**(Report)**

**Title:- Activity Recognition using Classification Models**

**Author:- Harvinder Singh Sethi**

**Contact:- harshsethi1996@yahoo.com**

**Table of Content**

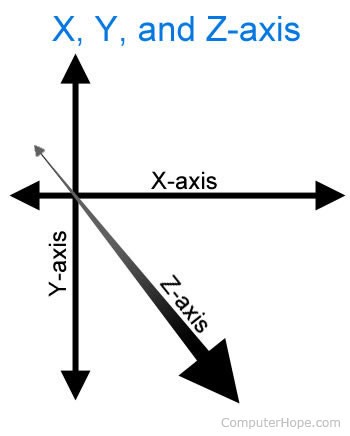
* Abstract
* Introduction
* Methodology
  1. Data Preparation
  2. Data Exploration
     1. Exploring Each column.
     2. Exploring in pairs.
  3. Data Modelling
     1. K-Nearest Neighbor
     2. Decision Tree
* Results Obtained
* Conclusions
* References

# ABSTRACT

Accelerometer are the device that is used in many devices and has a very vast extent. It is used to measure acceleration force with the help of newtons law. There are motion sensors in the accelerometer that detect forces like gravity and sudden change or movement. By using this device, the users can understand the state of the item or person and able to make some decisions, actions and investigations. Example accelerometer in our laptops also protects the hard drive when laptop falls. In this particular study, our goal is to detect a person’s actions (Labels) based on the reading or values of the three dimensions (X, Y, Z) of the accelerator that mounted on individual’s chest. We will compare and classify our data into categories of behavior and actions, then use two classification models K-Nearest Neighbor and Decision Tree and compare the result to decide which model works good and also to make future predictions.

# INTRODUCTION

The dataset collects data of 15 people, with the help of a wearable single accelerometer mounted on their chest. This is a tri-axis accelerator, which measures in three different directions. It records the X, Y, Z acceleration, X senses horizontal motion, Y senses vertical motion and Z senses diagonal motions. Each participant was asked to perform multiple of 7 activities. Please see the photo below for a visual explanation of these axis points.



