Predicting Human Activity Recognition Using Smartphone Data

University of Rochester

DSC383W Capstone

Author:

Yi Nei Charlene Lau

September 21, 2019

Introduction

People are constantly interacting with one another and being able to characterize the physical actions that people perform is a crucial task. The recognition of human activity is complex and involves motions and sensor signals produced when people perform various activities. Sensing technologies have been applied in experiments involving people performing activities such as walking, walking upstairs, walking downstairs, sitting, standing and laying. Human activity classification is analyzed in this project to explore the relationship between characteristic features and the activity performed. In summary, the project shows that we are able to predict human activities using collected data with a high accuracy.

Data Collection and Preparation

For this experiment, 30 volunteers between 19 and 48 years of age were asked to perform six different activities with smartphones tied to their waists. The volunteers were video recorded while they performed the activities and the data was reproduce from the recordings. With the built-in functions of the smartphone, the 3-axial linear acceleration and 3-axial angular velocity are recorded from the accelerometer and the gyroscope at a constant rate of 50Hz. Signals produced by the accelerometer and the gyroscope are preprocessed through noise filters. The timeframe for each piece of data was for a time of 2.56sec.

The acceleration signal was separated into body and gravity acceleration signals. A Butterworth low-pass filter was used for this purpose. Variables were calculated from both the time and frequency domains. Two more variables were added to the dataset: a subject index and an activity label. The full dataset thus contains 563 features with 10300 records. The number of records produced for each volunteer varied. On average, each volunteer had 343 records of data associated with them.

The train and test data was split into 70/30. The training set included 21 volunteers chosen at random, with the remaining 9 in the testing set.

Exploratory Analysis

In the dataset provided by UCI Machine Learning Repository, there were no missing values. 561 features were numeric and the target feature was categorical. The task was a classification task.

Exploratory analysis was performed by producing histograms and distributions for relevant features. The most commonly seen distributions were of three types shown in Figures 1, 2 and 3. Figure 1 tends towards a normal distribution. Figure 2 and 3 show skewed distributions that are negatively and positively skewed, respectively.

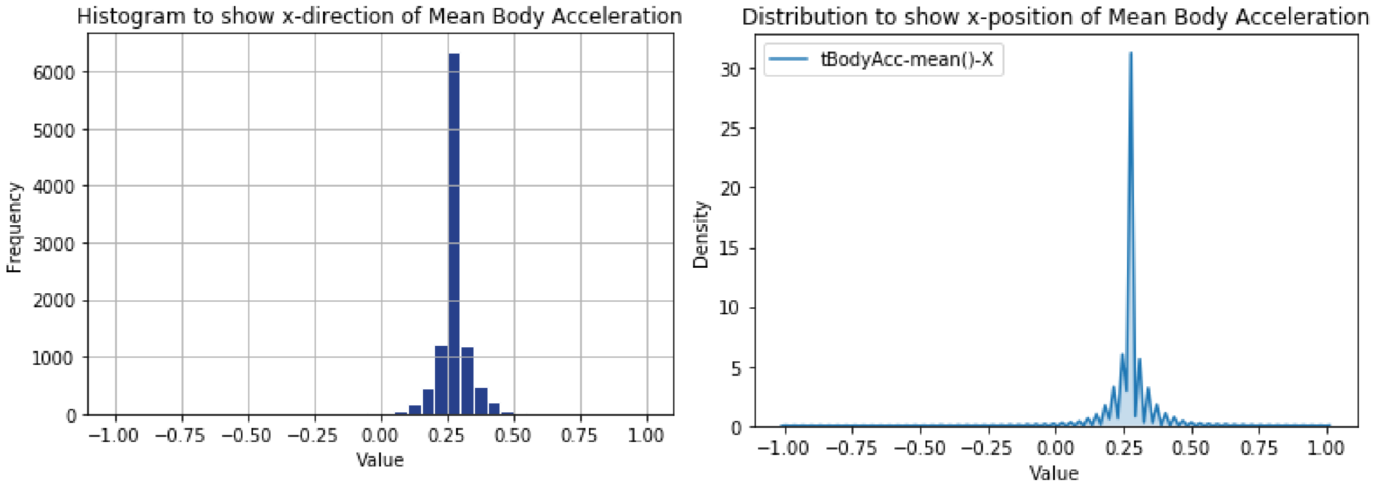


Figure 1. Histogram and distribution of Mean Body Acceleration in the x-direction

