



**UNIVERSITY OF SURREY**

**Department of Computer Science**

**MSc Dissertation**

**Human Activity Recognition via Data Analysis**

**A project supervised by Dr. Robert Granger**

**RAJ KANNAN CHENTHIL ARUMUGAM**

**URN: 6607133**

**I confirm that this is my work and that has not been earlier presented back as a portion of any assignment. The guidelines on plagiarism stated in the student manual have been understood and ensued in this report. I have further combined all references implemented in this project at the end.**

**-RAJ KANNAN CHENTHIL ARUMUGAM**

# **Abstract**

This project aims to enhance the experience of people indulge in physical activity

through the means of data analysis with the help of data derived by tracking the various

physical activities performed by the user and reporting out the activities with various

visual interfaces and various analytical mechanisms

Our project main motto is to create a multiclass classification model which would be trained upon the dataset with only six different activities and upon training it would be able to distinguish the activity given a test dataset, and all of which we mentioned was room of improvement with the possibility to embed the model into a backend application which powers a user faced mobile application.

The main goal of my element of the project was to build a predictive machine learning

model which can detect user activities upon proper training.

The inputs for the models are a training data set generated by various tracking devices

through volunteers.

The application aims to give user the prediction of the activity and visualisation of

various activities that user perform

Physical activities lead to healthy lifestyle. Another factor that is important is the amount

of food but at recent times people have a better accessibility to junk food and our favourite

food delivered at our doorstep, and the above factors are mitigated by recognising human

activity which is calculated by data analysis and reporting the activity in visual format

In order to prevent these complications, the user must ensure proper physical activity at

regular basis.

Unfortunately, there are many barriers that prevent people from engaging in physical

activities

This project also addresses the major barriers such as lack of motivation and helps to overcome unhealth lifestyle of humans

In doing so, a solution is developed with the aim of easing the process of tracking down

the activity

**Keywords :**

**Acknowledgments**

I would like to thank my supervisor, Robert Granger for his guidance and support throughout my dissertation. He not only helped me in finalizing my topic, but also gave me insightful ideas. He encouraged me to look at different aspects regarding my topic to get interesting results. He has been very supportive and stayed in contact throughout, which enabled me to complete work. I would like to thank him for sharing his knowledge and guiding me in the right direction.

I would also like to thank all my teachers for their countless efforts, my friends and family for all their support and care throughout my degree, especially during these difficult times caused by the pandemic, it would not have been possible to achieve all that I have without their encouragement.

I would like to thank my supervisor, Robert Granger for his direction and assistance during my dissertation. He has not only solely aided me in achieving my topic, but also gave me insightful ideas. He urged me to look at various perspectives concerning my topic to produce interesting outcomes.

He has remained highly supportive and stayed in touch, which allowed me to complete my work. I would like to appreciate him for yielding his experience and leading me in the correct direction.

I would furthermore wish to thank all my lecturers for their innumerable efforts, my friends and my parents for their care and help during my degree.

Contents

[**Abstract** 3](#_Toc78207046)

[**Introduction** 6](#_Toc78207047)

[**Literature Review** 7](#_Toc78207048)

## **Introduction**

**Overview**

With the advent of artificial intelligence into diverse fields especially into mobile

computing and desktop computing, devices had become smart and these smart

technologies can be incorporated into user’s daily life and can be embed into these

devices to track the user’s activity and guide them to be in good direction in order to

maintain healthy lifestyles. Our motto behind this project is to mostly encourage,

remind and guide users in good direction of healthy lifestyle.

As there is a famous saying ‘health is wealth’ In this generation of technology

millennium, people had a luxury to get everything delivered to their door with just a

movement of fingertips thereby discouraging users from involving any kind of physical

activities on top of that the variety of junk foods readily available to deliver making them

more calorific intake than they are burning down, and the final outcomes varies from

small health glitches like obesity to some serious anomalies like cardiac arrest etc.,

Our project is based on Multiclassification Problem on balanced dataset. This notebook

is to build a model that can predict whether a person is Laying, Standing, Sitting,

Walking, Walking\_upstairs, or Walking\_downstairs

**Objectives & Hypothesis**

**The main objective is to create a machine learning model of type Multi class classification; and training it on the dataset and running the tests and generate test results**

**We directly get into cleansing part to check anomalies like null values duplicates etc.,. and perform basic analysis on it**

**My aim or motto is to create a machine learning model of type**

**multiclass classification which will be able to distinguish six different activities.**

Initially, the details within the dataset arises from the measurements taken of the

accelerometer, gyroscope, magnetometer, and GPS of the smartphone.

In this project we are going to create a machine learning of type multiclass classification

which would be trained on balanced dataset and upon training would be able to predict

the user action with given accuracy.

This model should be input with a test data set of user activity and can predict and infer the user actions as an output

The first aim of the project is to do exploratory data analysis on the data set generated

by a group of 30 volunteers, each one of them perform six different activities. We

analyse the data to check whether it is balanced or imbalanced and generate

visualization on the raw data and removing duplicates and perform data cleansing

This report goes into various techniques of data cleansing and various technologies to

process the data like panda’s data frame to enhance the data integrity and reduce the

anomalies.

The second aim of the project is to create and train machine learning model based on

the problem of multiclass classification problem based on balanced data set. The proper

training model should be able to predict the activity of the user given the test input

Numerous other techniques were further analysed and are reviewed within the project.

The latest technique is assessed exerting an individual evaluation method and is

demonstrated to possess small fault or inaccuracy (i.e., over/under segmentation, False

Positives and False negatives).

In the end the project endorses what shall be appropriate that helps to enhance the

current project over the future.

**Abstract and introduction over**

### **Literature Review**

**The objective of this study is to analyse a dataset of human activities of about 30 volunteers and try to analyse the same and draw insights and predict the activity using Machine Learning. We also try to detect if we could identify the participants from their walking styles and try to draw additional insights. The potentials of such a study can be exploited to scenarios such as activity detection, monitoring persons for signs of fatigue, distinguishing one individual from another , with possible deployment in highly sensitive and secure workplaces etc.**

**To obtain the dataset for model training , experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING)**

**wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.**

**(Dataset is obtained from Kaggle**)

**Data Preparation : (used in another domain )  
The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain.**

**For each record in the dataset it is provided:**

**- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.**

**- Triaxial Angular velocity from the gyroscope.**

**- A 561-feature vector with time and frequency domain variables.**

**- Its activity label.**

**- An identifier of the subject who carried out the experiment.**

**----**

**An another approach is the novel energy efficient approach and it is used for the recognition of human activities that targets assisted living applications such as remote patient activity monitoring for the disabled and the elderly**

**The method exploits fixed-point arithmetic to propose a modified multiclass Support Vector Machine (SVM) learning algorithm, Random Forest model ,Decision tree model**

**Kernel SVM model ,Linear SVM model ,Logistic regression model with Hyperparameter tuning and cross validation – (Google and full )**

**with respect to the conventional floating-point based formulation while maintaining comparable system accuracy levels. Experiments show comparative results between this approach and the traditional SVM in terms of recognition performance highlighting the advantages of the proposed method. ( Read again for changes )**