# METHODOLOGY

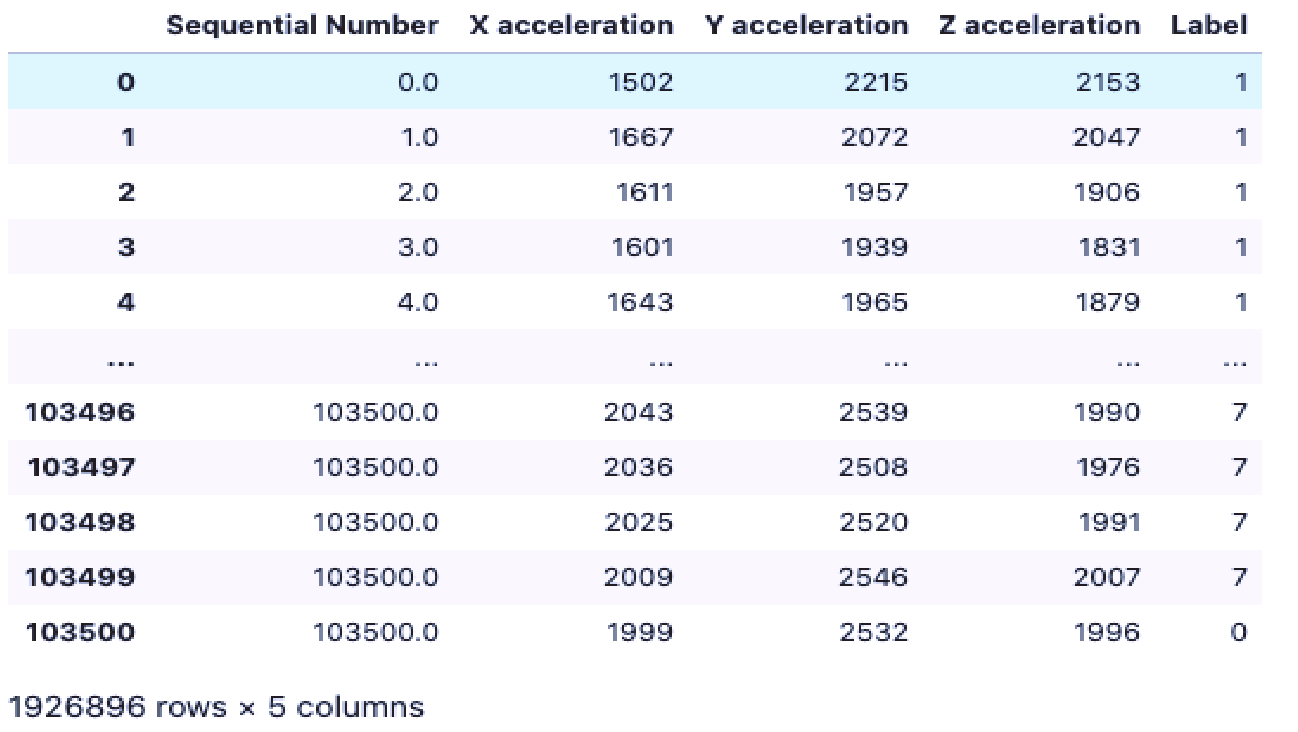
**Data Preparation**

I downloaded the Activity Recognition from Single Chest-Mounted Accelerometer Data Set from UCI

Machine Learning Repository. A center for machine learning and intelligent systems. The dataset is in CSV Format and it includes 15 separate csv files, each file represents 1 particular person’s accelerometers data. Each file contained 5 columns in order represents the readings for, X accelerator, Y accelerator, Z accelerator and last column is filled with numbers from 1 to 7 that present the motion reading of each row of data.

I have imported appropriate packages that are necessary to read the data file, then upload each of 15 CSV file and rename the column name of the data appropriately to make it more readable.

Then used the *pd.concat()* function to combine all 15 data into one dataset name data, this way we can look all data set in one file. Below shows example of the dataset and please see the following data to have a clear understanding of the number representation in column “Label”.



Labels are codified by numbers as follows:-

**1. Working at computer**

## 2. Standing up, walking and going up/downstairs

1. **Standing**
2. **Walking**
3. **Going up/downstairs**
4. **Walking and talking with someone 7. Talking while standing.**

After merging all 15 data, First I checked the data types to see whether they are appropriately defined then I checked to see if there are any missing or null value in the dataset, we found that there are no null or missing values, this is a very clean dataset. Therefore, all rows are filled with proper data and all columns have same length of data. Then I checked the unique value counts by applying value count function to Label(1-7) and I found that there exists a value 0 for label that is impossible value. As our data set is quite big, and the number of rows falling under “0” category are very small, therefore I decided to drop the rows that fall under category “0”, to make more sense of the data. Our goal is to use Machine Learning models to train machines to able to make predictions of a person’s action base on the combination reading from columns X, Y, Z accelerators. The prediction will be predicting the reading into 7 categories.

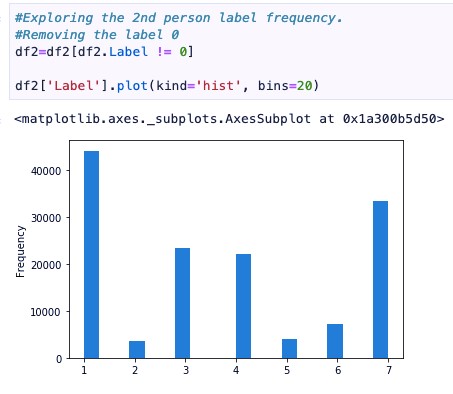
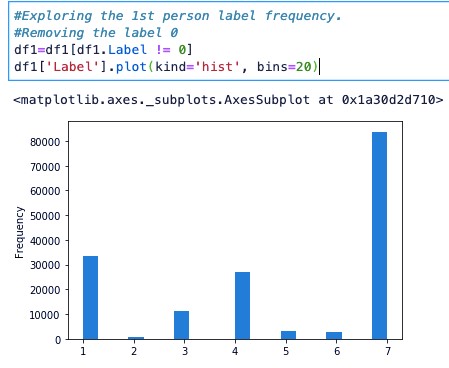
## Before total data was 1926896 After removing “0” total data is 1923177

After this process, I am able to identify my features and target, so I can easily decide that the columns including “X, Y and Z acceleration” can be labeled as “features” and the column named “Label” in dataset can be labeled “target’. Given that the goal is to predict and categorize readings of 7 categories, As the Classification models easily tries to draw some conclusion from the Input values given for training. I have chosen TWO Classification Machine Models, *K-Nearest Neighbor* and *Decision Tree* to train the machine.

