**ANALYZING HUMAN ACTIVITIES USING SMARTPHONE**

**DATA**

**Shravani Hariprasad**

**Simran Solanki**

**Sharanaya Akkone**

**Nikitha Sridhar**

BDA594-GEOG594: Big Data Science and Analytics Platforms

Ming-Hsiang Tsou

December 15, 2022

[**Website Link**](https://sites.google.com/view/humanella/home?authuser=1)

[**Video Link**](https://youtu.be/HQS7XUgEoCk)

**Problem Statement:**

These days almost everybody has a smartphone and all the smartphones have sensors and they have accelerometer and gyroscope. The ideology is how are we going to use these data for understanding patterns and for a good cause.

These days we are seeing a lot of unexpected medical emergencies even in young and middle- aged people and monitoring their activities can help in times of emergencies as we keep track of all their gestures and any unusual patterns can set an automatic alarm to some connected hospitals and this can save lives of people at the right time.

**Literature Overview**

“Comparing different supervised machine learning algorithms for disease prediction - BMC Medical Informatics and Decision Making.” *BMC Medical Informatics and Decision Making*, 21 December 2019, <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8>. Accessed 15 December 2022.

This paper offers a thorough evaluation of the relative efficacy of various supervised machine learning algorithm variations for disease prediction. Researchers can choose an appropriate supervised machine learning algorithm for their study with the help of this crucial information on relative performance.

Namini, Sima Siami, et al. “A Comparison of ARIMA and LSTM in Forecasting Time Series.” <https://www.researchgate.net/publication/330477082_A_Comparison_of_ARIMA_and_LSTM_in_Forecasting_Time_Series>.

With the recent creation of more advanced machine learning algorithms and techniques, such as deep learning, new algorithms are being developed to evaluate and forecast time series data. This article explores whether and how recently discovered deep learning-based algorithms for forecasting timeseries data, such as "Long Short-Term Memory (LSTM)", are superior to the established methods.

**Data:**

We have downloaded our dataset from HAR Dataset from the UCI dataset storehouse.

The UCI dataset store contains csv files and also .txt file which contains the data about the feature engineering done by experts. We have used both the csv files and also .txt files are converted and used to train the ML and DL models. The data is recorded against 30 participants, there daily activities are recorded through a smartphone mounted to their waist. The smartphone is configured in such a way that it can record data from two implemented sensors accelerometer and gyroscope.

The directors of the underlying study constructed the dataset for these time series by sliding a fixed-width window with a width of 2.56s over the series and performing feature generation. Since there was a 50% overlap between the windows, the points are evenly spaced (1.28s). This experiment was videotaped so that the data could be manually labeled.

They have recorded "3-axial linear acceleration" (tAcc-XYZ) from the accelerometer and "3-axial angular velocity" (tGyro-XYZ) from the gyroscope using the sensors (gyroscope and accelerometer) in a smartphone with various variations. In those metrics, the prefix "t" stands for time. 3-axial signals in the X, Y, and Z directions are represented by the suffix "XYZ."

**Data Labelling:**

In our dataset, the activities are represented as the numbers from 1 to 6 as their identifiers, as follows:

Walking as 1

Walking\_Upstairs as 2

Wallking\_Downstairs as 3

Sitting as 4

Standing as 5

Laying as 6

**Data Cleaning:**

We have checked and removed the duplicate values from both the train.csv and test.csv files. The next step was to check for any null values if they are present in both the files.

We have then converted the raw data to a data frame and saved it into the csv files.

All the feature names with unnecessary sighs are removed and have changed the feature names.

**Data Analysis and its Results:**

The dataset in split into test and train data, 30 subjects data was randomly split into train and test with 70 % split into train and 30% split into test data. Eg: 21 subjects were considered for training and 9 subjects were considered for testing.

**Models used:**

We have used both Machine Learning and Deep Learning models on out dataset.

We have used around 561 features that were featured by experts and have applied machine learning techniques on it.

