Sensors, Noise, and Walking Analysis

Project Report

CMPT 353

Spring 2022

Wan Ying Chan (301294873)
Saqib Hasib (301367893)
Gaurav Gupta (301368927)

Table of Contents

1 Introduction

In this project, we study the differences of step frequency amongst different factors such as weight, gender, shoe type, and activity (hiking, running, walking, and sitting) by collecting data using Physics Toolbox Sensor Suite, an Android based smartphone application that uses internal smartphone sensors to collect, display, record, and export .csv data files [1]. The rest of the report is divided as follows. In section 2, the report covers the methodology of our study on data collection and data processing. Section 3 shows our visualization of the results after data processing. In section 4, we present the analysis of our study from the results. Then, section 5 shows the conclusion of our findings of the study. Section 6 presents the problems and limitations we have encountered during our study of the project. Lastly, section 7 and 8 presents our references and accomplishments of this project experience.

2 Methods

This section explains our methodologies used for studying the project. We divided our methods into two parts: Data Collection and Data Processing, shown in section 2.1 and 2.2. Data collection explains the methods and tools used for collecting data. Data processing is further divided into three steps: Data Filtering, Classification methods using machine learning, and Linear regression, which is shown in section 2.2.1, 2.2.2, and 2.2.3.

2.1 Data Collection

Our group has used the Physics Toolbox Sensor Suite for Android smartphones to collect all of our data by attaching the phone onto the ankle of an individual. Then, using either the Roller Coaster or Multi Record data collection mode of the app, we are able to collect the g-Force, Linear Accelerometer, and Gyroscope of our walking data in a 20 seconds interval for each data. The process was then repeated on the other ankle of the individual.

Next, we chose a random sample of subjects who possessed different qualities to accurately model the relationship between step frequency and other metadata such as gender, weight, favorite activity of the individual, and shoe type. Hence, we are investigating whether an individual attribute such as gender, weight, favorite activity, and shoe type can impact an individual's frequency of step. For example, we would assume that an individual with a favorite activity such as hiking or running would likely have a higher frequency of steps compared to other activities such as sitting.

2.2 Data Processing

For this section, we have divided into a few steps to process our data: Data Filtering, Machine Learning Classification Methods, and Linear Regression Methods. The data filtering shows the process and methods of cleaning up data using the Extract-Transform-Load process. The Machine Learning Classification step shows the methods used to investigate the relationships between step frequency and other attributes such as gender, shoe type, and favorite activity. The Linear Regression step explains the process used to show the linear relationship between step frequency and weight.

2.2.1 Data Filtering

Using the raw data that we have collected, we have used the linear acceleration and angular velocity data collected in the x-, y-, z-directions to calculate the vector sum, which will produce the total linear acceleration and total angular velocity measured at each point in time.

Next, we transformed our data by using a low pass Butterworth filter since the raw data we collected using the smartphone sensors contained high frequency noise, and our goal was to eliminate noise using this technique. Hence, we have achieved removing noises from our data by applying the Python's library, SciPy, using the `.butter()` method, with an order of 2 and cutoff frequency of 0.02 half-cycles per sample. Then, the coefficients returned from

the `.butter()` method were then passed into the `.filtfilt()` method, which uses the same Scipy library. This is to use the coefficients to apply the filter on the total linear acceleration and angular velocity data.

After filtering the data, we have separated our frequency using the `.linspace()` method from NumPy Python's library to get evenly spaced intervals of frequencies. Then, perform Fourier transform on the filtered data and frequencies so that we can get average frequencies of steps on the transformed data.

2.2.2 Machine Learning Classification Methods

Classifiers are applied to show the relationship between frequency of steps and favorite activities, gender, and shoe types. The classifiers that we have used for this section are by using the Python's sci-kit learn library, where we include `GaussianNB()`, `KNeighborsClassifier()`, and `RandomForestClassifier()` to train, model and calculate the Bayes Model, k-nearest Neighbor, and Random Forest Model scores. These scores will then be outputted and used as information for us to compare the relationships for analysis.

2.2.3 Linear Regression Methods

Regression was used to show the relationship between frequency of steps and weight of an individual. This is to compare whether weight has an impact on the frequency of steps. The regression method that was used in this section is by using Python's SciPy library, where we include the `linregress()` method to train, model and calculate the linear least-squares regression of the data. Lastly, we calculate the p-value to show the linear regression score of weight and frequency of steps for analysis.

