IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

Link for collab file: IML_Project.ipynb

Link for ppt: link

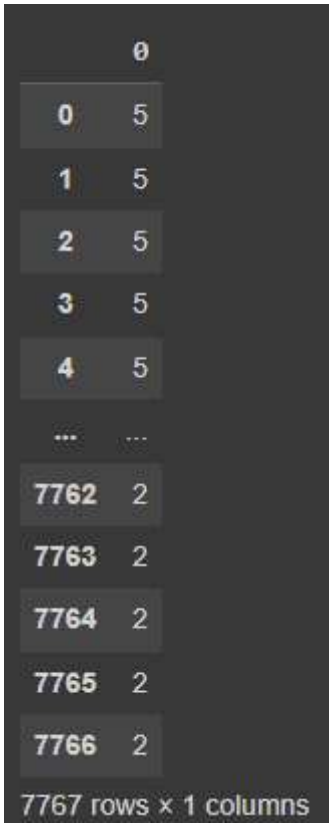
Given:-

- We were provided with raw data of acceleration and angular velocities of the moment of the arm in all 3 directions and were supposed to predict the activity performed by the user at a point in time.
- Normalized data- The features in the data were derived from accelerometer and gyroscope measurements.
- The data comes from accelerometer and gyroscope 3-axial raw signals, denoted as tAcc-XYZ and tGyro-XYZ. Signals were captured at a constant rate of 50 Hz.
- The raw signals underwent filtering to reduce noise. This involved using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz.
- The acceleration signal was further processed to separate body and gravity acceleration signals, resulting in tBodyAcc-XYZ and tGravityAcc-XYZ.

	0	1	2	3	4	5	6	7	8	9	...	551	552	553	554	555	556	557	558	559	560
0	0.043580	-0.005970	-0.035054	-0.995381	-0.988366	-0.937382	-0.995007	-0.988816	-0.953325	-0.794796	...	-0.012236	-0.314848	-0.713308	-0.112754	0.030400	-0.464761	-0.018446	-0.841559	0.179913	-0.051718
1	0.039480	-0.002131	-0.029067	-0.998348	-0.982945	-0.971273	-0.998702	-0.983315	-0.974000	-0.802537	...	0.202804	-0.603199	-0.860677	0.053477	-0.007435	-0.732626	0.703511	-0.845092	0.180261	-0.047436
2	0.039978	-0.005153	-0.022651	-0.995482	-0.977314	-0.984760	-0.996415	-0.975835	-0.985973	-0.798477	...	0.440079	-0.404427	-0.761847	-0.118559	0.177899	0.100699	0.808529	-0.849230	0.180610	-0.042271
3	0.039785	-0.011809	-0.028916	-0.996194	-0.988569	-0.993256	-0.996994	-0.988526	-0.993135	-0.798477	...	0.430891	-0.138373	-0.491604	-0.036788	-0.012892	0.640011	-0.485366	-0.848947	0.181907	-0.040826
4	0.038758	-0.002289	-0.023863	-0.998241	-0.986774	-0.993115	-0.998216	-0.986479	-0.993825	-0.801982	...	0.137735	-0.366214	-0.702490	0.123320	0.122542	0.693578	-0.615971	-0.848164	0.185124	-0.037080
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7762	0.048048	-0.042445	-0.065884	-0.195448	-0.278326	-0.219954	-0.282233	-0.305861	-0.357803	0.267874	...	-0.008381	-0.596760	-0.879026	-0.190437	0.829718	0.206972	-0.425619	-0.792292	0.238580	0.056020
7763	0.037639	0.006430	-0.044345	-0.235372	-0.302680	-0.232843	-0.322483	-0.354464	-0.345592	0.181271	...	0.209452	-0.404418	-0.684496	0.064907	0.875679	-0.879033	0.400219	-0.772288	0.252653	0.056252
7764	0.037451	-0.002724	0.021009	-0.218281	-0.378082	-0.076950	-0.304446	-0.400661	-0.193071	0.113141	...	0.237003	0.000207	-0.317314	0.052806	-0.266724	0.864404	0.701169	-0.779566	0.249121	0.047071
7765	0.044011	-0.004536	-0.051242	-0.219202	-0.383350	-0.081035	-0.310419	-0.380233	-0.201007	0.166671	...	0.069366	0.037919	-0.356579	-0.101360	0.700740	0.936674	-0.589479	-0.785603	0.246409	0.031700
7766	0.068954	0.001810	-0.080323	-0.269336	-0.366553	-0.147294	-0.377332	-0.360597	-0.255505	0.321881	...	0.002496	-0.400831	-0.742972	-0.280088	-0.007739	-0.056088	-0.616956	-0.783693	0.246785	0.042981