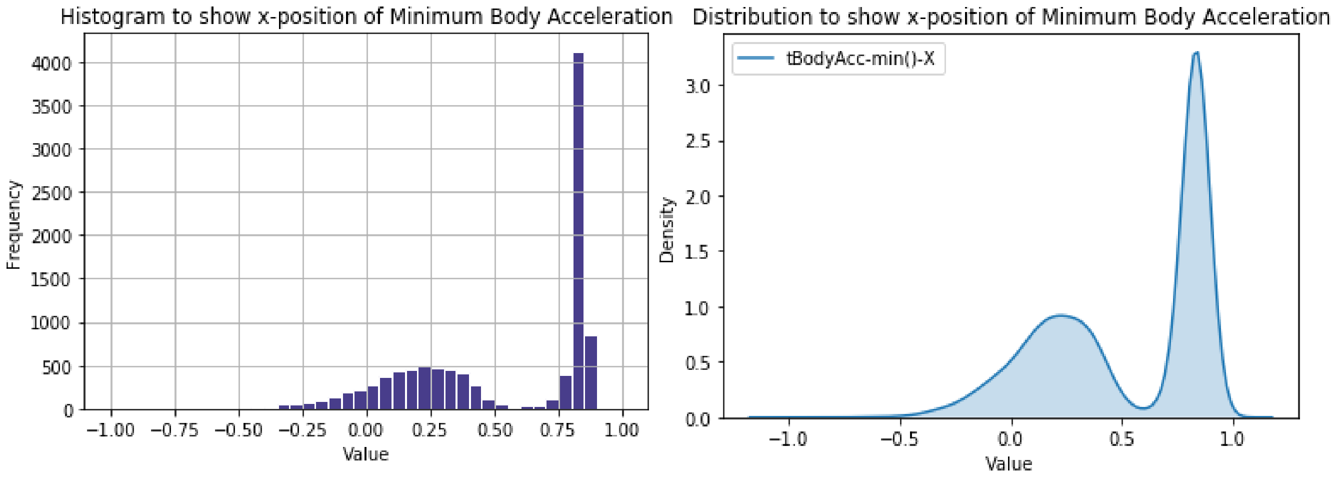


Figure 2. Histogram and distribution of Minimum Body Acceleration in the x-direction

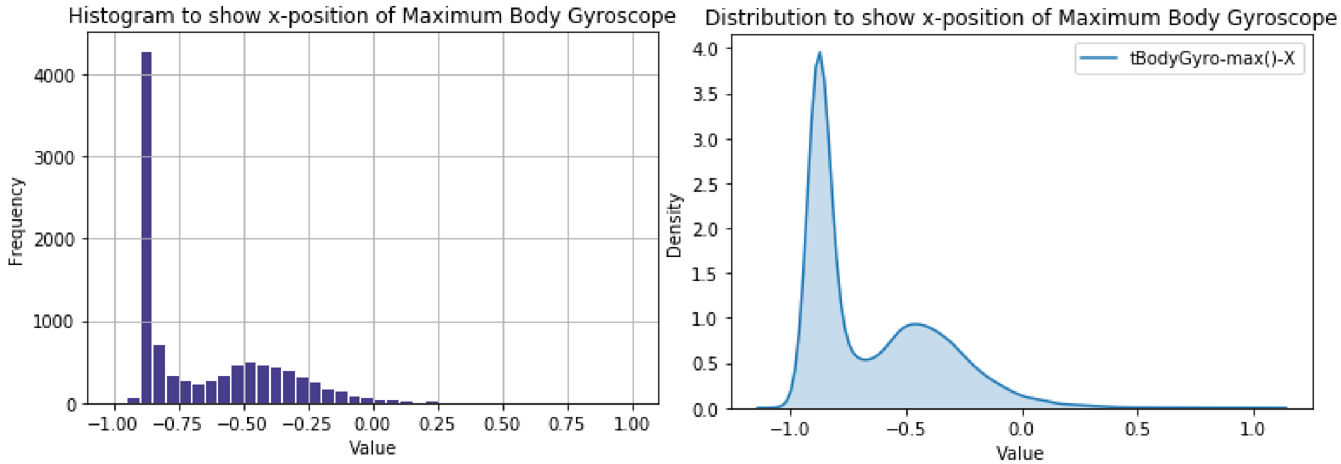


Figure 3. Histogram and distribution of Maximum Body Gyroscope in the x-direction

Figure 4 shows the correlation between the X, Y and Z directions for the Mean Body Acceleration feature. The six different activities are represented by the various colors. Values for laying are spread further out. The values for the other five activities are clustered closer together.

