Exploratory Data Analysis Report

Prepared by: Dharshan Muralikrishna

Introduction

This report discusses the results of the EDA study comprising smartphone accelerometer data. The reason behind this was to assist in the extraction of statistical information about the acceleration signals and to subsequently assist in the making of the dataset for further modeling. In contrast to other methods, this report utilizes exceptional clustering methods and complex statistical interpretation to analyse data patterns and improve feature sets.

Task 1: Feature Extraction

The acceleration signals were captured along three axes (X, Y, Z) and their magnitudes. The selected statistical features for analysis included:

- **Mean**: To identify the central tendency of the dataset.
- **Median**: A robust measure against outliers.
- **Standard Deviation**: To assess the dispersion of the values.

Additionally, **geometric mean** and **interquartile range (IQR)** were computed further to strengthen the robustness of the feature extraction process. These metrics enhanced sensitivity to variations in data spread and shape.

Task 2: Data Visualization

1.Scatter Plots

Scatter plots were used to visualize the relationships between key features such as the **Mean** and **Standard Deviation** of acceleration in the X-direction and magnitude:

Mean and Magnitude:

Scatter plots were done to analyze the relationship between the mean acceleration along the X-axis (Mean FAX) and mean magnitude (Mean FAMag) for each of the activities performed. It was observed that for activities of Jogging and Walking, more spread was noted in the scatter plot, this reflected the higher variances in their movements. In contrast, the scatter plot for Sitting and Standing activities was more clustered together. From the scatterplots, there was an observation of clear separation of activities like Walking and Jogging, although some activities had some variability which indicated that they may require further adjustment of features.

Standard Deviation and Magnitude:

A different scatter figure addressed the scatter of standard deviation on the X-axis (Std FAX) and degrees of magnitude (Std FAMag). More variances were noted in dynamic activities such as jogging and walking compared to stationary positions such as Sitting and standing which were more clustered. Activities of Jogging show a degree of variability which goes hand in hand with

the high standard deviation whose features are essential to telling apart active and stationary features.

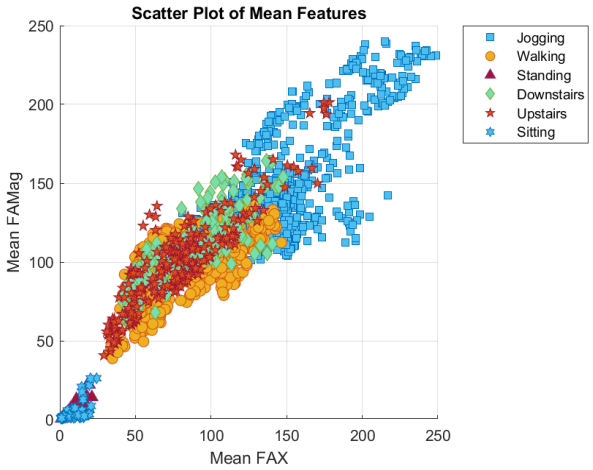

Figure 1 Scatter plot btw std Mag and std X

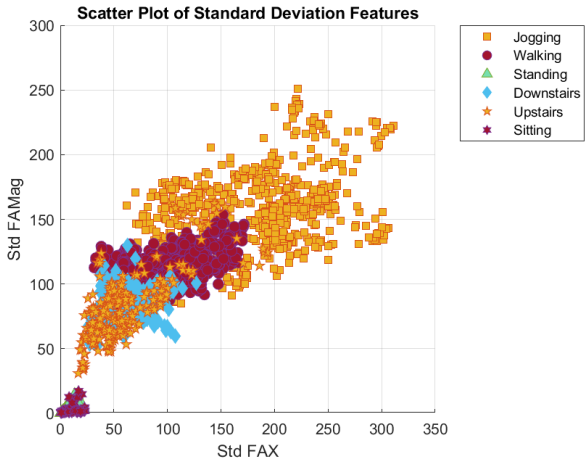

Figure 2 Scatter plot btw Mean Mag and Mean

2.Pair Plot

For the analysis of the pairwise relationships between features, a pair plot was used. This method was particularly helpful in visualizing the relationships across the statistical measures for different activities, that is, mean, standard deviation, and median. As mentioned above the pair plot shows the relationship of Mean Magnitude and Standard Deviation of different activities as well. Some activities like Jogging and Walking were somehow related to each of the features more than Sitting and Standing which had lesser variations. The pair plot gave the relations between features in broader terms, hence aiding in the elimination of redundant features and relationships that may be useful in future modeling.