**Logistic Regression:** Logistic Regression is a statistical linear model used for classification problems. In this model we use the logistic function to predict the possible probabilities of outcome using the logistic function. The logistic function can take any number of inputs and outputs values in the range between 0 to 1 and hence a suitable model for classification. To encounter the problem of overfitting regularization methods can be used we use the Ridge (L2) regularization. In ridge regularization the variance of the model is reduced andoverfitting can be overcome by penalizing the predictors that are too big.

**Cross Validation:** Cross Validation is used to validate the model performance. It is a good technique used for tuning the parameters of a model like regularization factor and other tolerance criteria’s. The model will be trained and then later be evaluated on the validation set. It can ideal to cv=5 or 10 which means folds that are run. The results from this can be used as optimum parameters for that particular model.

**Support Vector Machines:** Support Vector Machines are one amongst the Supervised Machine Learning Model that are used for classification problems. There are different SVM models here in our project we have used Linear SVM and Kernel SVM models.

**a. Linear SVC:** The linear SVC model fits all the data points and returns a best fit line or hyperplane that categories or divides your data. Later we can feed features to the hyperplane or classifier to know which predicted class it belongs to.

**Decision Trees:** Decision Trees can be used for both classification and Regression problems.

They are used for predicting the outcomes of the data. The nodes of the decision tree represents the attributes of the data. The edges represent the possible outcome values. Each branch from root to leaf nodes represents one classification rule.

**Random Forest:** Random forest model can be used for regression as well as classification problems. It is an ensemble method that selects the observations randomly that builds decision trees. Random Forest would give better and reliable improvement than the single classifiers.

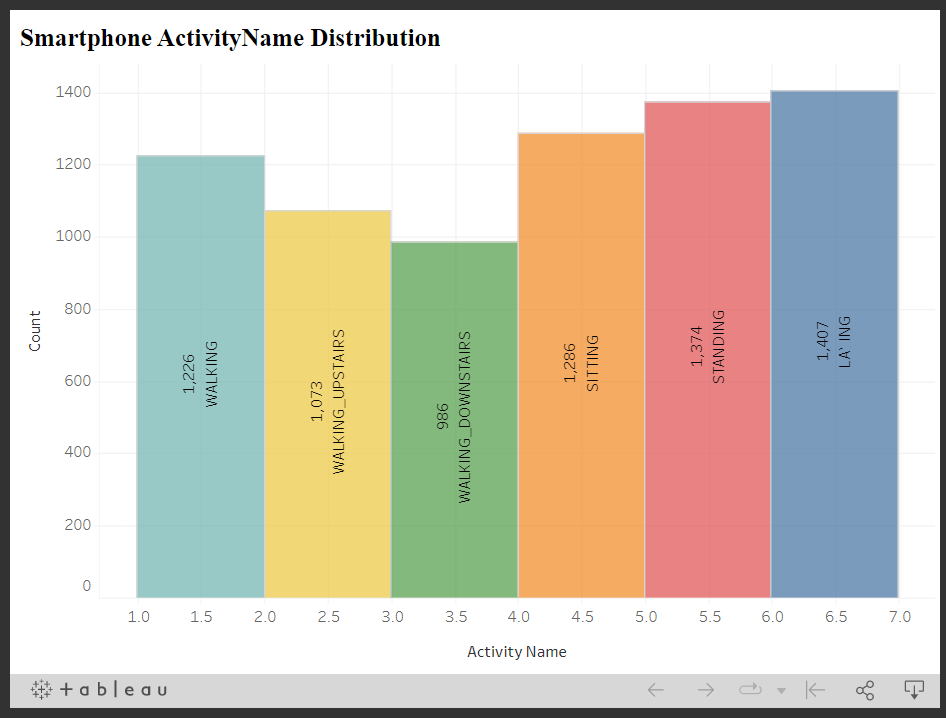
An algorithm would be used to fabricate every tree.

**Deep Learning Models:** We have applied Deep learning models in our project they are applied to the raw time series data.

