Human Activity Recognition using Smartphone Dataset Capstone Project

By: Mashlahul Afif

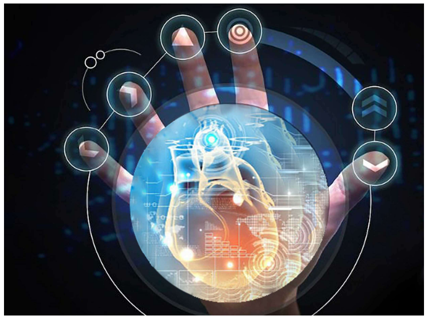
May 2019

Contents

- Background and Objectives
- About the Dataset
- Data Wrangling
- Understanding the Data
- Data Exploratory
- Machine Learning Classification Model
- Model Evaluation
- Variables’ Importance
- Conclusion

Background and Objectives

- **Explore**, **visualize**, and to **develop** human activity recognition using smartphones sensors' reading data
- **To what end? ➡ Human Activity based Apps**
  - Exercise assistance or healthcare apps
  - Augmented reality apps
  - Military
  - Security
  - Etc.



About the Dataset (1/2)

The Source

- Obtained from **UCI** Repository
- Collected from Galaxy S Smartphone's **Accelerometer & Gyroscope's Sensors** Reading
- **Experiments** by 30 volunteers (age 19-48 years)
- **6 Activities:**

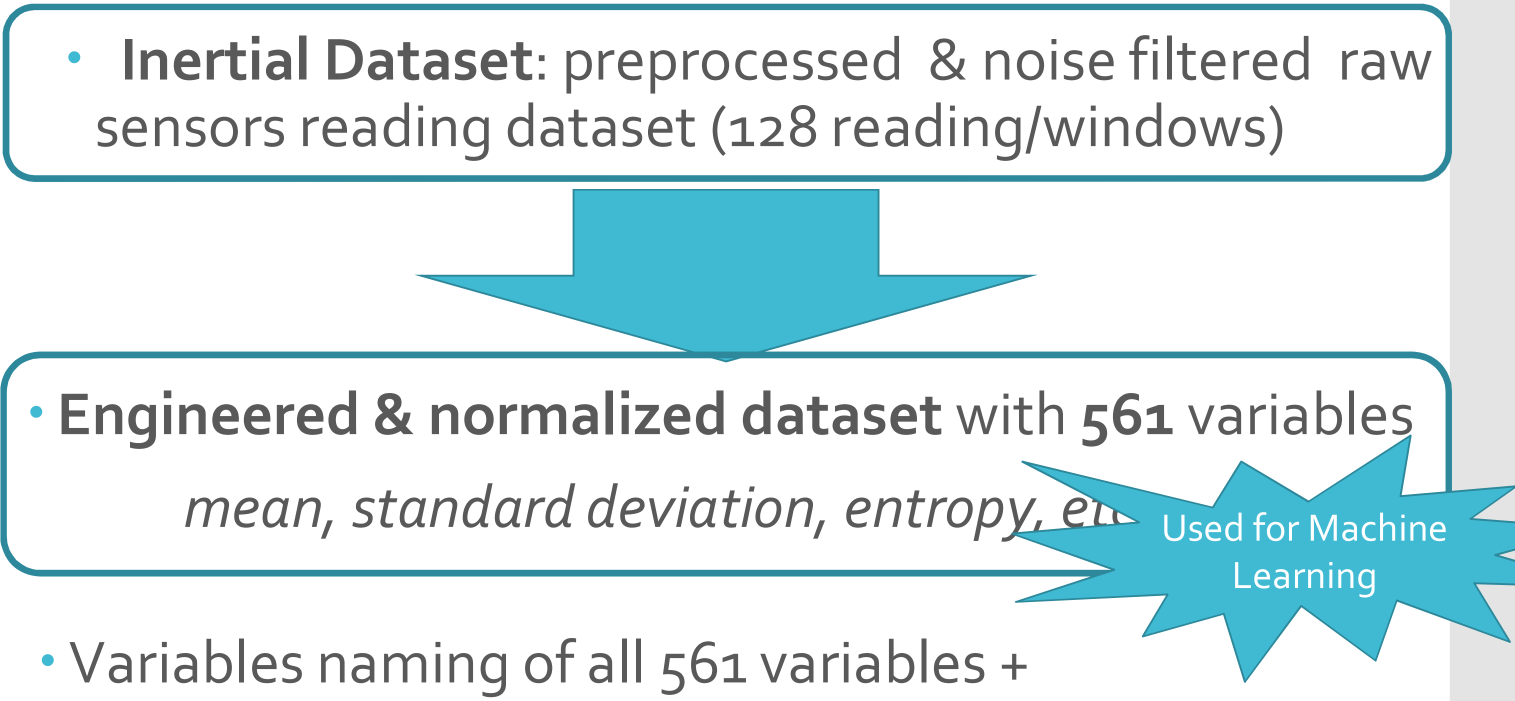
  STANDING, SITTING, LAYING, WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS

About the Dataset (2/2)

Main Dataset Composition

- **All dataset divided** : 70% training (7352 obs)

  30% test (2947 obs)



- **Inertial Dataset**: preprocessed & noise filtered raw sensors reading dataset (128 reading/windows)

- **Engineered & normalized dataset** with **561** variables

  mean, standard deviation, entropy, etc

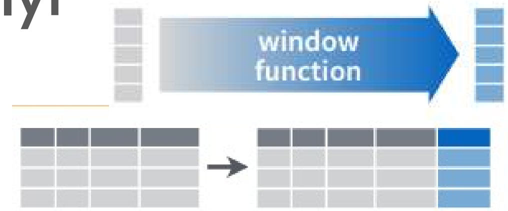
Used for Machine Learning

- Variables naming of all 561 variables +
- Human Activity Assignment for all 10000 observations above are **included in separate datasets**

Data Wrangling

1. Renaming the variables name by removing duplicates and replacing prohibited characters (space, bracket, etc)
2. Adding the Activity Label (WALKING, etc) for each observations.
3. Creating time stamp or index variables for data visualization purpose
4. Combining all the data
5. Later splitting the dataset into training and test for machine learning.