**Why this analysis is required as of today ?**

**Remote patient monitoring is nowadays allowing disabled and elderly patients a continuous health and well-being supervision while they perform regular activities throughout the day. Recent population benchmarks show that world population is aging rapidly. As an example, the projections of changes in population structure by main age groups in Europe are showing that by 2060 the elderly (namely people over 65 years) will be near 30% of its population. This represents an alarming growth of more than 70% of this age group, bringing new challenges to the research community, which aims to find beneficial alternatives for ensuring healthy living to the people.**

**Human Activity Recognition (HAR) is a research field that aims to identify the actions carried out by one or more subjects through the gathering and understanding of context information about the user state and its surrounding environment. This is done by the exploitation of environmental and on-body sensors, and distributed computing resources.**

**MC-HF-SVM allows to vary fixed-point number representation (number of bits) to control over model accuracy and complexity, leading to improvements in terms of both recognition accuracy**

**Several approaches have been previously proposed in literature for the recognition of human activities covering diverse application domains such as healthcare, smart homes, ubiquitous computing, ambient assisted living, surveillance and security**

**These approaches can be categorized according to many different criteria: by sensor type, which is reliant on the signals measured (e.g. inertial, vision-based and by sensor location, namely external sensing when sensors are located in fixed positions in the environment by modelling principle, which can be data- or knowledge-driven depending on whether the HAR models are built given pre-existing datasets or from the exploitation of prior knowledge regarding a particular domain**

**Machine Learning approaches that have been already applied for the recognition of activities include: Decision Trees and Support Vector Machines (SVMs) . Our approach exploits SVMs for the classification of activities similarly to other works which have successfully employed them Furthermore, they have shown to be effective in heterogeneous types of recognition such as in ……….**

**Different approaches have been explored for targeting this issue. The two most commonly used methods are:**

**Kernel & SVM**

**One-Vs-All (OVA) and OneVs-One (OVO), where each particular class is compared using a binary classifier against the rest of classes either all together (OVA) or one by one (OVO) to determine the most likely class for each new sample. In particular, we have selected the One-Vs-All (OVA) method and customized it to the fixed-point arithmetic case. The performance of the OVA approach is comparable to the OVO classification as it has previously been confirmed . Moreover its produced model needs less memory when compared against the OVO method, bringing up an advantage taking into account the limited resources available**

**Moreover, limited methods of modern classification are usually preferred for specific-purpose applications if they demonstrate similar performance to traditional processing Nowadays, it has become particularly interesting to retake these approaches and apply them in the development of ….. which are highly demanding in terms of energy consumption and system resources management.**

**This method was designed for binary classification problems by employing fixedpoint arithmetic in the feed-forward phase of the SVM classifier, with the purpose of allowing its use in hardware-limited devices. In this work, we adapt the model for the multiclass problem targeted towards HAR on smartphones.**

**---------------**

**In the last decades several works have been devoted to adapt Machine Learning (ML) approaches to specific hardware platforms and, in particular, to analyze the effects of parameter quantization on the training and feed-forward phases**

**Motivations for these activities are usually linked to application-specific requirements but also to the basic principle of the Statistical Learning Theory (SLT) where we have to search for the easiest model that correctly classifies the available data. The introduction of bit–based hypothesis spaces brings widespread benefits on the learning process of classifiers (i.e. classes of functions where models are described through a limited number of bits). This is due to the fact that reducing the number of bits largely influences the complexity of the hypothesis space which is a key issue in Machine Learning .**

**If we are able to reduce the complexity of the hypothesis space without affecting the ability of the algorithm to learn the function with low empirical error, in practice, we are able to learn more effectively**

**….**

**(Write a subtopic)**

**including multitasking and a variety of sensors, in addition to the basic telephony. The integration of these mobile devices in our daily life is growing rapidly, and it is envisaged that such devices can seamlessly monitor and keep track of our activities, learn from them and assist us in making decisions. Such assistive technologies can be of immense use for remote health care, for the elderly, the disabled and those with special needs, if there are autonomous and intelligent. However, currently, though there is good capacity for collecting the data with such smart devices, there is limited capability in terms of automatic decision support capability and making sense out of this large data repository. There is an urgent need for new data mining and machine learning techniques to be developed to this end. In this paper we propose a new scheme for human activity recognition using smart phone data, with potential applications in automatic assisted living technologies. Activity recognition systems aim to identify the actions carried out by a human, from the data collected from the sensors and the surrounding environment.**

**The current smart phones have motion, acceleration or inertial sensors, and by exploiting the information retrieved from these sensors, recognition of activities and events can be recognized. Automatic recognition of activities and events is possible by processing this sensor data with appropriate machine learning and data mining approaches. Rest of the paper is organized as follows.**

**The details of the publicly available activity recognition data set used in this work are described in Section 2. Section 3 discusses the relevant background work done in this area, and the proposed automatic activity recognition approach is discussed in Section 4. The experimental validation of the proposed approach is described in Section 5, and the paper concludes with some outcomes achieved from this work and the plan for future research**

**It is unrealistic, however, to expect in general home settings for people to wear them for their daily activities, because of their difficulty, time and convenience to wear them on daily basis, though, such elaborate setups can enhance the activity recognition performance.**

**Smart phones have an advantage because of their ease and convenience, along with the capability of multiple sensors on the phone, which can be exploited for activity recognition.**

**Appropriate machine learning and data mining methods need to be developed for processing these multiple sensor signals from smartphones for automatic and intelligent activity recognition. Though there have been several machine learning methods available7, 8, 9, 10, it is not clear, which algorithm can performs better for activity recognition with smartphones. If automatic activity recognition systems can be built based on intelligent processing of multiple sensor features on smart phones, it will be a great contribution to eHealth area, particularly for remote activity monitoring and recognition in aged care and disability care sector. In this article, we examine several new machine learning and data mining approaches based on decision trees and ensemble learning techniques including random forests and random committee and**

**compare them with traditional naïve Bayes classifier and unsupervised k-Means clustering approaches for processing smartphone sensor signals for activity recognition. The experimental evaluation of the proposed schemes with a publicly available smartphone activity recognition database 1 shows a significant improvement in recognition performance of proposed machine learning and data mining approaches, as compared to other methods proposed in the literature for smartphone based activity recognition. Next Section describes the details of machine learning techniques used for developing smartphone based automatic activity recognition system**

**…**

**(Not in literature review)**

**Although the use of numerous sensors could improve the performance of a recognition algorithm, it is unrealistic to expect that the general public will use them in their daily activities because of the difficulty and the time required to wear them. One drawback of the smartphonebased approach is that energy and services on the mobile phone are shared with other applications and this become critical in devices with limited resources. ML methods that have been previously employed for recognition include Naive Bayes, SVMs, Threshold-based and Markov chains [6]. In particular, we make use of SVMs for classification as it was also used in [7] and [8]. Although it is not fully clear which method performs better for AR, SVMs have confirmed successful application in several areas including heterogeneous types of recognition such as handwritten characters [9] and speech [10]. In ML, fixed-point arithmetic models have been previously studied [11, 12] initially because devices with floating-point units were unavailable or expensive. The possibility of retaking these approaches for AmI systems that require either low cost devices or to allow load reduction in multitasking mobile devices has nowadays become particularly appealing. Anguita et al. in [13] introduced the concept of a Hardware-Friendly SVM (HF-SVM). This method exploits fixed point arithmetic in the feed-forward phase of the SVM classifier, so as to allow the use of this algorithm in hardware-limited devices.**

**-----**

**End of LR**

**In this paper, we extend this model for multiclass classification. The SVM algorithm was originally proposed only for binary classification problems but it has been adapted using different schemes for multiclass problems such as in [9]. In particular, we have chosen the One-Vs-All (OVA) method as its accuracy is comparable to other classification methods as demonstrated [14], and because its learned model uses less memory when compared for instance to the One-Vs-One (OVO) method. This is advantageous when used in limited resources hardware devices.**

#### **Methodology & implementation**

#### **Software development Methodology** is the software development process that is followed In order to develop, test, and deliver the software in the defined deadlines while meeting quality standards and user expectations.