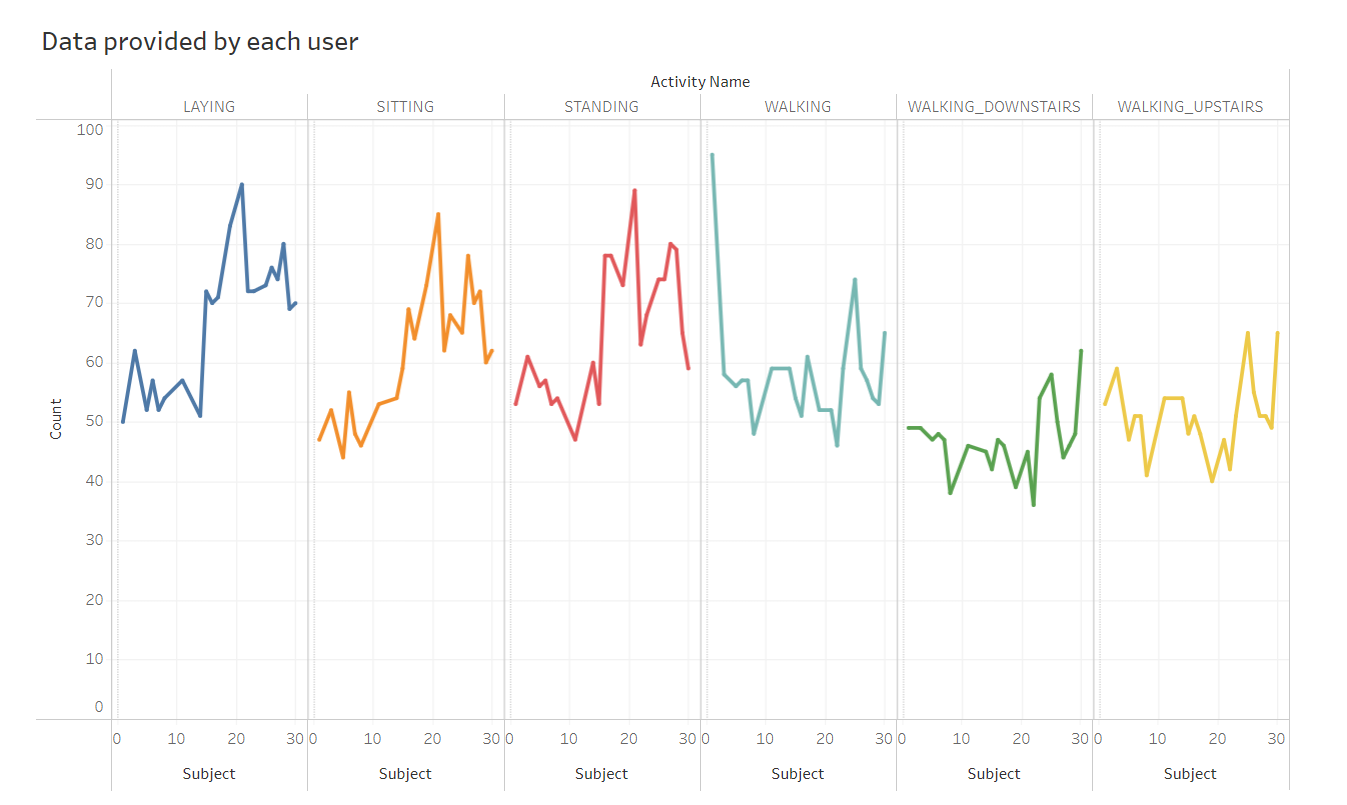
**LSTM Model:** Recurrent neural networks of the Long Short-Term Memory (LSTM) type can learn order dependence in sequence prediction issues. Aside from singular data points like photos, LSTM has backpropagation, making it capable of analyzing the complete sequence of data. This has uses in machine translation andspeech recognition, among others. A unique version of RNN called LSTM exhibits exceptional performance on a wide range of issues.

