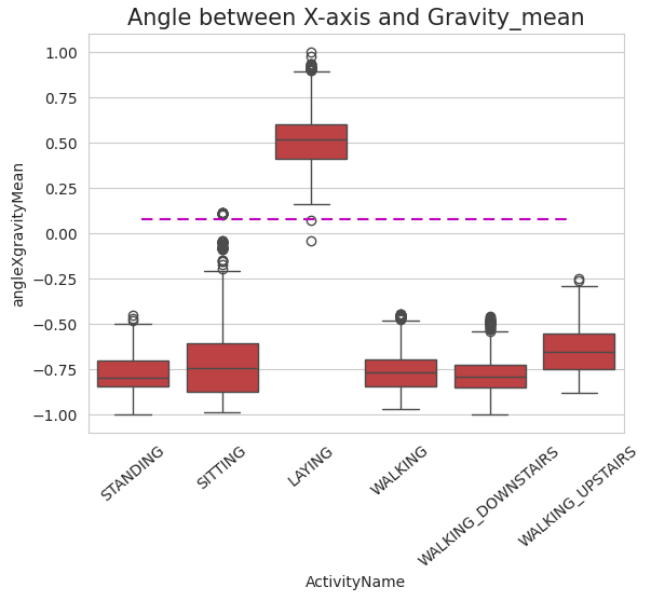

Fig. 6. Boxplot Distribution of Activities based on Angle X Gravity.

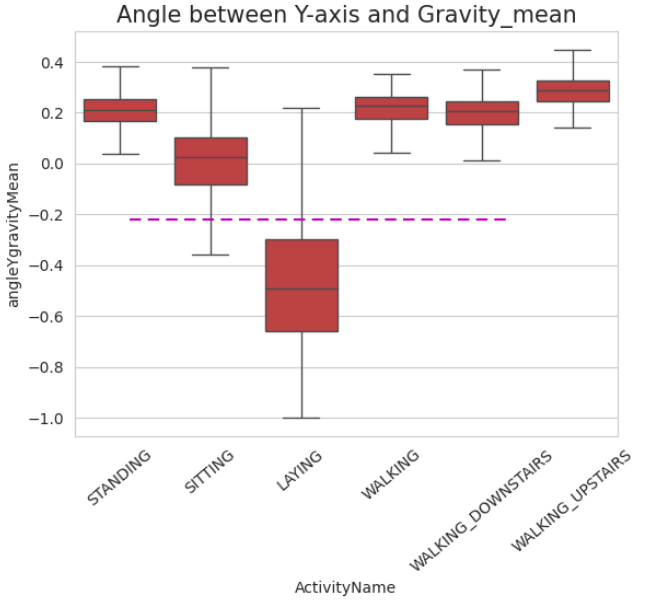

Fig. 7. Boxplot Distribution of Activities based on Angle Y Gravity.

Normalization checks confirmed that data values ranged between -1 and 1, indicating pre-scaled features, thus obviating additional normalization steps. Features were meticulously cleaned to remove extraneous characters, ensuring data quality and consistency.

III. Modelling

A. *Task 1*

K-Means Clustering Cluster Determination: Utilized the elbow method and silhouette scores across a range of 1 to 11 clusters. Optimal cluster number was identified as 2, resonating with the fundamental split between active and passive activities observed during EDA.