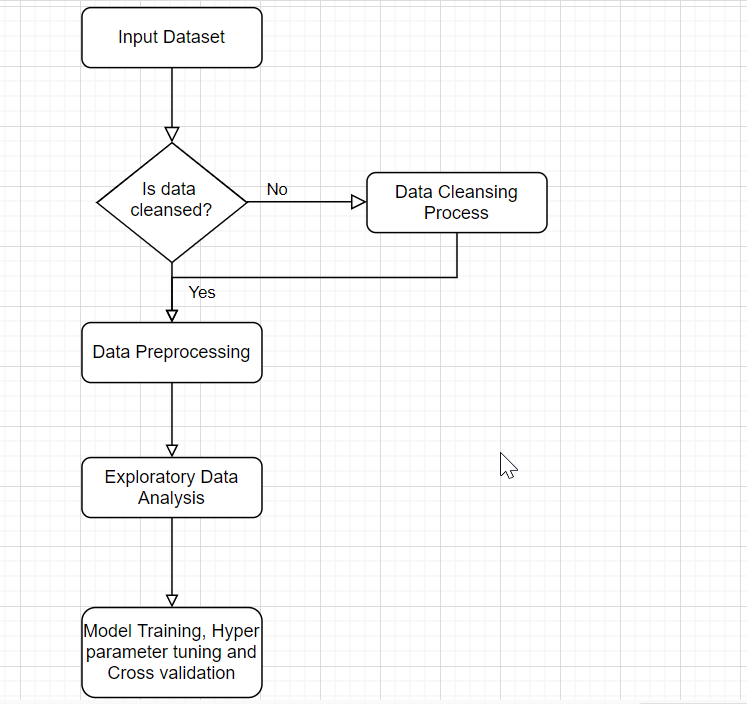
It is the process in which people divide the software development work into smaller independent, parallel/sequential steps to distribute among a team and enable members of a team to work independently and collaboratively in a parallel manner.

It also includes the requirements gathering and defining the features out of them and designing and implementation of the features. Its also technically called as **Software Development Lifecycle (SDLC)**

Based on the review of similar research and papers on **Multi-class classification Modelling On Balanced Dataset** the following roadmap has been drawn and implemented in **Agile** **Methodology**

Workload is broken down into sequence of independent steps which can be implemented either sequentially or in parallel in course of multiple sprints.

A flow chart describing the process flow involved in the developing of multiclass classification model is shown below in the figure



Below is the breakdown of the steps

**3.1. Obtaining Dataset**

**3.2. Data Cleansing and data preparation**

3.2.1. Removing outliers

3.2.2. Performing check for null values, NaN and duplicate values

3.2.3. Check for data imbalance

**3.3. Exploratory Data Analysis**

**3.4. Data Preprocessing**

**3.4.1. Categorical encoding of data variables**

**3.4.2. Normalization of data**

**3.4.3. Splitting and preparation of data set for testing and training**

**3.5. Models, Training, Hyperparameter tuning and cross validation**

**3.5.1. process of Logistic Regression**

**3.5.2. Naive Bayes**

**3.5.3. K-Nearest neighbour**

**3.5.4. Decision Tree**

**3.5.5. Random Forest**

**3.5.6. Support Vector Machine**

**Problem Specification:**

Human Activity Recognition is an example of Multiclass classification problem on Balanced Dataset. In this project we are aiming at building a model upon proper training can predict whether a person is performing any of the following activities.

LAYING

SITTING

STANDING

WALKING

WALKING UPSTAIRS

WALKING DOWNSTAIRS

**MultiClass Classification:**

A type of classification task which involves more than two classes.

Eg: Our task involves 6 classes where each of the class represents an activity performed by the user.

Multiclass classification involves making an assumption that none of the sample labels i.e test or train dataset can carry two labels i.e each user is allowed to perform a unique activity which has to be classified by the model, its also called as **Multinomial classification** as it involves multiple terms to classify.

**Balanced Dataset:**

Balanced Dataset refers to a dataset of multiple classes where all the classes are equally represented, whereas they are unequally represented to a greater degree in Unbalanced dataset

Eg: dataset which represents population statistics is usually balanced by gender because here the two classes representing male and female gets equal representation.

Dataset of application logs usually is un-balanced on as the successful login attempts usually overweigh that failure login attempts and both the classes has un equal representation.

**3.1. Obtaining Dataset:**

Input Dataset for this experiment is obtained from a publicly available repo

<https://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions>

This dataset is a Multivariate, timeseries dataset obtained by carrying out experiments on 30 volunteers falls within age group of 19-48 years, each of the volunteer referred as a subject performed a protocol of activities composed of 6 basic activities, This experiment also includes transitions of postures from static postures to dynamic and within static. Lay-stand, Lay-sit, Sit-stand etc., Participants are wearing a device like smartphones which captures 3-dimensional acceleration in linear direction and 3-dimensional angular velocity at a constant rate of 50Hz, using the accelerometer and gyroscope embedded within the device. The timeseries data generated was labelled to different categories. The obtained dataset was then randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

**Preparation and description of time series data:**

The time-series data obtained from the experiment contains 563 individual columns(features)

1. The features selected for this database come from the **accelerometer**and **gyroscope**3-axial raw signals **tAcc-XYZ**. These time domain signals (prefix **'t'** to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise.
2. Similarly, the acceleration signal was then separated into body and gravity acceleration signals **(tBodyAcc-XYZ and tGravityAcc-XYZ)**using another low pass Butterworth filter with a corner frequency of 0.3 Hz.
3. Subsequently, the body linear acceleration and angular velocity were derived in time to obtain **Jerk signals (tBodyAccJerk-XYZ**and **tBodyGyroJerk-XYZ)**. Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag)

jerk is the rate at which an object's acceleration changes with respect to time

1. Finally a Fast Fourier Transform (FFT) was applied to some of these signals producing

fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyGyro-XYZ, fBodyAccJerkMag, fBodyGyroMag, fBodyGyroJerkMag.

(Note the 'f' to indicate frequency domain signals).