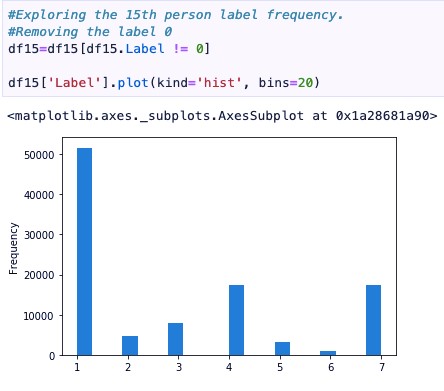
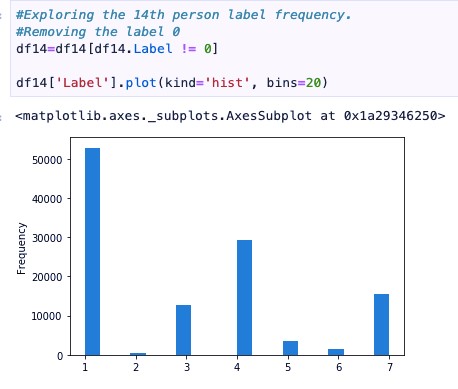
**Data Exploring**

For exploring the data, each column I have plotted many graphs like histograms, boxplots for X, Y, Z Acceleration grouped by Label, line graphs in groups and scatter matrix in my code. But here I will be discussing about only few of them.

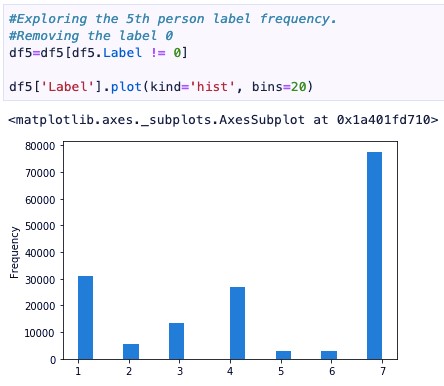
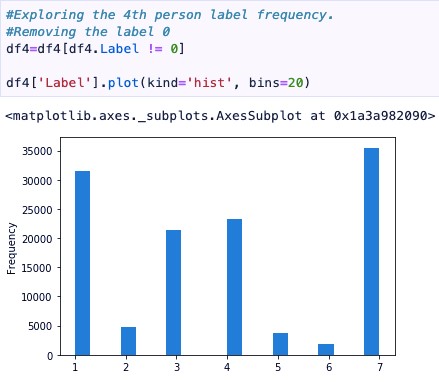
I have plotted histograms for each of the 15 people group by Label to find all of their Label frequency. While Exploring the First person’s Label frequency I found that 1st person is usually found to be “ Talking while Standing”. While second person is mostly “working on computer”.