7767 rows × 561 columns

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES



	0
0	5
1	5
2	5
3	5
4	5
...	...
7762	2
7763	2
7764	2
7765	2
7766	2

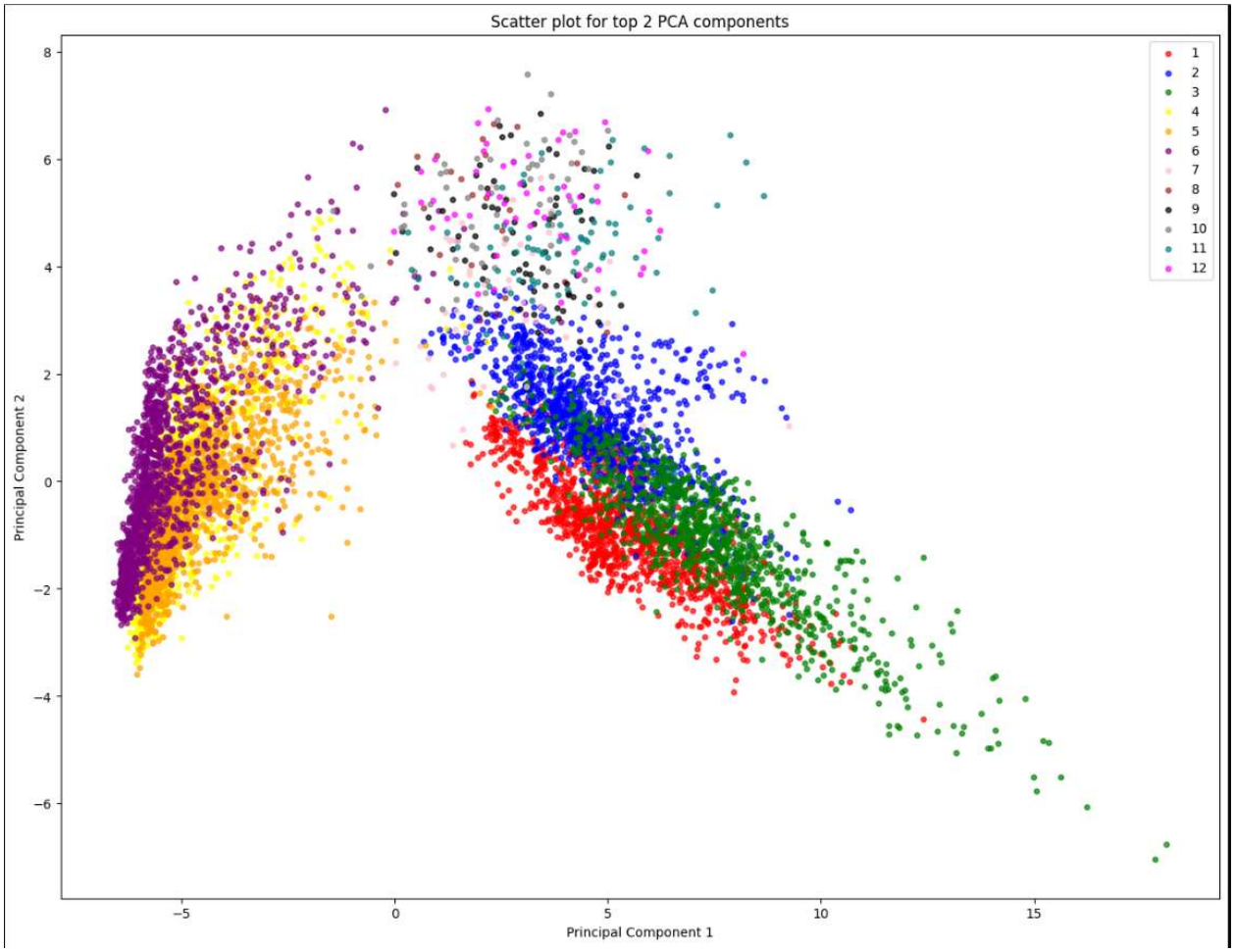
7767 rows × 1 columns

We started the project by plotting some basic in-class variance and applying PCA to find the number of components that have at least 80-90% variance. Then we used PCA and LDA to visualize the dataset.