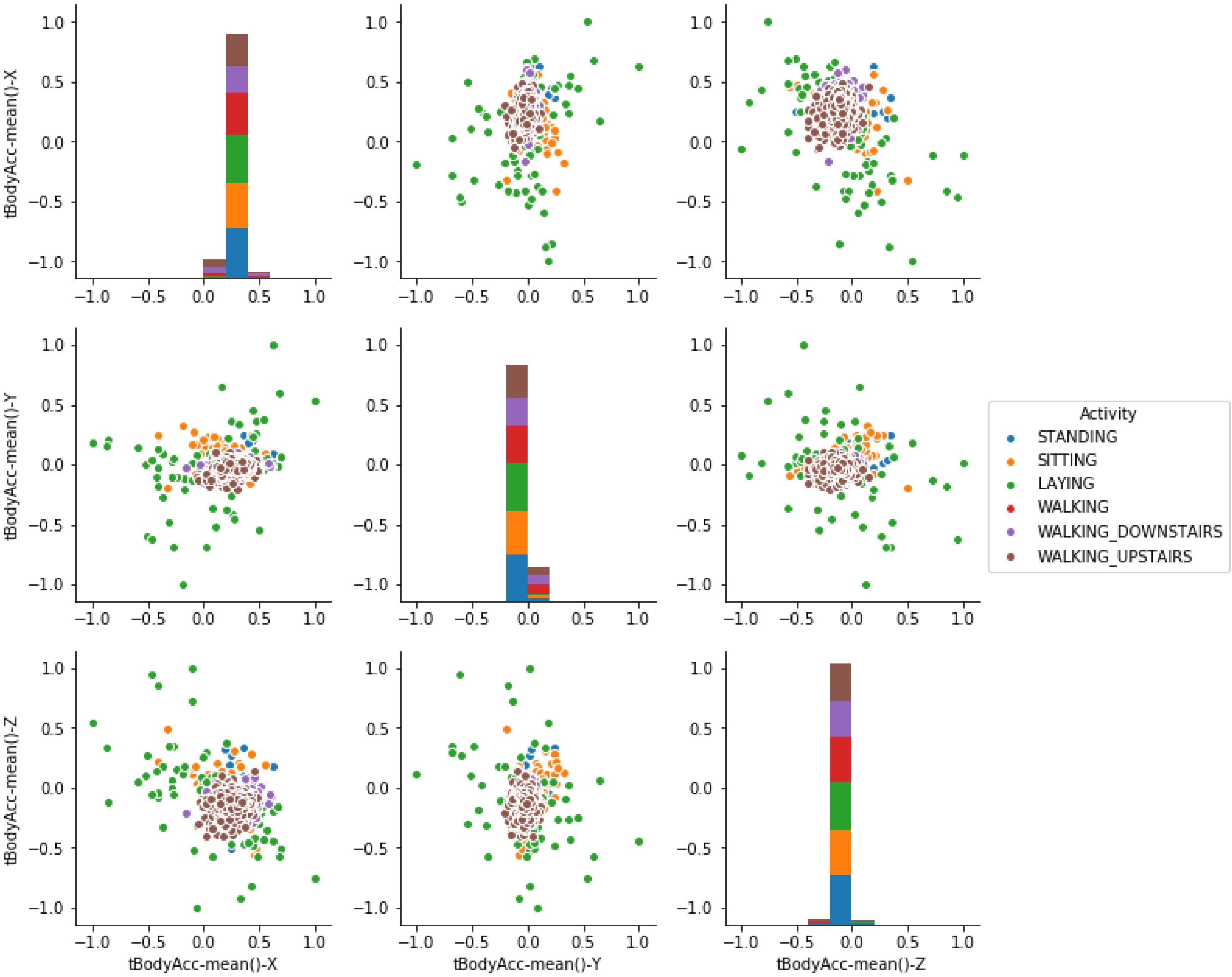


Figure 4. Pairplot showing the correlation between the X, Y and Z directions for Mean Body Acceleration by activity

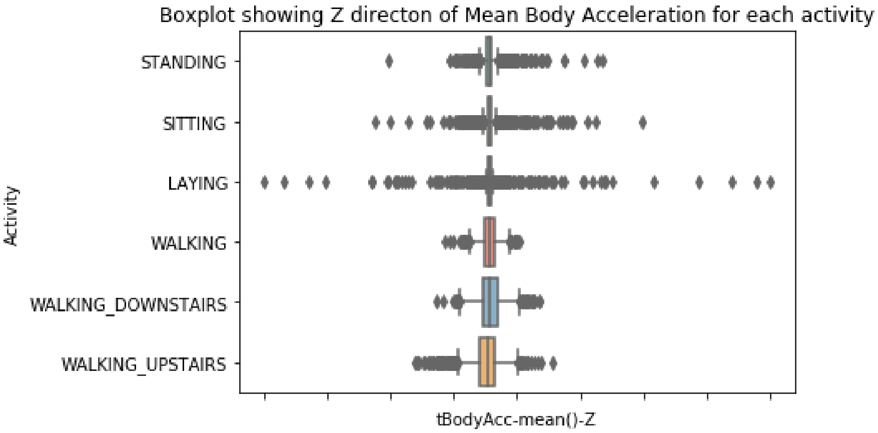


Figure 5. Boxplot of Mean Body Acceleration in the Z direction

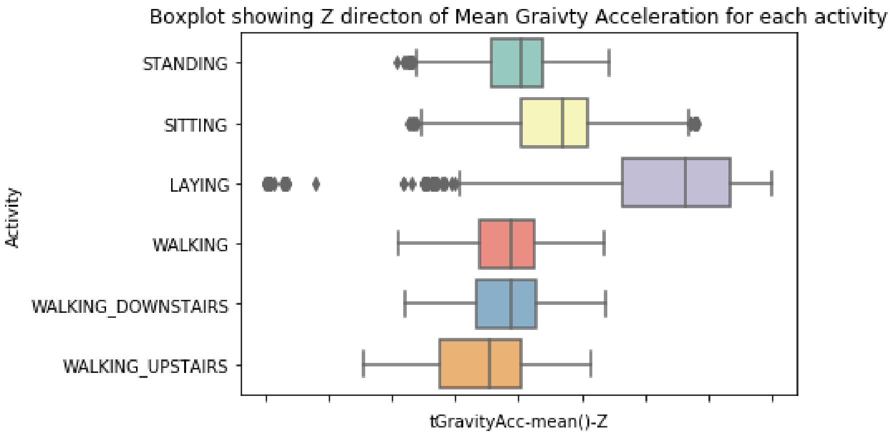


Figure 6. Boxplot of Mean Gravity Acceleration in the Z direction

Figures 5 and 6 show boxplot distributions of two different features. From the boxplots, the values for Mean Body Acceleration seem to contain many outliers and the outliers are especially prominent for the laying activity. Figure 6 also supports the claim since there are several outliers to the left of the boxplot.

