

Mid Term Report



Name: **sunkara venkata sreeram**
Roll No: **19125760094**

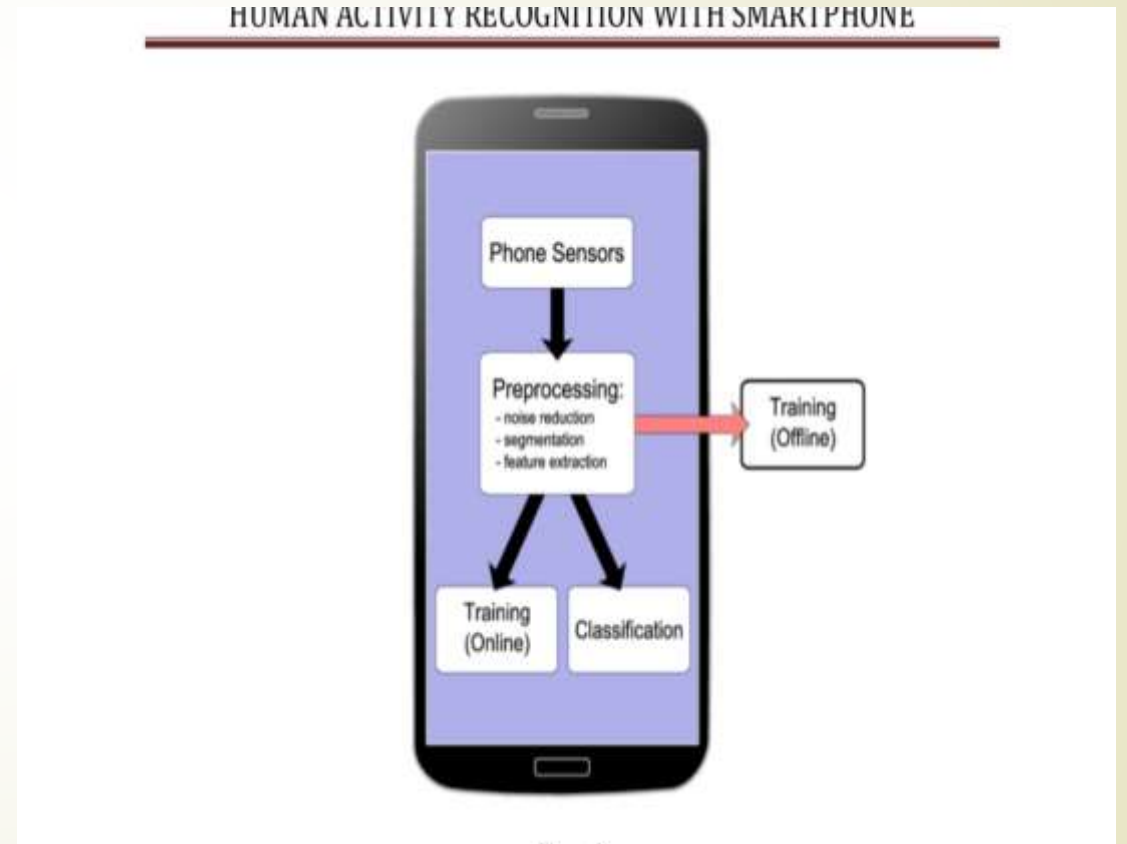
(Under the guidance of
Anantpadmanabh Divanji)



HUMAN ACTIVITY RECOGNITION WITH SMARTPHONES

Problem Statement

- The Human Activity Recognition with Smartphones data, which was built from the recordings of participants performing activities of daily living (ADL) while carrying a smartphone with an embedded inertial sensors.
- The goal is to classify activities into one of the six activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying) performed.
- In this project I am trying different models to make prediction easier and authentic.



Approach

Data Imputation

- Checking for duplicates.
- Checking for missing values.
- Checking for class imbalance.

Data Visualization

- Analysing tBodyAccMag-mean feature
- Analysing Angle between X-axis and gravity Mean feature
- Visualizing data using t-SNE

Modelling



Understanding the Data

- ▶ The dataset with a group of 30 volunteers with an age of 19-48 years. Every person performed six activities wearing a smartphone. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3- axial angular velocity at a constant rate of 50 HZ. Signals (Acceleration and Gyroscope) were pre-processed by applying noise filters and then sampled in fixed width sliding windows of 2.56 sec and 50% overlap.
- ▶ The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 HZ cutoff frequency was used. A vector feature was obtained by calculating variables from the time and frequency domain.
- ▶ The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity.
- ▶ Both the train and test data having 563 columns. Where 70% and 30% volunteers divided for training and testing data.