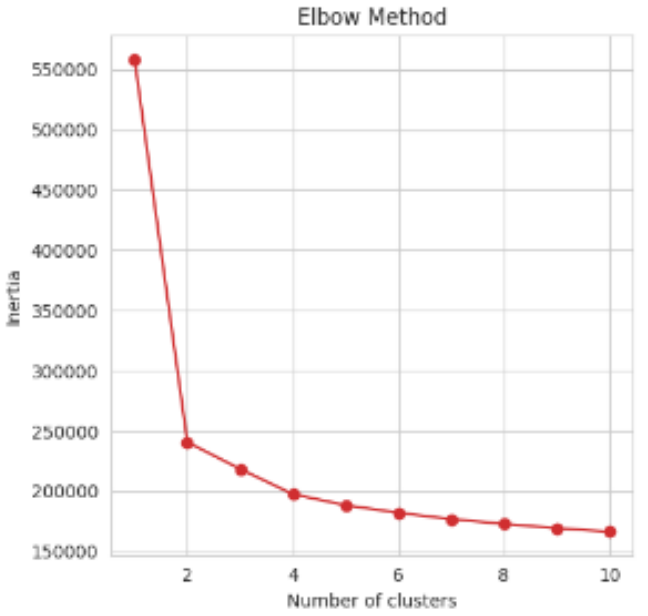

Fig. 8. Elbow Method K-Means.

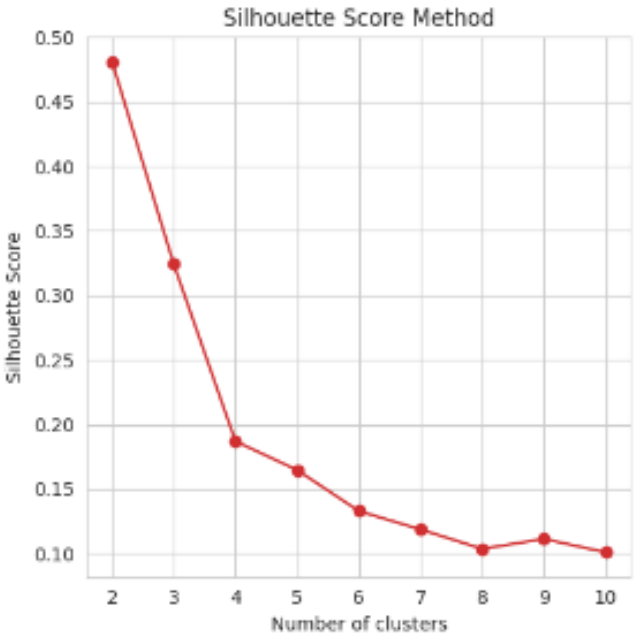

Fig. 9. Sillhouette Score K-Means.

DBSCAN Clustering Parameter Optimization: Determined the appropriate epsilon value through iterative experimentation, focusing on minimizing outliers (data points labeled as "-1"). An epsilon of 5.7585 emerged as optimal.

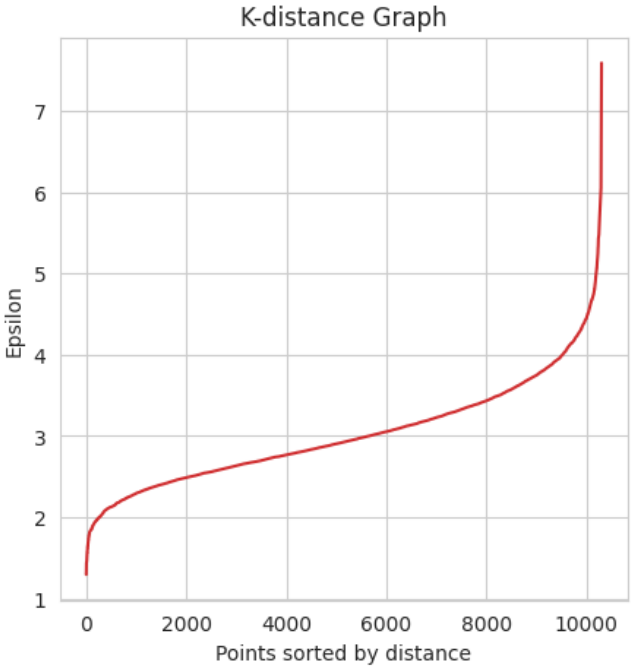

Fig. 10. Epsilon DBSCAN.

Visualizing the optimal K-Means and DBSCAN algorithms without dimensionality reduction using PCA and t-SNE and comparing them to the division of the original 6 activities, it can be clearly seen that the PCA visualization is more clumped together, whereas t-SNE shows distinct clumps of data. Moreover, stationary and moving activities can be easily divided into two hemispheres (left and right).