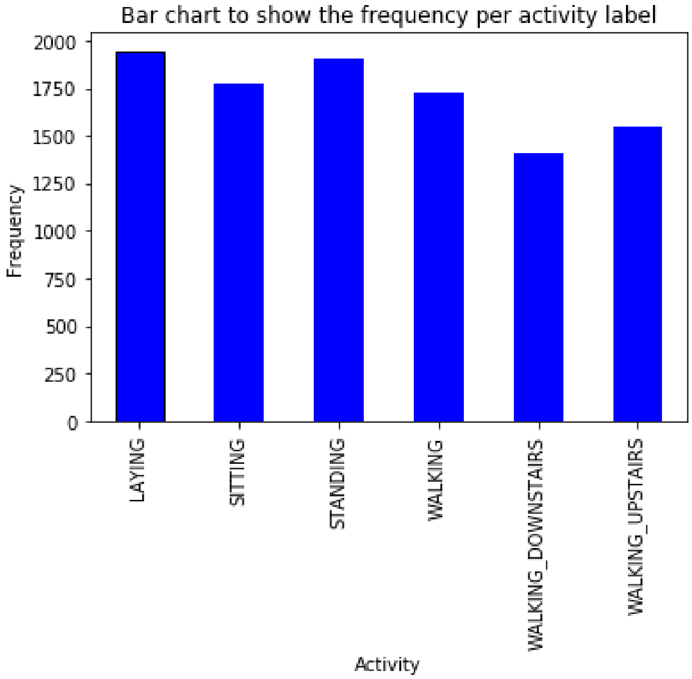


Figure 7. Bar chart to show the number of records for each activity

The six activity labels are evenly spread out and the potential risk of class imbalance is unlikely to cause issues in our dataset. Figure 7 shows that the laying activity appears the most in the dataset while walking downstairs appears least. Parallel coordinates shown in Figures 8-10 show that laying has a wider range of values when compared to other activity labels. Many of the extreme values (close to 1 and -1) seem to occur for the laying activity label.

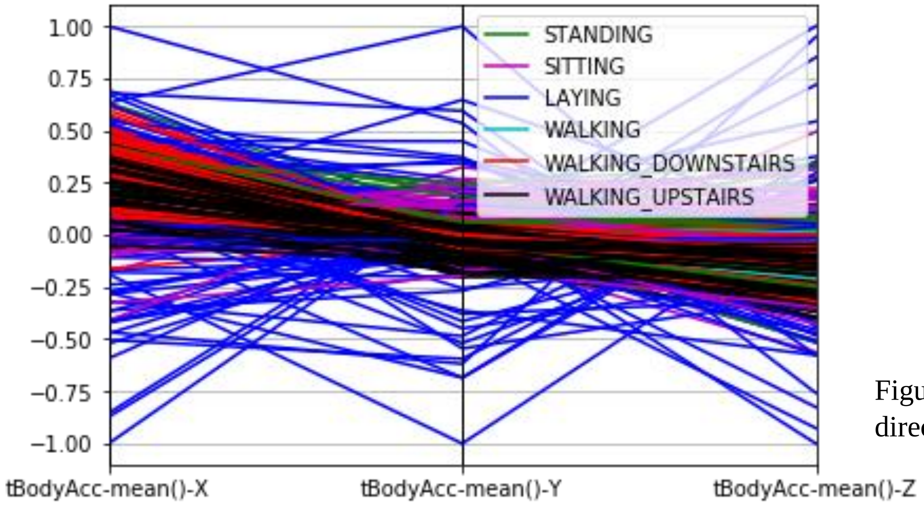


Figure 8. Parallel coordinates of X, Y and Z directions of Mean Body Acceleration