The dataset consists of a labelled column followed by some attribute fields:

- ▶ fBodyBodyGyroJerkMag
- ▶ fBodyAccJerk
- ▶ tBodyGyroJerk
- ▶ tGravityAcc
- ▶ fBodyGyro

Subject

- ▶ fBodyBodyAccJerkMag

Activity

- ▶ tBodyAcc
- ▶ tGravityAccMag

Angle

- ▶ tBodyGyro



Data Imputation

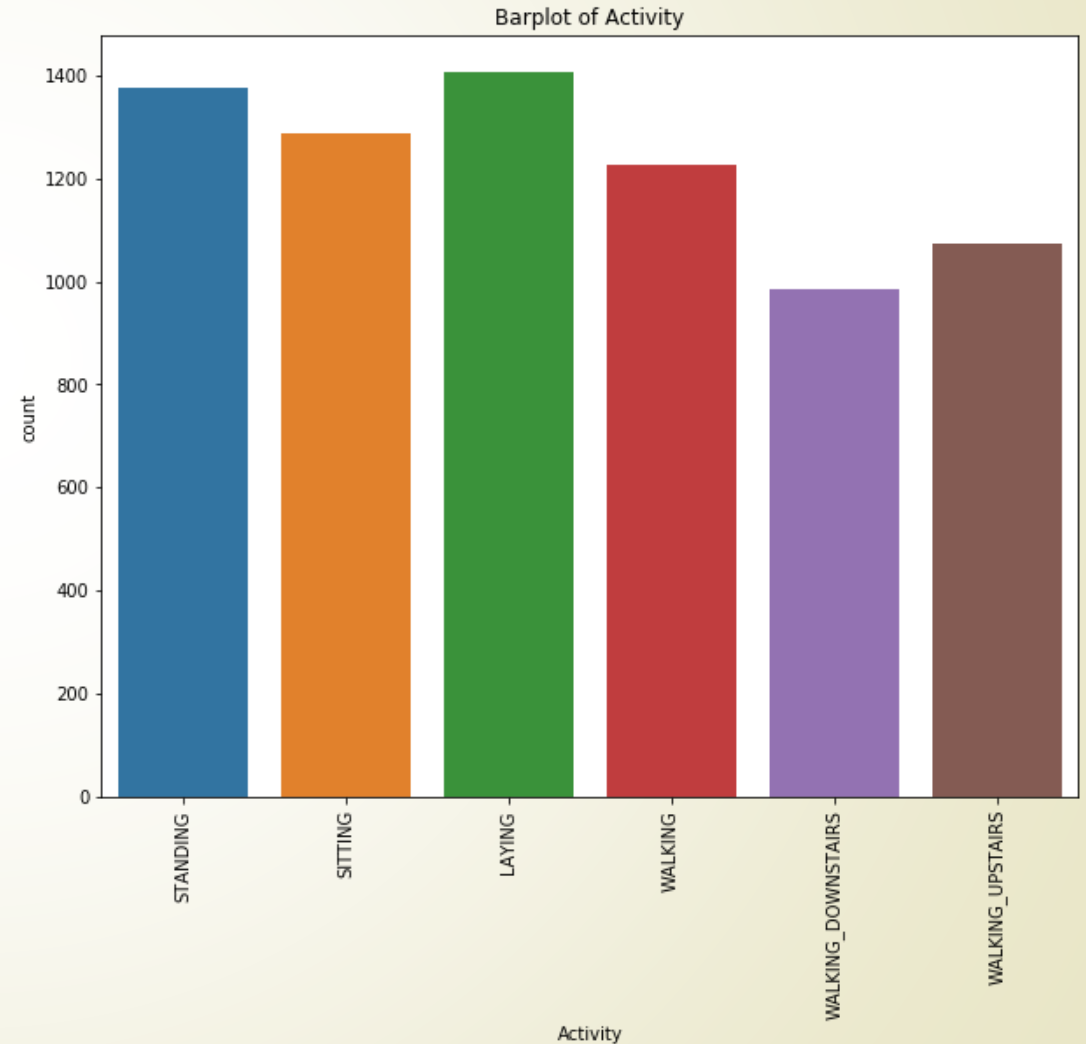
- ▶ By inspecting the data reveals that 561 continuous valued predictors are available, as well as ID variable describing the particular individual performing the activity.
- ▶ Latter variable will be omitted and interested in generalized activity recognition.
- ▶ Remaining predictors have already scaled to lie between -1 and 1.

❖ Checking for duplicates and missing values:

- Data was properly scaled and uniformly distributed. So, data having zero duplicates and null values.

❖ Checking for class imbalance:


- By using Bar plot the set having almost same number of observations across all the six activities. So this data does not have class imbalance problem.





