# Title: Prediction of the activity that a subject performs based in measurements obtained from the accelerometer and gyroscope of the Smartphones

## Introduction:

Recently, our lives are invaded by small mobile devices, known as smartphones. These devices are mobile mini-computers, they have an operating system that allows it to launch applications, including a set of applications to manage its contacts and address book, to create, edit or view different types of documents, to access or browse the Web, providing us telephony or messaging services, etc. Apart from these previous features, the most of the smartphones have currently begun to incorporate other features such as cameras, GPS and various types of sensors.

In this analysis, we used data obtained from the accelerometer [1] and gyroscope[2] sensor signals of the smartphones. The accelerometer and gyroscope sensors measure 3-axial linear acceleration and 3-axial angular speed. With these two sensors we can monitor the acceleration, the positions, the orientation, the rotation and the angular motion. All these data can be stored and used to recognize a user's activity. Referring us here to physical activities that a human being can perform daily such as walking, walking up, jogging, sitting, laying, etc.

The aim of this analysis consisted of perform a classification's task. We took a dataset with their attributes (acceleration, orientation,…) and its labeled variable (in this case is activity), and later we created various classification's models known as classifiers. To create these classification's models we can use various algorithms of classification. These algorithms use all available information of a dataset to help us to classify or predict that activity is performed by a human person.

To create models of classification (predictive model), we performed a first task that consisted of choosing different algorithms or techniques of classification, then we applied what is called cross-validation [3] for each algorithm or technique of classification, that is, we trained these algorithm with a set of training data that corresponds to several observations of our available dataset. The following task was tested our classification's algorithm to observe the accuracy, that is, if our predictive model can classify correctly a user's activity according to the acquired knowledge in the previous stage of training. This whole process is known as supervised learning [4].

## Methods:

### *Data Collection*

For this analysis we used a dataset on the Human Activity Recognition. This dataset were downloaded from coursera.org [5] in Data Analysis Course on March 03, 2013 using the R programming language. The data of this dataset is previously processed to make them easier to load into R, since the data was obtained from other raw data from the UC Irvine Machine Learning Repository [6] that has a dataset available about Human Activity Recognition [7], built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted Smartphone with embedded inertial sensors.

The dataset for this analysis contains 7352 observations and 563 variables. For each observation, there is a categorical or factor variable called “**activity**” (our labeled variable or class) that indicates the activity carried out by a human being, there are only six possible values for this variable: **laying, sitting, standing, walk, walkdown and walkup**. Too, there is another integer variable known as “**subject**” that is the identificator of the person that performed that activity. And finally, the rest of the 561 variables are numeric variables (quantitative) that contains features about time and frequency on triaxial acceleration (mean, standard deviation, energy, correlation, etc.) taken from the accelerometer, triaxial angular speed from the gyroscope, etc.

For more information about all these variables, you can find the features here in this compressed file [8]. This compressed file contains some interesting descriptive files that show information about the variables used in this dataset, all features and labeled variable or class.

### *Exploratory Analysis*

Exploratory analysis was performed by examining data and plots of the observed data. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, (3) check the name of variables that are syntactically correct and (4) identify possible different patterns among the different activities and so be able to distinguish when a user performs an activity or another.

Our predictive model [9] should be able to recognize patterns corresponding to every activity. Figure 1 shows the different patterns for different activities according to the analysis of acceleration X-axis. We can observe that there are different patterns according to what activity is carried out by a user.