These signals were used to estimate variables of the feature vector for each pattern:

1. **'-XYZ'**is used to denote 3-axial signals in the X, Y and Z directions.
   * tBodyAcc-XYZ
   * tGravityAcc-XYZ
   * tBodyAccJerk-XYZ
   * tBodyGyro-XYZ
   * tBodyGyroJerk-XYZ
   * tBodyAccMag
   * tGravityAccMag
   * tBodyAccJerkMag
   * tBodyGyroMag
   * tBodyGyroJerkMag
   * fBodyAcc-XYZ
   * fBodyAccJerk-XYZ
   * fBodyGyro-XYZ
   * fBodyAccMag
   * fBodyAccJerkMag
   * fBodyGyroMag
   * fBodyGyroJerkMag`
2. The set of variables that were estimated from these signals are:
   * mean(): Mean value
   * std(): Standard deviation
   * mad(): Median absolute deviation
   * max(): Largest value in array
   * min(): Smallest value in array
   * sma(): Signal magnitude area
   * energy(): Energy measure. Sum of the squares divided by the number of values.
   * iqr(): Interquartile range
   * entropy(): Signal entropy
   * arCoeff(): Autorregresion coefficients with Burg order equal to 4
   * correlation(): correlation coefficient between two signals
   * maxInds(): index of the frequency component with largest magnitude
   * meanFreq(): Weighted average of the frequency components to obtain a mean frequency
   * skewness(): skewness of the frequency domain signal
   * kurtosis(): kurtosis of the frequency domain signal
   * bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.
   * angle(): Angle between to vectors.
3. Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean tBodyAccMean tBodyAccJerkMean tBodyGyroMean tBodyGyroJerkMean

**3.2.1. Removing outliers**

As the multiclass classification has to be performed on balanced dataset, it has to be thoroughly verified for the presence of any outliers and then the outliers has to be taken off, after getting rid off the outliers we are left with the timeseries dataset squeezed between -1 to 1

**3.2.2. Performing check for null values, NaN and duplicate values**

The train and test datasets has to be loaded into pandas for easy exploratory analysis and data cleansing

Duplicate values within a dataset can be identified by calling **.duplicated()** method on the pandas dataframe.

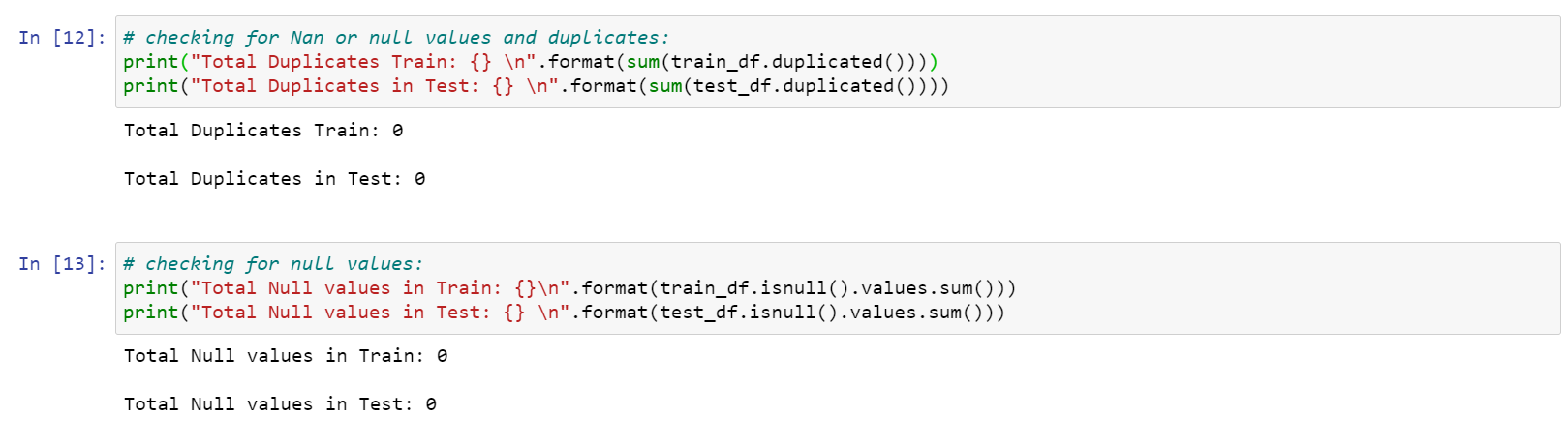
Null/NaN values can be obtained by **.isnull()** method of dataframe

Eg:

**train\_df.dulicated()** gives us the count of duplicated values in the train dataset

**test\_df.isnull().values.sum()** gives the count of total number of null values within the test dataset

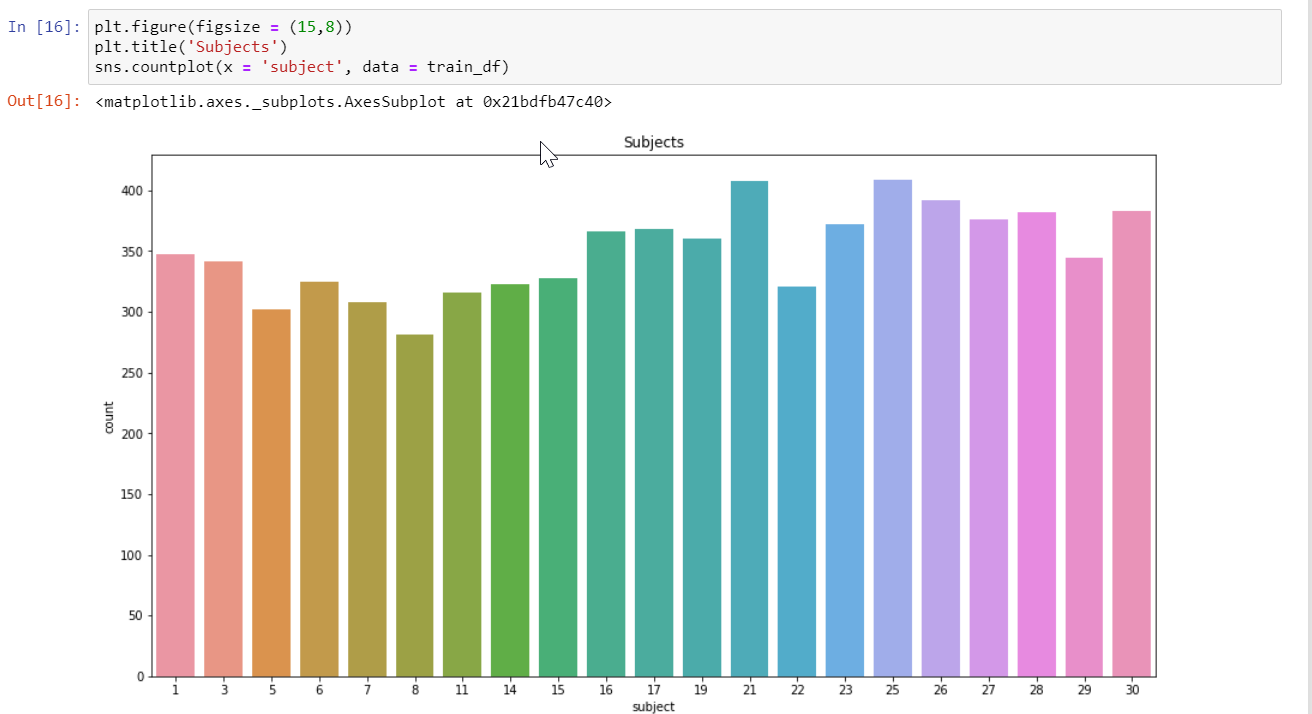
A snap of cleansed dataset with no duplicates and zero null values.