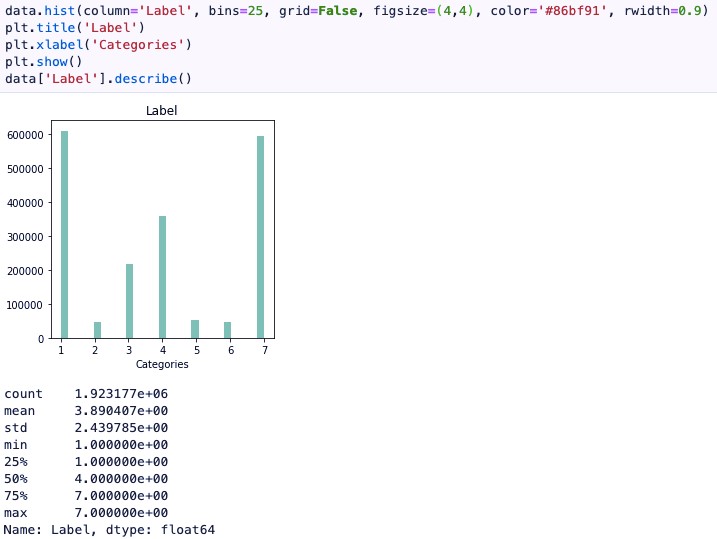
In the similar fashion, 3rd, 6th, 8th, 10th, 11th, 12th, 13th, 14th and 15th person are found to be mostly working on computer, as they have the highest bar height in Label = 1.



While 4th, 5th, 7th, 9th persons are mostly found to be talking while standing as they have the highest bar height in Label = 7



In the end of this I Plotted the histogram of whole data to see on an average of 15 people which Label is Highest and lowest. And whether data Labels are normally distributed or not, and also found the descriptive statistics of it. I found that mostly people are either of Label = 1 i.e.. Working at computer or Label = 7 i.e.. Talking while standing. Mean, standard deviation and maximum values can also been easily seen.

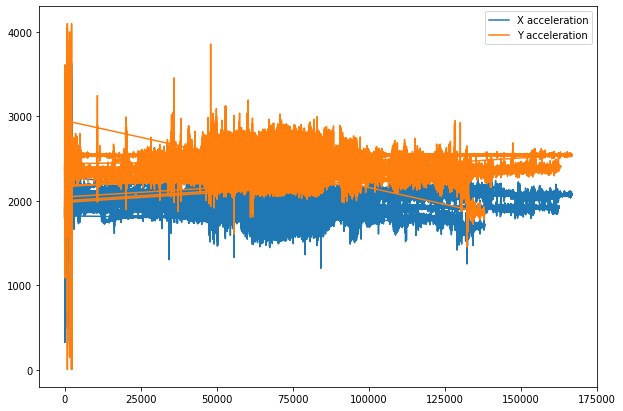


# Paired attributes

In this data there are 3 attributes that are set as our features and 1 target i.e.. Label, I have made 3 combinations of comparison using line graph to compare the results for discovering how a specific attribute perform in contrast to the other.

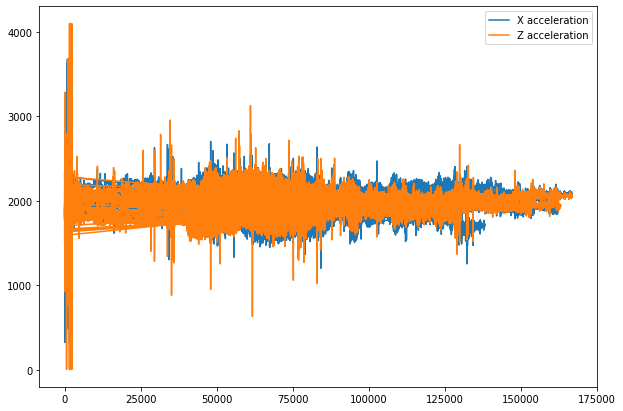
## X and Y Acceleration

Plotting the line graph for X and Y acceleration. We can see that as the Frequency for X increases the Y also increases in the similar fashion, and same goes for the decrease of X and Y.