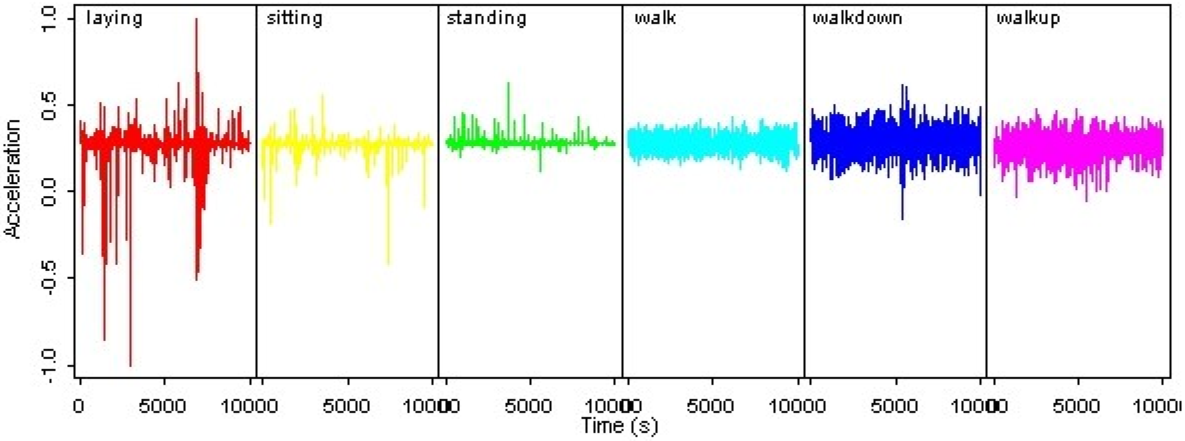

Fig 1. Acceleration X

Figure 2 shows the different patterns for different activities according to the analysis of acceleration Y-axis.

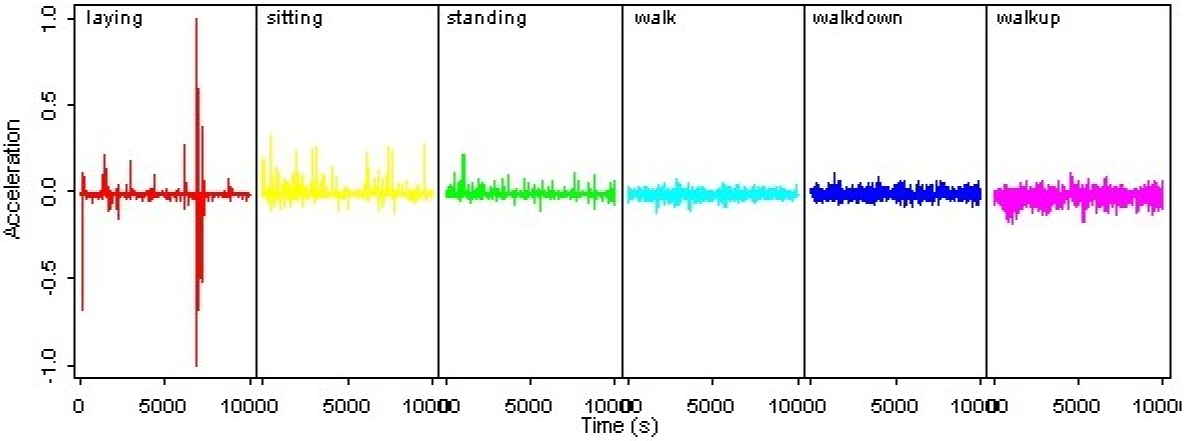

Fig 2. Acceleration Y

Figure 3 shows the different patterns for different activities according to the analysis of acceleration Z-axis.

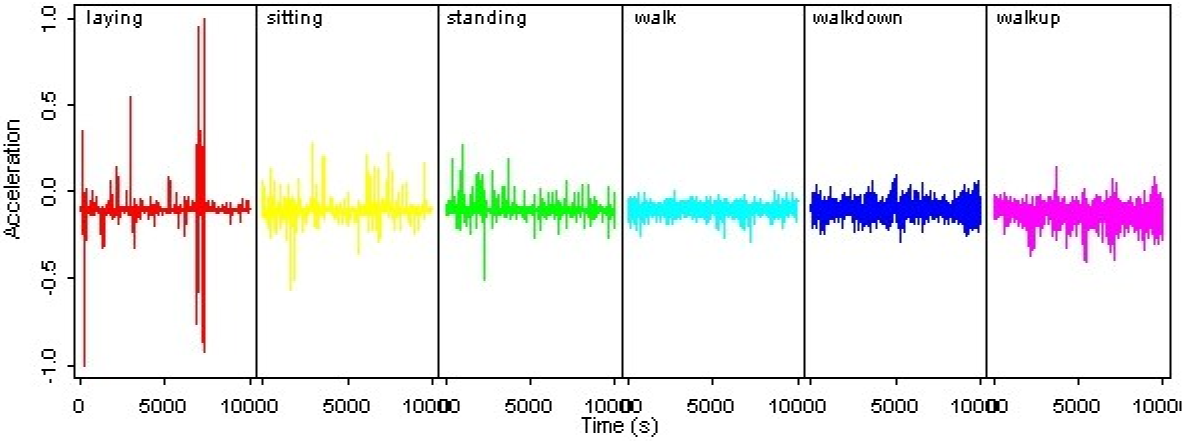

Fig 3. Acceleration Z

It's important keep in mind, if there are activities with common patterns, our predictive model will obviously have more difficulty to classify these activities correctly and therefore our model will have lower accuracy, that is, it has more difficulties to distinguish among activities.

