Human Activity Recognition using Smart Phones

Introduction:

To understand different human activities, experiments have been conducted on 30 volunteers to measure 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz through the accelerometer and gyroscope embedded in the smartphone attached to their waist. Based on the features captured, six different activities were identified (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) and manually labelled. The dataset has already been divided into 70% training and 30% testing data. The aim of this project is to perform an unsupervised clustering algorithm, DBSCAN on the dataset to identify all the six activities of humans. Since it is an unsupervised algorithm, all the training and testing have been combined to perform the clustering. The project also involves applying dimensional reduction techniques and parameter tuning methods.

Data Processing:

All the features have already been normalized between the range of (-1,1). In this section, the duplicate values and null values in the features were checked.

1. ***Duplicate values:***

   It is found that there are more than one feature that has the same column name. So the values of these columns are checked to see if these columns are duplicate.

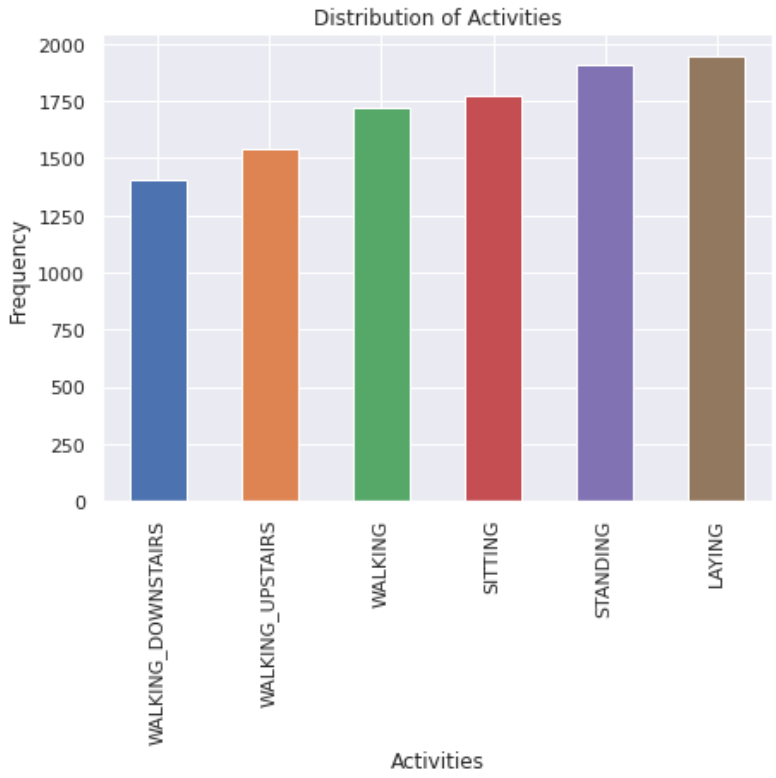
```
Number of features:  561
Number of unique features name:  477
```

	fBodyAccJerk-bandsEnergy()-41,48	fBodyAccJerk-bandsEnergy()-41,48	fBodyAccJerk-bandsEnergy()-41,48
0	-0.999640	-0.999729	-0.999814
1	-0.999814	-0.999685	-0.999769
2	-0.999906	-0.999627	-0.999626
3	-0.999930	-0.999846	-0.999735
4	-0.999929	-0.999769	-0.999688