3 Results

This section shows the visualization of the results after Data Filtering, Classification, and Linear Regression methods. Section 3.1 shows the linear acceleration before and after transformation of frequencies. Section 3.2 shows the classification scores for the three models: Bayes Model, k-nearest Neighbor (KNN) Model, and Random Forest Model. Section 3.3 shows the linear regression of weight and frequency steps for the overall data collected.

3.1 Frequencies and Linear Acceleration

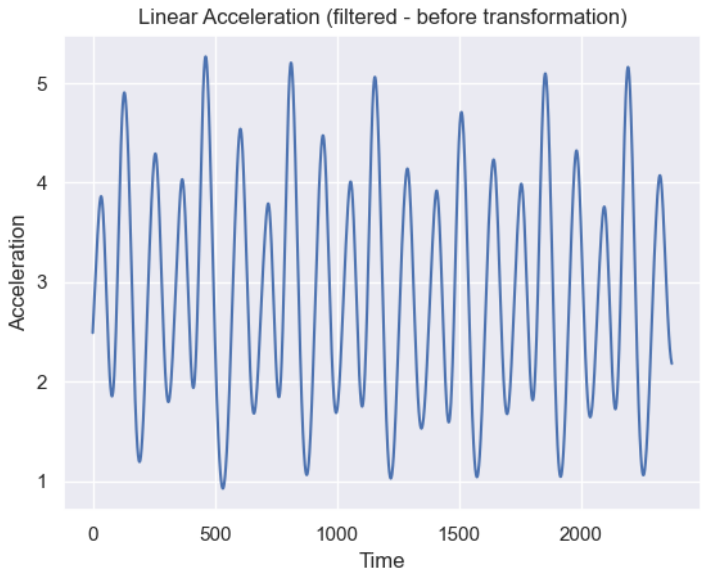

Figure 1: Time spectrum of the filtered linear acceleration before transformation

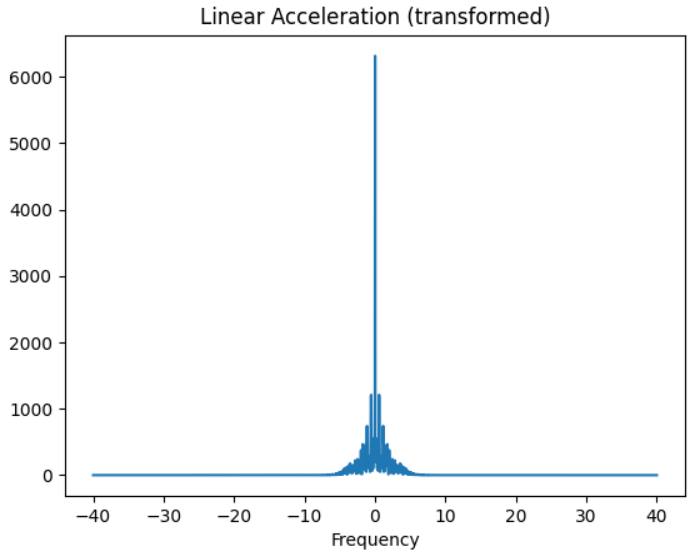

Figure 2: Frequency spectrum of the linear acceleration after transformation

3.2 Classification with Bayes, KNN, and Random Forest Model

Classifier	**Score**
Bayes Model	0.4327
KNN Model	0.2077
Random Forest Model	0.2827

Table 1: Summary of the classifiers scores for Favorite Activity vs Frequency of Steps

Classifier	**Score**
Bayes Model	0.6149
KNN Model	0.6500
Random Forest Model	0.6205

Table 2: Summary of the classifiers scores for Gender vs Frequency of Steps

Classifier	**Score**
Bayes Model	0.545
KNN Model	0.5366
Random Forest Model	0.5766

