Smartphone and Smartwatch Activity Prediction Using

Machine Learning

Nicholas Gurnard

Undergraduate Mechanical Engineering Student at University of California, Irvine

nk.gurnard@gmail.com

# Abstract

Fill in Later!!

# 1 Introduction

Smart phones and smart watches are wielded by millions of people in societies that all over the world. These gadgets contain powerful data collection hardware inside of them including acceleration sensors, cameras, location sensors, compasses, fingerprint sensors, and even heart rate monitoring sensors. The collection of these various sensor readings can give insight and research opportunities to many fields of study, including healthcare and exercise science, and much more.

Many companies are already taking this data into consideration, including Apple partnering with Johnson & Johnson in their Heartline Study [1]. The goal of the Heartline study is to analyze several variables from smart devices in order to create models to predict whether someone is at risk for stroke.

This project analyzes sensor readings from the WISDM smartphone and smartwatch activity and biometrics dataset [2]. The WISDM dataset includes data collected from 51 different subjects numbered 1600 to 1650, each of which performing activities such as walking, typing, clapping, eating soup, eating pasta, etc. Each user was instructed to perform each activity from a list of 18 different activities, listed in Table 1, for a total of 3 minutes. Each of the 51 subjects was equipped with a smartphone and smartwatch, which held accelerometers and gyroscopes inside of them that collected acceleration data in each of the coordinate directions x, y, z. The subjects were given a LG G watch as their smartwatch and one of three types of smartphones: Google Nexus 5, Google Nexus 5x, or Samsung Galaxy S5.

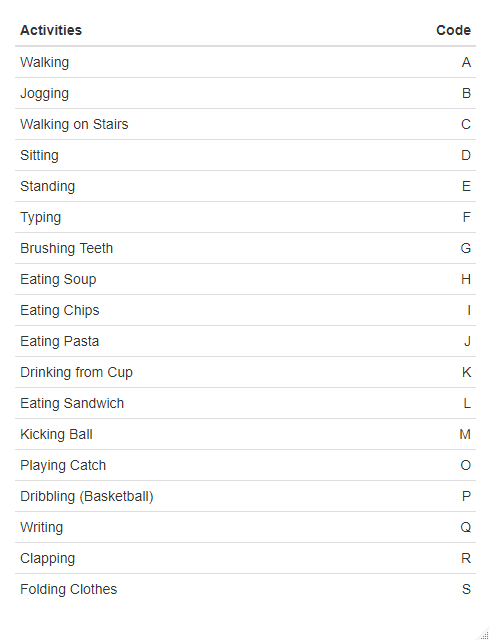


Table 1 - List of activities and corresponding code

Recognition of various activities from accelerometer and gyroscope data can help researchers identify causes for various diseases or medical issues. This project aims to use feature extraction techniques to filter the signal out of the noise and create a framework for predicting what activity a subject is performing.

# 2 Methodology

The WISDM dataset provides 15,630,426 raw accelerometer and gyroscope readings for each of the x, y, and z axes. For each reading there is an associated time stamp, activity code, and subject identifier. Each reading was sampled at 20Hz for both the smartwatch and the smartphones. There are effectively 4 types of readings available: phone accelerometer, phone gyroscope, watch accelerometer, and watch gyroscope readings (pa, pg, wa, wg).

## 2.1 Data Analysis

The data provided gave no indication as to which orientation each device was tested in. The directions x, y, and z could have been any direction in the 3D space the device occupies. In order to identify how the device was oriented relative to the ground, Figure 1 was constructed to show the acceleration for each of the three axes.