All of the above was done using **dplyr**



Understanding The Data (1/2)

Min Max Normalization

- Data is normalized : -1 to +1 range (min-max normalization)
- Only normalized in respect to same group of dataset (either training or test)

$$normalized\, x = \frac{x - max_{all\ x}}{max_{all\ x} - min_x}(new\, max_{all\ x} - new\, min_{all\ x}) + new\, min_{all\ x}$$

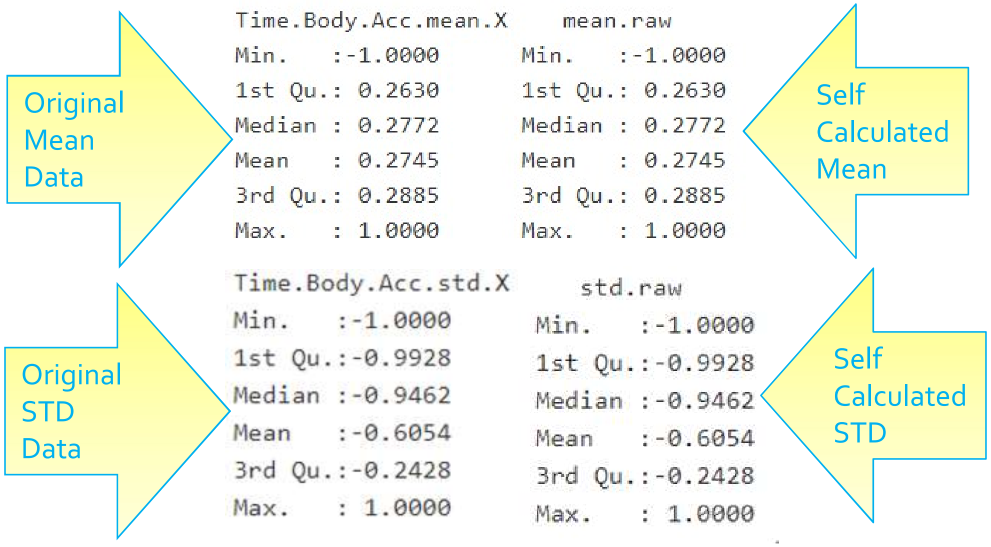
This means: each data is a **relative value**➡

- +1 = maximum value across all observations of 1 variable
- -1 = minimum value across all observations of 1 variable
- 0 = aprox. median value across all observations of 1 variable

Understanding The Data (2/2)

The Proof

- Self calculating the Mean and Standard Deviation from Inertial Signals Dataset (raw preprocessed sensors data)
- Normalizing above calculation using min-max formula
- Comparing with the same variable from original engineered dataset.
- Results:

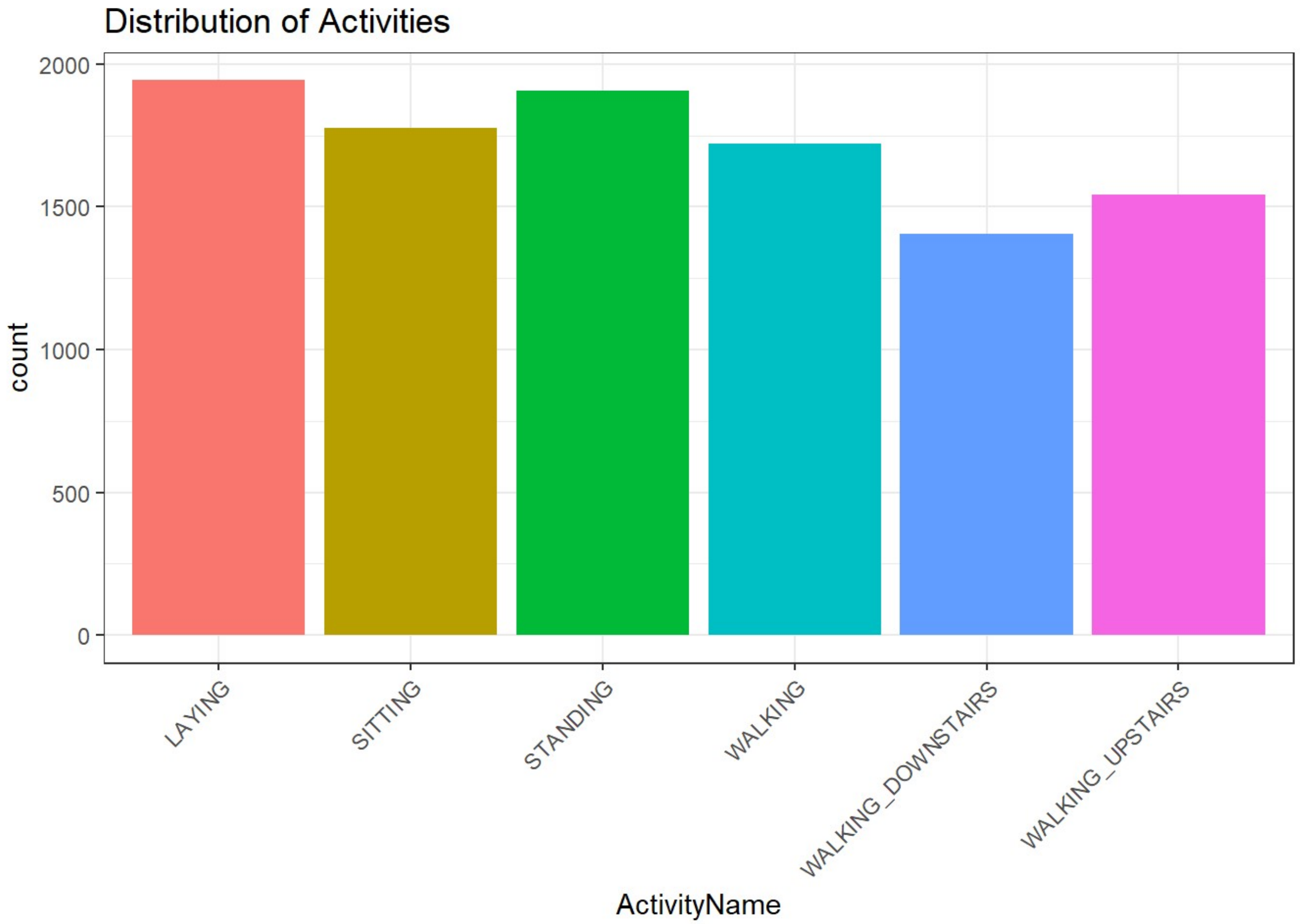


```
Time.Body.Acc.mean.X      mean.raw
Min.   :-1.0000       Min.   :-1.0000
1st Qu.: 0.2630       1st Qu.: 0.2630
Median : 0.2772       Median : 0.2772
Mean   : 0.2745       Mean   : 0.2745
3rd Qu.: 0.2885       3rd Qu.: 0.2885
Max.   : 1.0000       Max.   : 1.0000
```

```
Time.Body.Acc.std.X        std.raw
Min.   :-1.0000        Min.   :-1.0000
1st Qu.:-0.9928        1st Qu.:-0.9928
Median :-0.9462        Median :-0.9462
Mean   :-0.6054        Mean   :-0.6054
3rd Qu.:-0.2428        3rd Qu.:-0.2428
Max.   : 1.0000        Max.   : 1.0000
```