In the left hemisphere, sitting and standing are clumped together, with a better distinction in the t-SNE visualization. Laying is a separate group from the two, showing the potential for further clustering.

In the right hemisphere, Walking downstairs can be seen as a distinct cluster from the other moving activities. All these observations align with the predictions made in the EDA phase.

However, for the clustering algorithms, the curse of dimensionality limits their ability to perform clustering in such an efficient manner.

For K-Means, it is only capable of separating the stationary and moving activities with a sillhouette score of 0.48.

For DBSCAN, a significant proportion of the data is labelled as "-1", showing that it was unable to perform clustering on it. Besides the unlabelled data points, it was also only able to distinguish between stationary and moving activities. However, it performed worse than K-Means, getting a sillhouette score of 0.41.

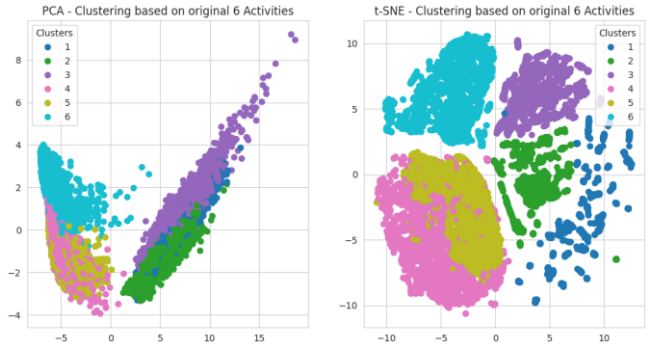

Fig. 11. Original 6 Activities Labelled on Dataset.

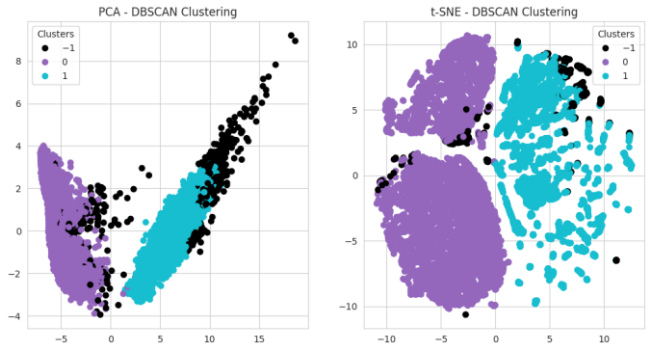

Fig. 12. DBSCAN Clustering Without Reduction.

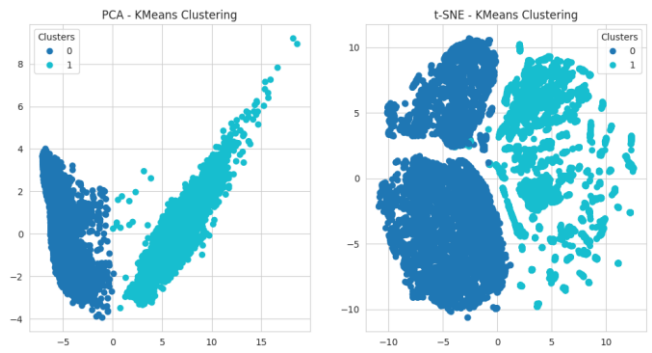

Fig. 13. K-Means Clustering Without Reduction.

B. *Task 2*

Technique Selection: PCA was chosen for dimensionality reduction to condense the feature space while retaining 83% of data variance, identified as a balance point ensuring minimal loss of information. The dataset dimensions were reduced from 561 to 15 principal components.