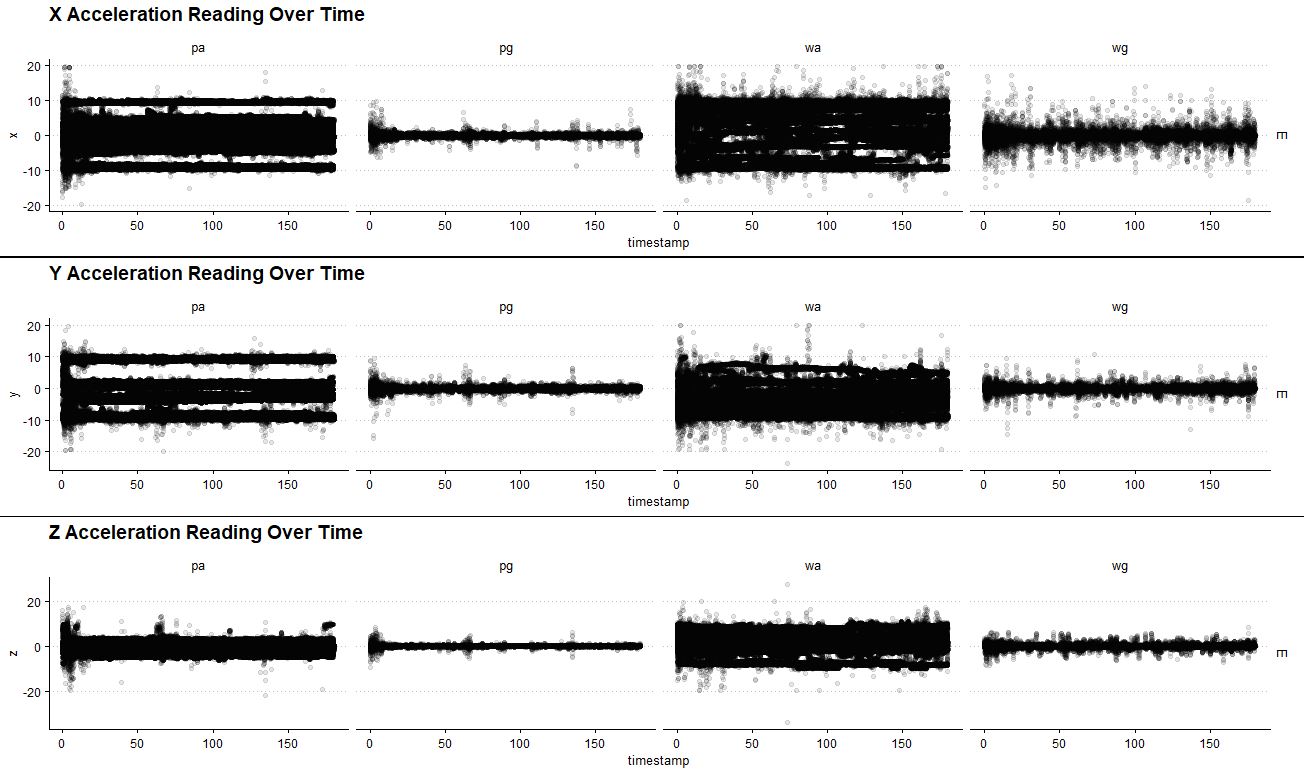


Figure 1 - Acceleration sensor readings for the 4 effective types of readings over time

For each axis, x, y, and z, it is clear from the figure that the orientation of each subjects’ phone and watch are not consistent. Upright orientation occurs when the acceleration matches that of the earth’s gravitational constant of magnitude 9.8 m/s2. This could be due to the fact that each phone has the sensors mounted differently, or upon manufacturing the orientation of the sensor isn’t consistent. Additionally, every subject could have wielded the devices at different angles when performing activities. Whatever the causation may be, it is clear that orientation cannot be taken into consideration without extensive data cleaning. Therefore, it is assumed that the device orientation is not a variable in the prediction of activities. If data were to be collected real time from millions of different people, the orientation would also not be consistent, so it is important to create a model that ignores that fact.

Figure 2 illustrates what a typical signal looks like from the jogging activity. It is clear that there is a lot of noise in the acceleration reading over time, so feature extraction is necessary to create an accurate prediction model.