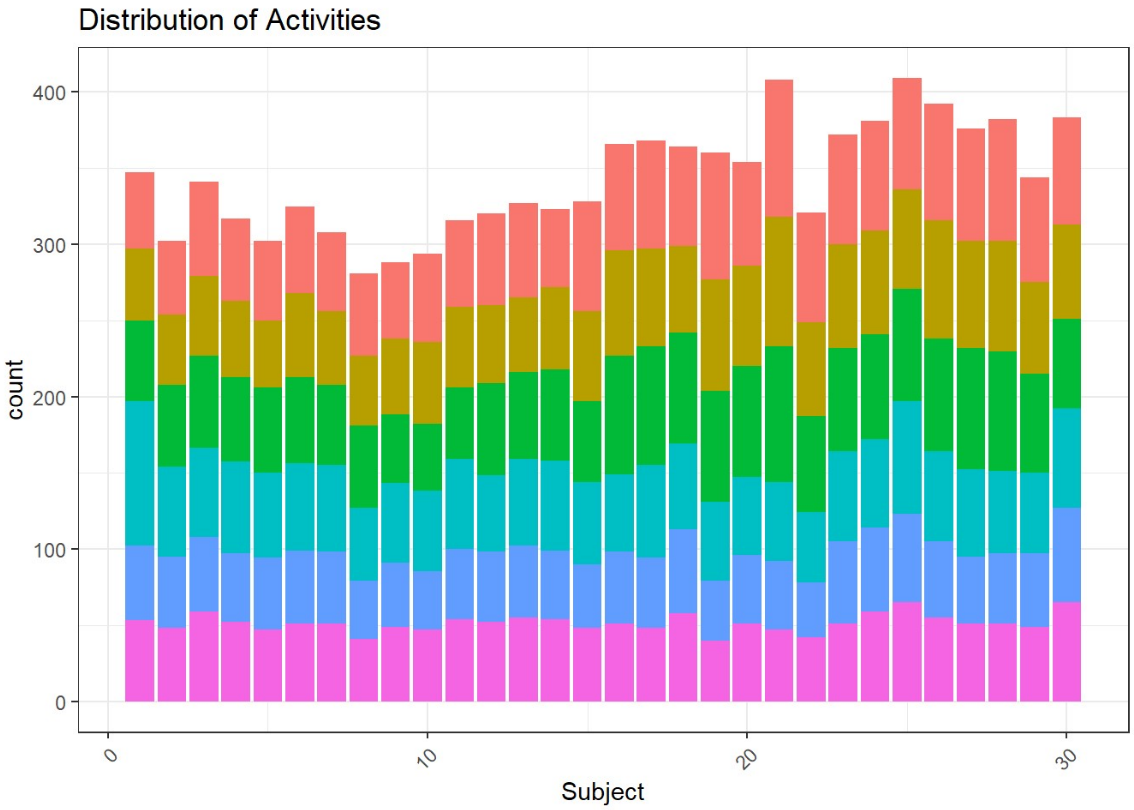
Data Exploratory (1/8)

Distributions for All Activities



Data Exploratory (2/8)

Durations of Experiment for each Subject



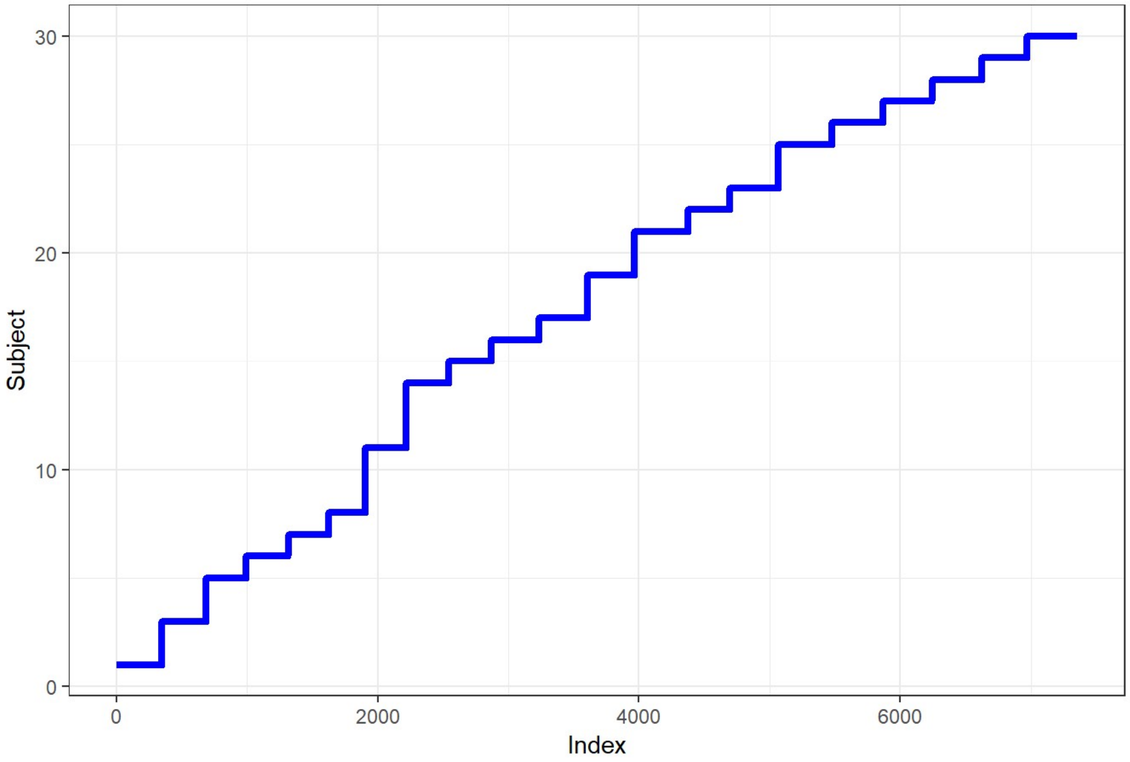
- All activities were not done in equal duration.
- Each subject also performs the experiment in different duration

Data Exploratory (3/8)