Exploratory Data Analysis:-

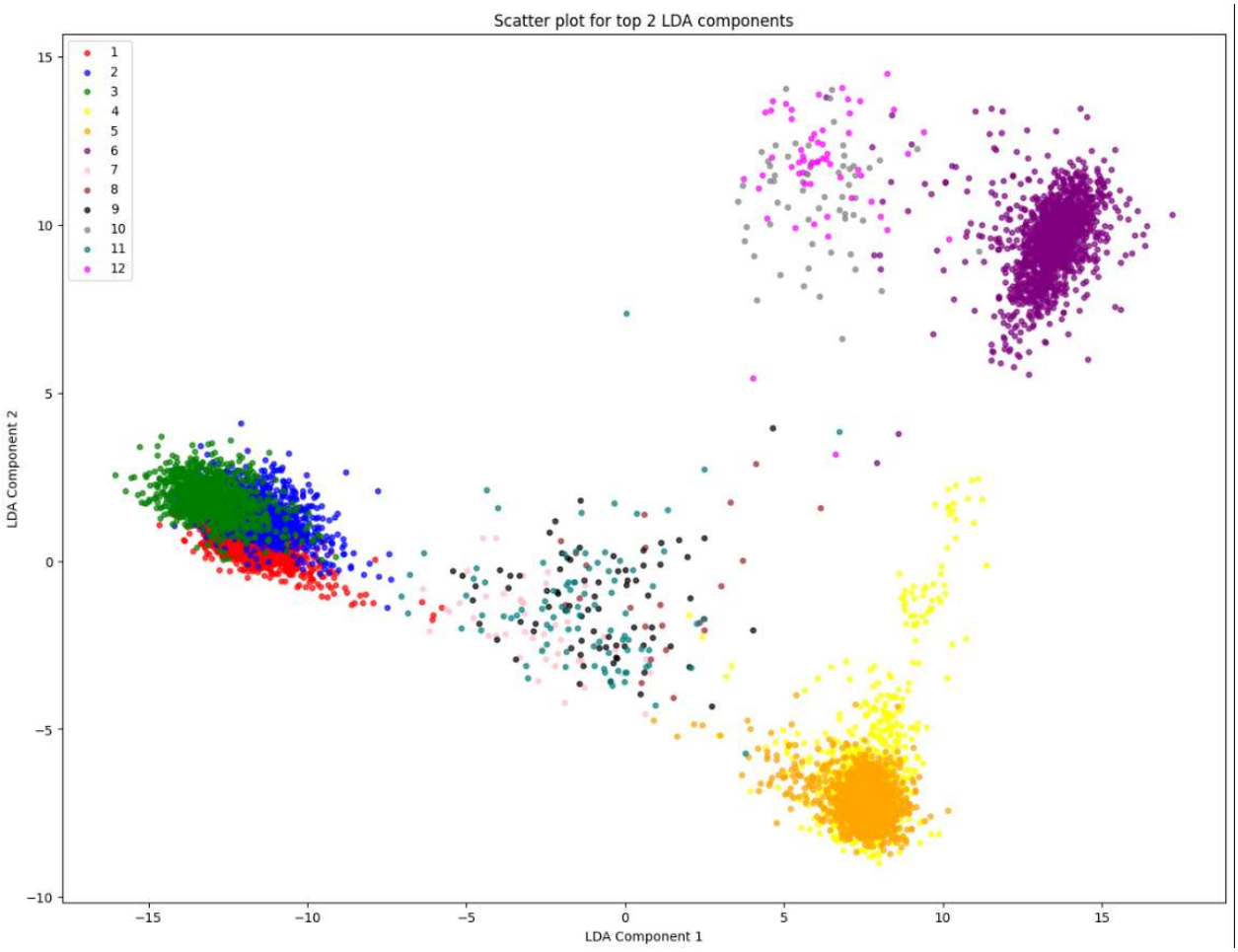
- *We used PCA for visualization:*

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES



- *We used LDA for Visualization:*

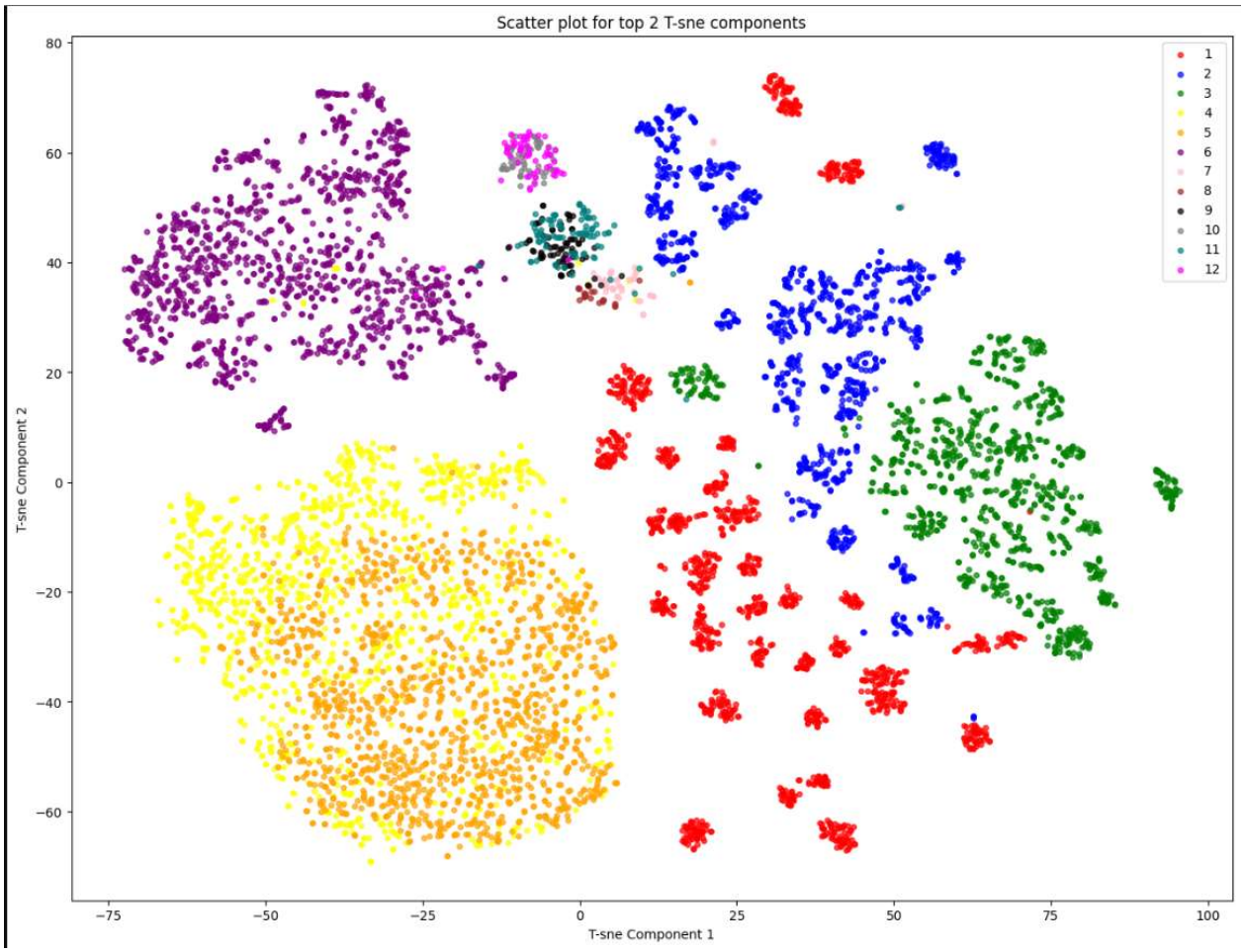
IML PROJECT REPORT-
SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES



- *We used T-SNE for visualization:*

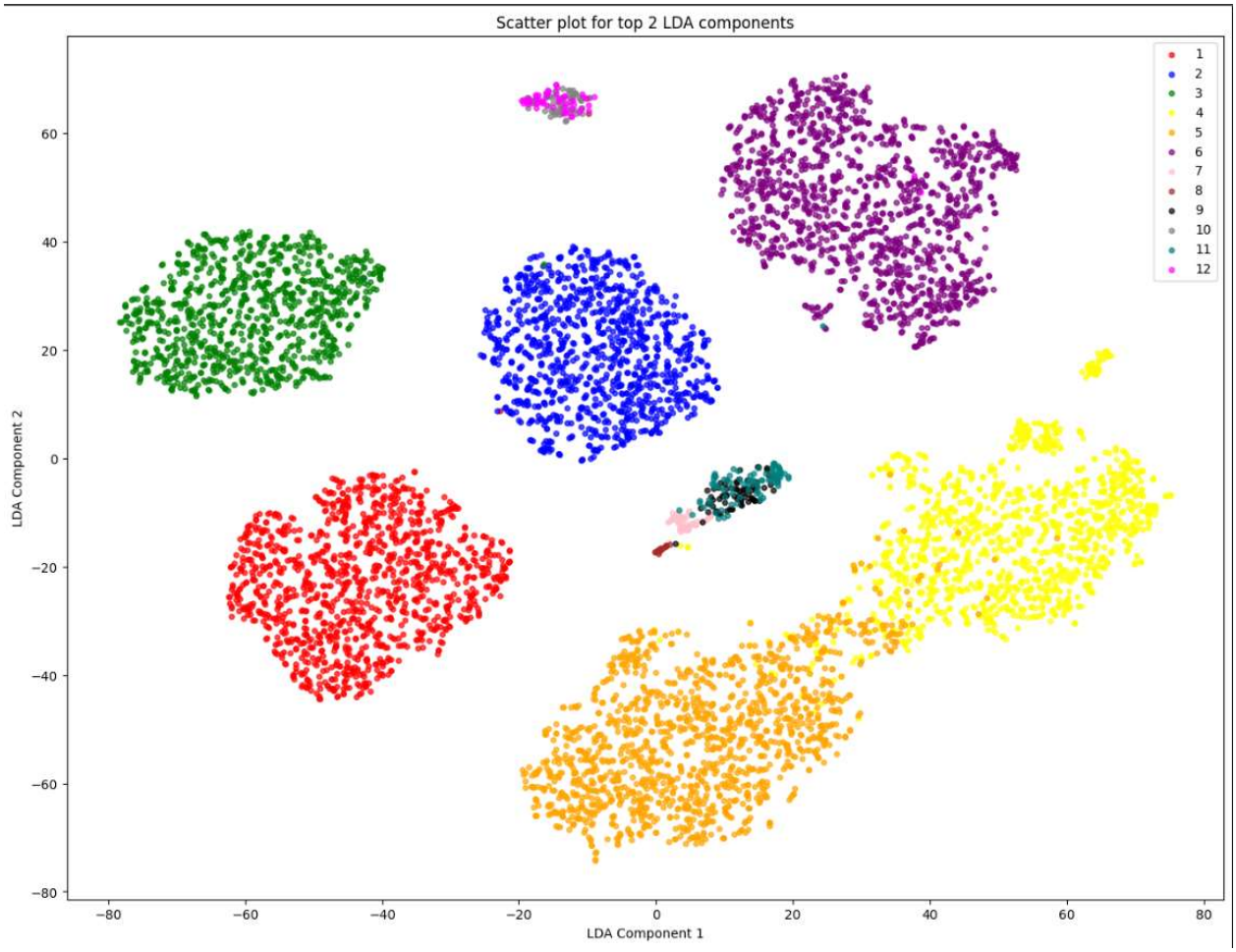
IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