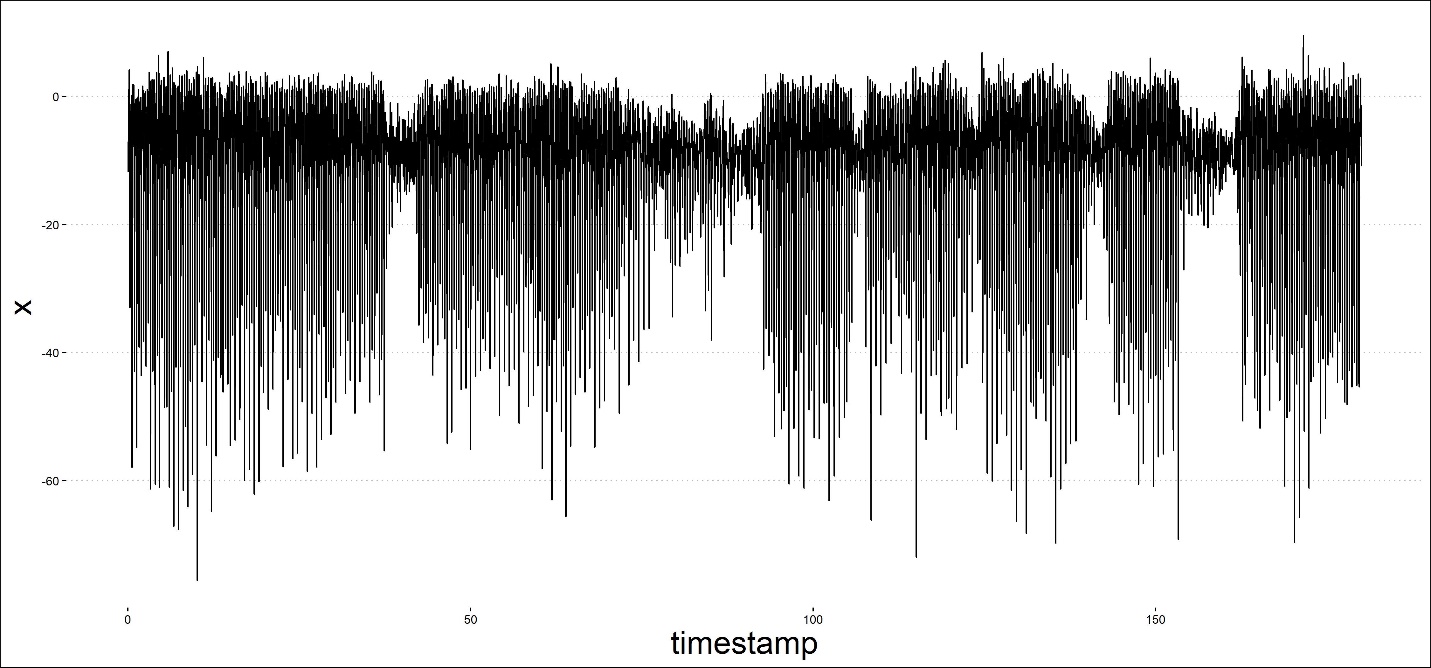


Figure 2 - Jogging signal from subject 1646: x acceleration over time

Upon data import, there was drift among the sensor readings that made it such that the readings were not sampled at exactly 20Hz. For the sake of simplification in order to save model training time, the time stamps were corrected to be exactly 50 milliseconds apart (20Hz).

After some subject analysis, it was discovered that every subject failed to perform every activity. The subject that failed were subjects 1607, 1609, 1616, 1642, and 1643. These users were not included in the training of the models.

## 2.2 Feature Extraction

Building a model to predict the activities with only the instantaneous accelerometer and gyroscope readings, AKA the raw data, is extremely inaccurate. Out of the 3 minutes each subject performed the activity, a total of 2 minutes was extracted from the activity. Because of uncertainty as to how the data was collected, the start and the end of each activity was removed in case the subject had recordings made while setting up and shutting down their devices, which then amounted to 2 minutes specified.

In order to extract a clearer signal for each activity, 19 different types of features were extracted from the raw data for each of the 4 effective types of readings (pa, pg, wa, wg). For the extraction of these features, a window of a size 10 seconds is considered. A 10 second window was chosen because 10 seconds allowed for an adequate number of cycles of the activity being analyzed. The window progressively rolls across the 2-minute span of the activity for each subject with an overlap of 50%. For example, if the first window covers the first 10 seconds of the 2-minute span, that would be rows 1-200. The following window would be rows 100-300 and the window would keep rolling until it reaches the end of the 2-minute span. A visual of how the window operation works is Illustrated in Figure 3.



Figure 3 - Rolling window visual with 50% overlap.  
Each column represents the whole 2-minute span.

The types of features extracted are listed below, with the number of features generated in braces:

* Mean {3} – The average sensor reading over the window. One for each axis
* Standard Deviation {3} – The standard deviation of the sensor readings over the window. One for each axis.
* Variance {3} – The variance of the sensor readings over the window. One for each axis.
* Time Between Peaks {3} – The time between the peaks of the sensor readings over the window. One for each axis.
* Average Resultant Acceleration {1} - The average resultant value of the sensor readings over the window. Found by taking square root of the sum of each instantaneous reading for each axis squared. Then, averaging those values over the window.
* Maximum {3} – The maximum sensor reading over the window. One for each axis.
* Minimum {3} – The minimum sensor reading over the window. One for each axis.

