Clustering Models on Human Activity Data

Main Objective

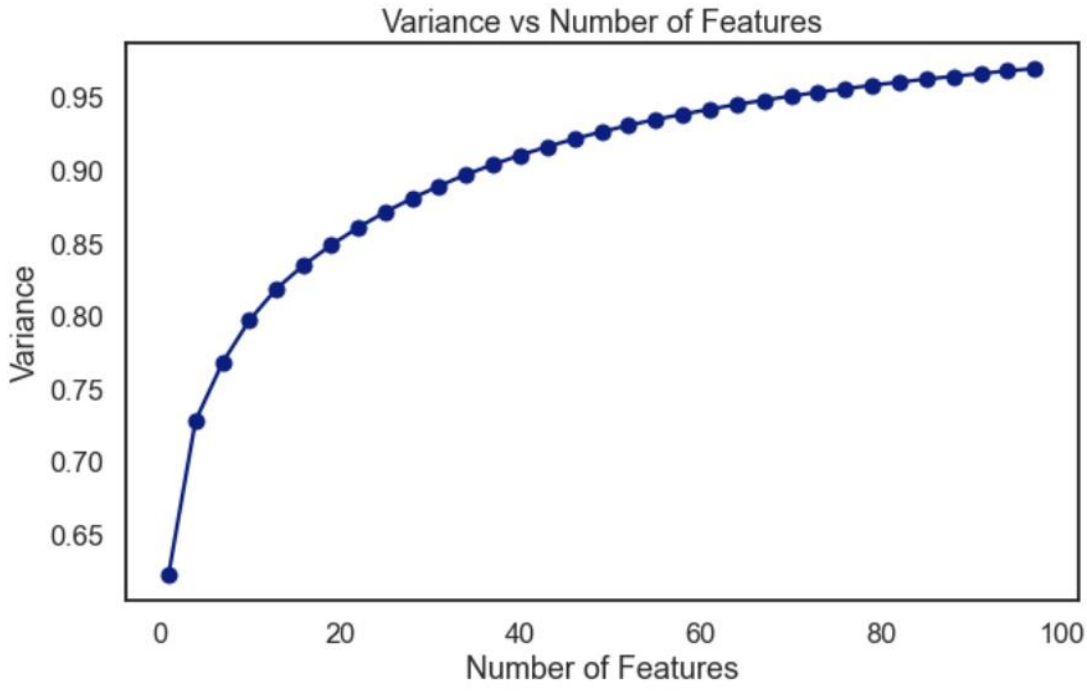
- In this report, we will analyze the human activity dataset collected from their smartphone activity
- The objective of this report is to find out the best clusters in which the various observations can be clubbed and to compare them with the actual human activity target column that is also present in the dataset
- We will use a variety of clustering models and figure out the one that is best suited to our needs.

Brief Description of the Dataset

- This dataset has 561 features and one target column.
- Since there are too many target features for us to list down all of them, we'll just give a broad idea of the features.
- The features describe, in detail, the physical state of a human through the data collected via their smartphones.
- The data describes statistical properties such as mean, standard deviation, minimum, maximum, etc., of various measurements like their acceleration along the three dimensions, the jerks associated with these accelerations, the gyroscopic angles as recorded by the smartphones, etc.
- All the features are of float type.
- The target column is categorical and takes 6 different values. They are 'Standing', 'Sitting', 'Laying', 'Walking', 'Walking Downstairs' and 'Walking Upstairs'.
- There are 10299 observations in this dataset.
- In our clustering models, we will drop the target column, and then we will compare our clusters with the target column.

Data Cleaning and Feature Engineering

- There are no empty cells in our dataset, so we move straight to looking at the outliers.
- We log transform the features with a high skew value.
- Subsequently, we define our outliers as those lying outside 1.5 times the inter-quartile range.
- In doing so, we find that we lose more than 90% of the rows. This is because the outliers in the various columns are non-overlapping.
- To solve this, we first reduce the number of dimensions using Principal Components Analysis (PCA). This will also modify the features so that they are linear combinations of one another, and so the outliers will be overlapping.
- Before using PCA, we scale our data using MinMax scaler which ensures that all values are between 0 and 1.
- We now perform a PCA using a variable number of features to find out the explained variance. The result is shown in the plot below