   Though the feature names are the same, these columns are not duplicate

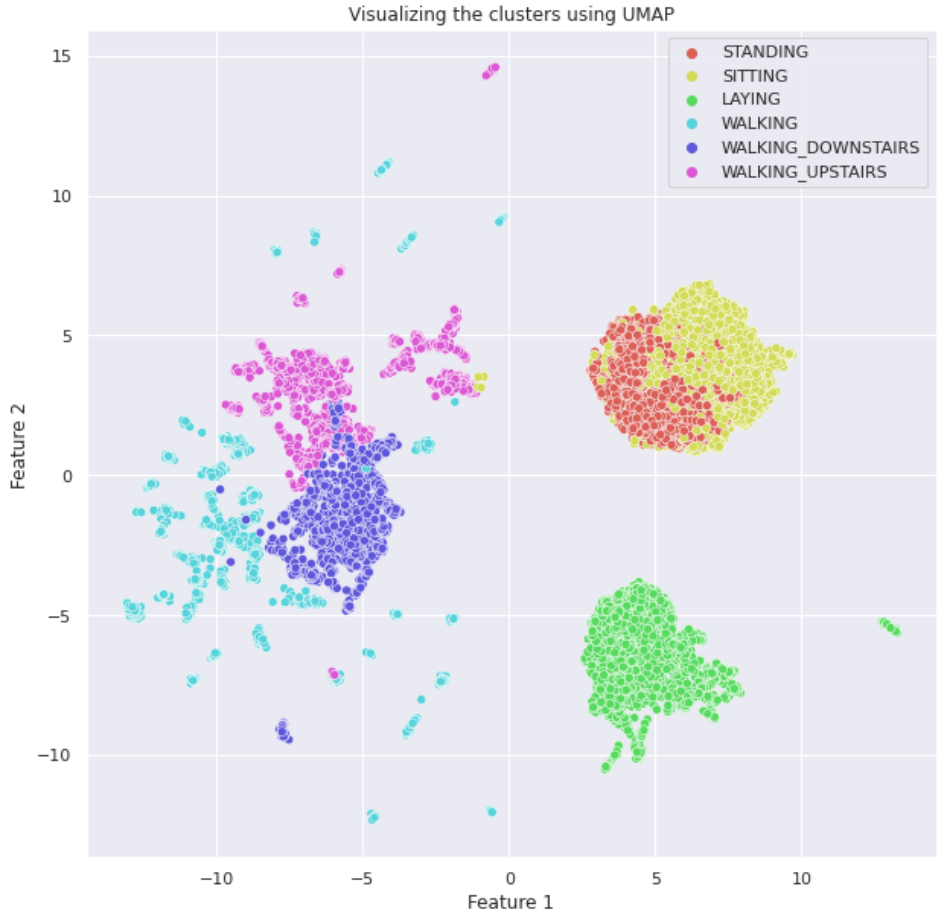
2. ***Null values:*** No null values are found in the features

Looking at the distribution of different activities



Laying and standing are the most performed activities.

Since there are many features in the data, it becomes hard to visualize all the columns. Therefore a dimensional reduction technique, umap has been applied to the data to visualize all the clusters.



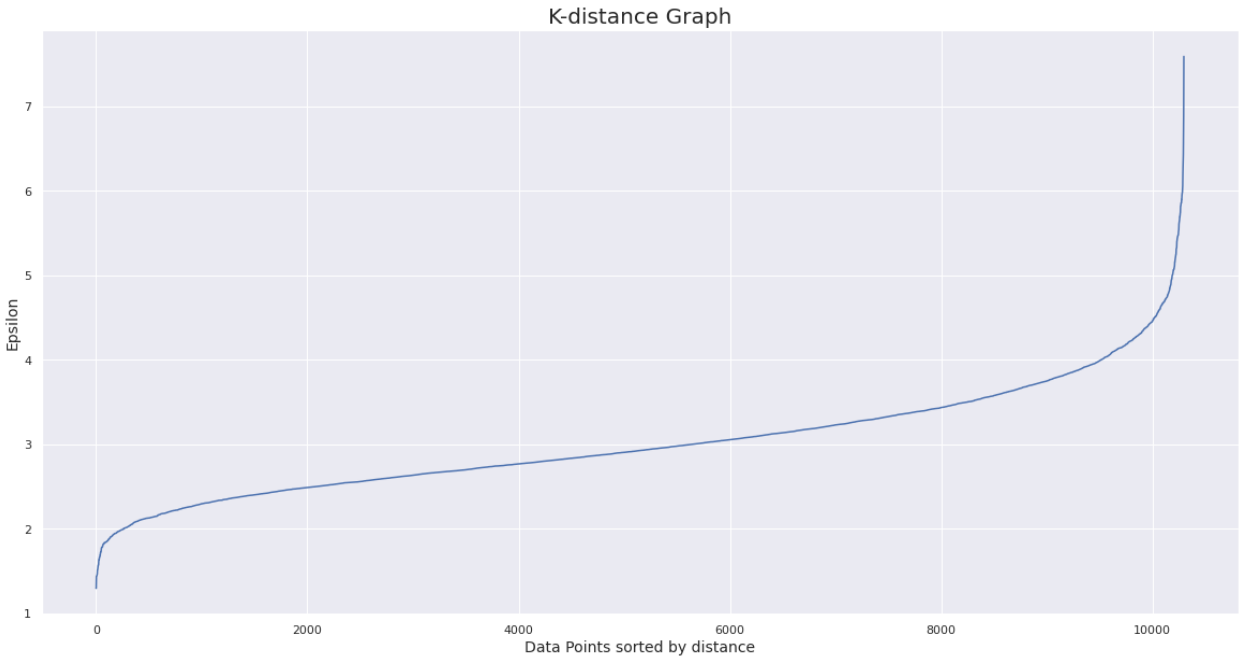
Modelling:

1. ***Clustering without Dimensional reduction***

   There are two main parameters used in DBSCAN:

   - **Epsilon (eps):** Represents the greatest possible distance between two samples for one to be considered in the neighbourhood of the other.
   - **Min_samples:** Reflects the number of samples (or total weight) in a neighbourhood for a point to be designated a core point.

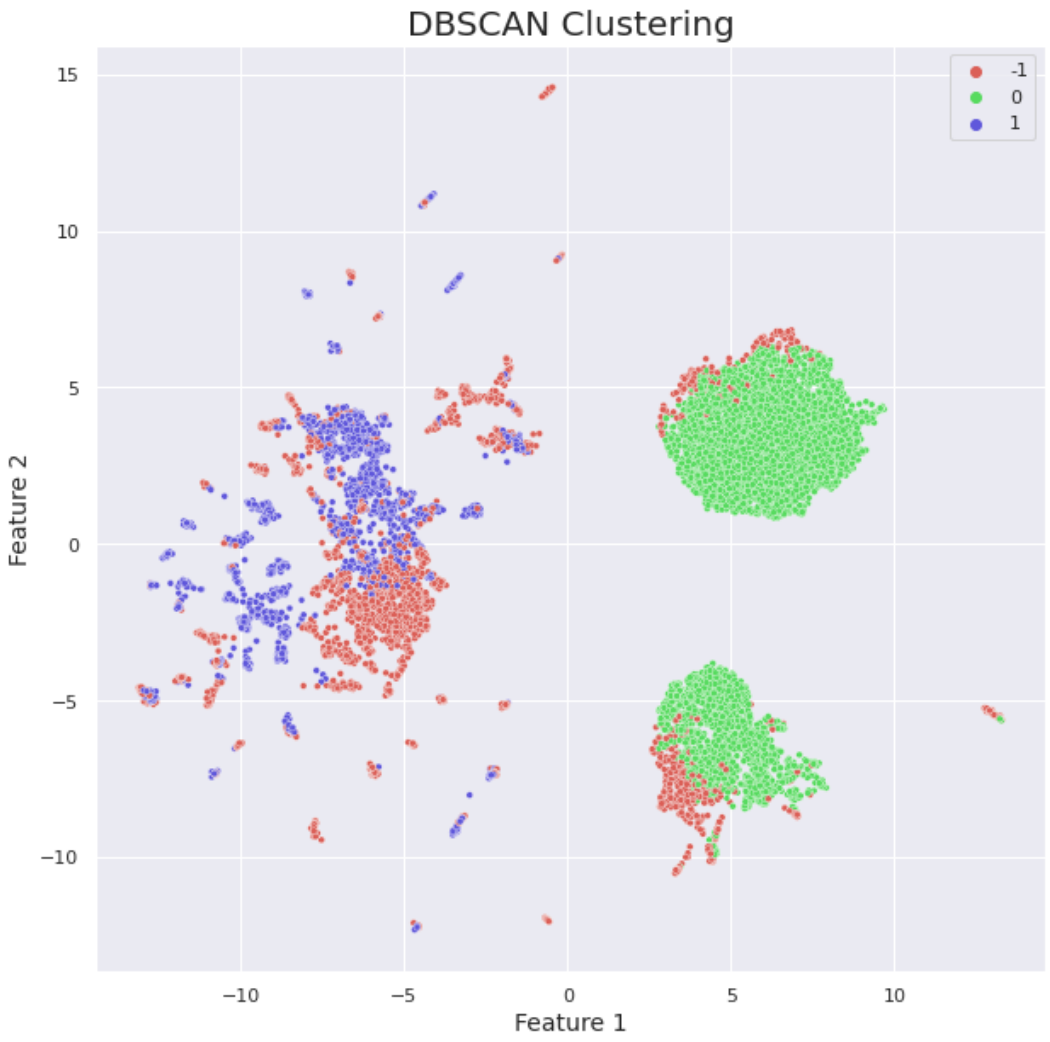
   To find the best epsilon value, a k-distance graph is plotted that visualizes the distance to the k = min_samples-1 nearest neighbour in descending order from biggest to smallest value. The best 'ε' value is found at the elbow.