### *Statistical Modeling*

To be able to classify the activity that is performed by a subject, we used various techniques or algorithms of classification to recognize and predict our labeled variable (activity). The techniques (classifiers) employed for this data analysis are the following:

- Decision Trees [10]
- CART [11]
- Bagging [12]
- RamdomForest [13]
- SVM [14]

We performed cross-validation for each of these previous techniques (classifiers). We also evaluated the performance, the accuracy and the error rate of these classifiers.

*Reproducibility*

All analyses performed in this manuscript are reproduced in the R markdown file samsungPredictive.Rmd [15].

**Note.** Due to security concerns with the exchange of R code, we don't submit code to reproduce analysis, in this data analysis,

**Results:**

As I said, the dataset for this analysis contains a total size 7352 observations with 563 variables. These observations are taken from 21 people. In Table 1, shows the number of examples per subject and type of activity, and also the percentage of total per activity from our dataset.

We found variables that have syntactically incorrect names, that is, the name of variables use incorrect character such as comma (","), brackets ("("),etc. , then it was necessary to have valid variable names and not duplicated in our dataset (or data frame). We observed to detect missing values in the dataset, and there weren't any missing values.

Our class or labeled variable was transformed from character variable to a factor variable with 6 levels: "**laying**", "**sitting**", "**standing**", "**walk**", "**walkdown**" and "**walkup**".

According to assignment, for this data analysis we used a training set that include the data from subjects 1, 3, 5 and 6 and a test set that include the data from 27, 28, 29 and 30. Table 2 shows the number of samples per activity that we used to perform the stage of training. And Table 3 indicates the number of samples per activity that we used to perform the stage of testing.

| id | laying | sitting | standing | walk | walkdown | walkup | Total |
|---|---|---|---|---|---|---|---|
| **1** | 50 | 47 | 53 | 95 | 49 | 53 | **347** |
| **3** | 62 | 52 | 61 | 58 | 49 | 59 | **341** |
| **5** | 52 | 44 | 56 | 56 | 47 | 47 | **302** |
| **6** | 57 | 55 | 57 | 57 | 48 | 51 | **325** |
| **7** | 52 | 48 | 53 | 57 | 47 | 51 | **308** |
| **8** | 54 | 46 | 54 | 48 | 38 | 41 | **281** |
| **11** | 57 | 53 | 47 | 59 | 46 | 54 | **316** |
| **14** | 51 | 54 | 60 | 59 | 45 | 54 | **323** |
| **15** | 72 | 59 | 53 | 54 | 42 | 48 | **328** |
| **16** | 70 | 69 | 78 | 51 | 47 | 51 | **366** |
| **17** | 71 | 64 | 78 | 61 | 46 | 48 | **368** |
| **19** | 83 | 73 | 73 | 52 | 39 | 40 | **360** |
| **21** | 90 | 85 | 89 | 52 | 45 | 47 | **408** |
| **22** | 72 | 62 | 63 | 46 | 36 | 42 | **321** |
| **23** | 72 | 68 | 68 | 59 | 54 | 51 | **372** |
| **25** | 73 | 65 | 74 | 74 | 58 | 65 | **409** |
| **26** | 76 | 78 | 74 | 59 | 50 | 55 | **392** |
| **27** | 74 | 70 | 80 | 57 | 44 | 51 | **376** |
| **28** | 80 | 72 | 79 | 54 | 46 | 51 | **382** |
| **29** | 69 | 60 | 65 | 53 | 48 | 49 | **344** |
| **30** | 70 | 62 | 59 | 65 | 62 | 65 | **383** |
| **Sum** | **1407** | **1286** | **1374** | **1226** | **986** | **1073** | **7352** |
| **%** | **19,14** | **17,49** | **18,69** | **16,68** | **13,41** | **14,59** | **100** |

