Applied Data Science 1

Clustring and Fitting

Submitted by: Bilal Ahmad #23035733

Link: Github Repository

Link: Dataset

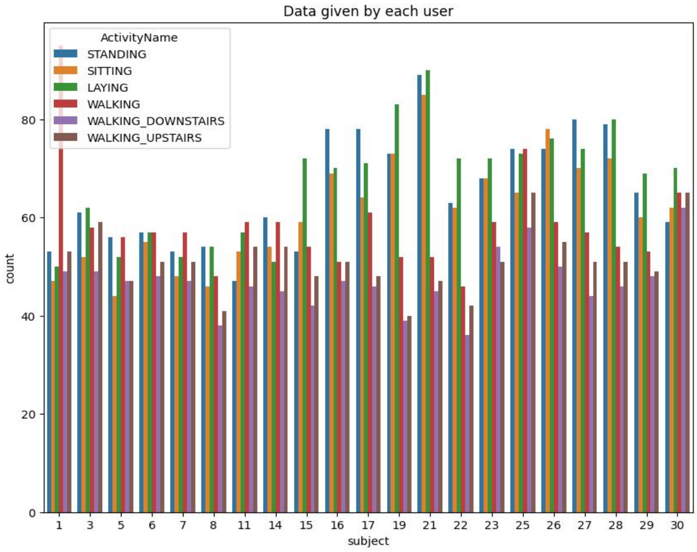
Introduction:

The problem we are trying to solve is to analyze and cluster human activity data collected from smartphone sensors. The dataset contains readings from accelerometers and gyroscopes for various activities such as walking, walking upstairs, walking downstairs, sitting, standing, and laying. The objective is to apply unsupervised machine learning techniques like clustering to group similar activities together based on patterns in the sensor data. In essence, we are aiming to discover underlying structures or patterns in the sensor data that differentiate between different human activities. By clustering similar activities together, we can gain insights into how sensor readings vary across different activities and potentially identify common characteristics or features that distinguish one activity from another.

Checking Dataset Balance:

- Purpose: To assess whether the dataset is balanced or imbalanced in terms of class distribution.

- Rationale: Imbalanced datasets, where one class is significantly more prevalent than others, can bias the results of clustering analysis. It's important to check the distribution of class labels or target variables to understand the underlying data distribution and potential biases. By identifying imbalances early in the process, appropriate sampling strategies or class weighting techniques can be applied to mitigate the impact of imbalanced data on clustering results.



Principal Component Analysis (PCA):

Purpose: PCA was applied to perform dimensionality reduction on the raw dataset.

Rationale: The raw dataset contained 561 features, which would make clustering computationally expensive and inefficient. By applying PCA, the dataset was transformed into a lower-dimensional space while retaining most of the information. This helped reduce the computational complexity of both the clustering algorithms while still capturing the maximum variance in the data. PCA transformed the 561 features into 151 principal components, which captured maximum variance in the data while reducing dimensionality significantly. This made the clustering algorithms more efficient.

Modeling

Two clustering algorithms were utilized for the analysis: K-Means and DBSCAN.

K-Means:

The K-Means algorithm was applied to the PCA-transformed training split as well as without PCA set with the variable number of clusters (K).

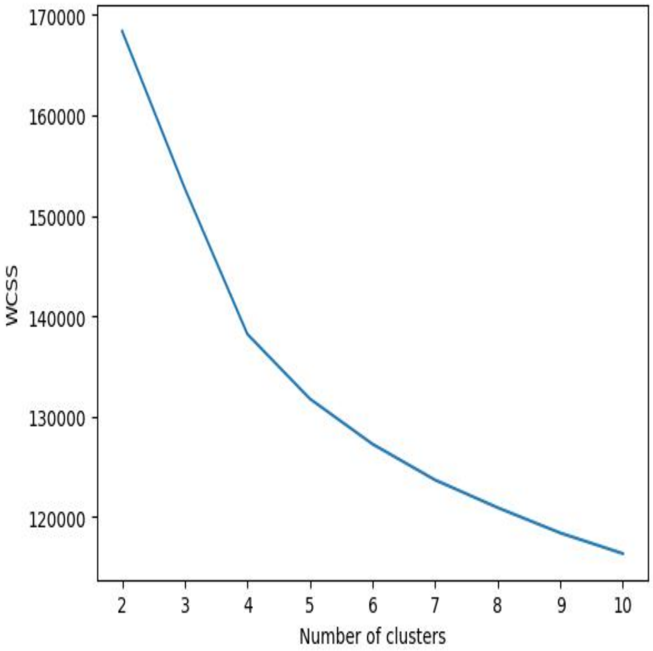
DBSCAN:

DBSCAN was run on the training split using mutiple epsilon values between 0.1 to 1.0 and min_samples parameter.