**3.2.3. Check for data imbalance**

As the classification takes on balanced dataset, the check for data balance has to be performed in order to verify that all the classes within the dataset re equally weighted or distributed which indicates that dataset is well-balanced

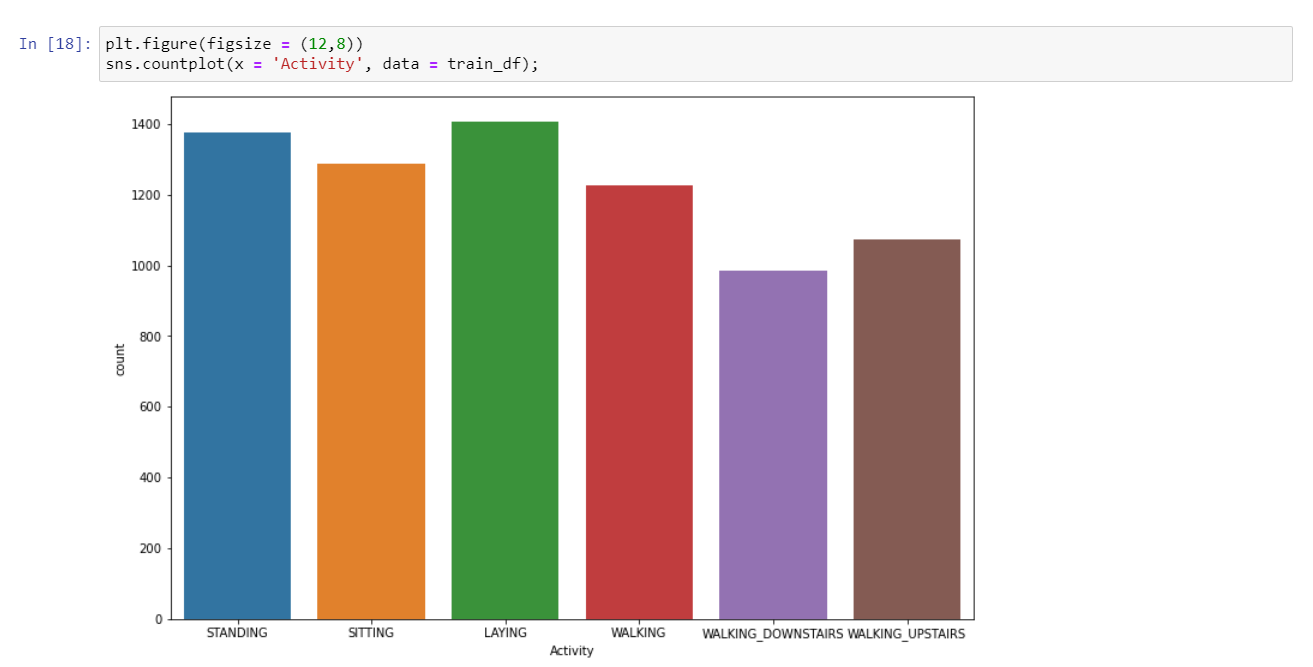
The following plot represents the data balance among the data collected from 30 volunteers each referred as a subject



The following plot represents the visualization of various activities performed by all the subjects.



And the following plot represents the data balance between 6 different activities performed which clearly indicates that our dataset is well balanced as each of the six different classes are equally distributed and none of the class is over-weighed or under-weighed either.



**3.3. Exploratory Data Analysis**

Exploratory data analysis is a step in which we perform basic analysis on the dataset in order to get the following insights into data

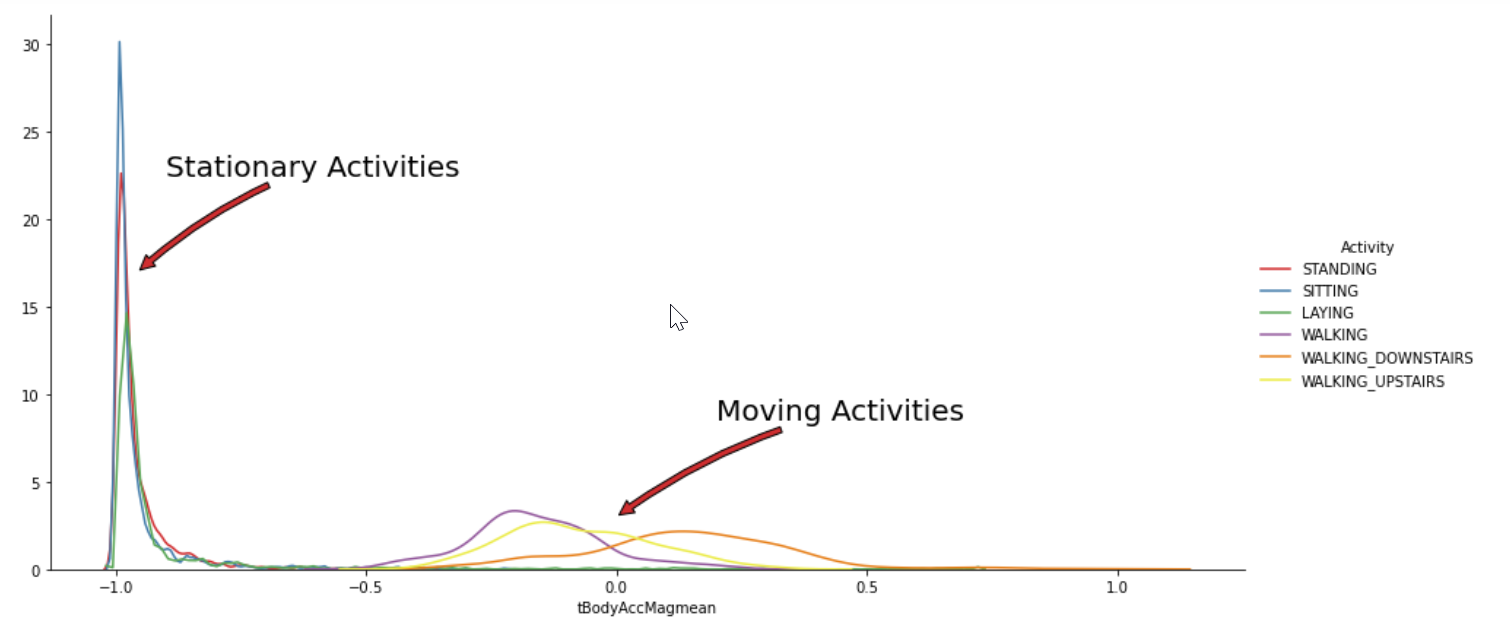
* 1. determine the main characteristics of the data.
  2. discover the patterns within the data.
  3. Spot the anomalies within the data.
  4. ­Check the assumptions or hypothesis on the data using some graphical representations.