(



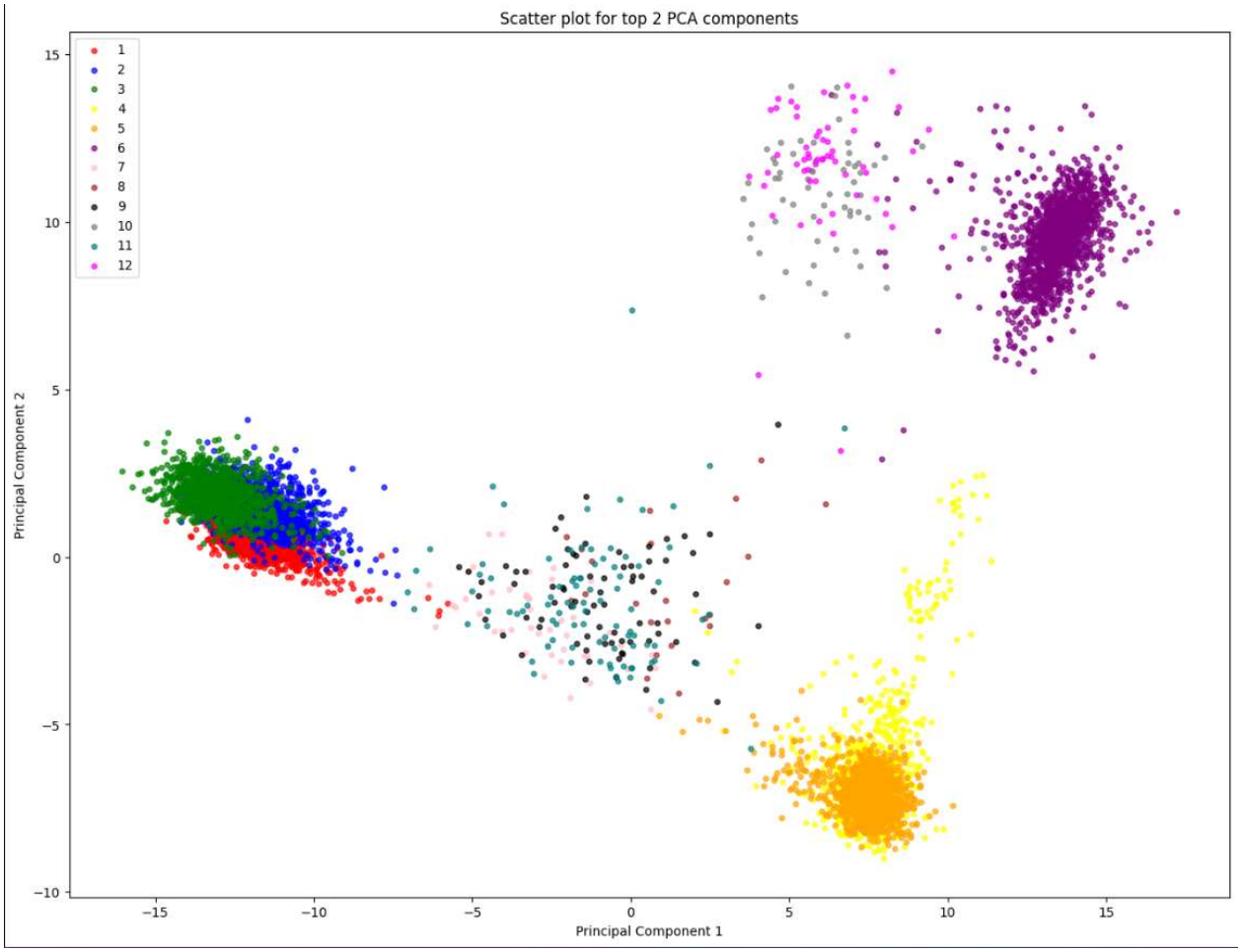
- *We used T-SNE on LDA data to form clusters:*

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES



- *We used PCA on LDA data:*

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES



Then we found 12 features from Sequential feature Selection (SFS) and got the following features:-

```
X_train_sfs = X_train[:, [1,37,40,41,50,53,69,201,274,450,508,518]]
X_test_sfs = X_test[:, [1,37,40,41,50,53,69,201,274,450,508,518]]
```

Feature names:

- tBodyAcc-Mean-2
- tBodyAcc-Correlation-1
- tGravityAcc-Mean-1
- tGravityAcc-Mean-2

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

- tGravityAcc-Max-2
- tGravityAcc-Min-2
- tGravityAcc-ARCoeff-5
- tBodyAccMag-STD-1
- fBodyAcc-Max-1
- fBodyGyro-MaxInds-3
- fBodyAccMag-Energy-1
- fBodyAccJerkMag-Max-1

We also found the features that gave maximum variance to the data:-

```
Index: 0, Value: 0.348251781514542
Index: 265, Value: 0.3320283401511527
Index: 73, Value: 0.3290288252399904
Index: 16, Value: 0.32880800254573894
Index: 39, Value: 0.32170019702568353
Index: 185, Value: 0.32068728637998606
Index: 34, Value: 0.3199224447588498
Index: 215, Value: 0.3188804258116586
Index: 232, Value: 0.31878489572760643
Index: 22, Value: 0.3181926796689048
```

Feature names:

- tBodyAcc-Mean-1
- fBodyAcc-Mean-1
- tGravityAcc-ARCoeff-8
- tBodyAcc-Energy-1
- tBodyAcc-Correlation-3
- tBodyGyroJerk-ARCoeff-1
- tBodyAcc-ARCoeff-10
- tBodyAcc-ARCoeff-10
- tGravityAccMag-Mad-1
- tBodyAccJerkMag-Energy-1
- tBodyAcc-ropy-1

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

Then we applied ADF test and KPSS test to find out the dataset given is stationary or not:-

We applied ADF test on the data which went through T-SNE on LDA

These are the results we get:

```
ADF Statistic: -15.1261
P-value: 0.0000
Critical Values:
1%: -3.4308
5%: -2.8617
10%: -2.5669
```