Fine-Tuning of K-Means:

The optimal number of clusters (K) for K-Means was determined using the elbow method:

The elbow point, indicating a significant decrease in within-cluster sum of squares, was observed at 4 clusters. This suggests that 4 is the ideal number of clusters for K-Means.

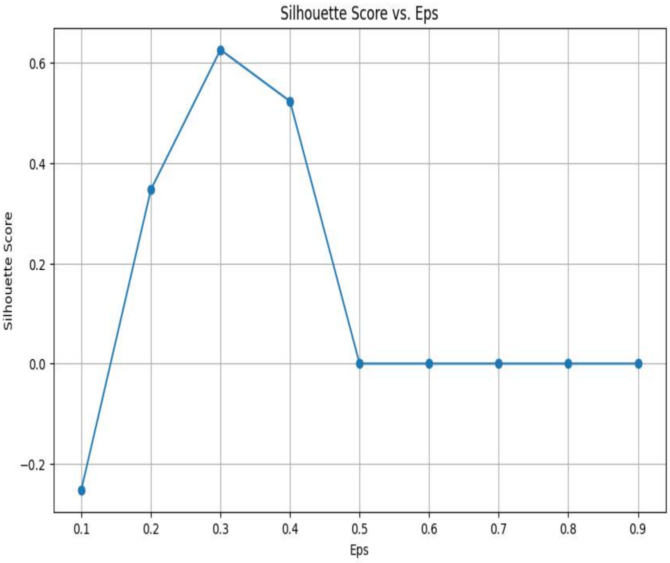


Fine-Tuning of DBSCAN:

The optimal epsilon value for DBSCAN was found through experimentation:

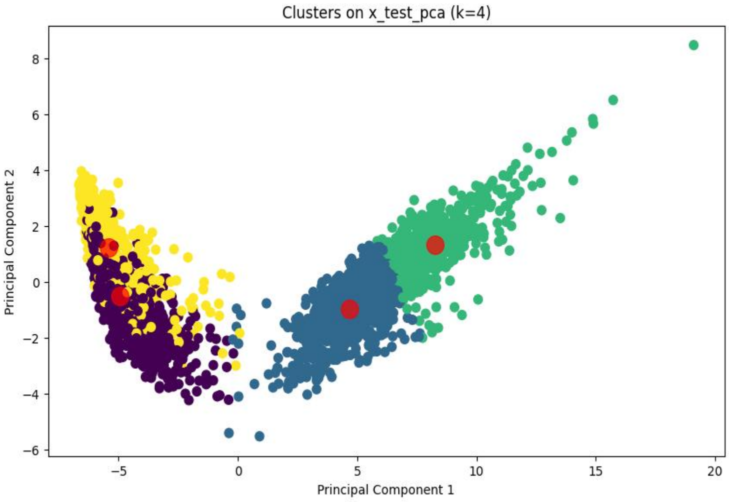
- Multiple values of epsilon were tested, and the silhouette statistic was calculated at each step.
- An elbow was observed in the silhouette statistic plot around an epsilon value of 0.3.
- The min_samples parameter was set to 5 based on experimentation.

Both models could identify the 6 true classes present in the dataset very accurately on the test split, with adjusted rand scores greater than 0.9.