Table 3: Summary of the classifiers scores for Shoe Types vs Frequency of Steps

3.3 Linear Regression

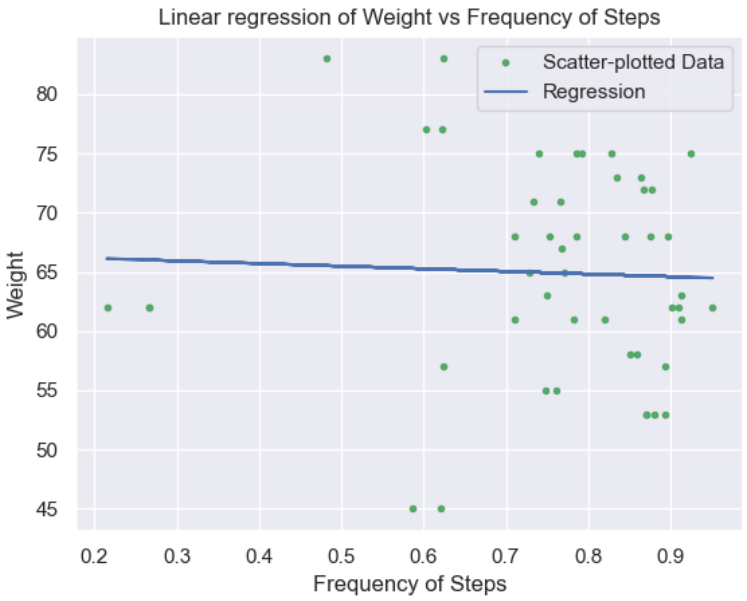

Figure 3: Linear Regression of Weight vs Frequency of Step

4 Analysis

4.1 Frequency Analysis

From Figure 1, we can see the time vs acceleration(m/s^2) distribution for linear acceleration before conduction fourier transformation. We used the butterworth filter to filter out the noise. We also removed the first and last 3 seconds of every dataset as to compensate for the fact that the user might stop walking while starting/stopping the recording. Upon applying the fourier transformation, we were able to get the distribution shown in Figure 2. Fourier transformation helped us to analyze the difference in the frequency of steps between each person and discover trends between the metadata such as weight, gender, favorite activity and shoe type.

After transforming the data using Fourier transform as shown in Figure 2, we have a distribution that is close enough to that of a normal distribution. However, the highest average frequency is at 0 steps per second, as this might be due to the user the leg which holds the recording device stops in place until the opposite leg moves forward. Moreover, following the Central Limit Theorem which states that if we take sufficiently large random samples, then the distribution will be approximately normally distributed. We can assume that our data is normally distributed.

4.2 Classification with Favorite Activity, Gender and Shoe Type

Based on the three tables in Section 3.2, we can see that the classifier scores for the relationship between gender and frequency steps are the highest, with Bayes Model score of 0.6194, k-nearest Neighbor score of 0.6500, and random forest model score of 0.6205. The lowest classifier score is from the relationship between favorite activity of an individual and frequency steps, with Bayes Model score of 0.4327, k-nearest neighbor score of 0.2077, and random forest model score of 0.2827. This means that the random forest model performs the best for relationship between gender and frequency steps, although there is not much difference compared to Bayes model score and k-nearest neighbor score. Since we ran each of the model on differently split data for 150 times, and averaged the scores on test datasets, which result in highest score at around 68%, we predict that we can have a higher chance of predicting gender vs frequency steps if we have more data to compute, which will bring the predictions up to a higher percentage.

The relationship for shoe types and frequency of steps have the second highest classifier score, but since all of the average scores are around 40% to 55%, this means that the chance of predicting shoe types and steps frequency are close to the chance of coin flips. Hence, it could mean that we may not have enough data to train and produce a higher score to predict the relationship.

The classifier scores for the relationship between favorite activity and frequency steps are the lowest as we predict that there is no relationship between the constraint being analyzed. Hence, all the classifier scores shown from Table 1 produced are consistently low.

4.3 Linear Regression between Weight and Steps Frequency

Figure 3 shows the relationship between weight and frequency of steps using overall data collected on linear regression. The linear regression model was run 150 times similar to running the classification models in section 4.2, which then produced the linear regression graph. However, based on the graph of Figure 3, we can see that the overall data are not linear, and the linear regression model produces poor results. Hence, we calculated the average p-value, which we get a score of 0.776. The p-value in this case is greater than 0.05, but it still does not provide us enough information to reject the null hypothesis. As such, this does not mean that the slope of linear regression is non-zero. Therefore, the linear regression model does not produce a good result in predicting the average frequency of steps based on an individual's weight.

5 Conclusions

In conclusion, the classification scores for frequency steps based on favorite activity, gender, and shoe type are performed low, with all scores less than 0.7. However, we can say that the scores are accurate for gender if we are only comparing the three types of attributes (favorite activity, gender, and shoe type) as the frequency steps for gender gives us the highest classification scores of 0.6149 to 0.6205 for all three models used. In addition, it produces the most consistent classifier scores compared to scores based on favorite activity and shoe types. Moreover, the linear regression showed that frequency steps based on weight using overall data do not give us a linear result since the p-value was above 0.05. Therefore, the linear regression produced a low result for us to predict the average steps frequency based on a person's weight.