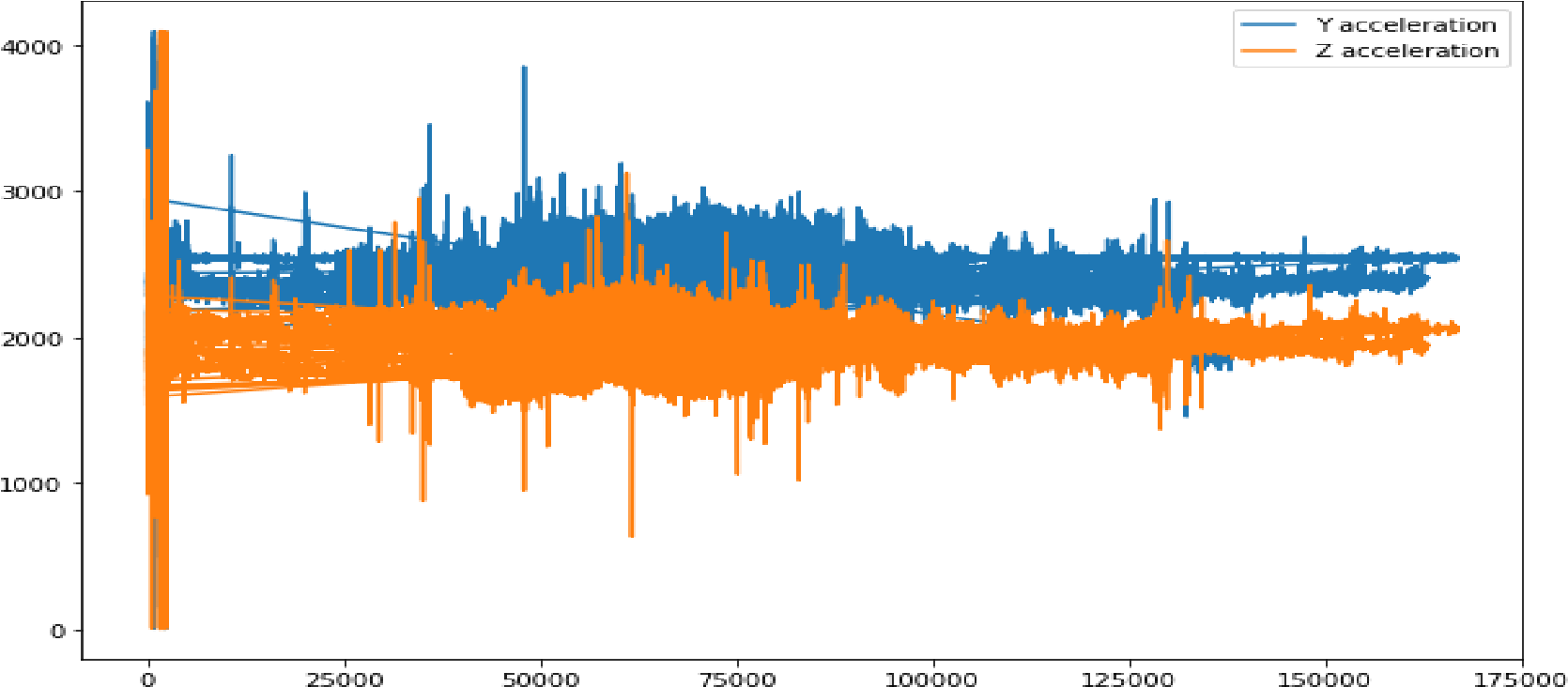
## X and Z Acceleration

Plotting the line graph for X and Z acceleration. We can see that as the Frequency for X increases the Z also increases in the similar fashion, and the values are approximately similar. And same goes for the decrease of X and Z.