When discusiing the LSTM model, we will use simple Raw data (in the ML model, we are using single orchestrated data created by an expert), but when we look at the outcome without any FE data, LSTM operates very well and got the highest 91% precision with 2 layers LSTM with optimization. Additionally, as we increase the number of LSTM layers and the number of tuning parameters, the cross-entropy value decreases and accuracy increases

**Data Visualization Results**

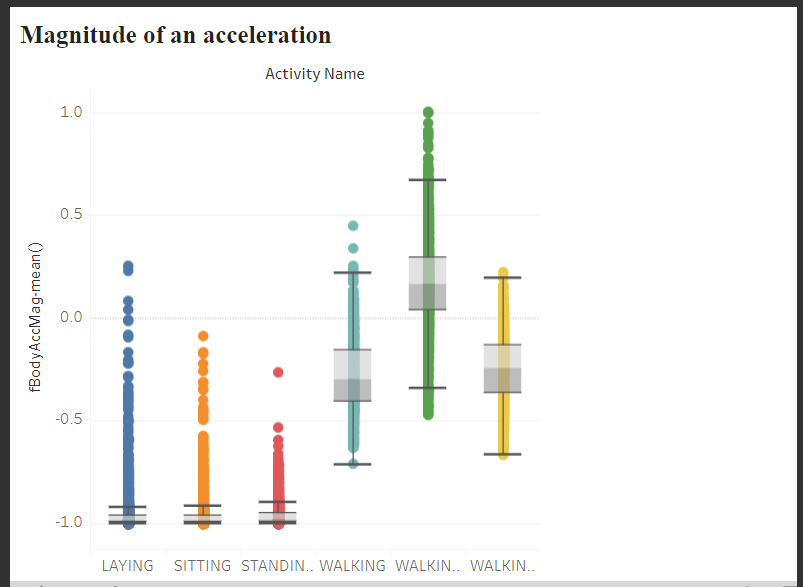
****

The smartphone activity distribution shows that the data is almost equally distributed and balanced, but laying takes the lead!



Almost all participants have more data about walking upstairs than downwards. The participants would take more time to walk upstairs, assuming an equal number of uphill and downhill trips.

Since we know that there are six different classifications, our main challenge is determining whether any data imbalances exist. We may conclude that the data is balanced after plotting the above graph.

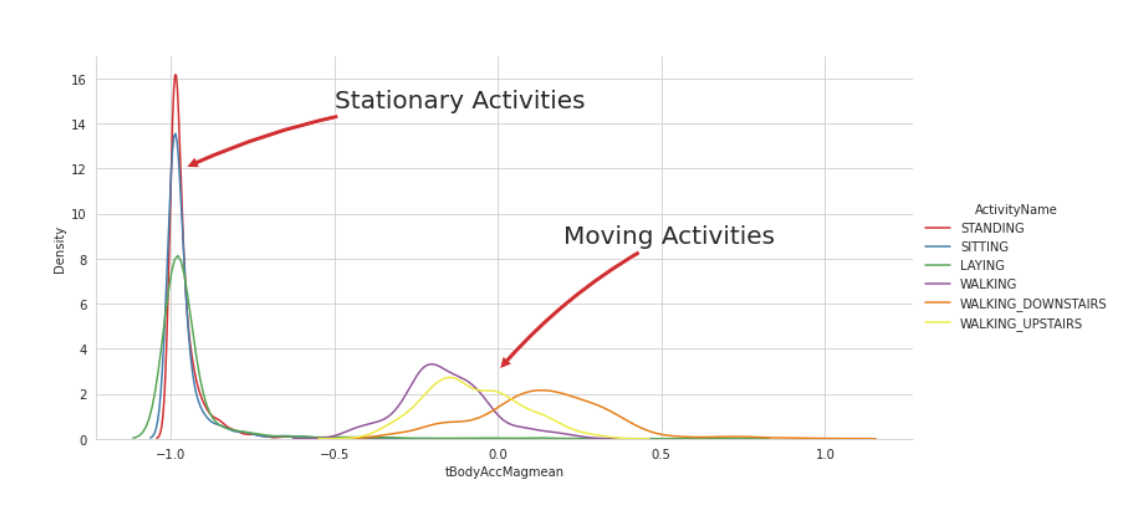


Acceleration magnitude determines what kind of activity a person does.