6 Limitations

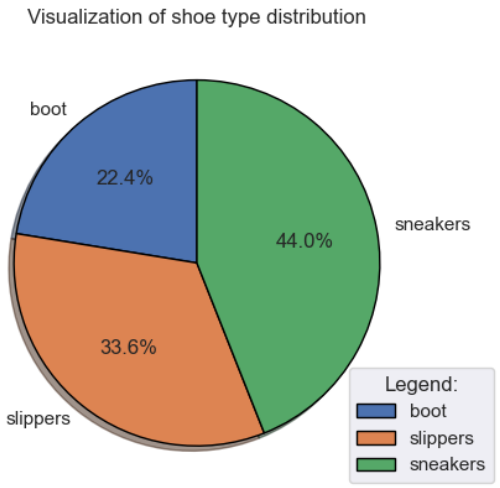

Figure 4: Distribution of shoe types from data collected

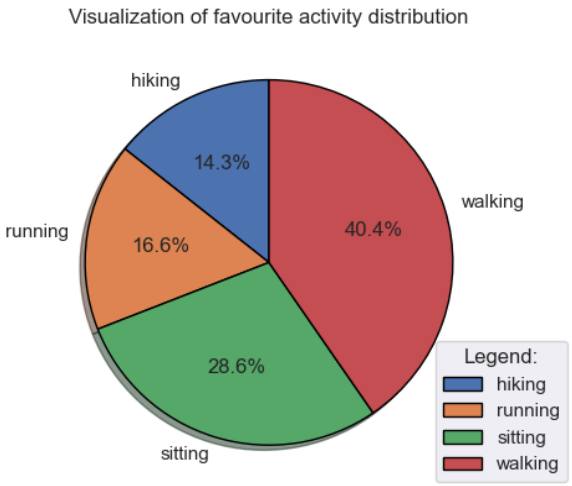

Figure 5: Distribution of favorite activity based on data collected

As we can see by the visualizations shown in Figure 4 and 5, the data collection for these specific attributes was not uniform. This means that the dataset that we have collected is imbalanced. Some classes were captured in majority such as sneakers in the shoe type distribution and walking in the favorite activity distribution. Whereas some were captured in minority such as the boot class and hiking/running. In retrospect, we would have collected more data to create a uniform sample for data consistency and accuracy. We also faced a challenge while collecting data using the roller coaster configuration of the app while placing the phone on the ankle, which resulted in the inconsistency of accuracy produced in our code to analyze data. If we had another device like a band or strapping device which we could securely place on our feet, we think that our data could be more accurate and precise.

7 References

[1] “Vieyra Software | Sensor & Generator Info", Vieyra Software, 2018. [Online]. Available: https://www.vieyrasoftware.net/sensors-sensor-modes. [Accessed: 23-Mar- 2022].

[2] “1. Supervised Learning.” Scikit, https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

[3] “Butterworth Filter Design and Low Pass Butterworth Filters.” Basic Electronics Tutorials, 24 July 2018, https://www.electronics-tutorials.ws/filter/filter_8.html.

[4] Hernandez, Julio, et al. “An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets.” Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2013, pp. 262–269., https://doi.org/10.1007/978-3-642-41822-8_33.

[5] “The Role of Probability.” Central Limit Theorem, https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html#:~:text=The%20central%20limit%20theorem%20states,will%20be%20approximately%20normally%20distributed.

8 Project Experience Summary

Wan Ying Chan

- Brainstormed with the group for topics to research on collected data.
- Researched and analyzed results from code into a report.
- Collected walking data to be used for analysis by experimenting with different smartphone sensors.

Saqib Hasib

- Implemented the classifiers and regression functions to show relationships between different types of data and frequency of steps
- Generated plot by creating multiple plot functions that are used in report to get a clear idea of the outcome
- Worked on developing the project structure and questions to answer using the data
- Collected data and their metadata for using in different classifications and regression on them

Gaurav Gupta

- Worked with the team to collect the data and model the questions for analysis
- Implemented the fourier transform to get the step frequency for the collected datasets
- Created the visualizations for the data distribution using pie graphs to comment on the sampling amongst different metadata.