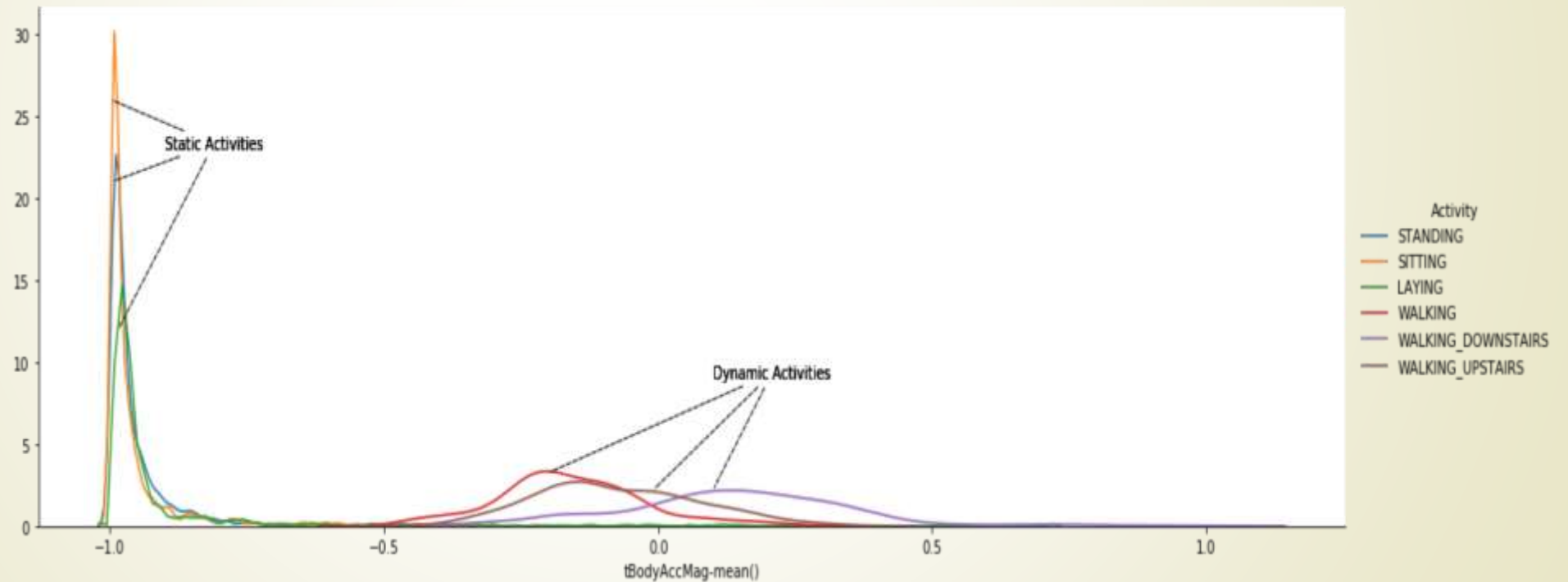
Data Visualization

Based on the common nature of activities broadly put them in two categories static and dynamic activities.

- ❑ Sitting, standing, laying can be considered as static activities with no motion involved.
 - ❑ Walking, walking-downstairs, walking-upstairs, can be considered as dynamic activities with significant amount of motion involved.
 - ❑ Let us consider `tBodyAccmag - mean()` feature to differentiate among these two broader set of activities.
- 

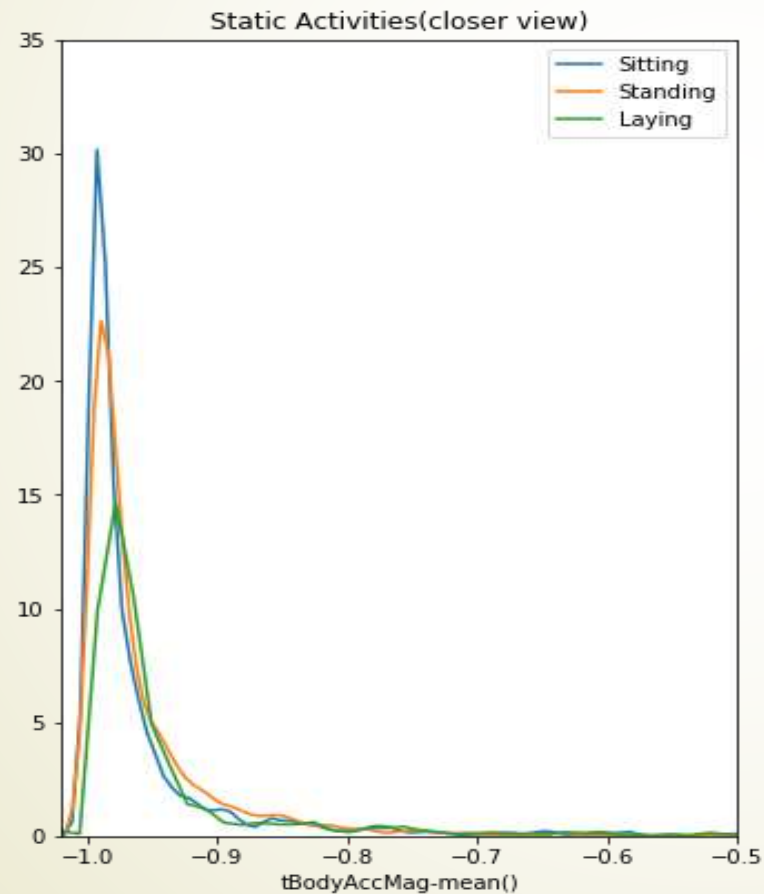
a)Analysing tBodyAccMag – mean feature

- By using the density plot we can easily come with a condition to separate static activities from dynamic activities.