The maximum, minimum, and time between peaks features are extremely sensitive to outliers. To correct for the outliers, the window was smoothed using a moving average. A small sub-window of size 10 readings, or 0.5 seconds, was used as the smoothing parameter that operates as another rolling window. The sub-window was intentionally designed to be very small to ensure that the data was not being smoothed too much. This way, important peaks are still kept with relatively the same amplitude, and the signals between activities are still distinct. The maximum, minimum, and time between peaks features were then calculated from this slightly smoothed signal.

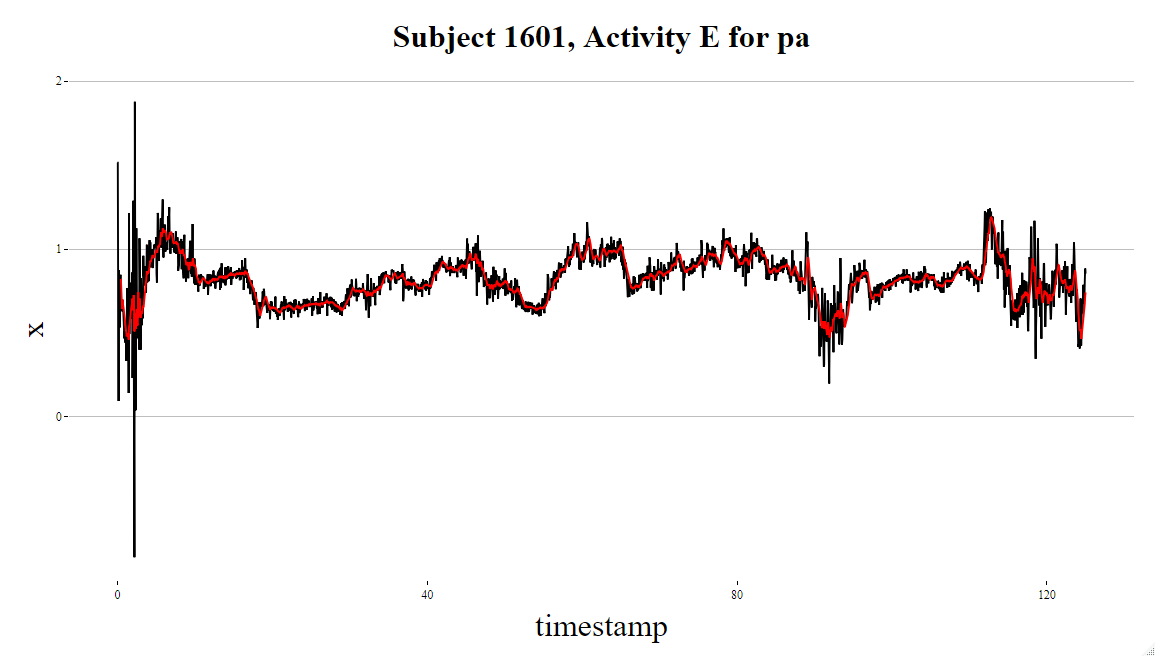


Figure 4 - Moving average visual to demonstrate close fit.  
Red Line: Smoothed using moving average with sub-window of size 0.5 sec.  
Black Line: original signal

It is important to note that, in general, smoothing signals is poor practice because it makes the various signals for each activity too similar to one another and confuses the predictive model [3]. However, since the smoothing was not too extreme, the model performed roughly 3% better on average when the data was smoothed.

In addition to the 19 extracted features, the activity label and the subject numeric identifier (1600-1650) were also included as variables for the model for a total of 21 variables.

Once the features were extracted, a correlation matrix was created to determine which variables were highly correlated, depicted in Figure 5. Note that the highly correlated features have a correlation coefficient close to magnitude 1.