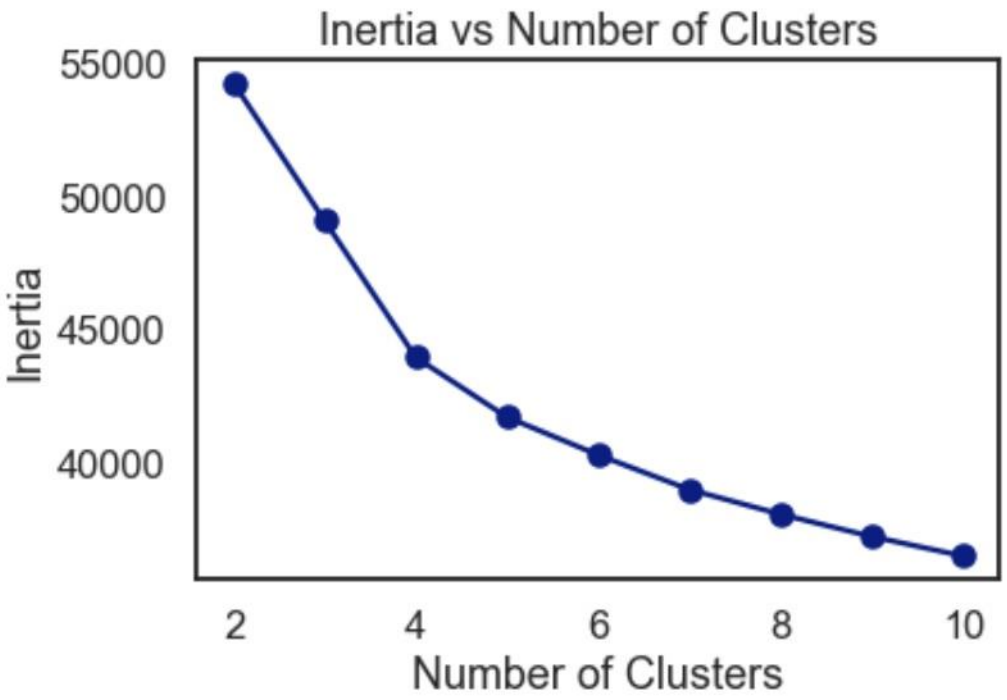
- We find that the explained variance approaches 95% for 70 features. We set the number of features to 100 for which the explained variance is just above 97%.
- Once again we look at the skew values and find that only one feature has a skew value greater than 0.75.
- Thus, we now remove the outliers that lie outside 1.5 times the inter-quartile range.

- This removes 200 outliers from our dataset and we have 10,099 rows remaining.
- Since the skew values are small and the mean of each of the columns has is approximately zero after PCA, we don't need to scale our data anymore before moving onto clustering models.
- Finally, we convert the target column, with which we shall compare our results into numerical values using the label encoder method.

Clustering Models

In this section, we will use various clustering algorithms to get information about the structure present in our dataset.

1. We start with the K-Means algorithm. The inertia of the models as a function of the number of clusters is shown in the plot below:



This graph suggests that there is a mild elbow present at n=4, although it is not a very prominent one. Moreover, we naively expected there to be 6 clusters because the target variable has 6 unique values. Thus, we will use a variety of n values for the number of clusters and see what information we can extract from these different values.

We first set n = 4 and look at the clusters formed and how they correspond to the activities in the target column in the table below.

Activity	Activity_pred_km4	number
LAYING	0	20
	3	1897
SITTING	0	5
	1	1688
	3	60
STANDING	0	6
	1	1896
WALKING	0	1074
	2	608
WALKING_DOWNSTAIRS	0	350
	2	983
WALKING_UPSTAIRS	0	1291
	2	221

This table shows the following relations between the clusters and the activity type (ignoring the outliers in the table):

i. Cluster 0 – Divided between walking, walking upstairs, and walking downstairs.
ii. Cluster 1 – Divided between sitting and standing.
iii. Cluster 2 - Divided between walking, walking upstairs, and walking downstairs (similar to Cluster 0).
iv. Cluster 3 - Laying

This shows us that only one cluster properly maps onto a corresponding activity while the other clusters include a combination of the various activities.

2. Next, we set n= 6 for K Means algorithm. This is an obvious choice because there are 6 unique values in the target column. In this case, we get the following table that shows the relationship between the clusters and activities in the target column.

Activity	Activity_pred_km6	number
LAYING	1	1752
	2	154
	3	11
SITTING	1	49
	2	479
	3	3
	5	1222
STANDING	2	692
	5	1210
WALKING	0	948
	3	547
	4	187
WALKING_DOWNSTAIRS	0	620
	3	162
	4	551
WALKING_UPSTAIRS	0	299
	3	1126
	4	87

This table shows the following relations between the clusters and the activity type (ignoring the outliers in the table):

i. Cluster 0 – Divided between walking, walking upstairs, and walking downstairs.
ii. Cluster 1 – Laying
iii. Cluster 2 – Divided between sitting and standing
iv. Cluster 3 – Divided between walking, walking upstairs, and walking downstairs.
v. Cluster 4 – Divided between walking, walking upstairs, and walking downstairs.
vi. Cluster 5 – Divided between sitting and standing.