**Table 1. Number of samples per subject and type of activity**

| Laying | sitting | standing | walk | walkdown | walkup |
|---|---|---|---|---|---|
| **55** | **50** | **57** | **64** | **49** | **53** |

**Table 2. Number of samples per activity for Training**

| laying | sitting | standing | walk | walkdown | walkup |
|---|---|---|---|---|---|
| **74** | **64** | **71** | **56** | **52** | **54** |

**Table 3. Number of samples data per activity fo Testing**

We performed the process of cross-validation for each of the previous classifiers using the set of training data and the set of test data were already previously specified.

The results obtained for the different classification techniques (predictive models) using the R programming language are presented in Table 4. In this table can be the accuracy of each classification technique per activity. The cells in bold and underlined indicate the best accuracy.

It is important take into account that we used all quantitative variables (561 variables) to predict the activity carried out by a subject in these 5 classification techniques. We must remember that if we have a lot of variables, the performance of the classification algorithm may be extremely affected, too a lot of these quantitative variables could add noise to classify correctly the activities, and others variables could not be interesting providing good information to distinguish among these activities. On the other hand, it will be very interesting, to perform a measure of how much the classifiers are overfitting[16].

In general, the most of the classification techniques used in this analysis have high levels of accuracy. But we can observe less accuracy for some activities and for some classification techniques.

**% Correctly Predicted**

| Model | Tree | CART | BAGGING | Random Forest | SVM |
|---|---|---|---|---|---|
| | library(tree) | library(rpart) | library(ipred) | library(randomForest) | library(e1071) |
| **laying** | **100,00** | **100,00** | **100,00** | **100,00** | **100,00** |
| **sitting** | 70,31 | 67,19 | 67,19 | **82,81** | **82,81** |
| **standing** | 85,92 | **88,73** | **88,73** | **88,73** | **88,73** |
| **walk** | 50,00 | 57,14 | 80,30 | **92,86** | **92,86** |
| **walkdown** | 84,61 | 86,54 | **94,23** | 86,54 | 86,54 |
| **walkup** | 85,19 | 85,19 | 87,03 | 96,30 | **98,15** |
| **All** | 79,34 | 80,80 | 86,25 | 91,21 | **91,52** |

**Table 4. Accuracies of the Classification Techniques**

The following tables (Table 5-9) show confusion matrices for each of classification techniques.

**Predicted Class**

| Actual Class | laying | sitting | standing | walk | walkdown | walkup |
|---|---|---|---|---|---|---|
| **laying** | **74** | 0 | 0 | 0 | 0 | 0 |
| **sitting** | 0 | **45** | 19 | 0 | 0 | 0 |
| **standing** | 0 | 10 | **61** | 0 | 0 | 0 |
| **walk** | 0 | 0 | 0 | **28** | 6 | 22 |
| **walkdown** | 0 | 0 | 0 | 0 | **44** | 8 |
| **walkup** | 0 | 0 | 0 | 1 | 7 | **46** |

**Table 5. Confusion matrix for the Decision Tree**

| | Predicted Class | | | | | |
|---|---|---|---|---|---|---|
| **Actual Class** | **laying** | **sitting** | **standing** | **walk** | **walkdown** | **walkup** |
| **laying** | **74** | 0 | 0 | 0 | 0 | 0 |
| **sitting** | 0 | **43** | 21 | 0 | 0 | 0 |
| **standing** | 0 | 8 | **63** | 0 | 0 | 0 |
| **walk** | 0 | 0 | 0 | **32** | 4 | 20 |
| **walkdown** | 0 | 0 | 0 | 0 | **45** | 7 |
| **walkup** | 0 | 0 | 0 | 1 | 7 | **46** |

**Table 6. Confusion matrix for the CART**

| | Predicted Class | | | | | |
|---|---|---|---|---|---|---|
| **Actual Class** | **laying** | **sitting** | **standing** | **walk** | **walkdown** | **walkup** |
| **laying** | **74** | 0 | 0 | 0 | 0 | 0 |
| **sitting** | 0 | **43** | 21 | 0 | 0 | 0 |
| **standing** | 0 | 8 | **63** | 0 | 0 | 0 |
| **walk** | 0 | 0 | 0 | **53** | 0 | 3 |
| **walkdown** | 0 | 0 | 0 | 0 | **49** | 3 |
| **walkup** | 0 | 0 | 0 | 1 | 6 | **47** |