We also applied KPSS test on the dataset which also checks whether the time series is stationary or not:-

```
KPSS Statistic: 3.15898
P-value: 0.01000
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
<ipython-input-81-4a347a018148>:4: InterpolationWarning: The test statistic is outside of the range of p-values available in the
look-up table. The actual p-value is smaller than the p-value returned.

  result = sm.tsa.stattools.kpss(X_train1d, regression='c')
```

```
KPSS Statistic: 0.34518
P-value: 0.01000
Critical Values: {'10%': 0.119, '5%': 0.146, '2.5%': 0.176, '1%': 0.216}
<ipython-input-82-c47995d57f9d>:4: InterpolationWarning: The test statistic is outside of the range of p-values available in the
look-up table. The actual p-value is smaller than the p-value returned.

  result = sm.tsa.stattools.kpss(X_train1d, regression='ct')
```

We got the results:-

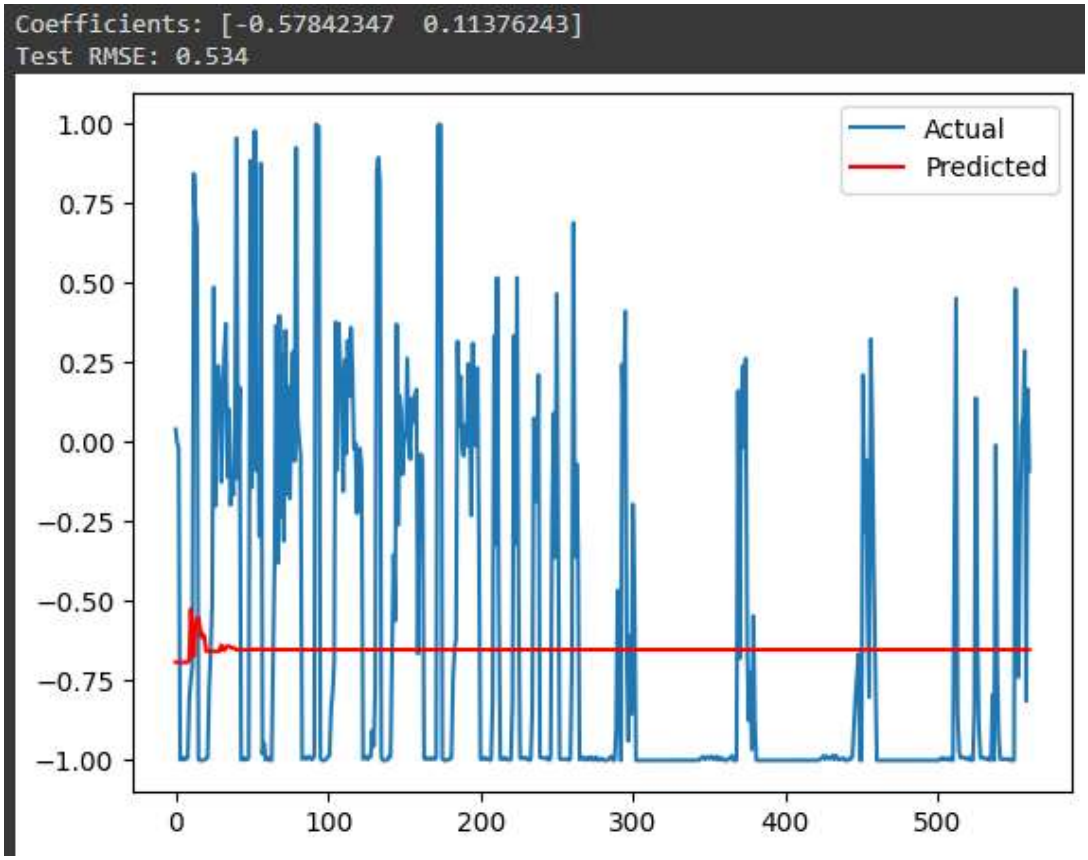
- Stationary from ADF test.
- Non-stationary from KPSS test.

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

- Reason for this contradicting result may be from the fact that the data fed to the ADF test was way too small to check any root whereas the KPSS test was applied to the whole dataset.
- Thus, we can infer that the given time series is non-stationary.