Order Subjects during Experiment

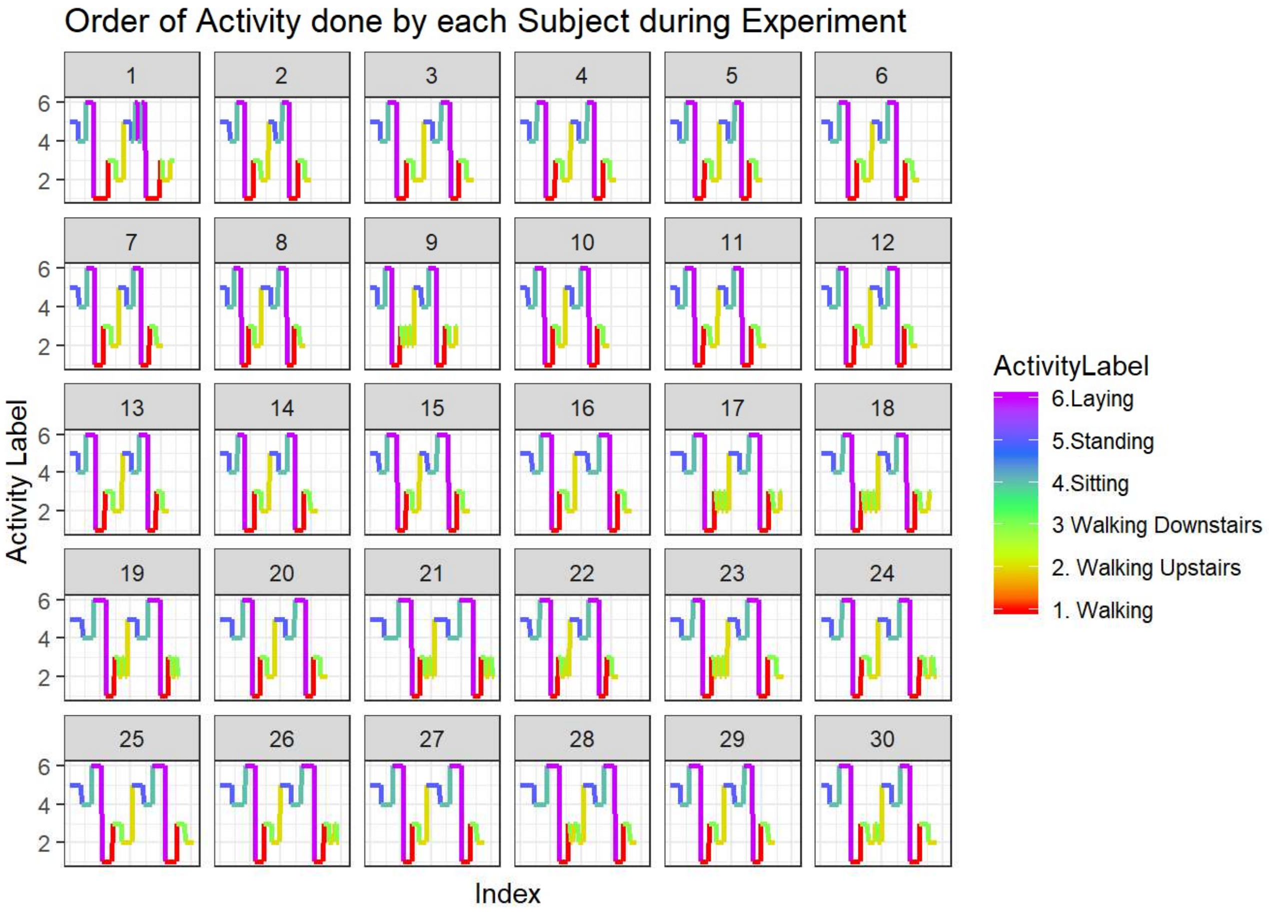
- Data has been sorted by the subjects of experiment
- No repetition of subject

Order of Subject During Experiment



Data Exploratory (4/8)

Order Activities Done by Each Subject



- All subjects performed the experiment by the following order:
- Standing ➡ Sitting ➡ Laying ➡ Walking ➡ Walking Downstairs ➡ Walking Upstairs

Data Exploratory (5/8)

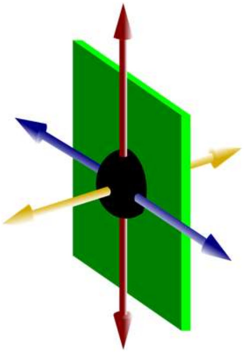
Time Series & Histogram Visualization of Variables

To Plot:

- Mean
- Standard Deviation
- Magnitude (Resultant of XYZ)

of **three main sensors** data:

- Body Motion from Accelerometer
- Body Gyroscope Reading
- Body Gravity from Accelerometer