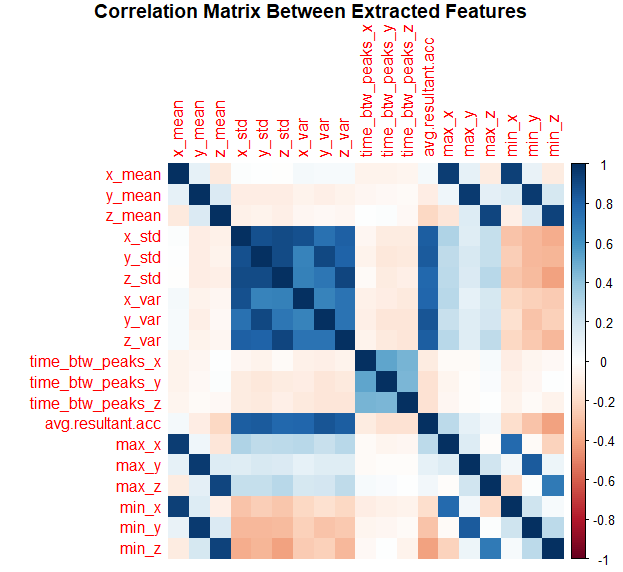


Figure 5 - Correlation matrix between extracted features

As expected, the variance and the standard deviation features are correlated because the standard deviation is simply the square root of the variance. In addition, a relatively strong correlation exists between the average resultant acceleration and the standard deviation/variance, as well as the minimum/maximum and the mean. In the creation of the model, it is important to recognize what is correlated in order to avoid confusing the model thus decreasing prediction accuracy.

## 2.3 Building a Model

The random forest classification algorithm was the main predictive model used to identify subject activity. A random forest is powerful for this particular data set because there is little need for model interpretability, only high performance. There is no clear indication of clustering, which is the motivation for choosing random forests over a simpler, faster, and more interpretable model such as KNN. Random forests are bagged decision tree models that randomly choose a specified number of *m* predictors as split candidates from a full set of *p* predictors. Each split can only use one of the *m* predictors and a fresh set of *m* predictors is taken at each split [4]. This leads to decorrelation of the trees within the forest thus leading to lower variance. The predictors in this case will be the extracted features.

For each of the 4 effective reading types (pa, pg, wa, wg), There were 3-4 random forests made. The difference between each of the random forests is outlined as follows:

1. RF1 - Random Forest predicting all activities at once.
2. RF2 - Random Forest predicting only activities where the corresponding smart device is of importance.
   1. Phone activities: walking (A), jogging (B), walking on stairs (C), sitting (D), standing (E).
   2. Watch activities: typing (F), brushing teeth (G), eating soup (H), eating chips (I), eating pasta (J), drinking from a cup (K), eating sandwich (L), kicking ball (M), playing catch (O), dribbling (P), writing (Q), clapping (R), folding clothes (S)
3. RF3 - Random forest predicting all activities at once and eating activities are combined as one category.
   1. Eating activities: H, I, J, L
4. RF4 - Random forest predicting only activities where the corresponding smart device is of importance and eating activities are combined into one category.

In the creating of each random forest, the highly correlated features identified from Figure 5 were removed as variables. The variance was removed instead of the standard deviation because it resulted in marginally less accuracy for each random forest. The minimum, maximum, average resultant acceleration, and mean features were all reincluded in the random forests because the removal of any combination of those features resulted in a decrease in accuracy that was greater than 5%. Additionally, the value of *m* was tuned to get the maximum possible accuracy.

The features were then split into model training data and model validation (test) data. The first 75% of features for each of the 4 types of readings was used as the model training data, and the remaining 25% as the test data. This allowed for the model to be trained on most users while intentionally leaving out several users to then validate the model.

## 3 Results and Discussion

The performance of each random forest for each reading type is outlined in Table 2. Note that RF4 for the phone has no entries because there are no eating activities where the phone was considered of importance.

For each random forest model, the performance was assessed by prediction accuracy, Cohen’s kappa coefficient, and a 95% confidence interval. The accuracy column displays what the percent agreement is between the predictions made from the test data and the actual test data. The kappa coefficient is a number that lies between 0 and 1 that gives insight as to how many of your correct predictions may have resulted by pure chance. A kappa closer to 1 indicates that there is less chance that the prediction accuracy was due to chance [5]. Lastly, 95% confidence interval is the interval in which the model accuracy will fall given a new validation set with 95% confidence.