## Y and Z Acceleration

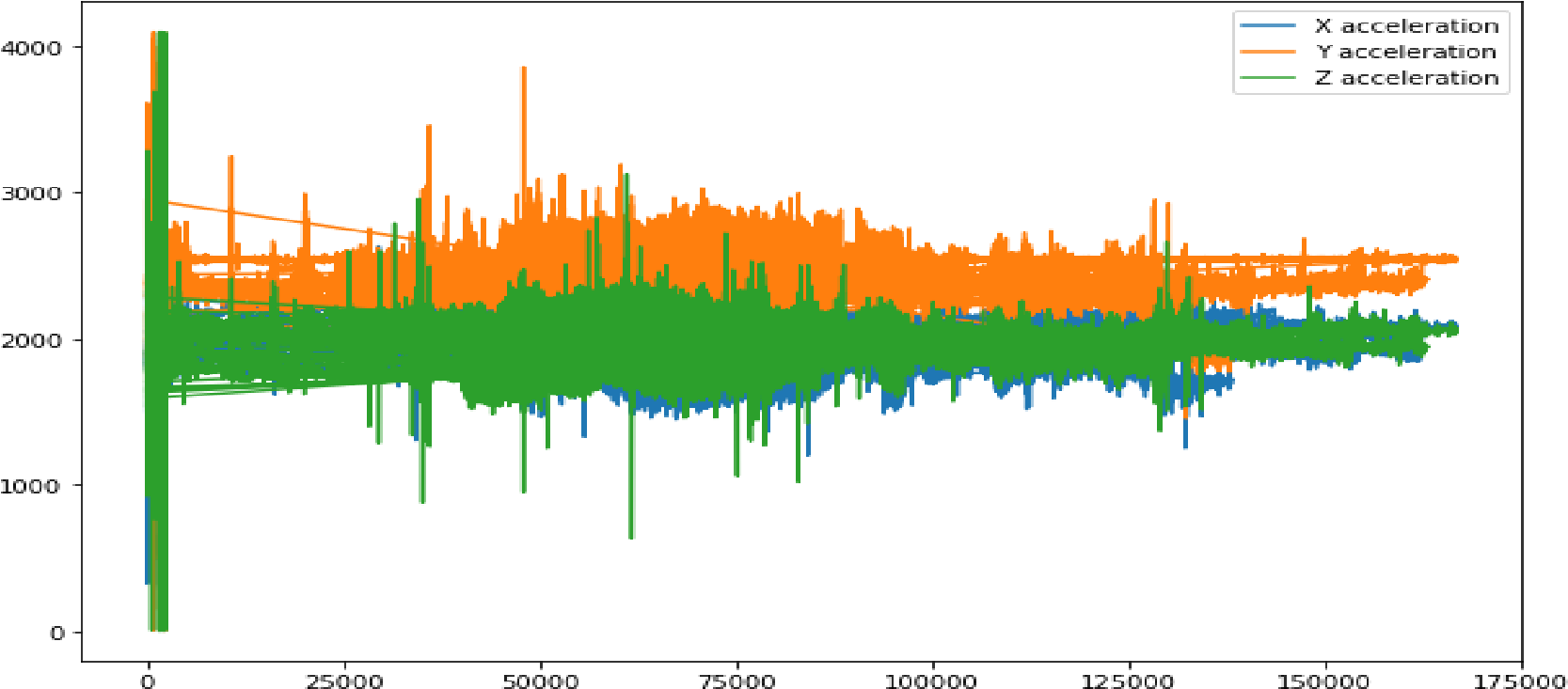
Plotting the line graph for Y and Z acceleration. We can see that Y have a higher value of frequency compare to Z and as the Frequency for Y increases the Z also increases in the similar way, and same goes for the decrease of Y and Z. Overall X and Z axis both seem to fluctuate with each other.



### X ,Y and Z Acceleration

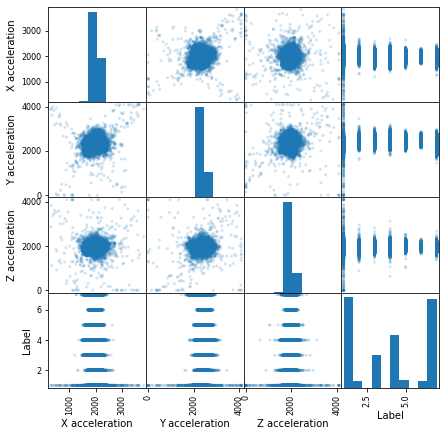
Plotting the line graph for features to understand and compare the x, y , z acceleration together. This line graph compare all three features. From this graph we can see the Y acceleration line have the highest value of hertz due to the placement of the accelerator.

We can assume that Y axis will have an increase (standing up) as the first step. x and z values are not as high as y axis because the position of the accelerator. Z will likely to increase or decrease when a person is going up and down of the stairs or sit and stand or even walk. x will likely to scale when a person is moving during sitting down and moving to another position.



## X, Y, Z Acceleration and Label

A scatter matrix (pairs plot) compactly plots all the numeric variables we have in a dataset against each other one. In Python, this data visualization technique can be carried out with many libraries but if we are using Pandas to load the data. we can use the base scatter matrix method to visualize the dataset. scatterplot to see the correlation either positive or negative.