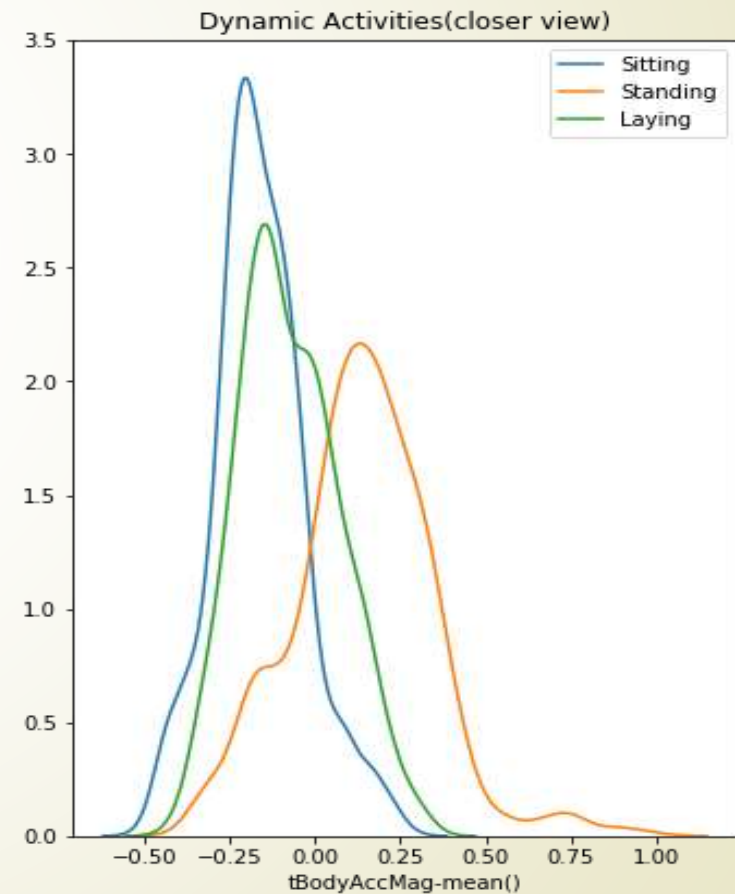


Density plot for Static and Dynamic Activities

Static Activities

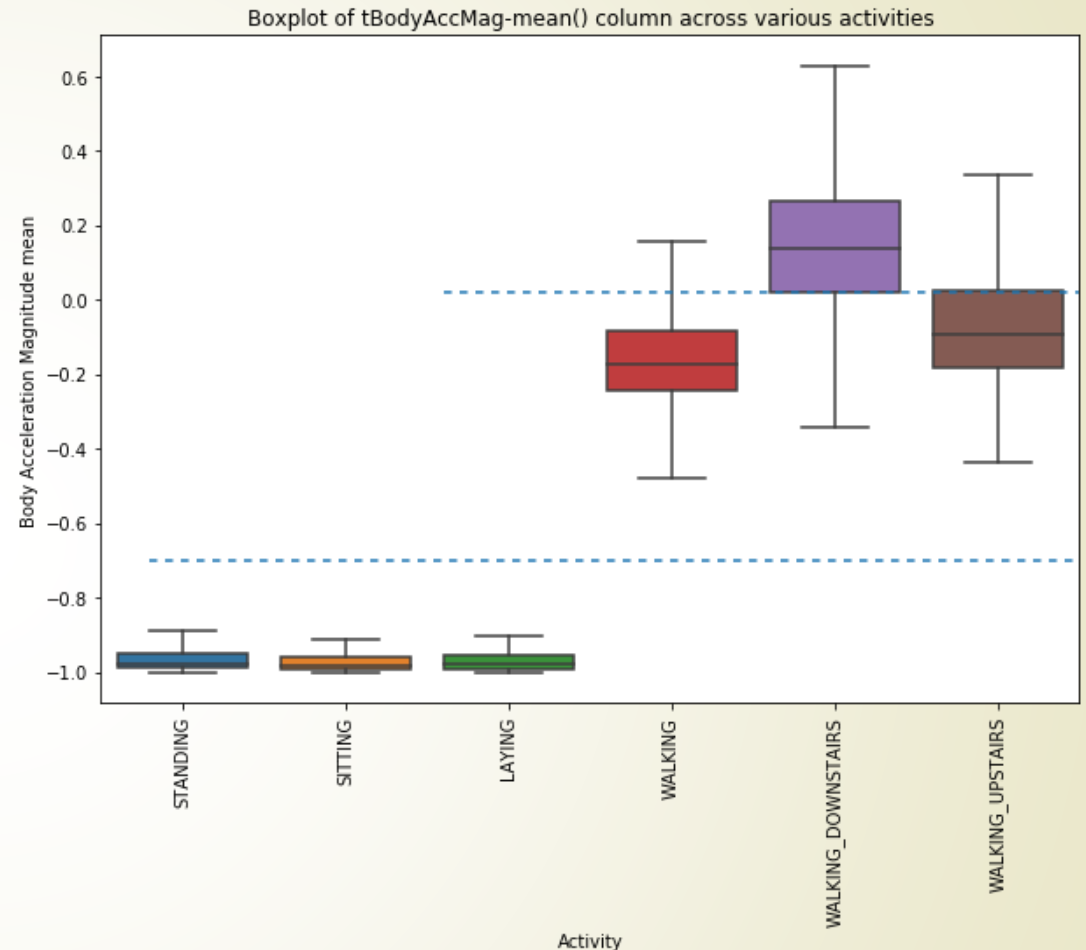


Dynamic Activities



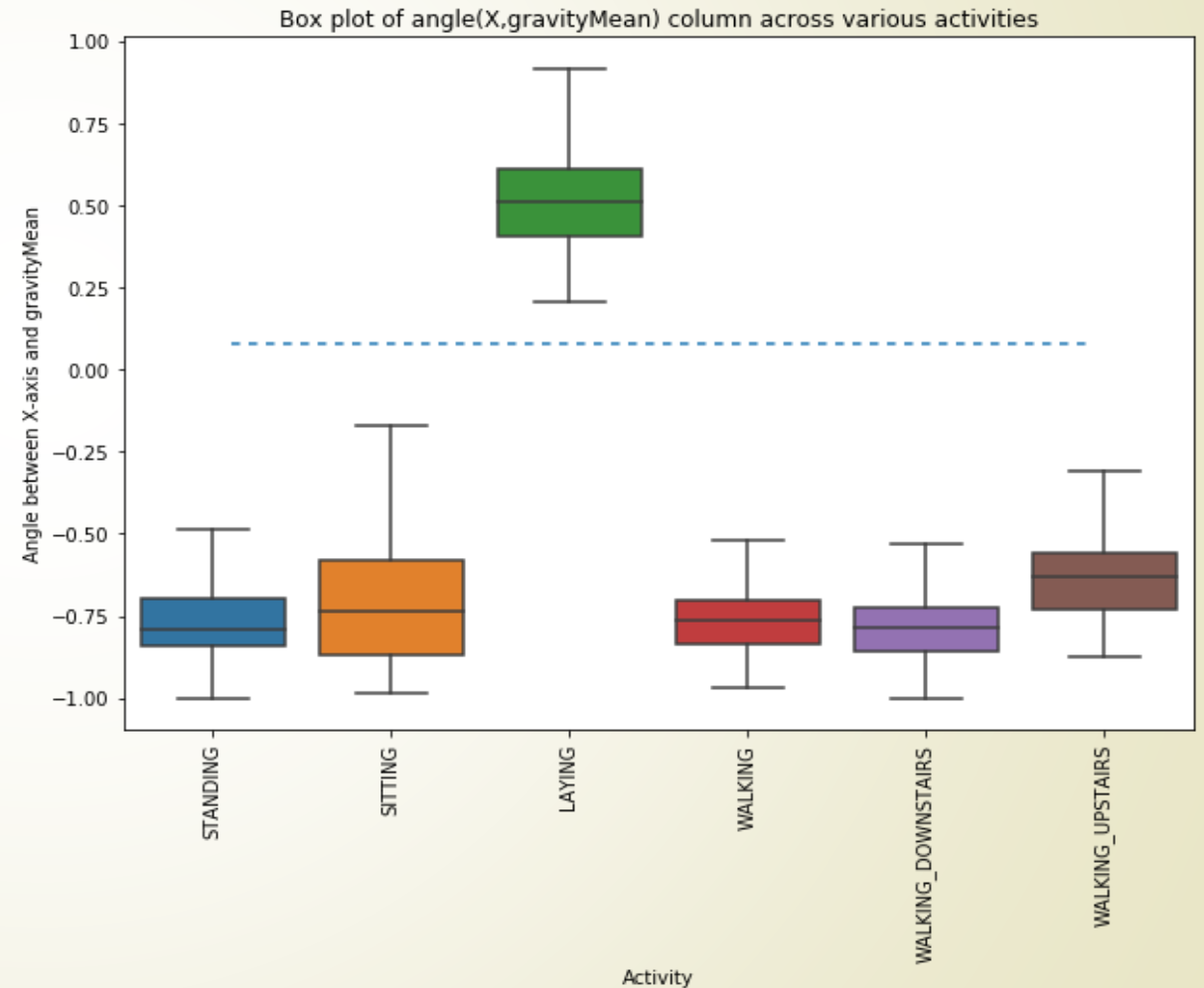
Box plot

- Density plots can be represented using boxplots.
- plot the boxplot of body acceleration magnitude $\text{mean}(\text{tBodyAccMag} - \text{mean}())$ across all the six categories.
- Easily separate Walking-Downstairs activity from other users using boxplot.
- But still 25% of walking-downstairs observations are below 0.02 which are misclassified as others. so this condition makes an error 25% in classification.



