Topic: Sensors Noise and Walking

Angus Fan
301317306
angusf@sfu.ca
Phong Le
301303290
lephongl@sfu.ca

Introduction

Throughout the semester, the "core" of data science has been explored through lectures, quizzes and weekly exercises. In this project, such techniques were integrated and data science ideas were explored . The necessary and common collection of steps of data analysis were followed as shown in figure 1. First the question and topic were chosen. Specifically this project explores the topic of sensors and noise associated with walking. Using our smartphones as a sensor (3D accelerometer), data was acquired from

7 participants. Next the data was cleaned and prepared for analysis. Finally this project will demonstrate the findings and present the results found.

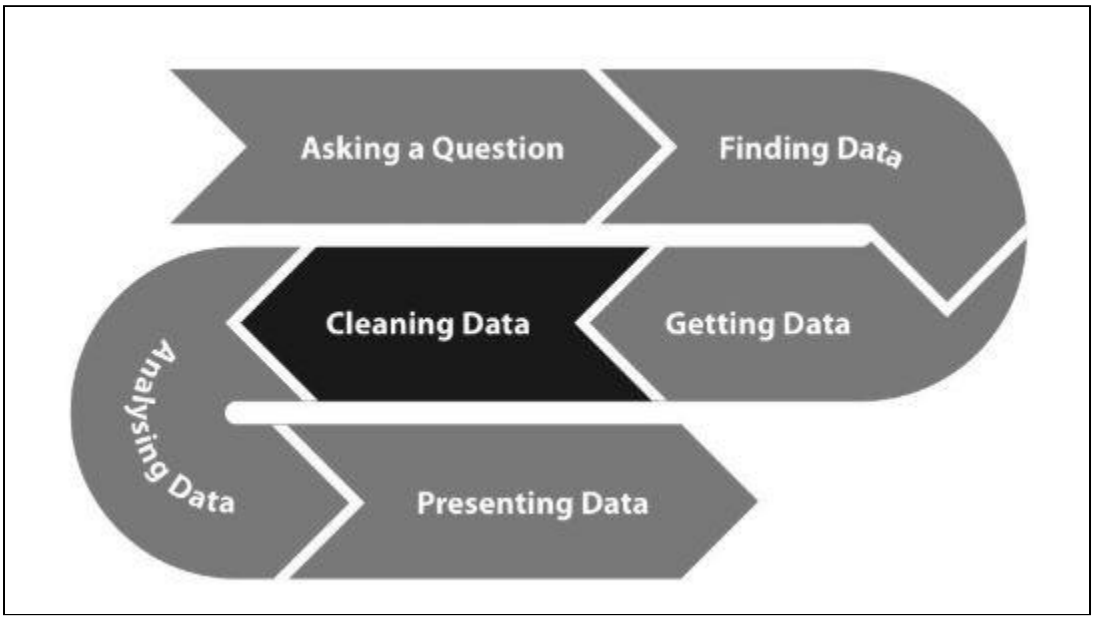

Figure 1, the process we followed for the project

Problem Addressed

As stated in the introduction, the main topic of this project was to explore sensors and noises associated with walking. Collecting data using an accelerometer app for iOS (figure 2), this project explored the following question:

> How does walking pace (steps/minute or similar) differ between people? Does it vary by age, gender, height, …?

More specifically, how can we determine if there are indeed different walking paces between people? Which feature or features can draw conclusive statistical results based on the data? And how can we validate those tests using machine learning?

Data

Aforementioned, the data for this project was collected by an app from the apple store called accelerometer (https://apps.apple.com/ca/app/accelerometer/id499629589). This app allows the user to measure acceleration (change in velocity) in all three axes while plotting real time charts while having hertz and gravity adjustable to the users liking. To collect the data, an iPhone was strapped onto a person's ankle on the inner

side of the right leg with tape. Before they started, we instructed the person to walk from a range of 30 seconds to 5 minutes to reduce the pressure of having to fulfill a certain time requirement, possibility affecting their gait. Then, the person would then walk outdoors with the app at a 20Hz refresh rate to gather sufficient data points in a CSV file. This process was repeated for seven different people of different ages, height and gender.

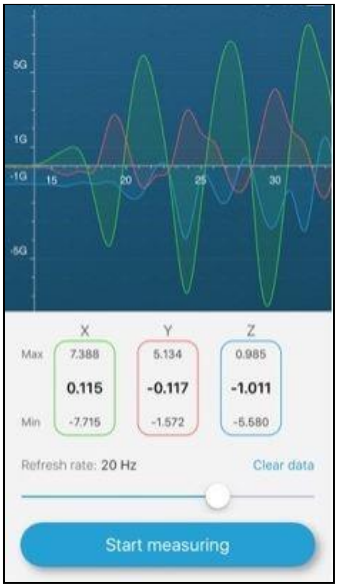

Figure 2, interface of our app

Cleaning The Data

Since the data collected from a cheap phone sensor will always contain a lot of noise, the data must be treated through a filter to reduce the amount of noise produced. Following the suggestion from the projects page, the data was put through Scipy's Butterworth filter which is a type of signal processing filter to flatten the frequency response as much as possible.