**Model Selection**

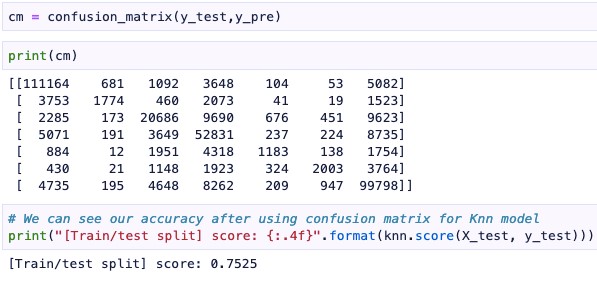
Supervised machine learning algorithms is a commonly used technique to train machines using data(features) to learn the relationships between set of input and output variables. I chose to use supervised machine learning model technique for this dataset as I have the features and targets. I understand my problem statement well and believe that classification models can work well to fit the classification model by applying mixture of the features as pairs to predict outputs(target label). I applied this dataset to two different type of classification models, K-Nearest Neighbors (KNN) and Classification And Regression Tree (CART) also often known as Decision Tree.

## 1) K-Nearest Neighbors

K-nearest neighbors is a simple machine learning Non- parametric algorithm commonly used to classify data with known categories and regression as well. First, we will import the appropriate packages into our Jupyter Notebook and separate our data into training set and testing set. In this study, I have decided to reserve 20% of the data to testing and 80% of the data to training. Then we have to randomly tune our parameter of K value to the most suitable value to give us the highest accuracy rate. Given that our data set is very large, we can try for higher values for our K. In this model, I have tuned all the other parameters with different values as well such as weights='uniform', algorithm='auto', but I find the result of accuracy do not affect much in my tuning process with K values at different numbers as well, but I decided to set my K value at 25 due the highest accuracy result I got. This indeed means, that our prediction will be categories into a category base on 25 closest to its position. After determining the K value, we validate our model with validation technique to find the accuracy of this model.

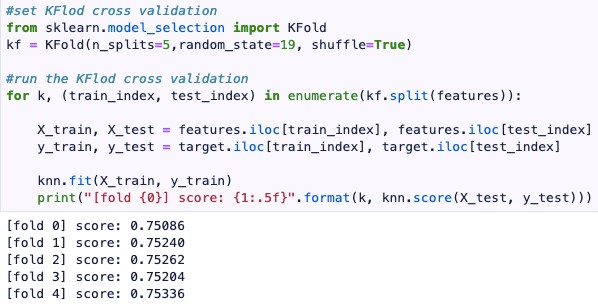
## Confusion Matrix KNN

Below is the result after validating the KNN model with Confusion Matrix, accuracy of model comes out to be **75.25%**



## K-Folds KNN

Below is the result of validating the KNN model with K-Folds, accuracy of model = 75.34% in fold 4

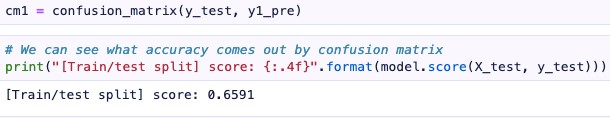


## Classification And Regression Tree

Decision Tree algorithm are referred as Classification and Regression Trees (CART). It is one of the most commonly used tools to supervised machine learning methods in many industries as well as in large scale organizations. It allows us to use both categorical data and numeric datasets into one try to calculate the probability of each category. We will now be using Decision tree on this data set.