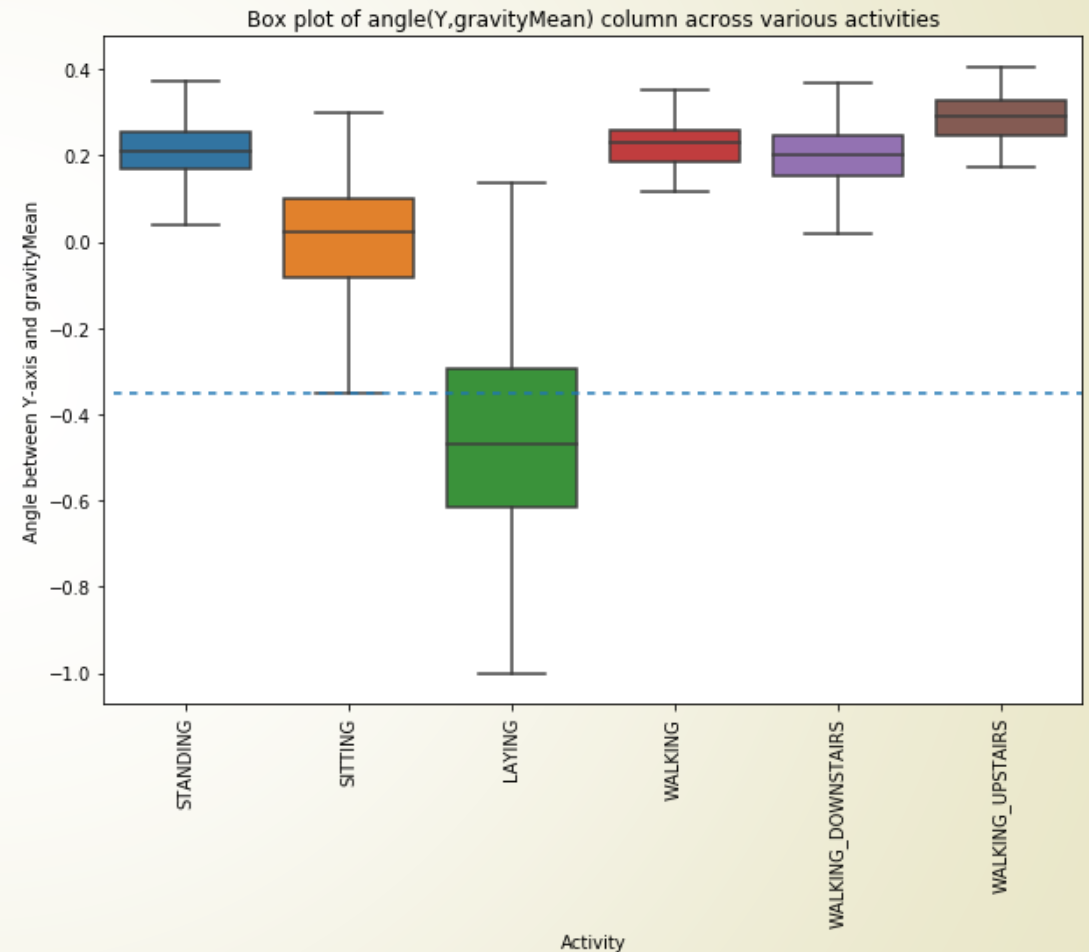
b)Analysing angle between X-axis and gravity mean feature

- From the boxplot we can observe that angle (X-gravity mean) perfectly separates **LAYING** from other activities.