<u>We then applied different time-series model and plotted the result:-</u>

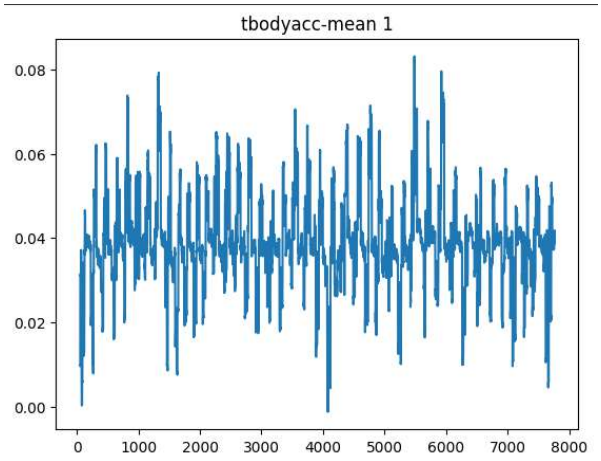
Auto-Regression:-

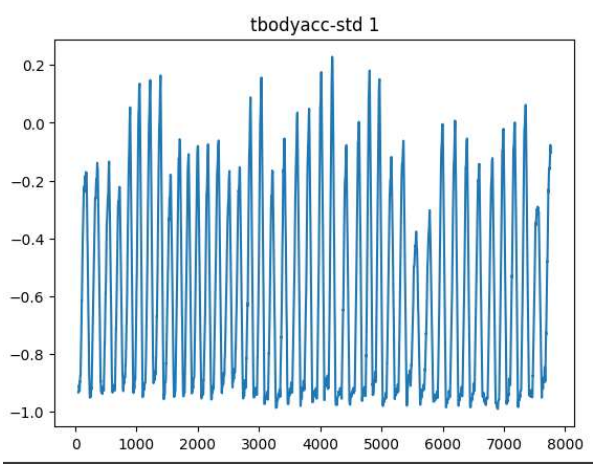


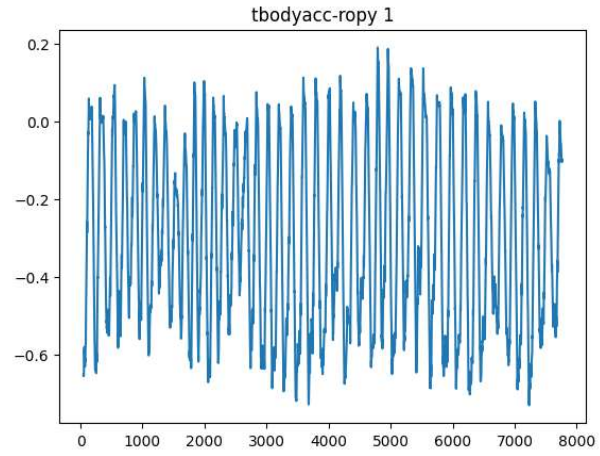
MOVING AVERAGE:-

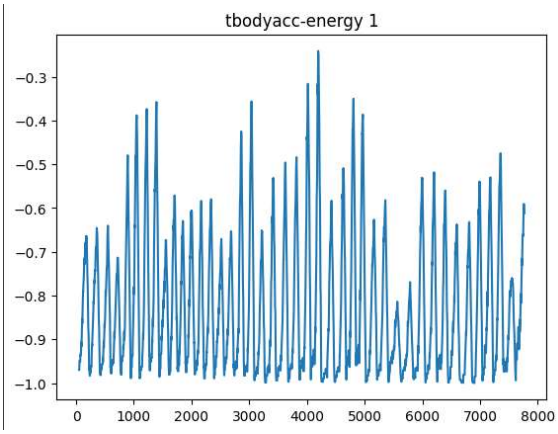
- Moving averages of some features are given below

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES





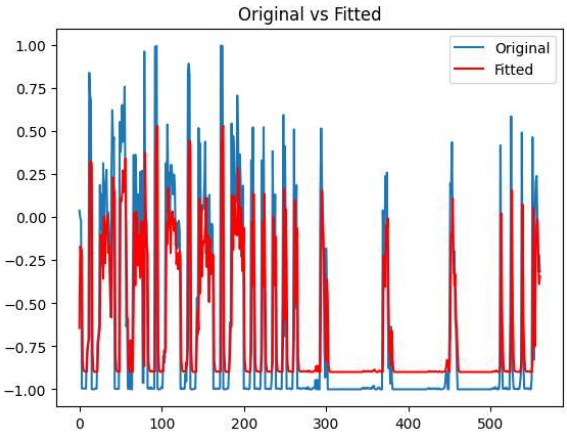


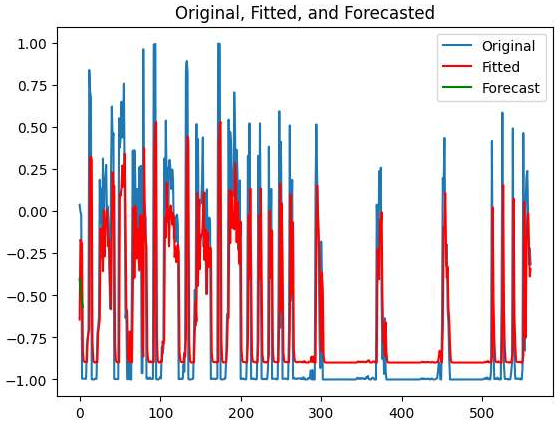


AUTO-REGRESSIVE and MOVING AVERAGE:-

- Order = (1,1)
- Mean Squared Error (MSE): 0.15514070387873785
- Root Mean Squared Error (RMSE): 0.39387904727052675

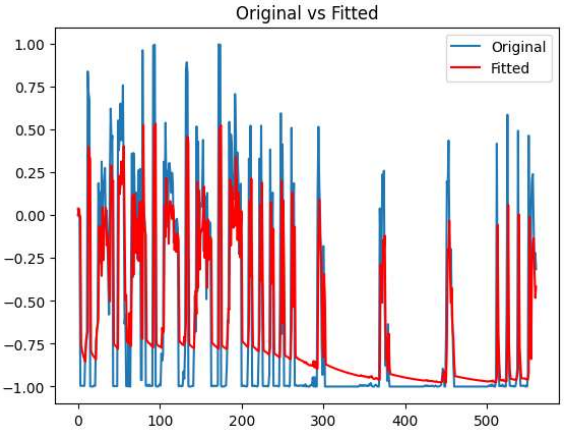
IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