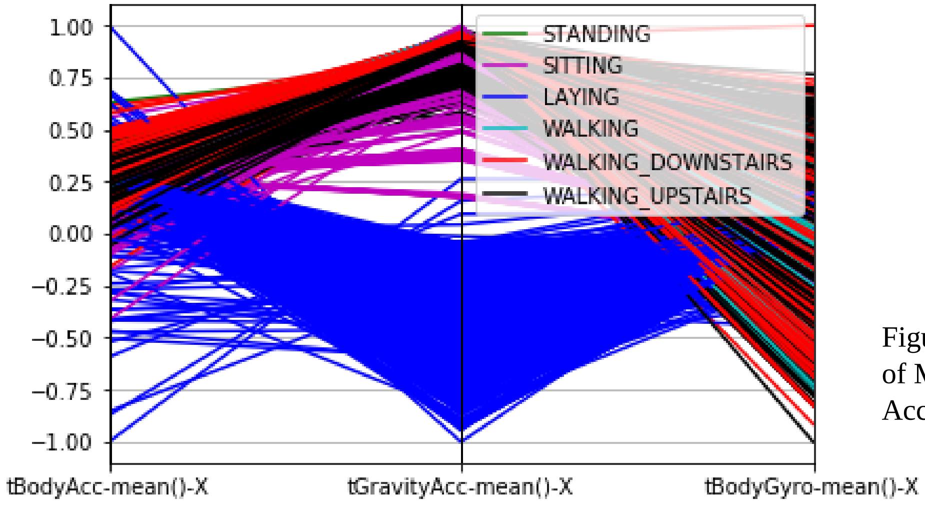


Figure 9. Parallel coordinates of X direction of Mean Body Acceleration, Mean Gravity Acceleration and Mean Body Gyroscope

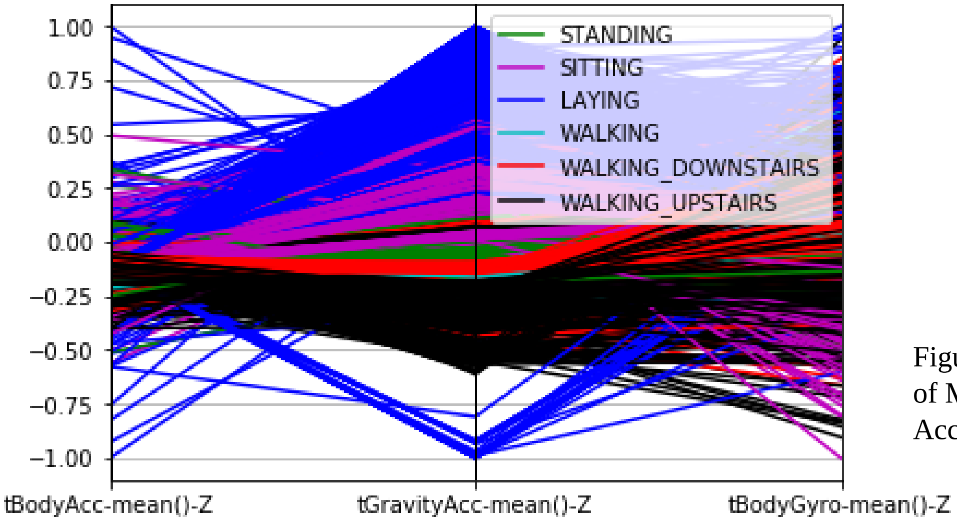


Figure 10. Parallel coordinates of Z direction of Mean Body Acceleration, Mean Gravity Acceleration and Mean Body Gyroscope

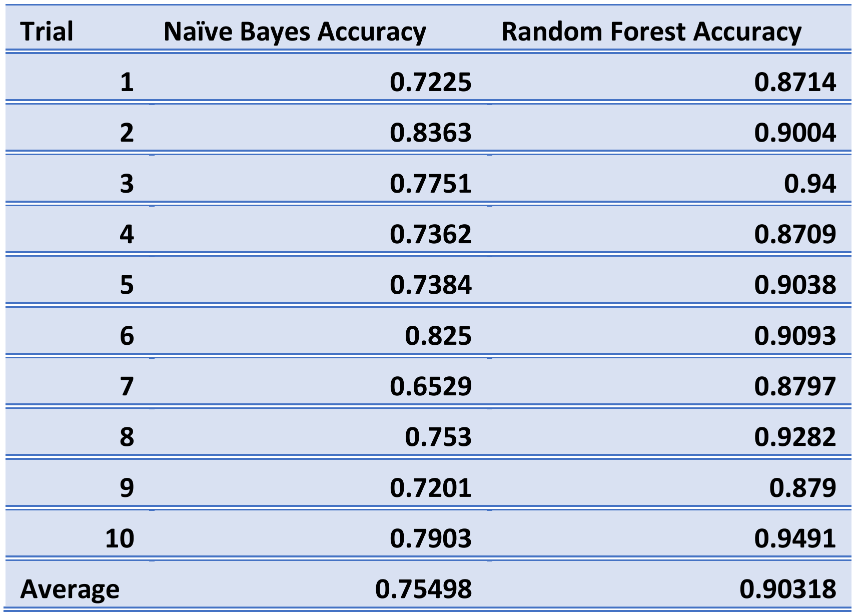
Methods

There were two objectives for this project. The first objective was to choose two predictive models for all 561 features. Within the classification algorithms, the models chosen were the Naïve Bayes and Random Forest classifiers. The Naïve Bayes model in particular took the Gaussian form, since the data being worked with was continuous values. Bayes is a good choice for a model if we are able to assume conditional independence among the variables in the dataset. Random Forest builds multiple decision trees and merges them together. It is mostly trained with the bagging method, and the combination of learning models may increase the accuracy of the model. Another advantage of the Random Forest is that if there are enough trees in the forest, overfitting will be an unlikely event. (Donges, 2018)

The second objective was to reduce the set of features by performing feature selection in order to achieve the desired accuracies of 80% and 90%. The method use for this objective was to select a variable number of best features from the dataset and to apply a model to the new subset of features. The model chosen for this objective was the Random Forest model.