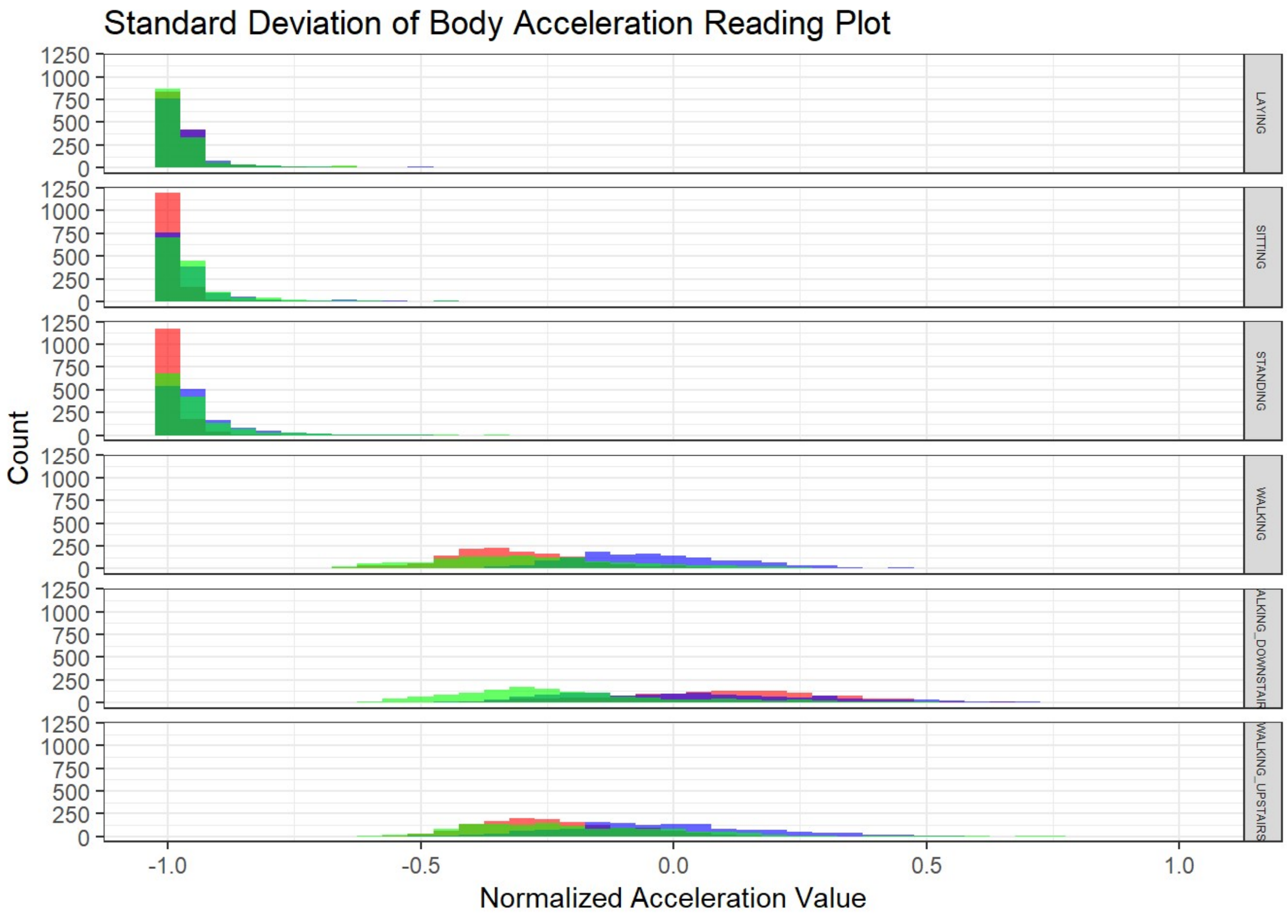


This Photo by Unknown Author is licensed under CC BY-SA

Only the most meaningful plot will be discussed in this presentation

Data Exploratory (6/8)

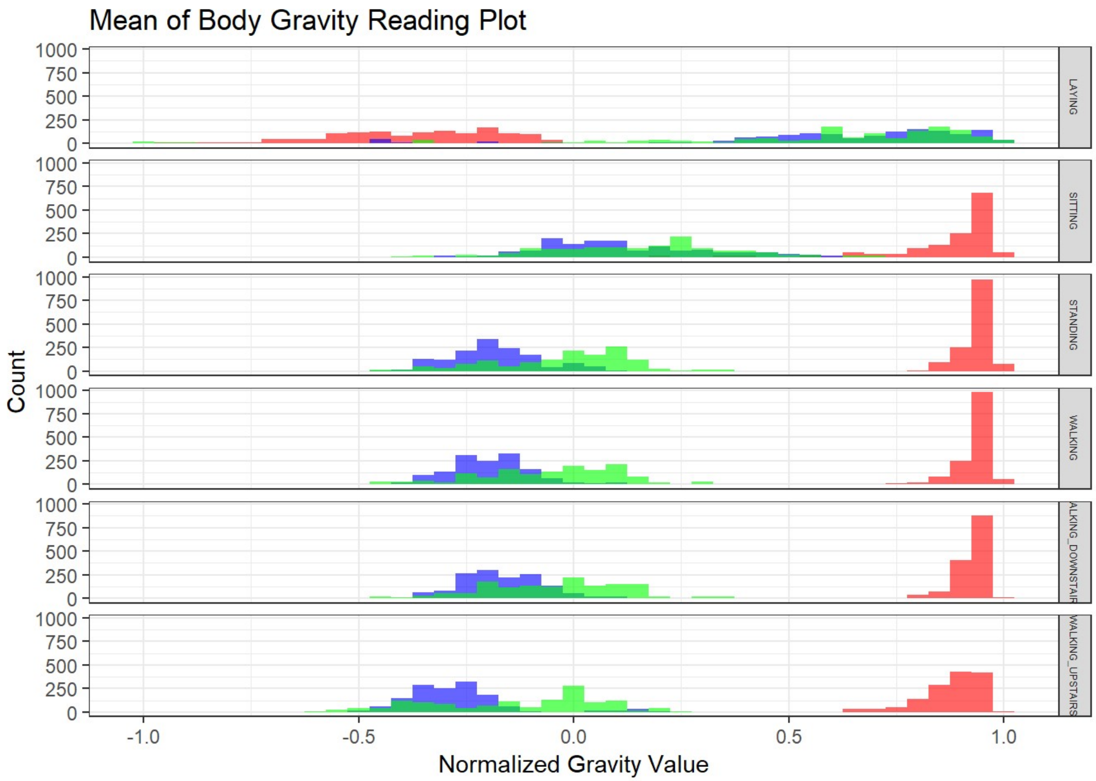
Moving vs. Stationary Activities



- All stationary activities' standard deviation at minimum value(-1) ➡ Almost zero variance in raw sensors data
- Moving activities are more equally distributed

Data Exploratory (7/8)

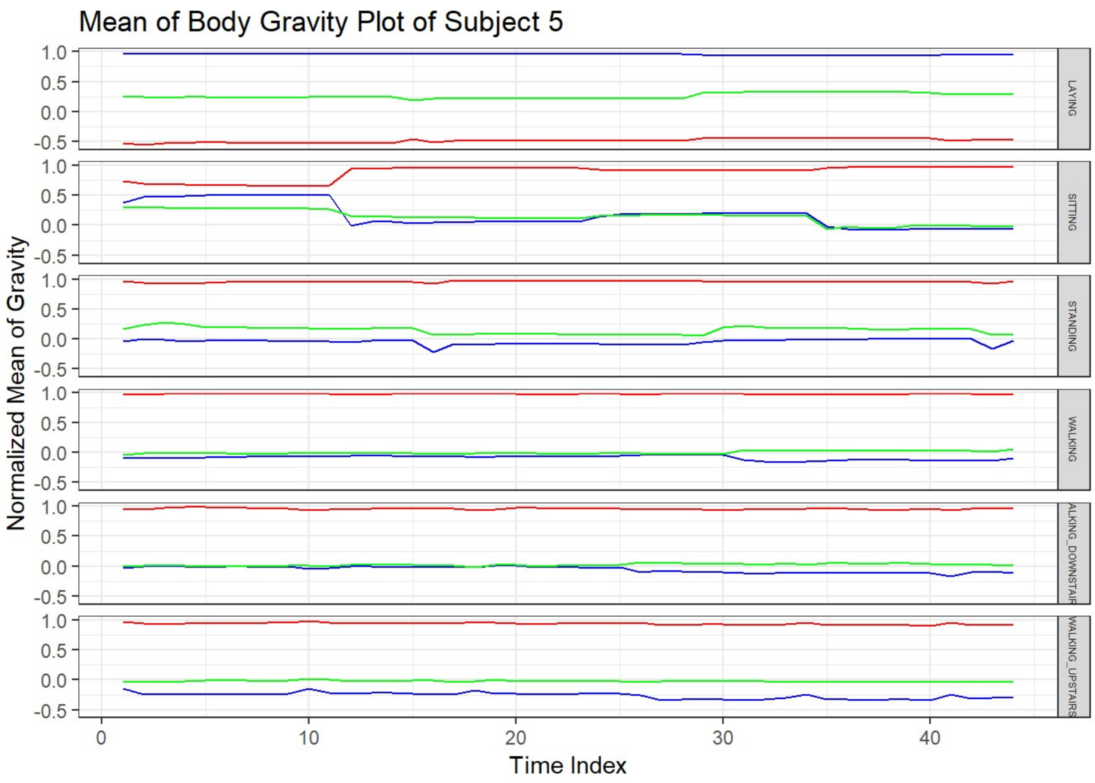
Laying vs. other Activities



- Mean body gravity for laying is more equally distributed to all values
- Also, x axis value of mean body gravity is relatively lower.