In this section we performed exploratory analysis to distinguish Static and Dynamic tasks

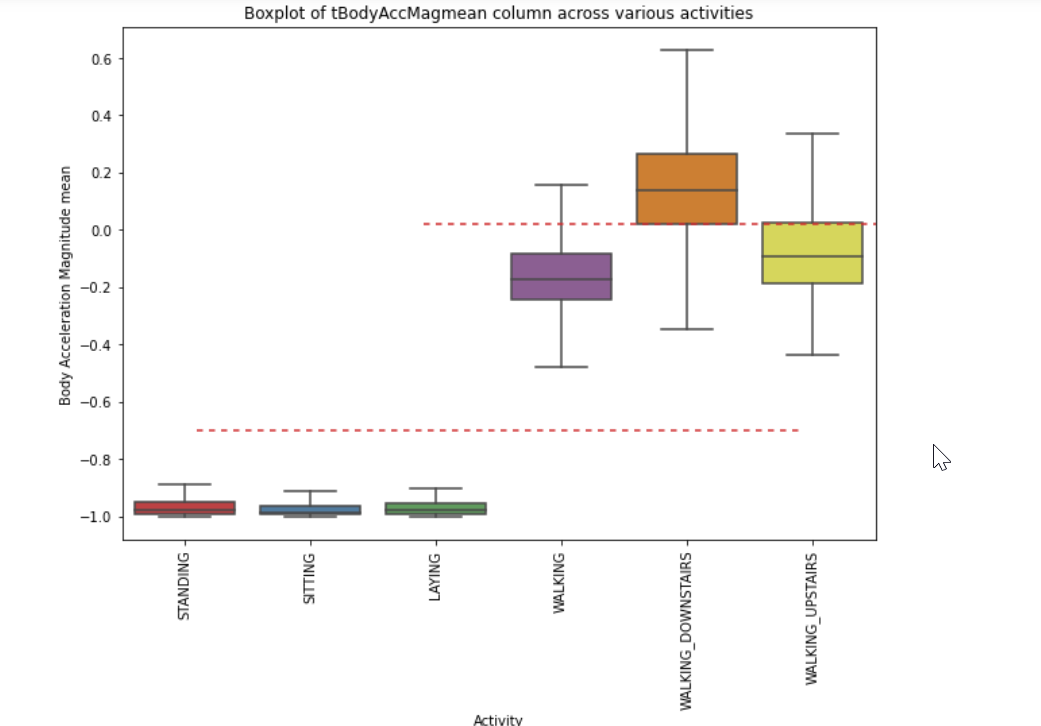
Static activities like sitting, laying, standing doesn’t not involve any motion of the subject and can be easily distinguished from the Dynamic activities like Walking, Walking Up Stairs, Walking Downstairs which involves motion of the subject.

Exploratory analysis on motion info will be significant in distinguishing and determining these two types of activities

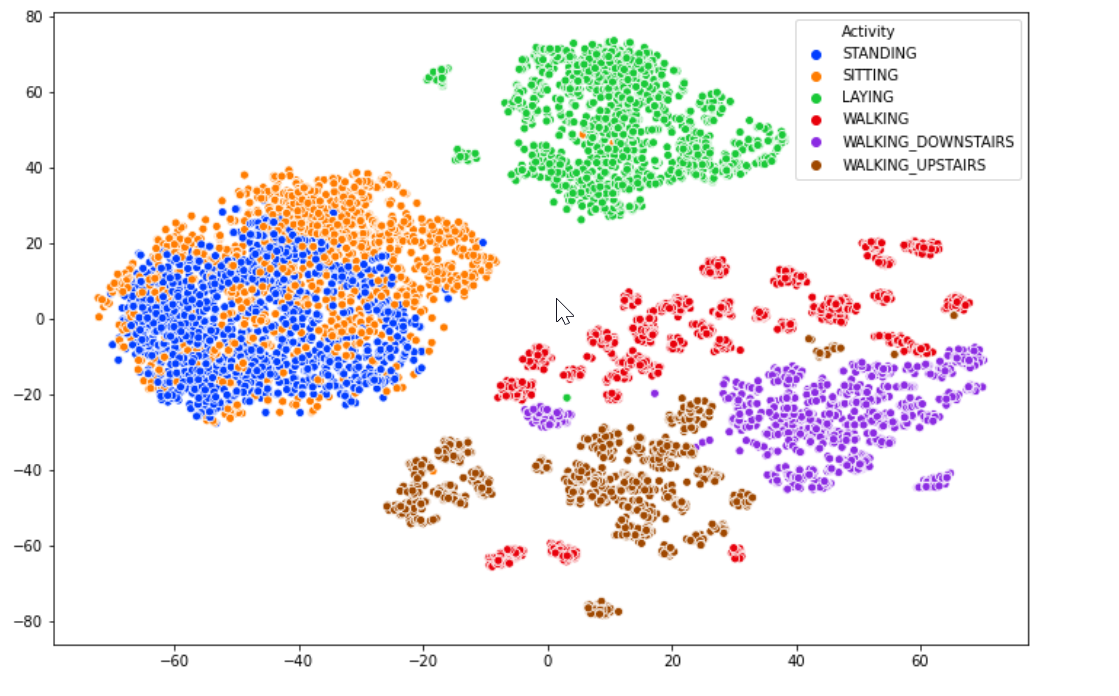
Below is the seaborn distplot which gives the insights into distribution of all the activities on variable **tBodyAccMagmean** which clearly gives thedistribution of the static and dynamic activities



A Boxplot on variable **tBodyAccMagmean** gives the insight of the activities and clearly differentiate static activities which have lower values from dynamic activities which usually have higher values



Below is the scatterplot on the activities which describes the correlation among these activities and based on the visualization following observations has been made

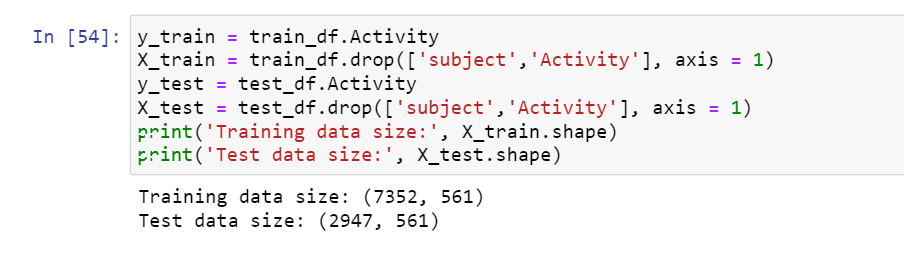


Observations made by interpreting the above scattered plot are:

1. Laying is a completely different activity which exhibits more vertical (Y-axial) gravitational pull.
2. Walking, Waking upstairs, Walking downstairs are the most interrelated or co-related activities and these three activities are clustered together as shown above.
3. Sitting and Standing are the correlated activities and are mostly associated with each other while transitioning from one activity to another and are clustered together as shown in the above figure

**3.4. Data Preprocessing**

Once dataset is cleansed and performed with the exploratory analysis it is split-up into two datasets training dataset and testing dataset below snippet shows the size distribution of train and test datasets