Analysis and Results

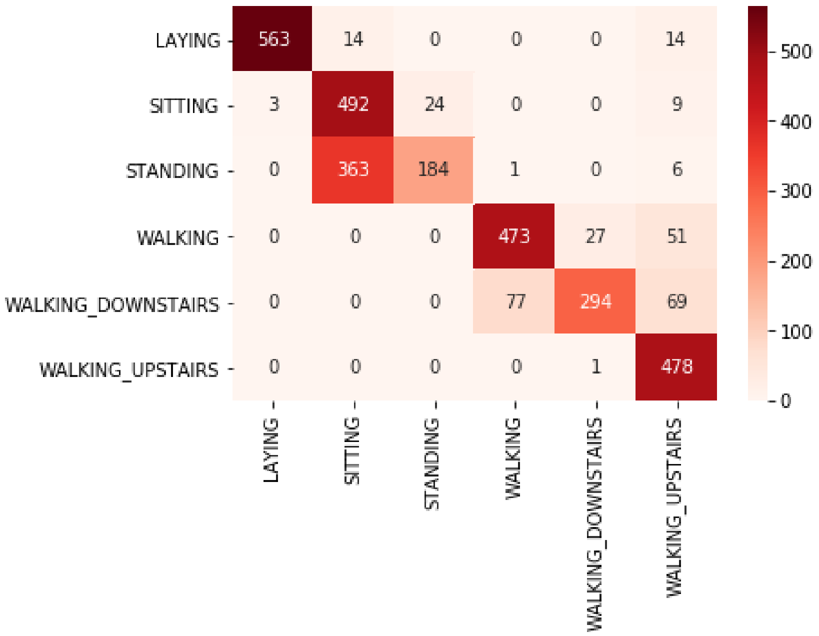
Table 1. Accuracy of the two predictive models over ten trials



Trial	Naïve Bayes Accuracy	Random Forest Accuracy
1	0.7225	0.8714
2	0.8363	0.9004
3	0.7751	0.94
4	0.7362	0.8709
5	0.7384	0.9038
6	0.825	0.9093
7	0.6529	0.8797
8	0.753	0.9282
9	0.7201	0.879
10	0.7903	0.9491
Average	0.75498	0.90318

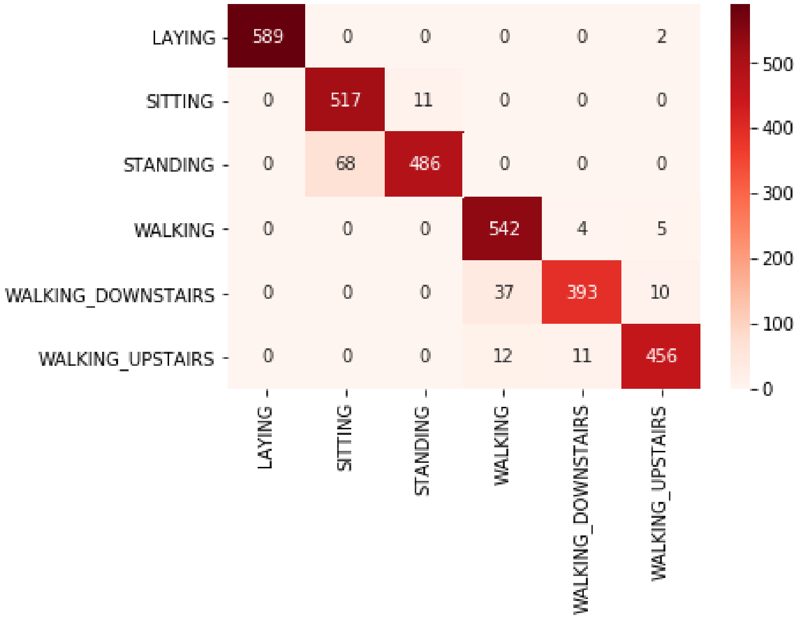
Each model was applied ten times to new sets of training data for each trial. The average accuracy given by the Naive Bayes was 75.5% and 90.32% for the Random Forest model as shown in Table 1. The Naive Bayes classifier assumes conditional independence among attributes which contributes to loss in accuracy. In a practical setting, dependencies exist among variables and Naive Bayes would not be the best model for this dataset. Random Forest provides better accuracy in this case.