This shows us that only one cluster properly maps onto a corresponding activity, namely cluster 1 maps to laying. Cluster 0, 3 and 4 are divided between walking, walking upstairs and walking downstairs. Cluster 2 and 5 are divided between sitting and standing.

3. We now use Hierarchical Agglomerative Clustering. We start by setting number of clusters equal to 4 to compare our results with those obtained from K Means algorithm with 4 clusters. We get the following table that shows the relationship between the clusters and the activities in the target column

		number
Activity	Activity_pred_ag_4	
LAYING	0	1917
SITTING	0	106
	1	1
	2	1646
STANDING	0	18
	2	1884
WALKING	1	1390
	3	292
WALKING_DOWNSTAIRS	1	108
	3	1225
WALKING_UPSTAIRS	1	1395
	3	117

This table shows the following relations between the clusters and the activity type (ignoring the outliers in the table):

i. Cluster 0 – Laying
ii. Cluster 1 – Divided between walking and walking upstairs.
iii. Cluster 2 – Divided between sitting and standing
iv. Cluster 3 – Divided between walking downstairs (primarily), walking and walking upstairs.

This shows us that only one cluster properly maps onto a corresponding activity, namely cluster 0 maps to laying. The other clusters are once again divided between various activities.

4. We now use Hierarchical Agglomerative Clustering. We use ward linkage and set the number of clusters to 6 to compare it o the result we got for K Means. We get the following table that shows the relationship between the clusters and activities in the target column.

   This table shows the following relations between the clusters and the activity type (ignoring the outliers in the table):

		number
Activity	Activity_pred_ag	
LAYING	2	1590
	4	327
SITTING	0	1
	2	9
	3	857
	4	97
	5	789
STANDING	3	1010
	4	18
	5	874
WALKING	0	1390
	1	292
WALKING_DOWNSTAIRS	0	108
	1	1225
WALKING_UPSTAIRS	0	1395
	1	117

i. Cluster 0 – Divided between walking and walking upstairs.
ii. Cluster 1 – Divided between walking downstairs (primarily) and walking and walking upstairs.
iii. Cluster 2 – Laying
iv. Cluster 3 – Divided between sitting and standing
v. Cluster 4 – Divided between laying (primarily) and sitting
vi. Cluster 5 – Divided between sitting and standing

Once again we find that laying directly maps onto a unique cluster while other clusters are a combination of various activities.

Best Model

We tried four models and each of them gave us different results. Overall, the two K Means models mixed the clusters with the activities more than the two Hierarchical Agglomerative Clustering models did. Specifically, the agglomerative model with 6 clusters reveals interesting details about the dataset, so we will use that as our final model and derive insights into the dataset from this model.

Key Findings

- We conclusively find that laying contains data that is very different from the other 5 activities.
- While there is some sub-structure even in laying, it does not share much in common with any another activity.
- Sitting and standing are divided between clusters and this shows that the data for these two activities has some commonalities.
- Walking downstairs can be approximated with a unique cluster with a few outliers that represent either walking or walking upstairs.
- Walking and walking upstairs share a lot of structure in common with one another.
- Some of these findings are rather intuitive. Laying is a unique activity compared to all the other 5 and thus the data for it forms a unique cluster.
- Sitting and standing are both static activities and the data for these two activities form overlapping clusters.
- Walking downstairs, while being somewhat similar to walking and walking upstairs, is significantly different to be treated as its own unique cluster.
- Walking and walking upwards are rather similar activities and their clusters are highly overlapping.

Next Steps

While we tried a variety of clustering scenarios on our dataset to find the most informative one, there are multiple other ways in which we could have approached this problem. We could have used the DBSCAN or the Mean shift algorithm to find clusters in our dataset. Moreover, in our data

engineering step, we could have allowed for different number of remaining features in our dataset post PCA and compare the results as a function of the number of clusters.

Since the models found significant overlapping between certain activities, it appears that we needed more data to properly cluster our dataset. This might be because the original dataset had a rather large number of features (561) and the number of data required to form effective models rises exponentially with the number of features. The limited number of rows could have prevented us from getting proper clusters in our dataset and only a larger dataset and reveal the extent to which this might be a problem.

We attach the python code used to prepare this report as an appendix. The code is not necessary to follow the report but can be used to verify any statements made in the report.