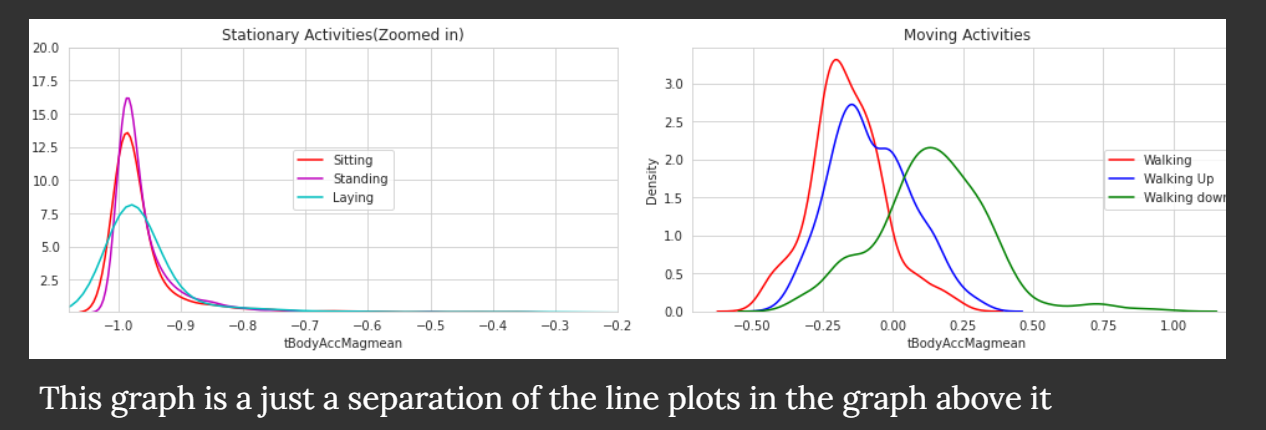
if the value is < -0.8 - standing, sitting, or laying

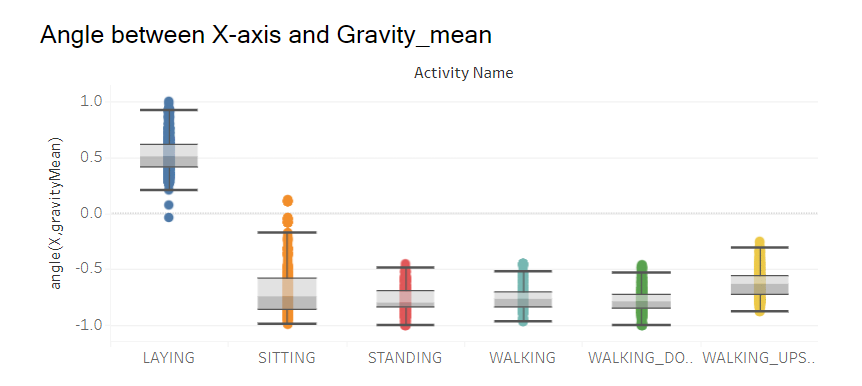
if the value is > -0.6 - walking, walking down, or walking upstairs

if the value is > 0.0 - walking downstairs

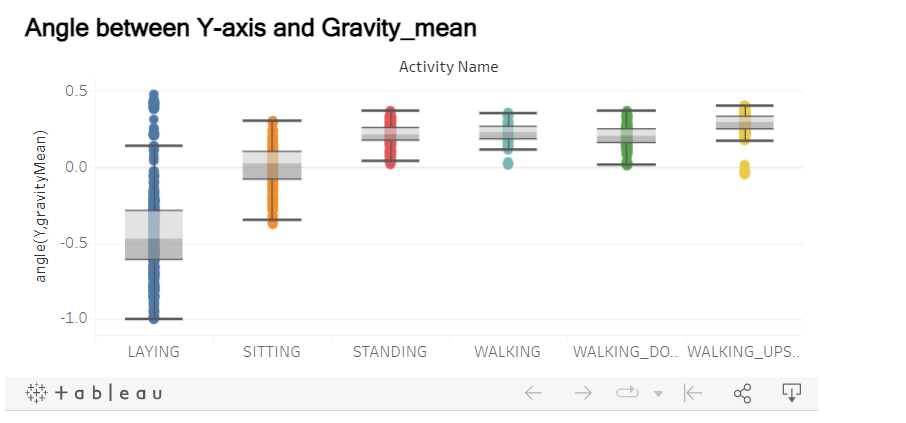


To better comprehend the static and dynamic activities of humans, we are utilizing the "tBodyAccMagmean" (tBody acceleration magnitude feature mean value) function to produce the graph.