   The best epsilon value, in this case, is found at: **4.9**

   The rule for choosing the min_sample is that min_samples>= number of dimensions +1
   Min_samples chosen are: 1000

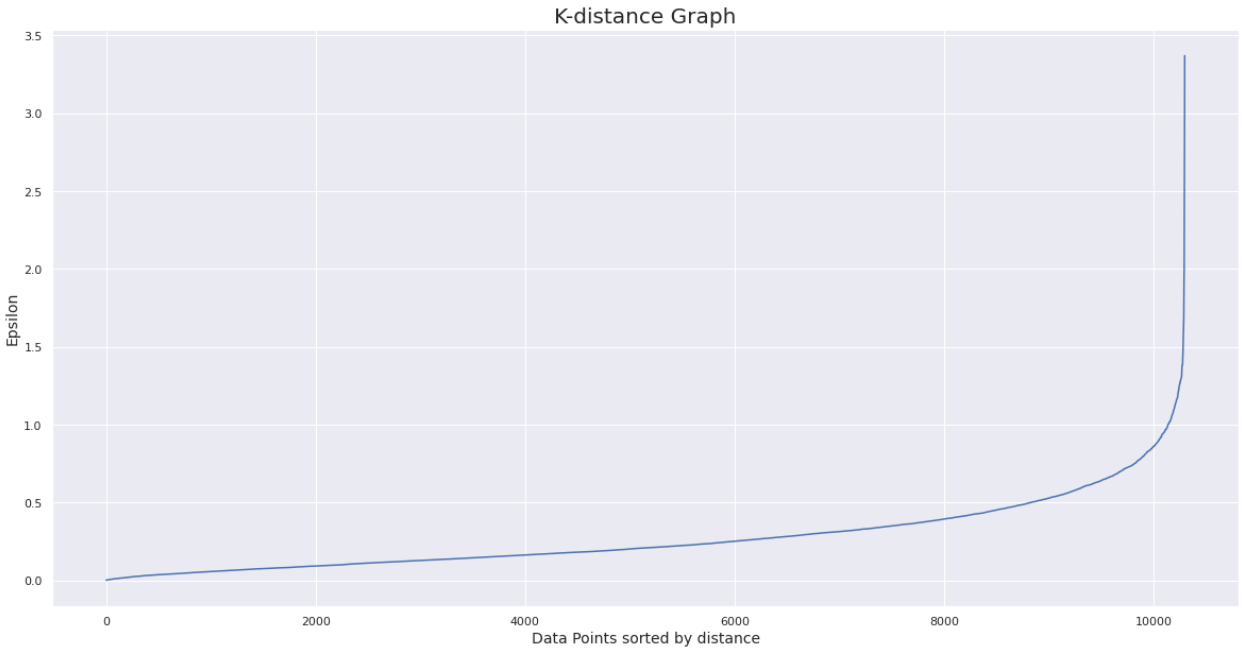
   After fitting the tuned parameters and all the features in DBSCAN, the cluster labels formed by the model are applied to the umap reduced data, to visualize the cluster formed by DBSCAN.