After the data is cleaned, a fourier transformation is applied to the cleaned data to obtain the frequency of steps. A fourier transformation decomposes a function into its composing frequencies. Similar to the Butterworth filter, this function belongs to the Scipy library.

The fourier transformed data is then processed through a stepFrequency function which utilizes scipy's fftfreq, which contains the N frequency bin centers in cycles per

second, where N is the number of acceleration data entries recorded. This value is absolute valued to obtain the peak with the largest magnitude, representing the approximate frequency of the sinusoid as an index. The return value is multiplied by the refresh rate, 60seconds/min, and by 2 to get the number of steps per minute accounting for both feet.

Before the data is able to show us any valuable information, the readings must be unbiased. To do this, the indices from the CSV file must start and end where the acceleration values for the three axes demonstrate walking patterns, so we appropriately shaved the edges of the data off.

Techniques for Analysing

Statistical Tests

After cleaning the data, values for step frequency, steps per minute and normal/levene tests' p-value for acceleration are obtained. A one way ANOVA (analysis of variance) test was used to determine if the means of any of the groups differ. This was done to check if any of the experiment results obtained previously are significant. The data collected from each person was independent because we approached each participant separately to do the experiment. As for the normality and variance assumptions, they failed under the normal and levene test. However, since the number of frequency data points gathered was large, we can apply the Central Limit Theorem here and reasonably approximate that the data follows a normal distribution, therefore satisfying all ANOVA test requirements. We also applied a Mann-Whitney U Test to see if any groups sorted higher than the other on average.

Machine Learning

We used Decision Trees, Naive Bayes, and KNN Classifiers to train and test on the data provided by "walkdata.csv". We were interested in confirming our statistical results of which feature was the most accurate in predicting slow, medium, or fast paced walkers.

Results

As stated in the techniques for analyzing section, after the Butterworth filter, fourier transformation and unbiasing the data, we obtain the step frequency, steps per

minute and normal/levene tests' p-value for acceleration. The raw data versus the filtered data for the three axes of acceleration can be found in figure 3. In figure 4, the overview results of walking data are shown.

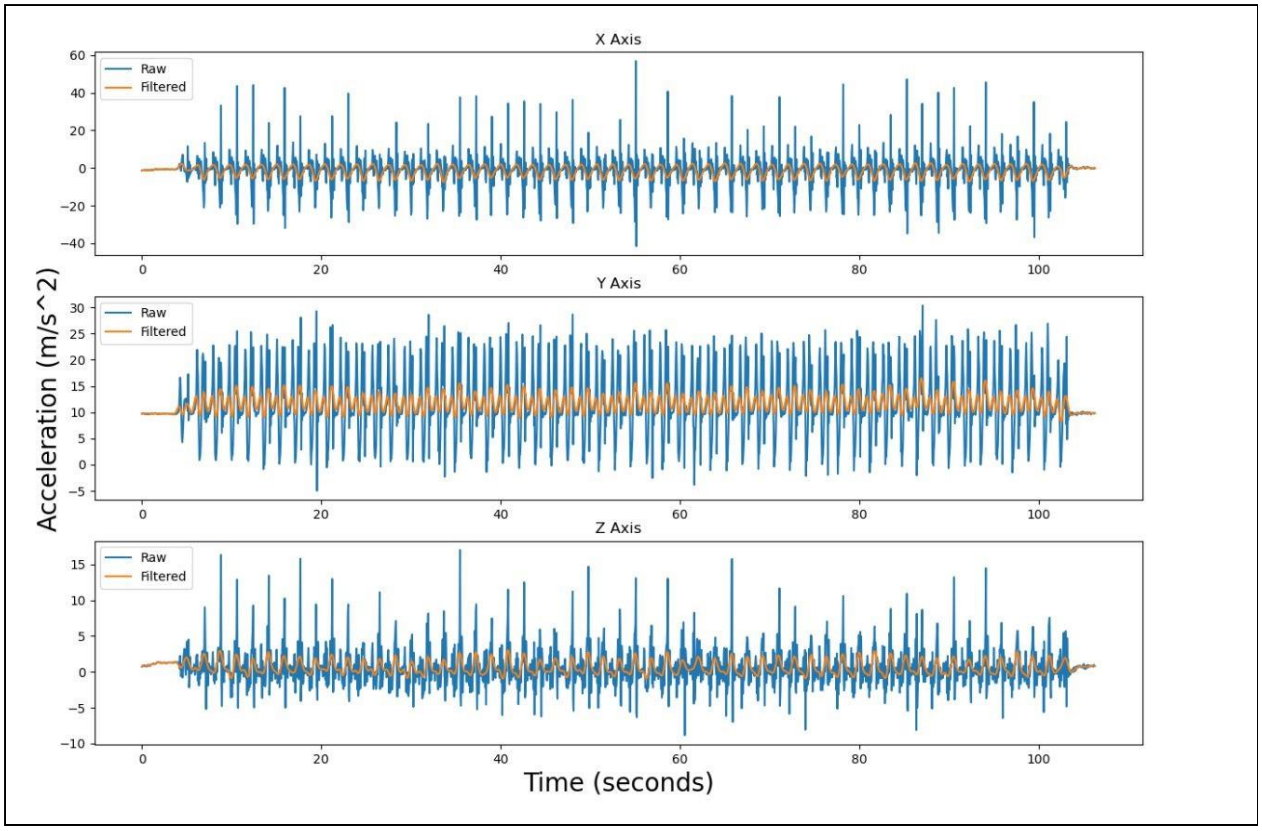

Figure 3, graph created from TJ-walk.py and saved as "TJ-acceleration.png"