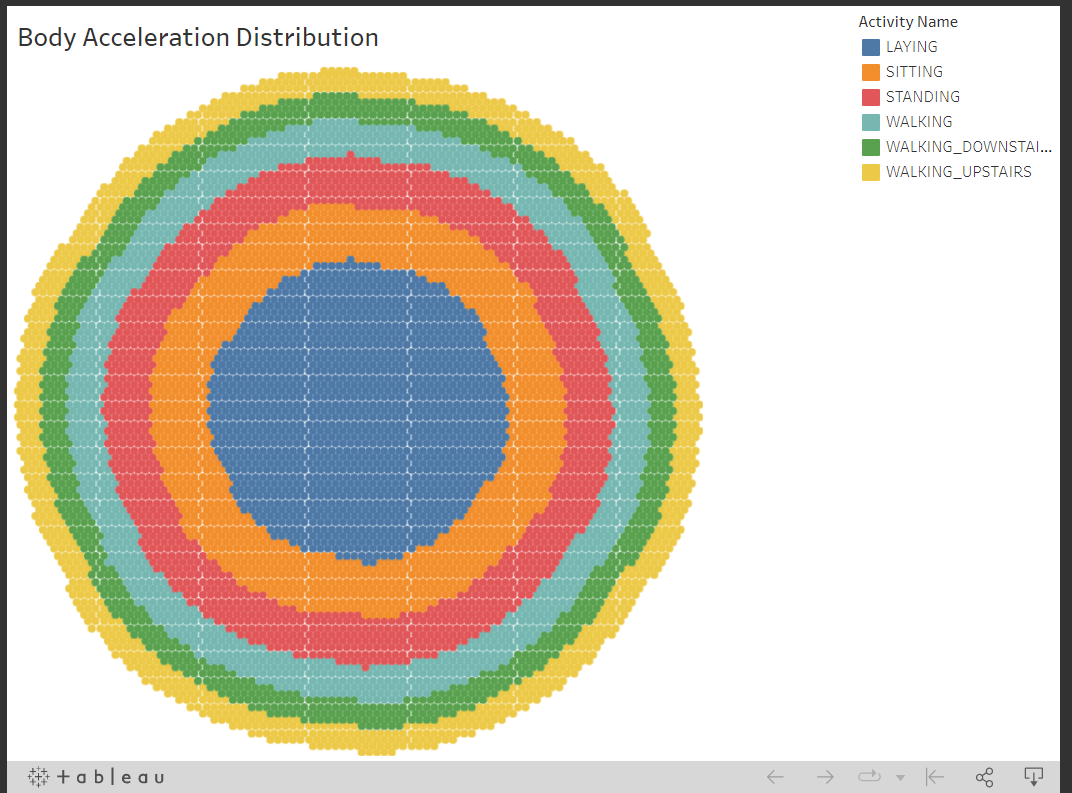




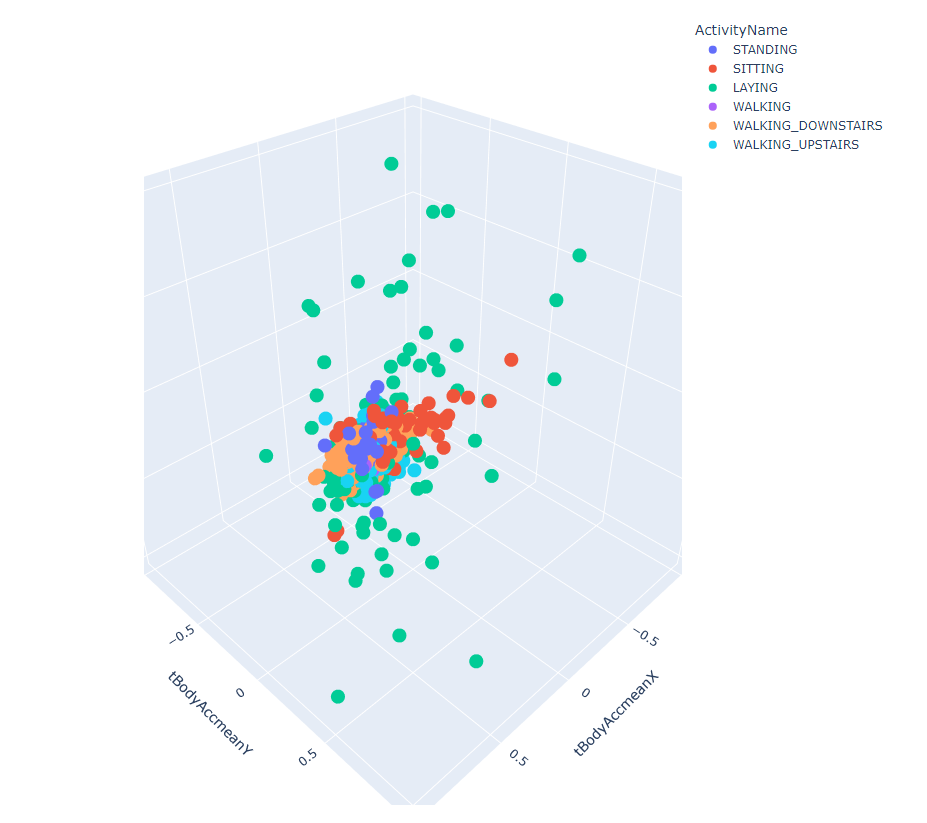
Activity is laying if angleX, gravityMean is greater than 0.



With just one if-else statement, we can categorize every data piece associated with the Laying activity.



This "dart board" represents how the data points which is the distribution of the different activities are stacked in a 2D plot.



This is a 3-D plot of the points that represent the six different activities suspended in the 3 axes - X, Y & Z



We have used a visualization tool in Python which uses a statistical method called "t-sne", that basically separates data points for classification, and helps us understand complex, high-dimensional data by reducing and projecting it in a low dimensional space such as in 2D or 3D.

The Visualization shows how the change in perplexity affects the classification.

**Results and Conclusion**

| **Models** | **Accuracy %** | **Error %** |
| --- | --- | --- |
| Logistic Regression | 95.83 | 4.174 |
| Linear SVC | 96.47 | 3.291 |
| SVM | 96.27 | 3.733 |
| Decision Tree | 86.46 | 12.45 |
| Random Forest | 92.6 | 7.16 |

We used the Linear SVC model because it has the highest accuracy among the models that we have tested.

**Deep Learning LSTM Model Comparison**

| **Model Name** | **Hyperparameter Tunning** | **categorical\_crossentropy** | **Accuracy** |
| --- | --- | --- | --- |