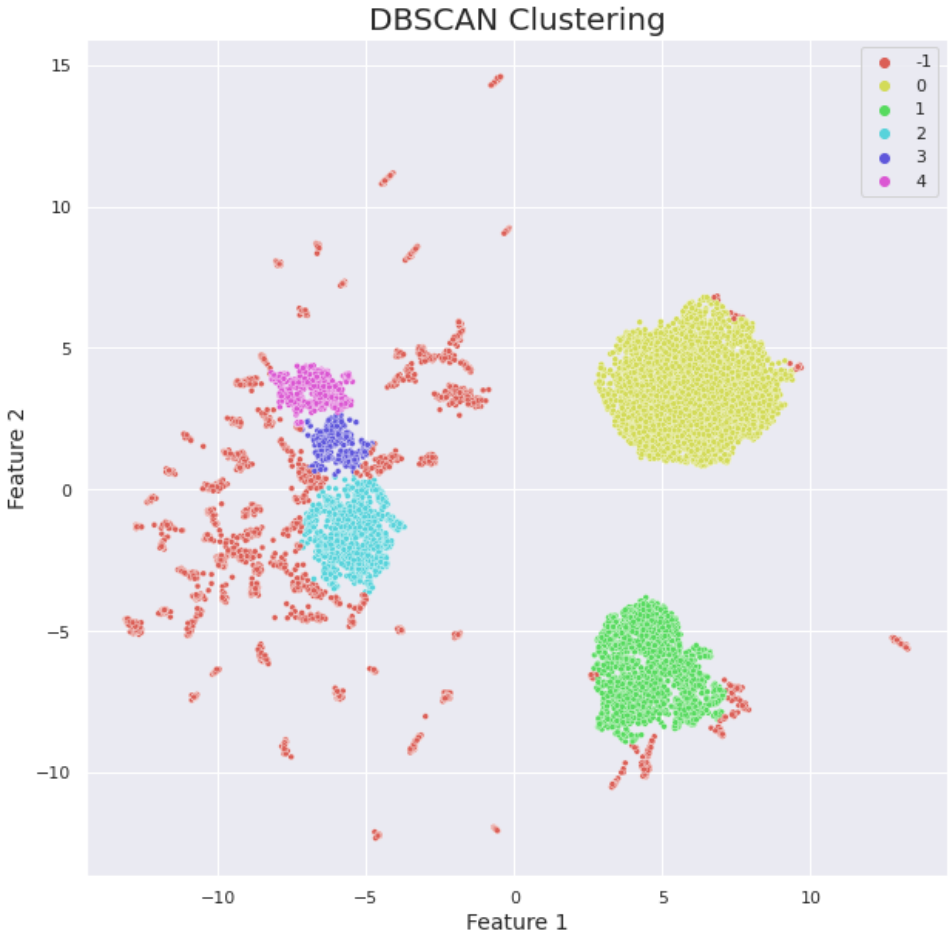
The number of clusters formed in this case: 3

2. ***Clustering with Dimensional reduction***

To select the best epsilon value, a k-distance graph has been plotted with the reduced features



The best epsilon value is found at: 0.08 and min_samples: 179



The number of clusters formed is: 6

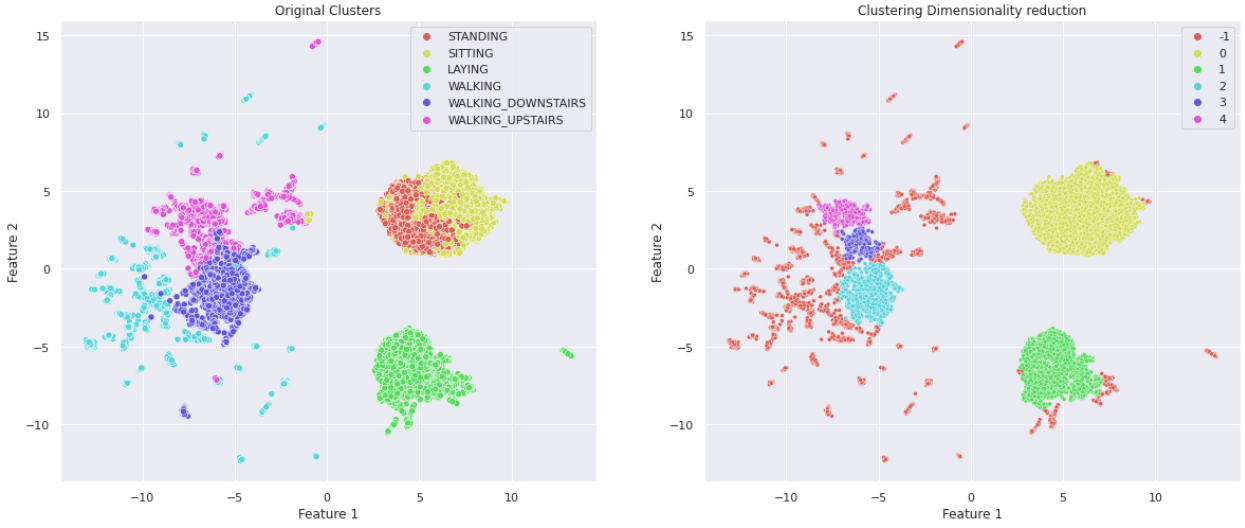
Comparing the two models:

1. ***In terms of clustering:***

The model without dimensionality reduction identified 3 clusters only.

The above figure compares the original clusters to the clusters of the model without dimensionality reduction. The model clustered standing, laying and sitting as one cluster. The model was able to distinguish waking downstairs and walking upstairs properly but couldn't identify walking as a separate cluster.

The model with dimensionality reduction identified 6 clusters.

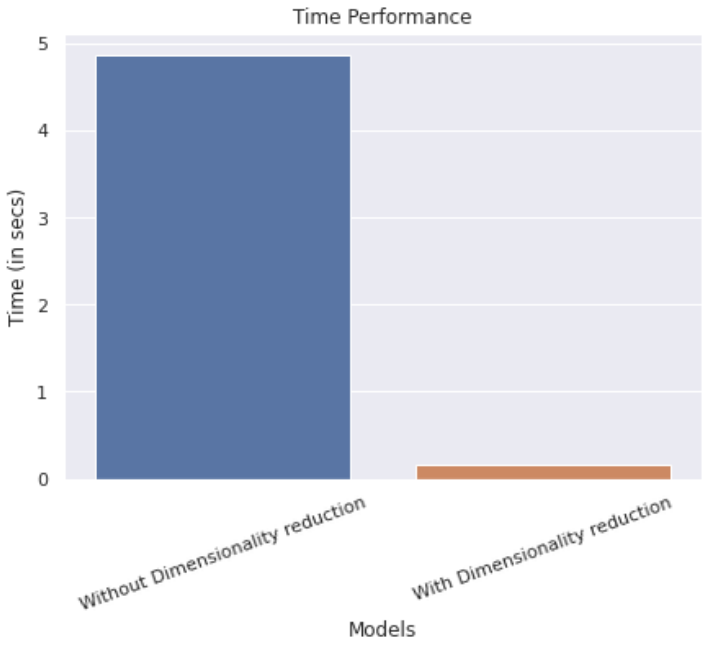


The above figure compares the original clusters to the clusters of the model with dimensionality reduction. Unlike the previous model, this one was able to identify (sitting, standing) and laying as different clusters. But couldn't identify standing in the sitting cluster. It perfectly-identified walking, walking downstairs but further identified an extra cluster in walking upstairs.

Overall when both the models are compared in terms of clustering, the model with reduced dimension features performs the best.

2. ***Time performance***

Time taken for clustering without dimensionality reduction: 4.87 seconds
Time taken for clustering with dimensionality reduction: 0.15 seconds



From the above graph, it is clear that in terms of time computation the model with reduced dimensions performs better. It is almost 3 times better than the other model

Comparing the clusters as well as the time computation, the model with reduced features performs the best.

Bottlenecks and challenges:

- Even though the k-distance graph helps choose the best epsilon value, in this project, the elbow value in the graph as epsilon was not good enough. Since the dataset was large, epsilon with a larger value seems to work better.
- When visualizing the actual clusters, it is observed that sitting and standing clusters are quite overlapped, which becomes difficult for the model to cluster them separately.

Conclusion:

Overall, the DBSCAN algorithm was able to cluster different human activities to a satisfactory extent with fewer outliers. The dimensionality reduction helped not only reduce the computational time but also formed better clusters. With good parameter tuning, the supervised algorithm was able to cluster the data points well.