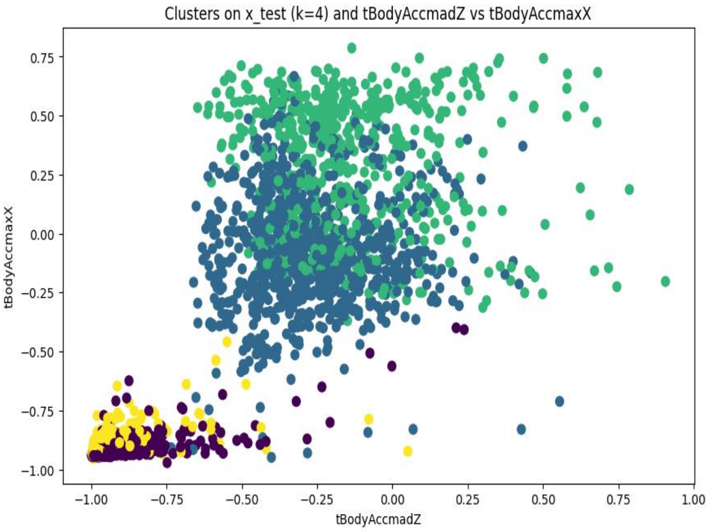


Cluster Visualizations:

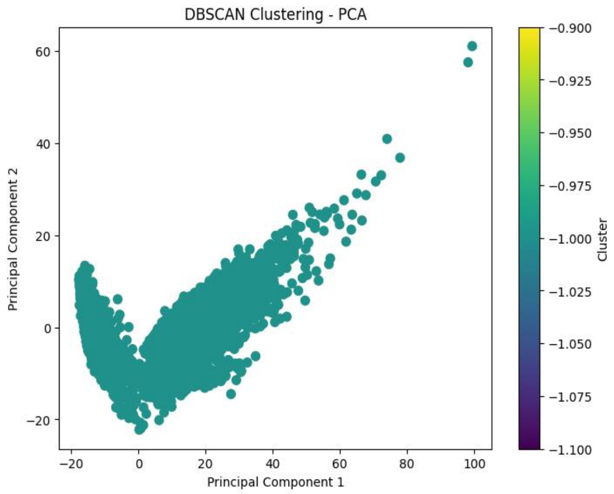
KMeans with PCA:



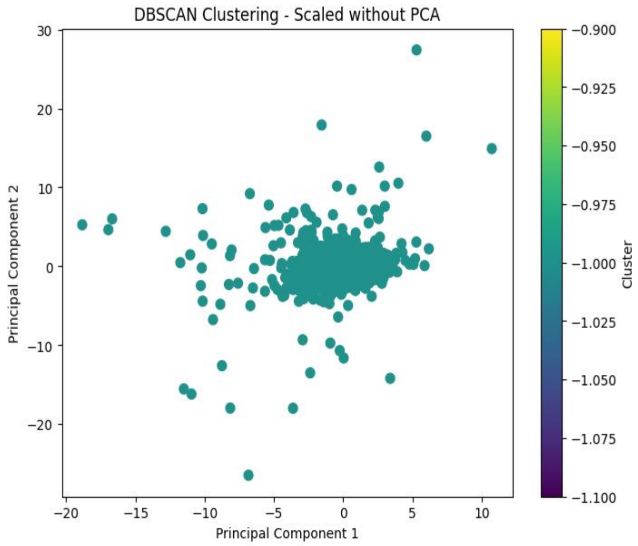
KMeans without PCA:



DB SCAN with PCA:



DB SCAN without PCA:



Dimensionality reduction and visualizations:

High Dimensionality and Impractical Visualization: With 561 features, visualizing pair plots or scatter plots in the original feature space became so impractical and computationally expensive. The sheer number of dimensions makes it challenging to represent the data in a comprehensible manner, so we converted it to only 2 clusters using the principal component.

Limited Human Perception and Overplotting: Human perception is limited, and it becomes increasingly difficult to interpret data in high-dimensional spaces. Additionally, without dimensionality reduction, plotting data points directly in their original feature space can lead to overplotting, where data points overlap and hinder each other, making it challenging to collect meaningful patterns or clusters.

Computational Complexity and Time-Consuming Process: Visualizing high-dimensional data directly requires significant computational resources and time. Generating pair plots or visualizations for 561 features would be computationally intensive and time-consuming, hindering the efficiency of the analysis.

Conclusion:

The clustering analysis of human activity data from smartphone sensors aimed to group similar activities together based on patterns in accelerometer and gyroscope readings. The identified clusters represent distinct patterns of sensor data associated with different human activities, such as walking, walking upstairs, walking downstairs, sitting, standing, and laying. Each cluster captured the unique sensor signatures corresponding to specific activities, allowing for the differentiation and classification of these activities based on their sensor data patterns.