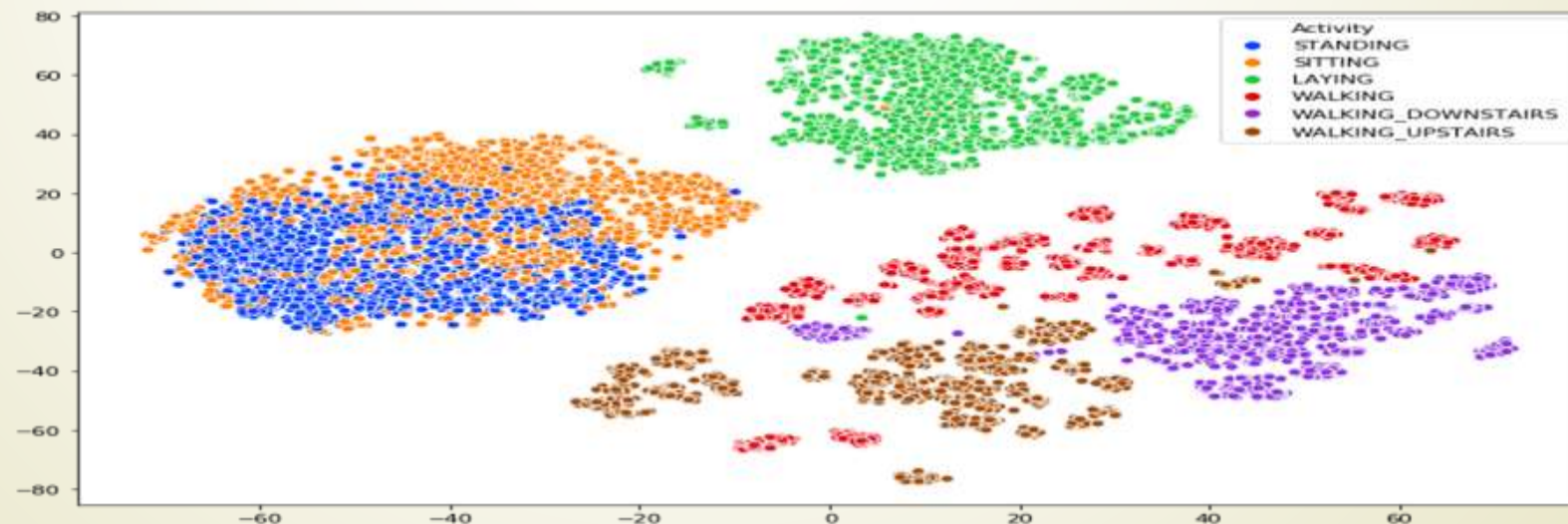
c) Analysing angle between Y-axis and gravity mean feature

- Similarly, using angle between y-axis and gravity mean we can separate **LAYING** from other activities but again it leads to some classification error.



d) Visualizing data using t-SNE

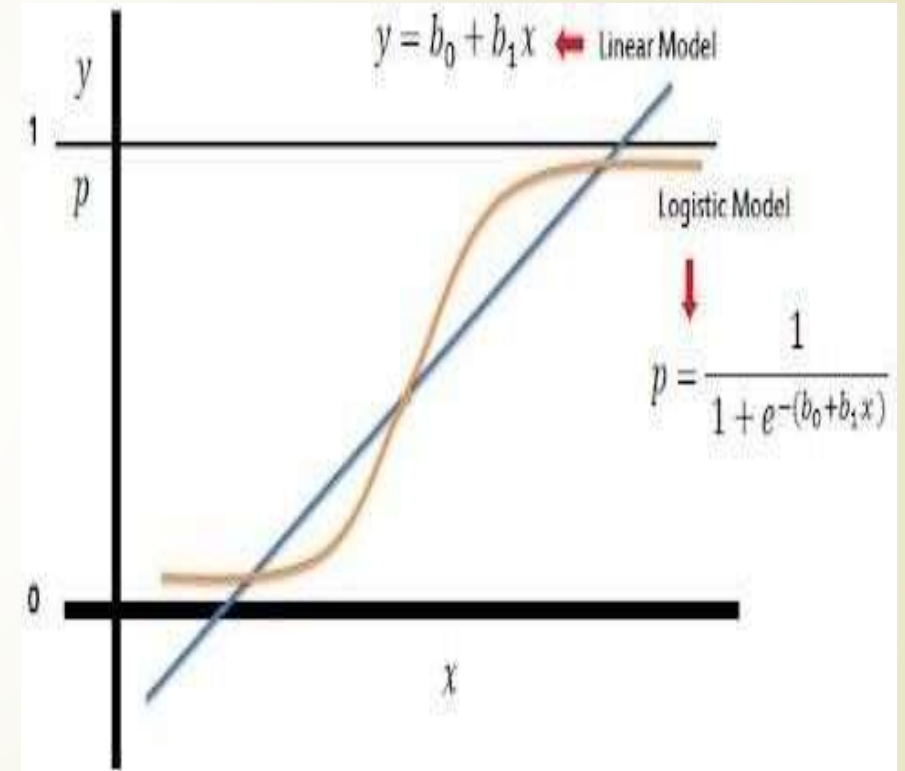
- Using t-SNE can be visualized from an extremely high dimensional space to a low dimensional space and still it retains lots of actual information.
- Training data has 561 unique features, using t-SNE visualize it to a 2D – space.
- Using two new- components obtained through t-SNE we can visualize and separate all six activities in a 2D- space.