Figure 11. Confusion Matrix given by Naïve Bayes Model



	LAYING	SITTING	STANDING	WALKING	WALKING_DOWNSTAIRS	WALKING_UPSTAIRS
LAYING	563	14	0	0	0	14
SITTING	3	492	24	0	0	9
STANDING	0	363	184	1	0	6
WALKING	0	0	0	473	27	51
WALKING_DOWNSTAIRS	0	0	0	77	294	69
WALKING_UPSTAIRS	0	0	0	0	1	478

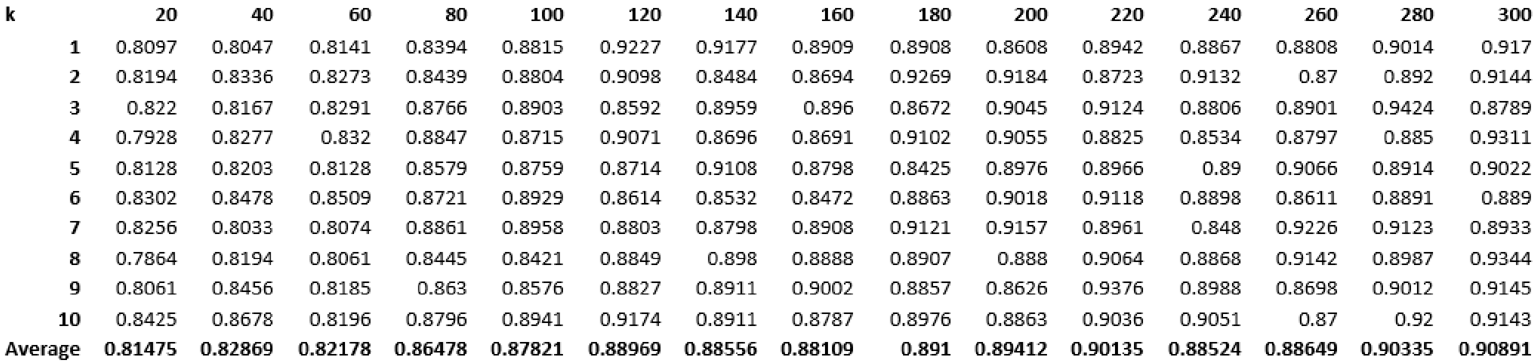
Figure 12. Confusion Matrix given by Random Forest Model



	LAYING	SITTING	STANDING	WALKING	WALKING_DOWNSTAIRS	WALKING_UPSTAIRS
LAYING	589	0	0	0	0	2
SITTING	0	517	11	0	0	0
STANDING	0	68	486	0	0	0
WALKING	0	0	0	542	4	5
WALKING_DOWNSTAIRS	0	0	0	37	393	10
WALKING_UPSTAIRS	0	0	0	12	11	456

The confusion matrices in Figures 11 and 12 are for one of the trials. The Naive Bayes classifier mostly confused sitting with standing. It predicted that the volunteer was sitting when he/she was actually standing for a large number of cases. Walking downstairs was also incorrectly predicted for several cases and it could not accurately tell between walking, walking upstairs and walking downstairs. This is also true for the Random Forest model, although, altogether the number of incorrect predictions were fewer than the Naive Bayes model.

Table 2. Accuracies of varied number of features over ten trials



k	20	40	60	80	100	120	140	160	180	200	220	240	260	280	300
1	0.8097	0.8047	0.8141	0.8394	0.8815	0.9227	0.9177	0.8909	0.8908	0.8608	0.8942	0.8867	0.8808	0.9014	0.917
2	0.8194	0.8336	0.8273	0.8439	0.8804	0.9098	0.8484	0.8694	0.9269	0.9184	0.8723	0.9132	0.87	0.892	0.9144
3	0.822	0.8167	0.8291	0.8766	0.8903	0.8592	0.8959	0.896	0.8672	0.9045	0.9124	0.8806	0.8901	0.9424	0.8789
4	0.7928	0.8277	0.832	0.8847	0.8715	0.9071	0.8696	0.8691	0.9102	0.9055	0.8825	0.8534	0.8797	0.885	0.9311
5	0.8128	0.8203	0.8128	0.8579	0.8759	0.8714	0.9108	0.8798	0.8425	0.8976	0.8966	0.89	0.9066	0.8914	0.9022
6	0.8302	0.8478	0.8509	0.8721	0.8929	0.8614	0.8532	0.8472	0.8863	0.9018	0.9118	0.8898	0.8611	0.8891	0.889
7	0.8256	0.8033	0.8074	0.8861	0.8958	0.8803	0.8798	0.8908	0.9121	0.9157	0.8961	0.848	0.9226	0.9123	0.8933
8	0.7864	0.8194	0.8061	0.8445	0.8421	0.8849	0.898	0.8888	0.8907	0.888	0.9064	0.8868	0.9142	0.8987	0.9344
9	0.8061	0.8456	0.8185	0.863	0.8576	0.8827	0.8911	0.9002	0.8857	0.8626	0.9376	0.8988	0.8698	0.9012	0.9145
10	0.8425	0.8678	0.8196	0.8796	0.8941	0.9174	0.8911	0.8787	0.8976	0.8863	0.9036	0.9051	0.87	0.92	0.9143
Average	**0.81475**	**0.82869**	**0.82178**	**0.86478**	**0.87821**	**0.88969**	**0.88556**	**0.88109**	**0.891**	**0.89412**	**0.90135**	**0.88524**	**0.88649**	**0.90335**	**0.90891**

For the second objective, a univariate feature selection was used to perform feature selection and the best features were selected according to ANOVA F-values. The chi squared method was not used since it is normally used for non-negative data. Various number of features were tested for ten trials and results showed that it would take fewer than 20 features to achieve 80% accuracy and around 280 features to achieve 90% accuracy. From the first objective, the Random Forest model has 90.32% accuracy with all 561 features. By reducing the number of features to 300, we are still able to achieve 90.89% accuracy. Table 2 shows the difference in accuracies using the Random Forest model for a number of increasing features in increments of 20 up to 300 features.