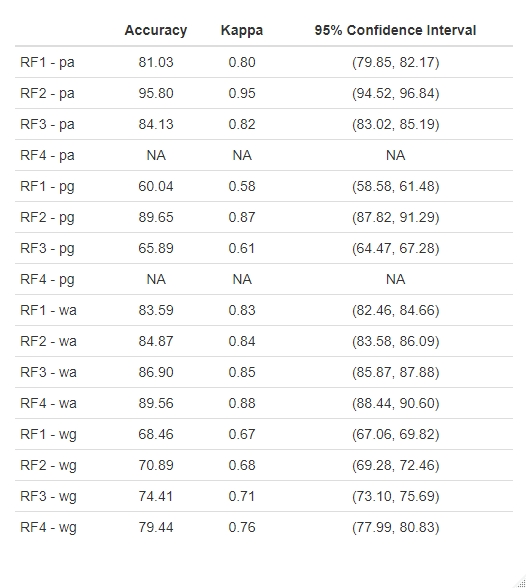


Table 2 – Performance of each random forest model

When comparing the features from the accelerometer readings against the gyroscope readings, the accelerometer outperformed the gyroscope. This is likely due to the fact that gyroscopes and accelerometers are inherently different in how they collect acceleration data. An accelerometer measures the rate of change of velocity an object, while a gyroscope maintains its orientation by allowing the freedom of rotation and then measuring rotational changes in velocity [6]. Therefore the, rate of change of the velocity linearly is a more important factor than any rotational movements.

As expected, the phone features performed significantly better when considering only the activities where the phone is of importance (RF2-pa). A maximum classification accuracy of nearly 96% was achieved when using only the phone accelerometer data in the RF2 model. In addition, this accuracy wasn’t due to chance because of a kappa value of 0.95 and is extremely close to 1. The watch features performed slightly better to the phone features when trying to predict all activities at once. However, only marginal improvements were made when considering only activities where the watch is of importance (RF2-wa) against predicting all activities at once (RF1-wa). So, even after filtering out phone activities, the model struggled to reach a result as impressive as RF1-pa. The causation for this absence of drastic improvement is because of how similar the eating activities were to each other.

The differentiation between activities H, I, J, and L, the eating activities, was the most difficult challenge when considering any subset of data with the given features extracted. About a 5% increase in accuracy was observed between models RF1 and RF3 for each of the 4 types of readings when combining all eating activities. Activity K, drinking from a cup, was equally as difficult to distinguish from each of the eating activities. However, it was intentionally left separate because a subject eating doesn’t necessarily imply drinking and vise-versa. If it were to be included in the combination of eating activities, the accuracy would likely increase by a few percent. Table 3 shows more performance measures for RF1-pa to demonstrate the difficulty of differentiating between eating activities/drinking.

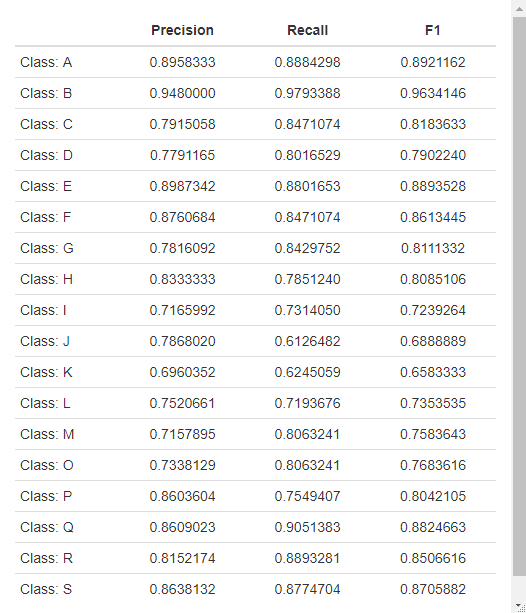


Table 3 - Extra performance measures for RF1-pa

For activities H, I, J, K, and L, the recall and F1-score are much lower in comparison to the other activities. The F1-score is the harmonic mean of the recall and the precision, where a higher F1-score indicates better results. Therefore, differentiating between tasks that entirely different is much easier than trying to distinguish eating chips vs. eating pasta.

After each model was created and assessed, a variable importance plot was created to identify the impact each variable had in the training of the model, shown in Figure 6. Each variable importance plot was slightly different when training each forest, however they are largely very similar.