| LSTM With 1\_Layer(neurons:32) | Done | 0.47 | 0.9 |
| LSTM With 2\_Layer(neurons:48, neurons:32) | Done | 0.39 | 0.9 |
| LSTM With 2\_Layer(neurons:64, neurons:48) | Done | 0.27 | 0.91 |

We applied the LSTMs as follows:

a. 1 layer of LSTM

b. 2 layers of LSTM with more hyper tuning of parameters

| **TRUE** | **Laying** | **Sitting** | **Standing** | **Walking** | **Walking\_Downstairs** | **Walking\_Upstairs** |
| --- | --- | --- | --- | --- | --- | --- |
| Laying | 536 | 0 | 1 | 0 | 0 | 0 |
| Sitting | 3 | 420 | 68 | 0 | 0 | 0 |
| Standing | 0 | 122 | 410 | 0 | 0 | 0 |
| Walking | 0 | 0 | 0 | 467 | 28 | 1 |
| Walking\_Downstairs | 0 | 0 | 0 | 1 | 419 | 0 |
| Walking\_Upstairs | 0 | 0 | 0 | 0 | 29 | 442 |

**Limitations and Challenges**

1. The data size is small and limited
2. It is challenging to assess results for each activity independently because the data points for standing and sitting are very clustered in our visualization
3. The possibility of a misclassification can lead to serious consequences

**Future Development Plan**

1. Design an app for our “Ultimate Human Activity Recognition Model”
2. Use the app we launched to directly get sensor data from the user's smart phone.
3. Predicting what a person is doing based on their movements and hand crafting features using sensors and also add the feature to track limping as well which could be useful in healthcare sector