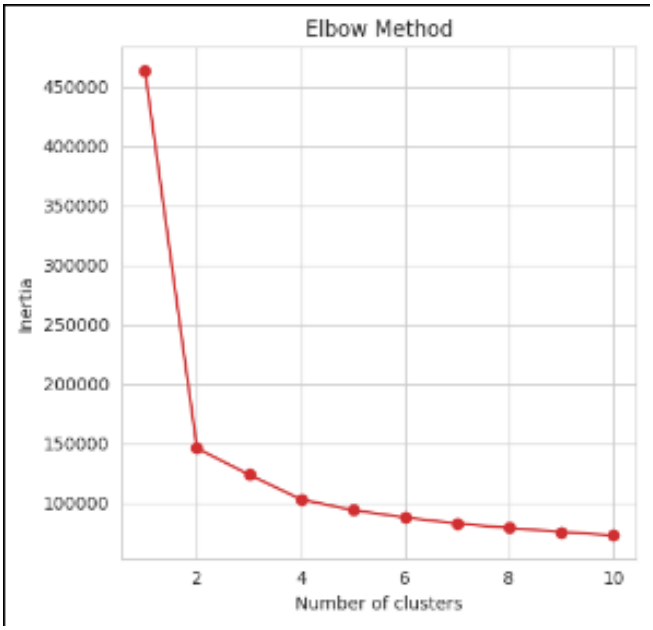

Fig. 14. Elbow Method K-Means.

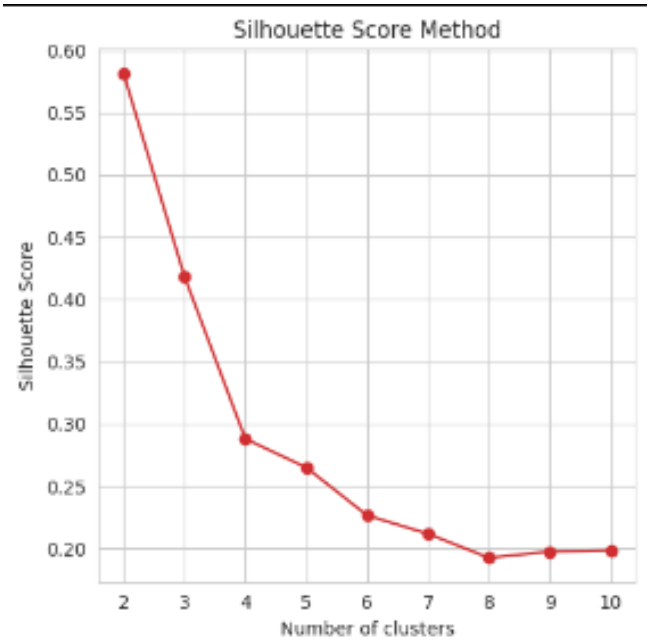

Fig. 15. Sillhouette Score K-Means.

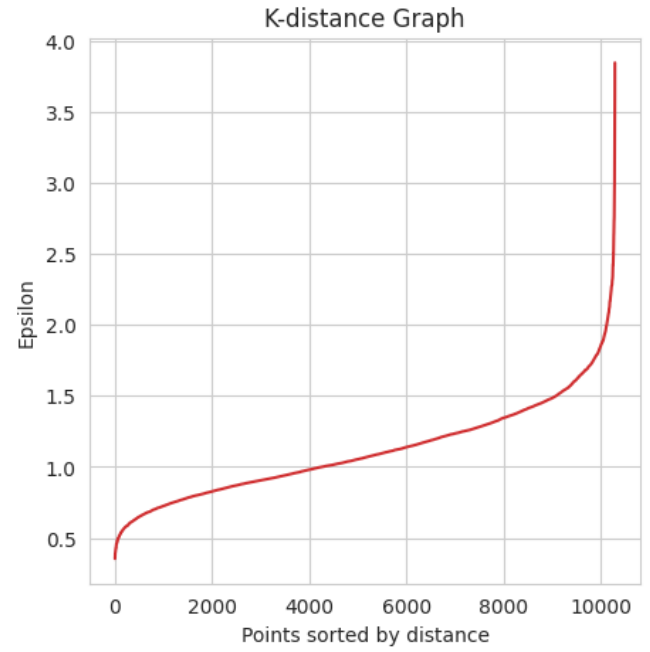

Fig. 16. Epsilon DBSCAN.