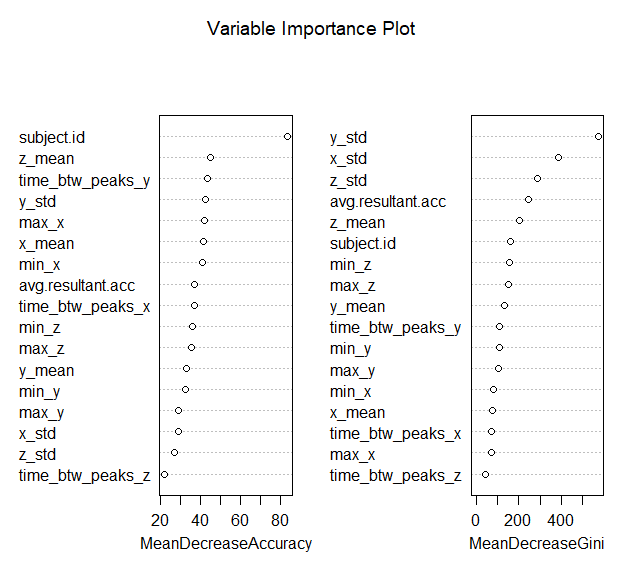


Figure 6- Variable importance plot for RF3-pa

The feature that was most critical in the performance of every model was the subject identifier, as demonstrated by the left-hand panel of Figure 6. What matters most when analyzing a single activity is the style in which each activity is performed. For example, the features of one user while walking is entirely different than another user while walking despite the fact that they are performing the same exact activity.

Every feature in every model contributed significantly to the overall performance of the random forest models thus signifying the simplistic features that were chosen are appropriate when predicting subject activity.

## 4 Conclusions

This project sought to predict subject activity when extracting only simple features from smartphone/smartwatch accelerometer and gyroscope sensor readings and identify what features are most important. Accurately predicting subject activity based solely off of one of many sensors in a smart device can help researches identify patients at risk of various ailments, help medical professionals monitor exercise and rehabilitation routines, or even aid in law enforcement when identifying what a suspect was doing at the time a crime was committed. Using the WISDM smartphone and smartwatch activity and biometrics dataset, several random forest models were used to predict activity and assess variable importance.

Instead of trying to predict users off of instantaneous sensor readings, simple features can help a prediction model filter the signal out of the noise. The single most important feature when predicting subject activity is knowing which subject is physically performing the activity. The style in which each activity performed is critical information when trying to train a model that has to interpret several people performing the exact same activity. In addition, separating activities by device is

XXXXXXXXXXXXXXXXX FINSIH XXXXXXXXXXXXX

## 5 Acknowledgements

I especially thank professor Donald Dabdub for his inspiration in me pursuing this project. Thanks to my peers Felix Slothower, Jonathon Palafoutas, Brian Fox, Courtney Rowe, Patrick Youssef, and Eric Abdulaziz in provoking new ideas on next steps for the project and useful discussions.

## 6 References

# References

|  |  |
| --- | --- |
| [1] | H. Apple and J&J, "You’re Invited to Join the Heartline Study," 12 3 2020. [Online]. Available: https://www.heartline.com/?utm\_source=google&utm\_medium=cpc&utm\_campaign=GO-USA-ENG-PS-Heartline-BC-BM-RN-Brand\_Core\_Desktop&utm\_content=Core&utm\_term=%2Bheartline&&gclid=Cj0KCQjwu6fzBRC6ARIsAJUwa2T52bRBCzfKL9E50dGMRUt-Lf0OTzp7Vxmwzvs3x\_hhrzsnbY\_cttwaAqOd. |
| [2] | G. M. Weiss, *WISDM Smartphone and Smartwatch Activity and,* Bronx, New York, 2019. |
| [3] | D. Dabudb, *Professor at UC, Irvine,* Irvine, 2020. |
| [4] | D. W. T. H. Gareth James, An Introduction to Statistical Learning, 2017. |
| [5] | A. J. Viera and J. M. Garrett, "The Kappa Statistic - Columbia University," May 2005. [Online]. Available: http://web2.cs.columbia.edu/~julia/courses/CS6998/Interrater\_agreement.Kappa\_statistic.pdf. [Accessed March 2020]. |
| [6] | Shawn, "Seed Studio," 24 December 2019. [Online]. Available: https://www.seeedstudio.com/blog/2019/12/24/what-is-accelerometer-gyroscope-and-how-to-pick-one/. [Accessed March 2020]. |

## 