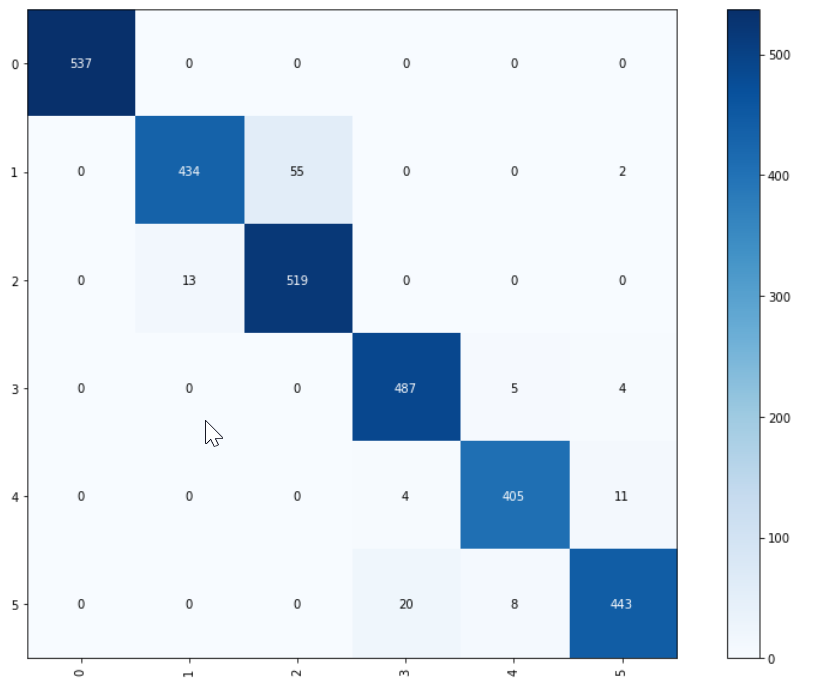
* 1. **Models, Training, Hyperparameter tuning and cross validation**

A set of multi class classification models were developed and trained, these models were imported as a python classes from ski-kit learn and trained and then plotted the **Confusion matrix** to evaluate each model performance and finally the accuracy table where each row represents the accuracy of each model is presented.

* + 1. **process of Logistic Regression**

Logistic Regression or Logistic model is a statistical model which is used to model the probability of a certain class in a multi-class classification.

Below is the confusion matrix for the results of Training, hyperparameter tuning and cross-validation of logistic regression model.

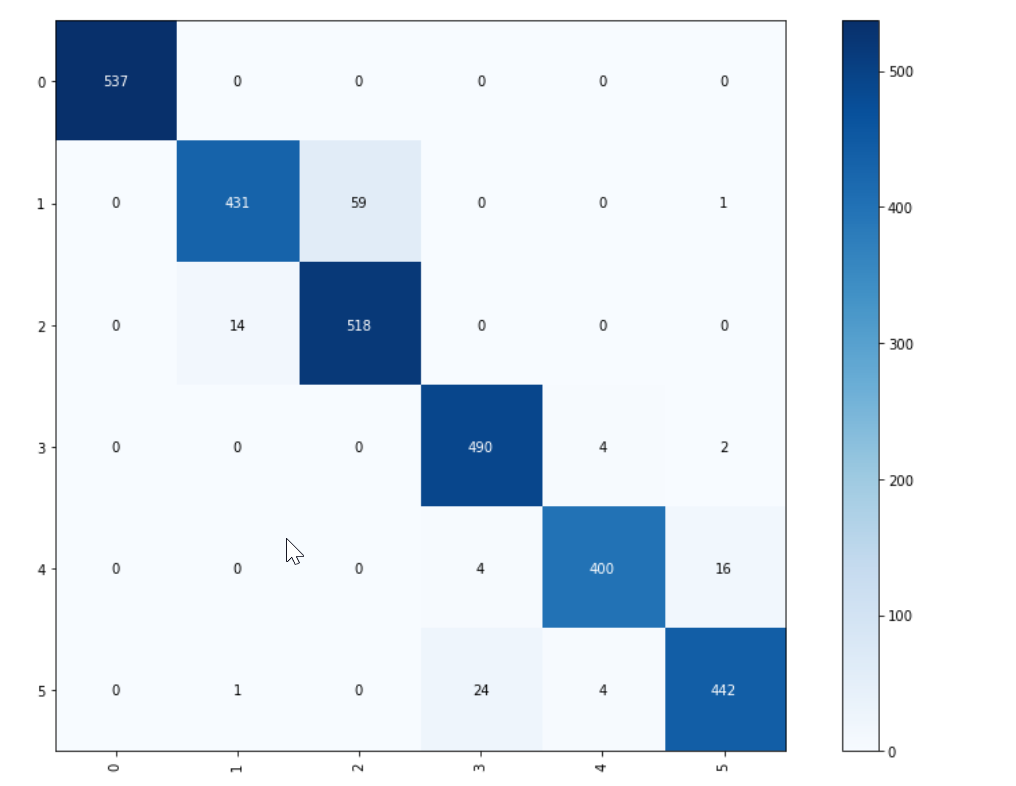


* + 1. **Naive Bayes**

These classification models apply bayes law on the models which then able to apply classifiers to the test data and produce the results.

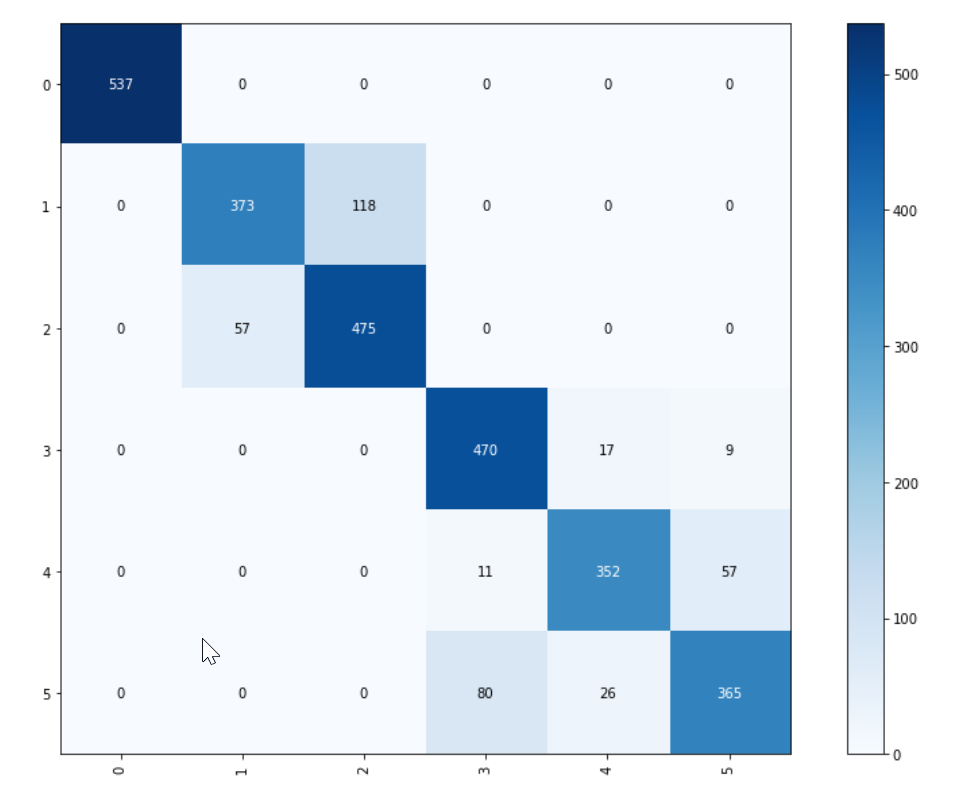
* + 1. **K-Nearest neighbour**

K-nearest neighbour (KNN) algorithm is a **simple, supervised machine learning algorithm** that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.