**Table 7. Confusion matrix for Bagging**

| | Predicted Class | | | | | |
|---|---|---|---|---|---|---|
| **Actual Class** | **laying** | **sitting** | **standing** | **walk** | **walkdown** | **walkup** |
| **laying** | **74** | 0 | 0 | 0 | 0 | 0 |
| **sitting** | 0 | **53** | 11 | 0 | 0 | 0 |
| **standing** | 0 | 8 | **63** | 0 | 0 | 0 |
| **walk** | 0 | 0 | 0 | **52** | 0 | 4 |
| **walkdown** | 0 | 0 | 0 | 0 | **47** | 5 |
| **walkup** | 0 | 0 | 0 | 0 | 2 | **52** |

**Table 8. Confusion matrix for Random Forest**

| | Predicted Class | | | | | |
|---|---|---|---|---|---|---|
| **Actual Class** | **laying** | **sitting** | **standing** | **walk** | **walkdown** | **walkup** |
| **laying** | **74** | 0 | **0** | 0 | 0 | 0 |
| **sitting** | 0 | **53** | **11** | 0 | 0 | 0 |
| **standing** | 0 | 8 | **63** | 0 | 0 | 0 |
| **walk** | 0 | 0 | **0** | **52** | 0 | 4 |
| **walkdown** | 0 | 0 | **0** | 0 | **47** | 5 |
| **walkup** | 0 | 0 | **0** | 0 | 1 | **53** |

**Table 9. Confusion matrix for SVM**

We observed that the classification techniques identify correctly **laying** (100%). It appears much more difficult to distinguish between sitting and standing, and also to distinguish between walk, walkdown and walkup.

The Bagging, Random Forest and SVM are classifiers that require more computing and memory resources, and therefore more classification time than Tree and CART.

## Conclusions:

In this analysis, we employed various classification techniques to obtain different predictive model. The SVM classifier algorithm achieved the highest levels of accuracy for this analysis (91,52% accuracy). It will be recommendable to increase the number of observations. It will be aslso recommendable to increase the samples for the set of training data, and the samples for the set of test data, and observe if the accuracy increases or decreases. On the other hand, there are some problems to detect patterns in some activities relating with others, because there are a lot of similar patterns among the different activities and then the classifier doesn't classify correctly.

## References

[1] Accelerometer
http://en.wikipedia.org/wiki/Accelerometer. Accessed 03/04/2013
[2] Gyroscope
http://en.wikipedia.org/wiki/Gyroscope. Accessed 03/04/2013
[3] Cross Validation
http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29. Accessed 03/10/2013
[4] Supervised Learning
http://en.wikipedia.org/wiki/Supervised_learning. Accesed 03/05/2013
[5] Dataset of Human Activity Recognition Coursera
https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda. Accessed 03/03/2013
[6] UC Irvine Machine Learning Repository
http://archive.ics.uci.edu/ml/. Accessed 03/06/2013
[7] Dataset of Human Activity Recognition Using Smartphones Data Set
http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.
Accessed 03/06/2013
[8] File of Human Activity Recognition UCI
http://archive.ics.uci.edu/ml/machine-learning-databases/00240/UCI%20HAR%20Dataset.zip. Accessed 03/06/2013
[9] Predictive Modelling
http://en.wikipedia.org/wiki/Predictive_modelling. Accessed 03/10/2013

[10] Tree Learning
http://en.wikipedia.org/wiki/Decision_tree_learning. Accessed 03/10/2013
[11] CART
http://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees. Accessed 03/10/2013
[12] Bagging
http://en.wikipedia.org/wiki/Bootstrap_aggregating. Accessed 03/10/2013
[13] Random Forest (RF)
http://en.wikipedia.org/wiki/Random_forest . Accessed 03/10/2013
[14] Support Vector Machine (SVM)
http://en.wikipedia.org/wiki/Support_vector_machine . Accessed 03/10/2013
[15] R Markdown Page.
http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 03/06/2013
[16] Overfitting
http://en.wikipedia.org/wiki/Overfitting. Accessed 03/10/2013