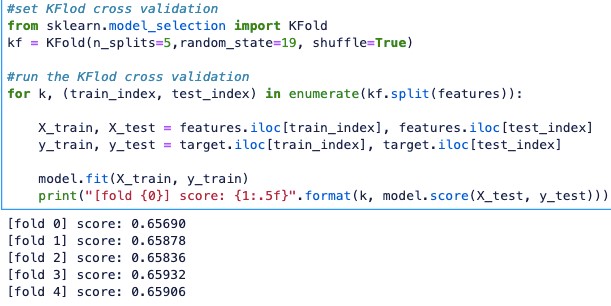
## Confusion Matrix Decision Tree

Below is WHEN validating the KNN model with Confusion Matrix, result of model comes to be **65.91%**



## K-Folds Decision Tree

Below is the result of validating the KNN model with K-Folds, result of model **65.93% in fold 3**

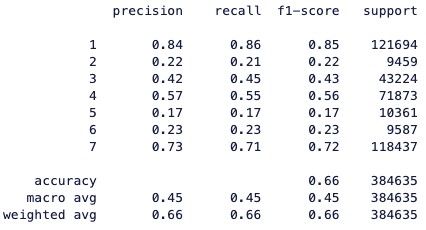
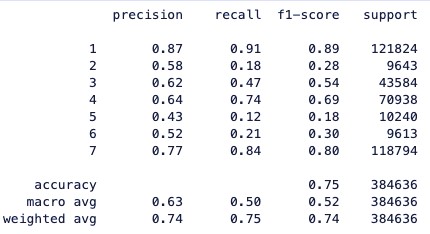


# RESULTS

## KNN VS Decision Tree Classification Report

Classification Report specifically prints out the result of each model, it is good tool used for comparing models and visualization of model results more deeply. From the reports, we can see that the accuracy of KNN is about 75% and the accuracy of Decision Tree is about 66%. I believe the K-Nearest Neighbor Classification Machine Model is the most suitable out of the two due to the higher and more accurate accuracy with both of the validation methods. It is a good algorithm that categorizes data point in the selected category base your K = 25 value selection, and measures the nearest 25 points to determine the category of point it belongs in. Please see report below to see the classification report comparison.

**Precision** shows accuracy of prediction while, **Recall** shows the result of positives that were accurately identified. **F1-Score** determines the weighted average of Precision and Recall.



**Classification Report for KNN. Classification Report for Decision Tree DISCUSSION**

## K-Nearest Neighbor

Advantages & Disadvantage- Easy to understand and apply the model but the K value Tuning is not that easy as the accuracy is mostly dependent on that so should be carefully selected otherwise we can waste our time also.

## Decision Tree

Advantages & Disadvantage- Easy to read and understand for any layman person also, no processing data but the chances of overfitting the models are high in decision tree algorithm.

# CONCLUSION

In conclusion, in this study I found that Y accelerator will have the highest value due to the placement of accelerator. In order to perfume, standing, stand and talk to someone, walking, going up and down stairs each person will require to have change in their vertical positioning. Therefore, the fluctuation result of Y axis should be quite high, due to the fact that it is the fundamental step for many actions from the list. With Z accelerator being the lowest value out of the three, again it is due to the placement of the accelerator. Z accelerator is used to measure diagonally, therefore likely a person will have a higher diagonally movement is during the time of going up and down the stairs not as often standing or working on the computer. By applying our data collected from 15 individual participants to different supervised machine learning model, I have decided K-Nearest Neighbor is more suitable for our dataset, due to the result of higher accuracy compared to result of Decision Tree.

# REFERENCES

1. Ren, Y., 2020. *Practical Data Science*.

Archive.ics.uci.edu. 2020. *UCI Machine Learning Repository: Activity Recognition From Single ChestMounted Accelerometer Data Set*. [online] Available at:

<https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+ChestMounted+Accelerometer#> [Accessed 9 June 2020].

1. "Feature Engineering Made Easy", *O’Reilly Online Learning*, 2020. [Online]. Available:

https://www.oreilly.com/library/view/feature-engineering-made/9781787287600/fd260382-2ea4-41b597c2-03e9ff8c9afd.xhtml. [Accessed: 10- Jun- 2020].

1. Contributor, R., 2020. *Accelerometers: What They Are & How They Work*. [online] livescience.com. Available at: <https://www.livescience.com/40102-accelerometers.html> [Accessed 9 June 2020].
2. Definitions, H. and Hope, C., 2020. *What Is Horizontal?*. [online] Computerhope.com. Available at:

<https://www.computerhope.com/jargon/h/horizont.htm> [Accessed 9 June 2020].