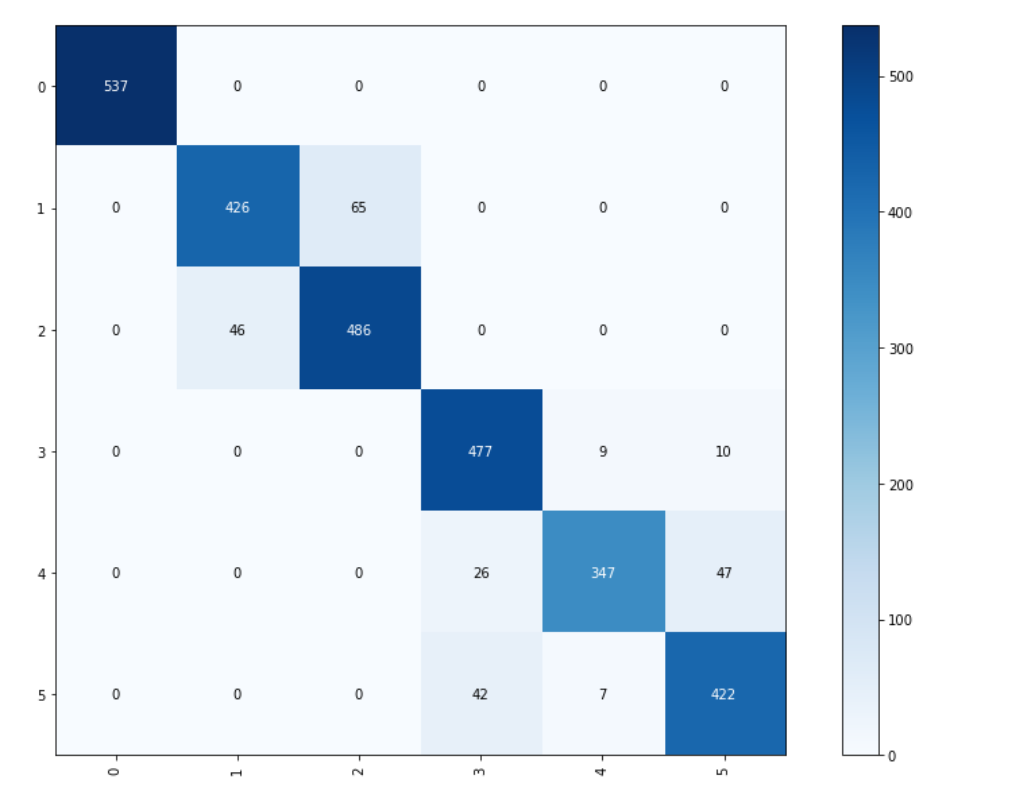
* + 1. **Decision Tree**

Decision Trees are a **type of Supervised Machine Learning** (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves

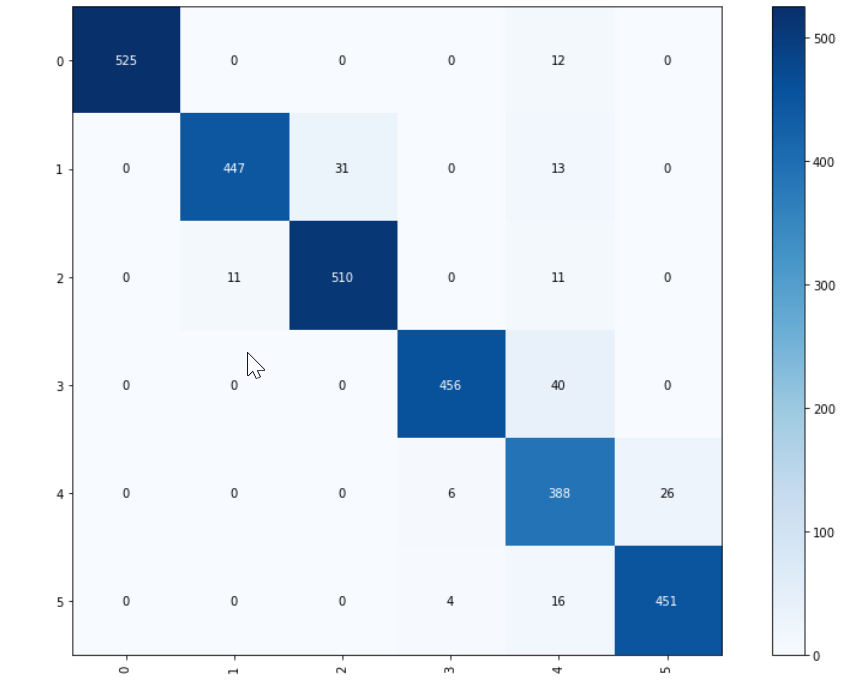
****

* + 1. **Random Forest**

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). In this post we'll learn how the random forest algorithm works, how it differs from other algorithms and how to use it.



**3.5.6. Support Vector Machine**

****

Each model is trained upon the training dataset sample and then the tuned for Hyperparameter and then validated against

Implementations

Types of error detected under implementations

Examination of (one) Important method

Errors – Sources

Evaluation and Testing

Proposed evaluation methods

Comparing or testing methods

Integration , System and Regression , compatibility testing

Development Methodologies

Analysis

Requirement gathering

Functional & Non function requirement

Design

Use case diagram

**Above are the topics and subtopics chosen from the samples.  
I hope you can put the content in relation to the above topics under Methodology and Implementation**

**Try to put as much content as you can and further, we can add after checking few reference papers that are listed in Kaggle link from where we took the dataset**

Results & Discussion

(To be discussed later)

**(Can be included in a different domain )**

**About Data**

**The analysis have been executed alongside a grouping of 30 volunteers inside an age**

**bracket of 19-48 years. Everyone are made to perform six activities (WALKING,**

**WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING). Using its**

**embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-**

**axial angular velocity at a constant rate of 50Hz. The trails possess video-recorded to**

**designate the data manually. The acquired dataset has been incidentally subdivided into**

**two sets, where 70% of the participants was nominated for inducing the training data**

**and 30% the test data.**

**The sensor signals were pre-processed by seeking noise filters and then measured in**

**fixed-width gliding windows of 2.56 sec and 50% overlap (128 readings/window). The**

**sensor acceleration signal, that possess gravitational and physique movement element,**

**that is discrete and managed on a Butterworth low-pass filter towards bodywork’s**

**acceleration and gravity. The gravitational energy is presumed through low frequency**

**elements , therefore a filter was used. One and all , the direction of the features was**

**acquired by evaluating variables in distinction to time and frequency fields.**

**Another factor to be examined is due to the quick increase in mobile phone consumption in**

**evolving countries at present among the common population, comprising of those living in**

**outlands.**

**Apps and the availability of mobile device. In 2010, only about 4,000 health-related app**

**were available, and now more**

**The possibilities to improve the quality of healthcare, leads to healthier lifestyles**

**This section layouts a detailed initiation of the problem interpretation, together with**

**expressing the target and purpose of the project. The section finalises with a concise of the outline of the rest of the report.**