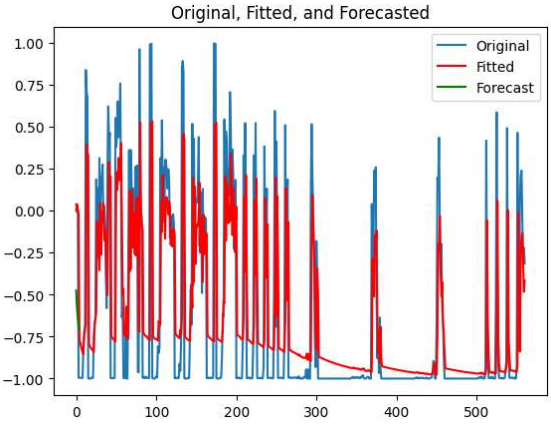




AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE:-

- Order = (1,1,1)
- Mean Squared Error (MSE): 0.15281820689155806
- Root Mean Squared Error (RMSE): 0.39091969366042184

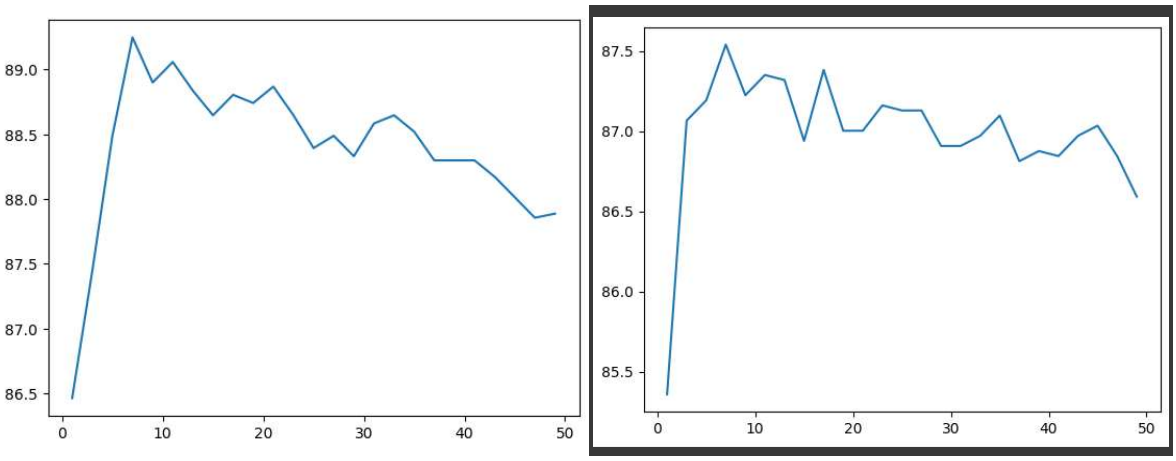




<u>After this we applied some machine learning models to predict the test data and got the following accuracy with the original data(at first) and the data we get after sequential feature selection (at second):-</u>

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

KNN:-



- Maximum accuracy is at k=7
- Original data accuracy: 89.25%
- SFS data accuracy: 87.54%

LOGISTIC REGRESSION:-

- Original data accuracy: 94.56%
- SFS data accuracy: 86.18%

SVM:-

- With linear kernel:
  - Original data accuracy: 95.193%
  - SFS data accuracy: 87.54%
- With rbf kernel:
  - Original data accuracy: 93.675%
  - SFS data accuracy: 88.204%
- With polynomial kernel:
  - Original data accuracy: 94.56%
  - SFS data accuracy: 88.362%

IML PROJECT REPORT- SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

DECISION TREE:-

- Original data accuracy: 80.42%
- SFS data accuracy: 80.39%

Multi Layer Perceptron:-

- Original data accuracy: 94.12%
- SFS data accuracy: 87.73%

RANDOM FOREST CLASSIFIER:-

- Original data accuracy: 91.43%
- SFS data accuracy: 86.5%

XGBOOST:-

- Original data accuracy: 15.68627450980392%
- SFS data accuracy:15.68627450980392%

LIGHT GRADIENT BOOST:-

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	0.74	0.72	0.73	496
2	0.68	0.72	0.70	471
3	0.79	0.68	0.73	420
4	0.78	0.74	0.76	508
5	0.80	0.75	0.77	556
6	0.98	0.89	0.93	545
7	0.11	0.39	0.18	23
8	0.06	0.20	0.09	10
9	0.25	0.44	0.32	32
10	0.19	0.36	0.25	25
11	0.39	0.33	0.36	49
12	0.13	0.30	0.18	27
accuracy			0.73	3162
macro avg	0.45	0.50	0.46	3162
weighted avg	0.77	0.73	0.75	3162

Original Data

	precision	recall	f1-score	support
1	0.88	0.95	0.91	496
2	0.88	0.86	0.87	471
3	0.93	0.88	0.91	420
4	0.78	0.75	0.76	508
5	0.79	0.81	0.80	556
6	1.00	0.99	1.00	545
7	0.50	0.61	0.55	23
8	0.39	0.70	0.50	10
9	0.62	0.50	0.55	32
10	0.48	0.48	0.48	25
11	0.54	0.51	0.53	49
12	0.50	0.37	0.43	27
accuracy			0.85	3162
macro avg	0.69	0.70	0.69	3162
weighted avg	0.86	0.85	0.85	3162

SFS Data

IML PROJECT REPORT-
SMARTPHONE BASED RECOGNITION OF HUMAN ACTIVITIES

Conclusion:-

- We were provided with the aforementioned dataset and different models were applied on the dataset.
- Exploratory Data Analysis(EDA) was performed.
- Some basic models such as PCA, LDA, t-SNE etc. were applied as well as some time series models such as ARMA, ARIMA were also applied.
- Some experiments were also done by combining two models to improve the accuracy.

- It was found that SVM with linear kernel gave the best results with an accuracy of 95.16