Data Exploratory (8/8)

Distinguishing Walking Upstairs



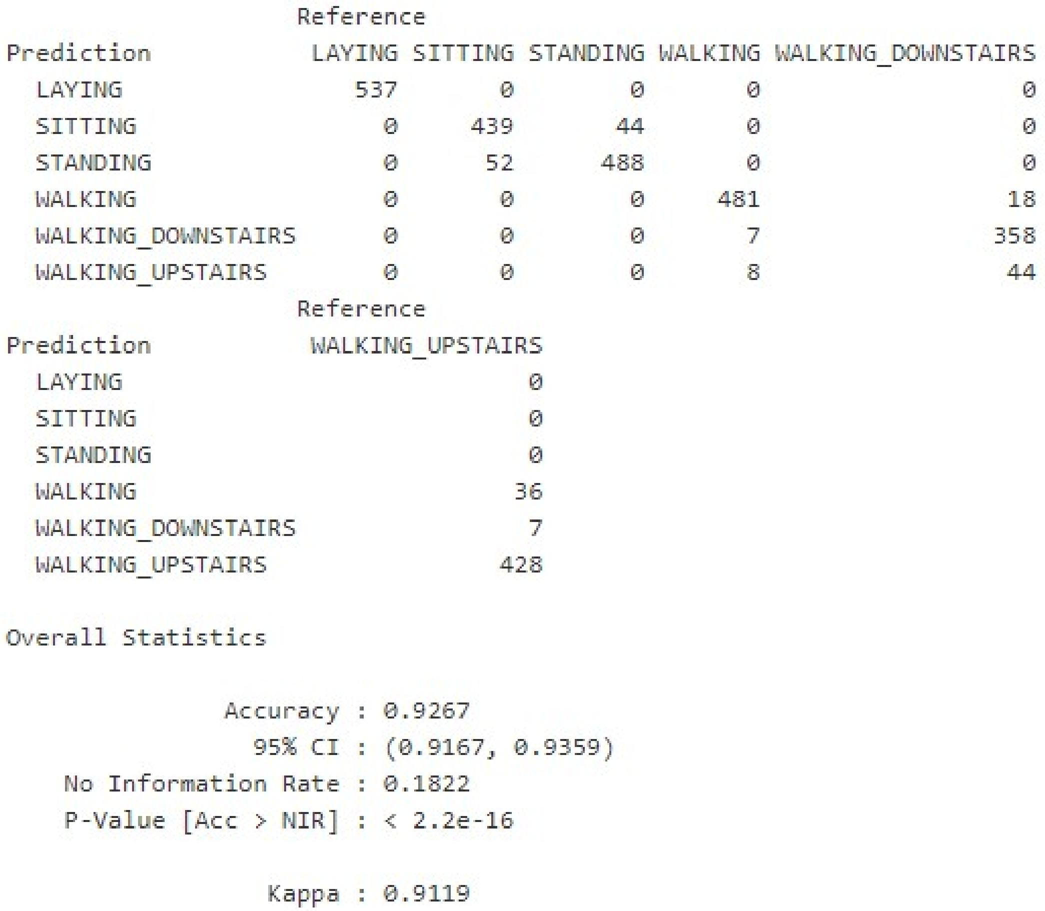
- Mean body gravity for walking upstairs is relatively lower compared to other moving activities

Machine Learning Classification Model

- Random Forest Multiclass Classification
- Tuned using Caret package.
- Number of variables randomly sampled as candidates at each split (mtry)=24.
- Number of trees to grow (ntree)=500.

Model Evaluation (1/2)

Confusion Matrix

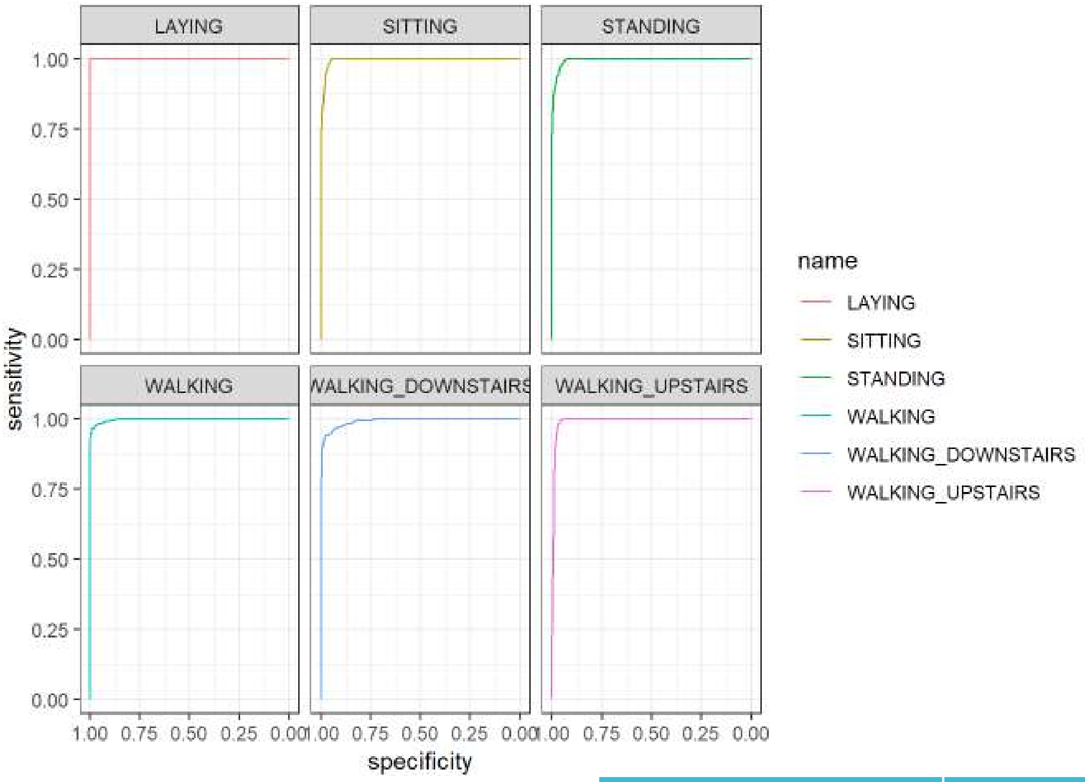


```
                    Reference
Prediction           LAYING SITTING STANDING WALKING WALKING_DOWNSTAIRS
  LAYING                537       0        0       0                  0
  SITTING                 0     439       44       0                  0
  STANDING                0      52      488       0                  0
  WALKING                 0       0        0     481                 18
  WALKING_DOWNSTAIRS      0       0        0       7                358
  WALKING_UPSTAIRS        0       0        0       8                 44
                    Reference
Prediction           WALKING_UPSTAIRS
  LAYING                            0
  SITTING                           0
  STANDING                          0
  WALKING                          36
  WALKING_DOWNSTAIRS                7
  WALKING_UPSTAIRS                428

Overall Statistics
                                          
               Accuracy : 0.9267          
                 95% CI : (0.9167, 0.9359)
    No Information Rate : 0.1822          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.9119          
```