The scatterplot in Figure 13 summarizes the results achieved by varying the number of features using this method. The general trend shows a general upward trend. The correlation is positive and generally speaking, as the number of attributes increase, the accuracy increases.

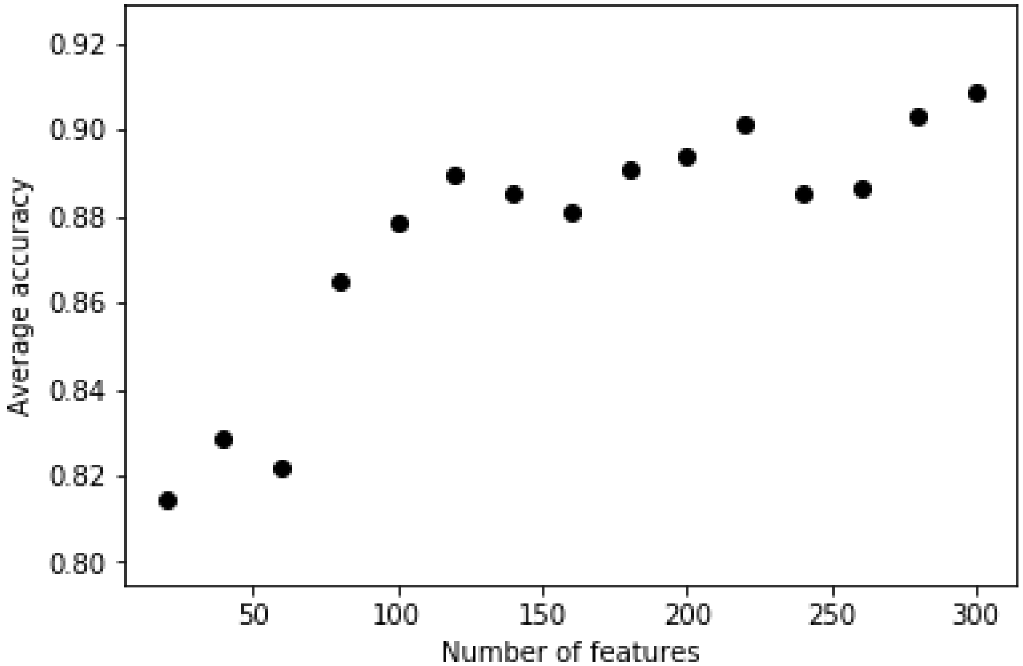


Figure 13. Scatterplot showing the number of features again average accuracy

Conclusion

The prediction of human activities with an accuracy as high as 90% can be achieved by varying the number of features in the set. The findings are able to classify whether a person is sitting, standing and laying, although the chosen models do not provide very good accuracy for differentiating between walking, walking upstairs and walking downstairs. Through data collection with smartphones and their sensor signals, we can almost accurately predict the activities of a new person.

References

1. Train/Test Split and Cross Validation in Python. URL: https://towardsdatascience.com/traintest-split-and-cross-validation-in-python-80b61beca4b6 (Accessed: 16 September 2018)

2. Using a smartphone acceleration sensor to study uniform and uniformly accelerated circular motions. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172014000200015 (Accessed: 13 September 2018)

3. Simple Graphing with IPython and Pandas. URL: http://pbpython.com/simple-graphing-pandas.html (Accessed: 13 September 2018)

4. High-Dimensional Data Visualization. URL: https://brage.bibsys.no/xmlui/bitstream/handle/11250/2454370/16474_FULLTEXT.pdf?sequence=1&isAllowed=y (Accessed: 17 September 2018)

5. Visualization with Seaborn. URL: https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html (Accessed: 17 September 2018)

6. Data Science with Python: Intro to Loading, Subsetting, and Filtering Data with pandas. URL:

https://towardsdatascience.com/data-science-with-python-intro-to-loading-and-subsetting-data-with-pandas-9f26895ddd7f (Accessed: 17 September 2018)

7. What is Feature Selection. URL: https://machinelearningmastery.com/an-introduction-to-feature-selection/ (Accessed: 18 September 2018)

8. Feature Selection in Python with Scikit-Learn. URL: https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/ (Accessed: 18 September 2018)

9. Feature importances with forests of trees. URL: http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html#sphx-glr-auto-examples-ensemble-plot-forest-importances-py (Accessed: 19 September 2018)

10. The Random Forest Algorithm. Donges. 2018. URL: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. (Accessed: 20 September 2018)