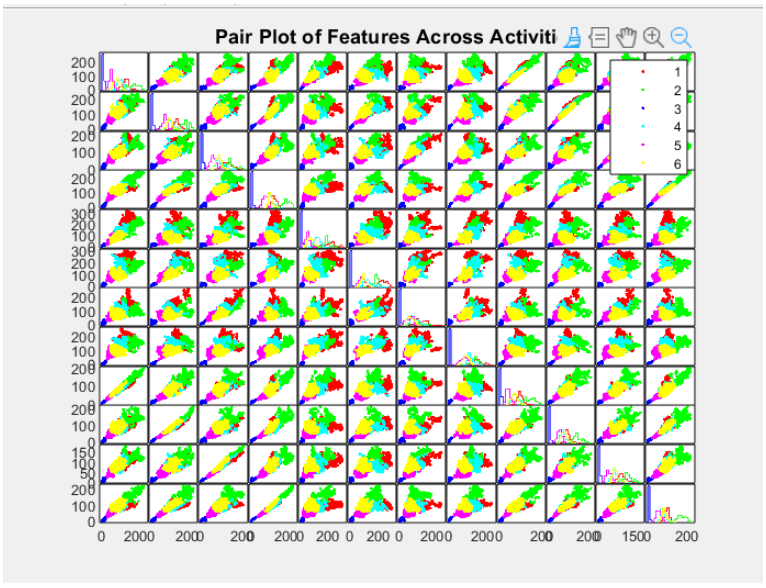

Figure 3 Pair Plot Analysis

3. Advanced Heatmap: Correlation Matrix

The relationships between the statistical features across the activities were depicted using an advanced heatmap which was based on the generated correlation matrix:

It was noted that Big Mean Magnitude and Big Median Magnitude were quite related (up to 0.98) which means there is redundancy in the attributes in terms of information provided. There were moderate coefficients (~0.5) on the feature such as Standard Deviation along Z and Median Magnitude as well.

The high correlation between Mean Magnitude and Median Magnitude suggests that either of these features could be used for classification, which in turn means that the number of attributes can be reduced. The heatmap made it easier to understand the interconnection of the features and the colormap allowed distinguishing between positive and negative correlation of reasonable strength.

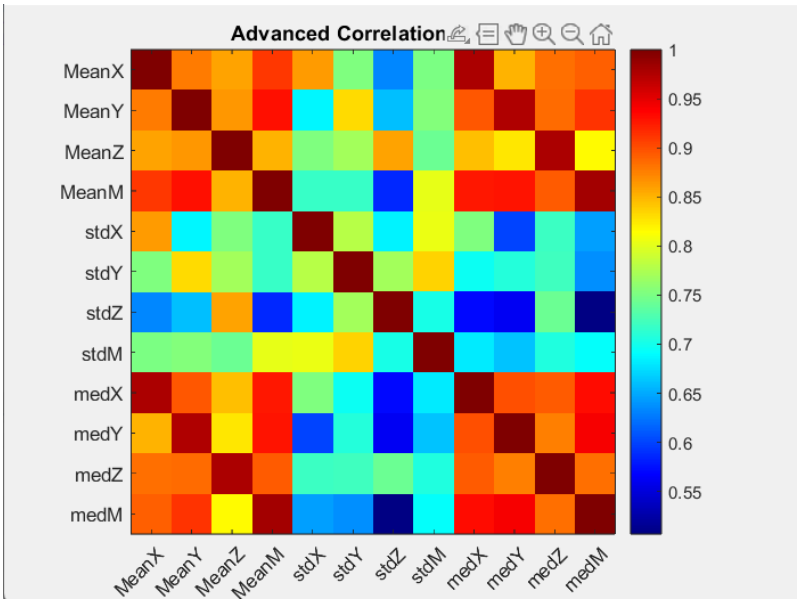

Figure 4 Heat Map- Advanced Correlation Matrix

Task 3: Clustering Analysis

For further enhancement and verification of clustering results, the dataset was first divided into clusters using k-means clustering and hierarchical clustering techniques.