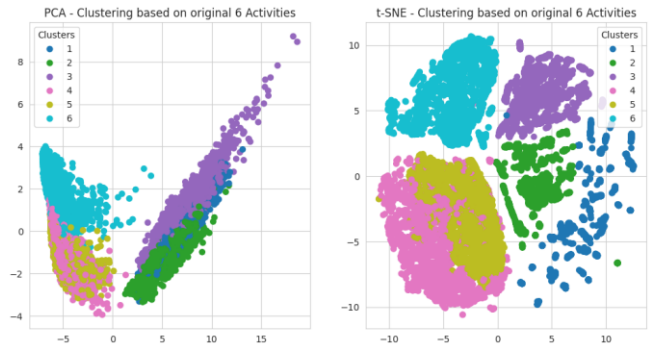

Fig. 17. Original 6 Activities Labelled on Dataset.

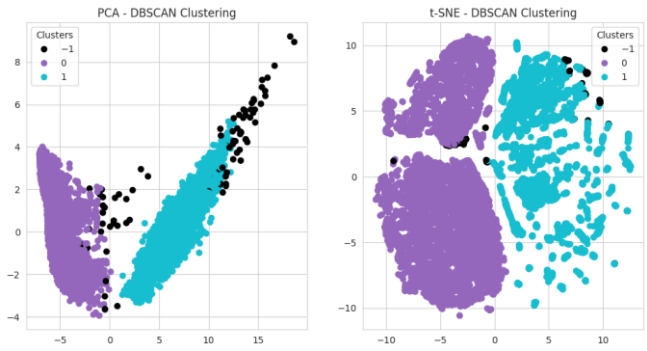

Fig. 18. DBSCAN Clustering With Reduction.

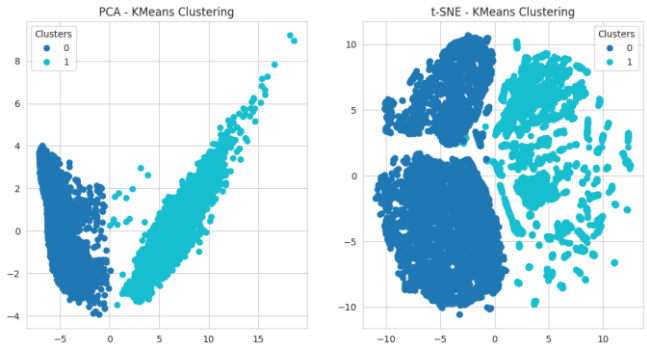

Fig. 19. K-Means Clustering With Reduction.

Impact Analysis: Post-reduction, DBSCAN exhibited improved performance with fewer outliers, enhancing the silhouette score. K-Means' performance remained stable, underscoring its robustness to dimensionality.

C. Task 3: Cluster Visualization

Visualization Techniques: Employed PCA and t-SNE for 2D reductions facilitating intuitive graphical interpretations. While PCA visualizations appeared more clustered, t-SNE provided clearer separations aligning with anticipated active and passive activity distinctions.

IV. Conclusion

The identified clusters principally delineated stationary versus mobile activities, echoing the intrinsic physical dichotomy in human motion. The activity clusters represent foundational behavioral patterns, crucial for applications needing activity inference without explicit labeling.

Scientific Insights: The division between stationary and mobile activities was the most salient feature, with finer distinctions (e.g., different types of walking) proving more elusive due to overlapping sensor signatures.

Bottlenecks and Solutions: High dimensionality posed significant challenges, particularly for DBSCAN, which initially struggled with outlier differentiation. Dimensionality reduction via PCA effectively mitigated this, improving clarity and reducing computational burden without substantially compromising the data's informative value.

In sum, this analysis underscores the potential of leveraging unsupervised learning for extracting meaningful insights from complex, high-dimensional datasets, particularly in the context of human activity recognition using ubiquitous sensing devices.