Model Evaluation (2/2)

ROC and AUC



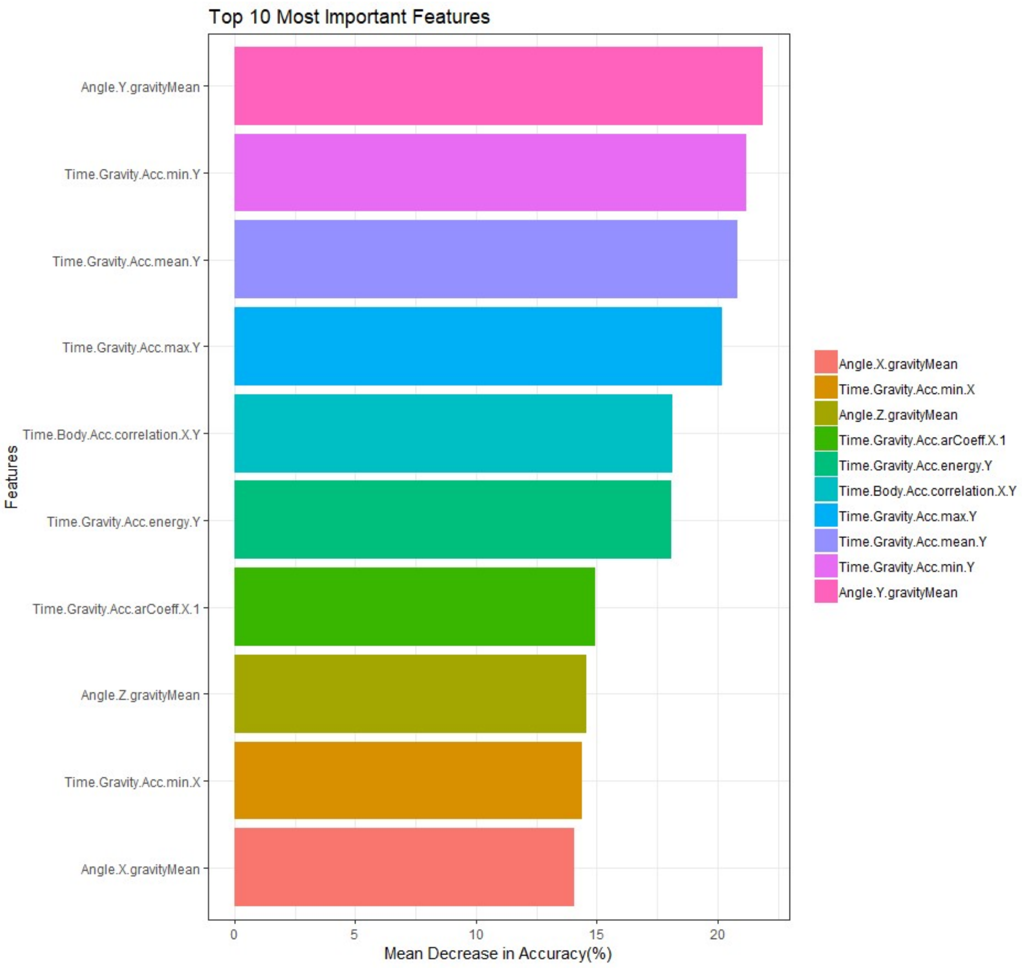
Area Under the Curve is close to 1

➡ very capable in distinguishing between classes

Class	Area Under The Curve
LAYING	1.000
SITTING	0.995
STANDING	0.996
WALKING	0.997
WALKING DOWNSTAIRS	0.992
WALKING UPSTAIRS	0.992