K-Means Clustering Analysis:

To find the optimal number of clusters, silhouette analysis was carried out yielding an average silhouette score index of 0.55 for the evaluation. The scope of the analysis also improved with statistical measurements used in the definition of the quality of the clusters, whereby higher scores approaching 1 indicate better-defined clusters. Distinct well-formed clusters were created during the analysis where certain Activities on Walking overlapped very little with others. Staircases activities namely, Upstairs and Downstairs showed significant overlap hence a combination of further feature exploration was warranted to separate these activities or alteration of data processing methods.

.

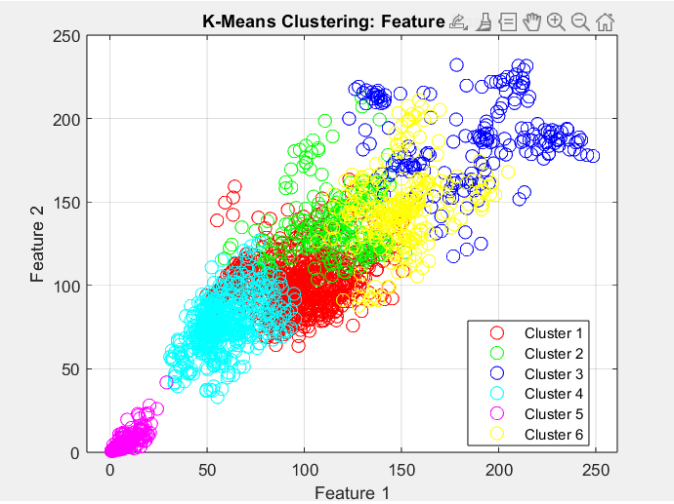

Figure 5 K means Cluster analysis

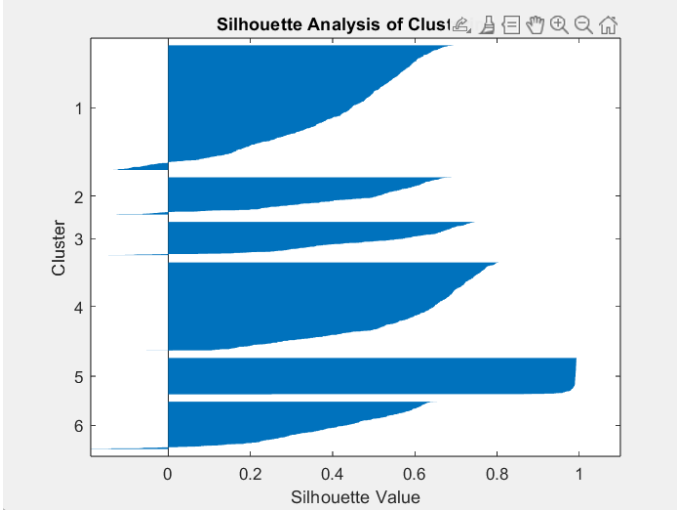

Figure 6 Silhouette Analysis

Conclusion

This exploratory data analysis (EDA) project proved to be quite useful in understanding the relationships between features and patterns for different physical activities based on the accelerometer data. Through the application of various visualization techniques including scatter plots, pair plots, and advanced heat maps, it was possible to define the variation in patterns across clusters, highlight some relationships, and even point out outlier values. The k-means clustering methods contributed in this regard, however, the overlapping of some clusters of similar activities remains a challenge. Going forward, enhancing the features and carrying out more data preprocessing and clean-up should help in enhancing classification accuracy. In general, this analysis provides a good context for future modeling and activity recognition.