MODELLING

➤ Logistic Regression

- Used for classification problems.
- Based on probability concept.
- Limit between 0 and 1.
- Used for binary variables.
- Ex: yes or no.
- By using this formula we can calculate the predicted values.



➤ Linear SVM

- Used for both linear and non – linear problems and work well for many practical problems.
- The goal of the svm algorithm is to create the best line.

➤ Kernel SVM

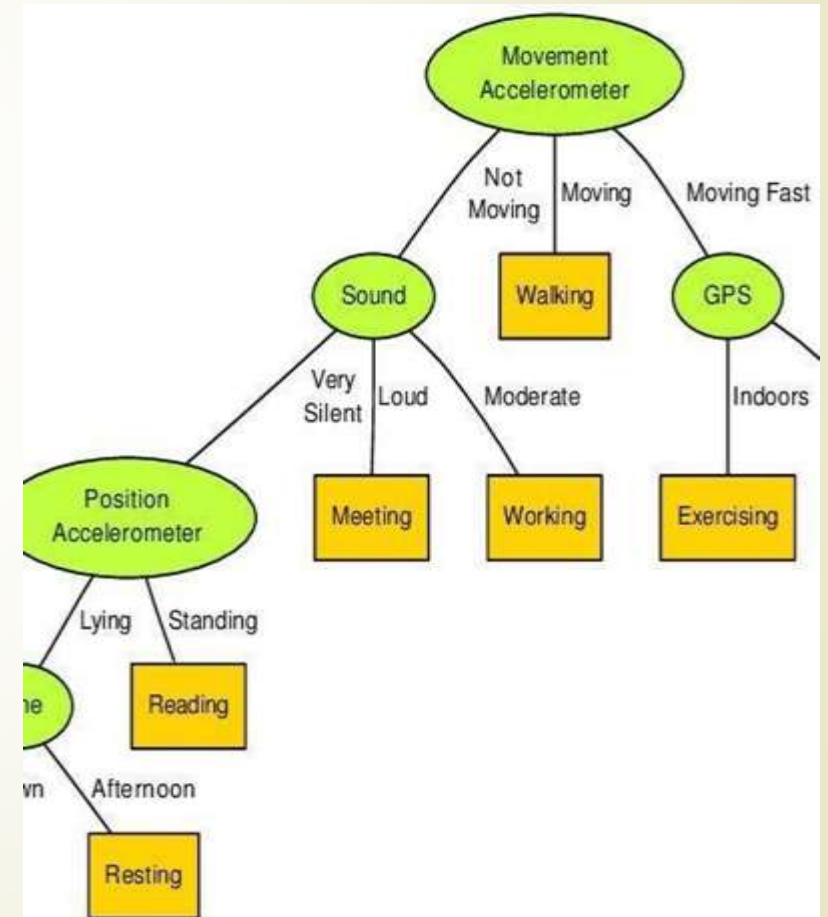
- Used for pattern analysis best known is svm.
- Introduced for sequence data, graphs, text, images, as well as vectors.

By using formulas we can calculate both linear and Gaussian svm.

Kernel Functions	Formulae
Linear kernel	$K_s(\chi, \chi_i) = \chi \cdot \chi_i$
Quadratic kernel	$K(\chi, \chi_i) = (\chi \cdot \chi_i + 1)^2$
Polynomial kernel	$K(\chi, \chi_i) = (\chi \cdot \chi_i + 1)^p$ where $p = 3$ is the order of the polynomial
RBF kernel	$K(\chi, \chi_i) = e^{-\frac{ \chi - \chi_i ^2}{\sigma^2}}$ where $\sigma = 1$ is the width

➤ Decision Tree

- Flow chat like structure.
- Used for supervised learning.
- It is a non-parametric for both classification and regression.
- Discrete values for classification trees.
- Continuous values for regression trees.
- Model to be simple and explainable.





➤ Random Forest

- It is a classification algorithm.
- It uses bagging and feature randomness.
- Achieve low prediction error.
- Average of all selecting trees.
- Random forest always wins in terms of accuracy.

$$MSE = \frac{1}{N} \sum_{i=1}^N (f_i - y_i)^2$$

Where N is the number of data points,
 f_i is the value returned by the model and
 y_i is the actual value for data point i .

RESULTS

<u>MODELS</u>	<u>ACCURACY</u>
Logistic Regression	96.19995249
Linear SVM	96.84424838
Decision Tree	87.24126230
Random Forest	92.33118420
Kernel SVM	94.16355615


As we know that the algorithm gives more value that is the best accuracy model, by seeing we can clearly state that Linear SVM model has high accuracy than the other models.



Conclusion

From this project how machine learning can help to study data from sensors which are already present in most smartphones.

➤ Insights of the project will show:

- Nature of the candidate.
 - Activity Duration and Patterns of individuals.
 - Movement Disability and illness.
 - Person Fatigue or not.
- 



THANK YOU