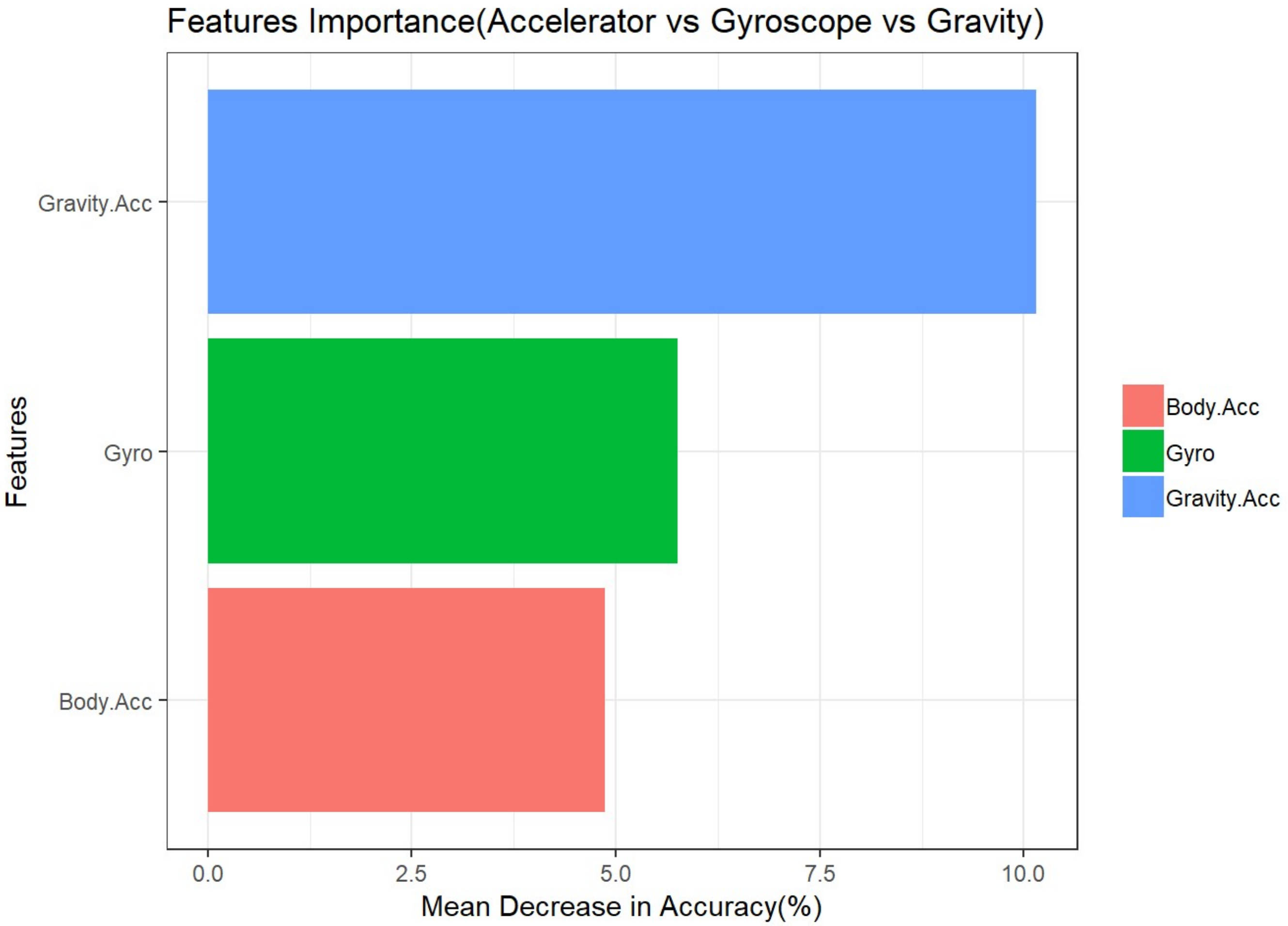
Variables' Importance (1/3)

Top 10 Most Important Features



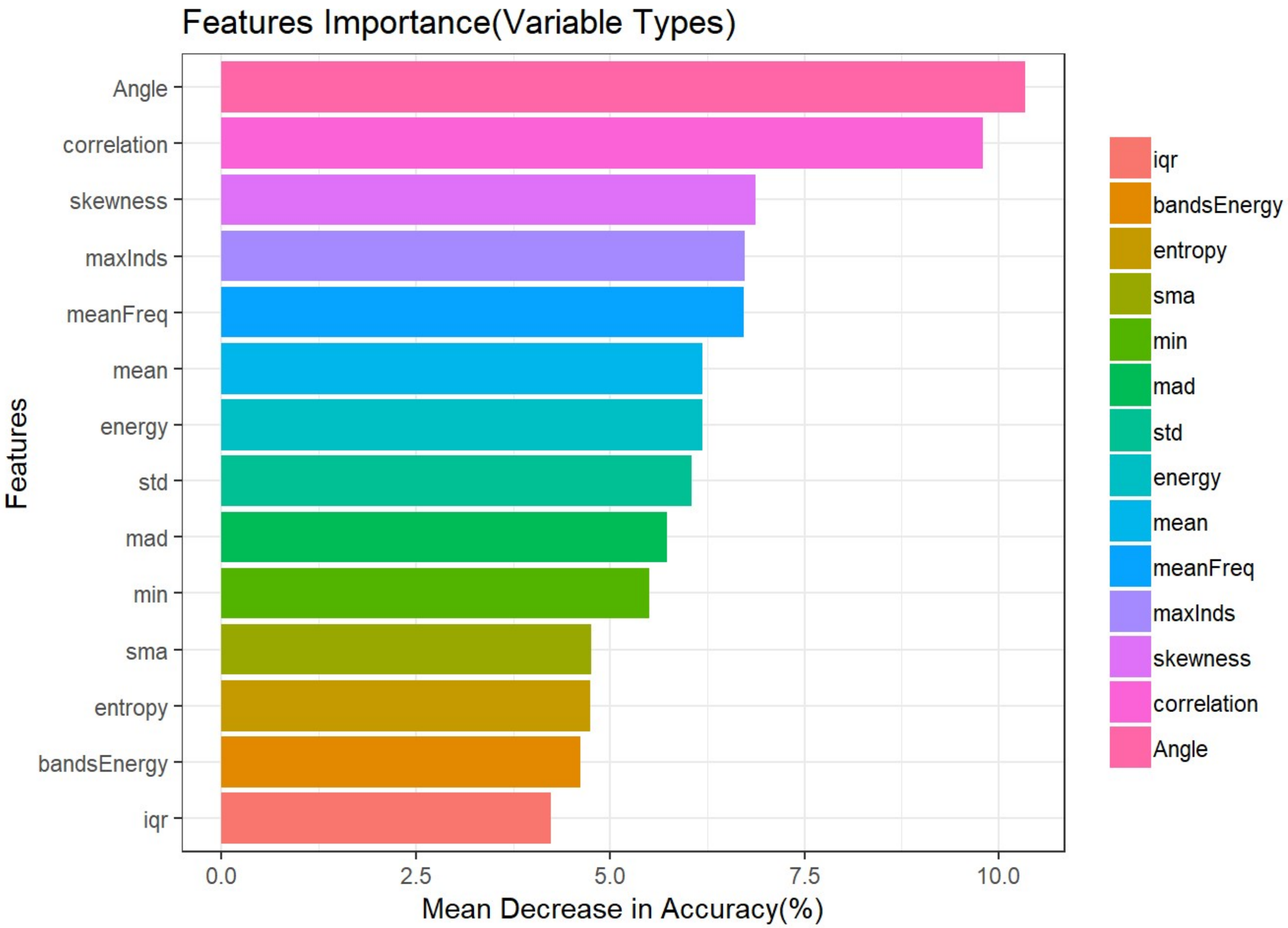
Variables' Importance (2/3)

Three Sensors ' Type's Reading Comparison



Variables' Importance (3/3)

Variables' Type Comparison



Conclusion

- Random forest model (mtry=24 and ntree = 500), achieved Accuracy of 92.7%, and the Kappa value of 91.2%. These are considered quite high.
- Area Under the Curve (AUC) very close to 1, which indicates that our model is very capable in distinguishing or classifying between classes of activities.
- Among all 561 variables Time.Body.Acc.correlation.X.Y is the most significant variable.
- Between Body Acceleration, Body Gravity, and Body Gyroscope, body's gravitational components from accelerometer is the most important.
- Among all engineered variables, Angle and correlations are the most important by large margins if compared to the other variables