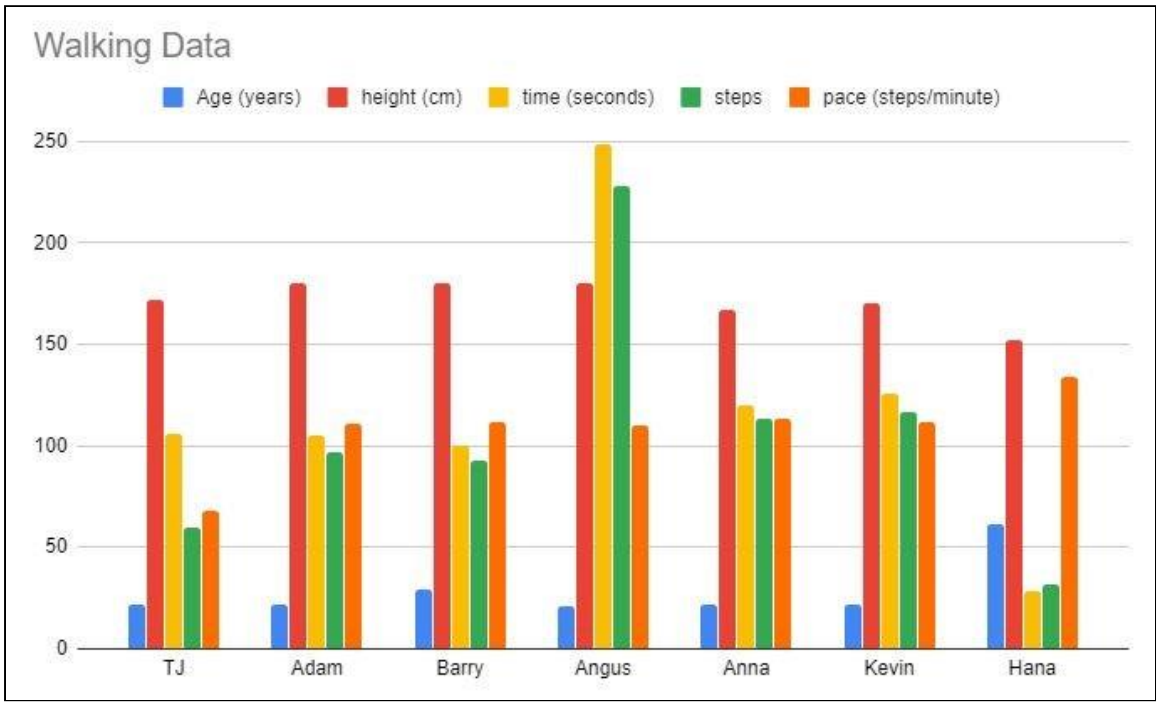

Figure 4, walking data

The ANOVA test was performed with each person as a group and their sum of acceleration frequencies. This resulted in 0.0 due to a numeric underflow which confirms that the walking pace between people are indeed different and motivates for more tests to continue. Since this report focuses on how the walking paces differ with respect to a person's features and not so much of whose walking pace is different, a Post Hoc analysis was not required.

The following table resulted in running all the subject's python files outlined in the README instructions (figure 5). This summary of the data was used in the Mann-Whitney and Machine Learning components.

name	age	gender	height	steps_min	time	steps	pace
TJ	22	male	172	67.7647059	106.25	120	67.7647059
Hana	61	female	152	133.565217	28.75	64	133.565217
Angus	21	male	180	109.98995	248.75	456	109.98995
Barry	29	male	180	111.655828	99.95	186	111.655828
Kevin	22	male	170	111.650099	125.75	234	111.650099
Anna	22	female	167	113.047103	119.95	226	113.047103
Adam	22	male	180	110.909957	104.95	194	110.909957

Figure 5, summary of subject's data in "walkdata.csv"

The next step was to perform the Mann-Whitney U-test to determine whether age, gender or height impact walking pace. The height cutoff between the two groups

“short” and “tall” was an average of the height column in figure 5. Observations are independent and walking pace can be sorted by magnitude. The p-values from the test showed that age and height are inconclusive towards affecting walking pace, at 0.105 and 0.0558 respectively. On the other hand the p-value for gender was 0.0407 which means gender affects walking pace. We can see in the graphical representation for this test in figure 6 that females on average walk faster than males.

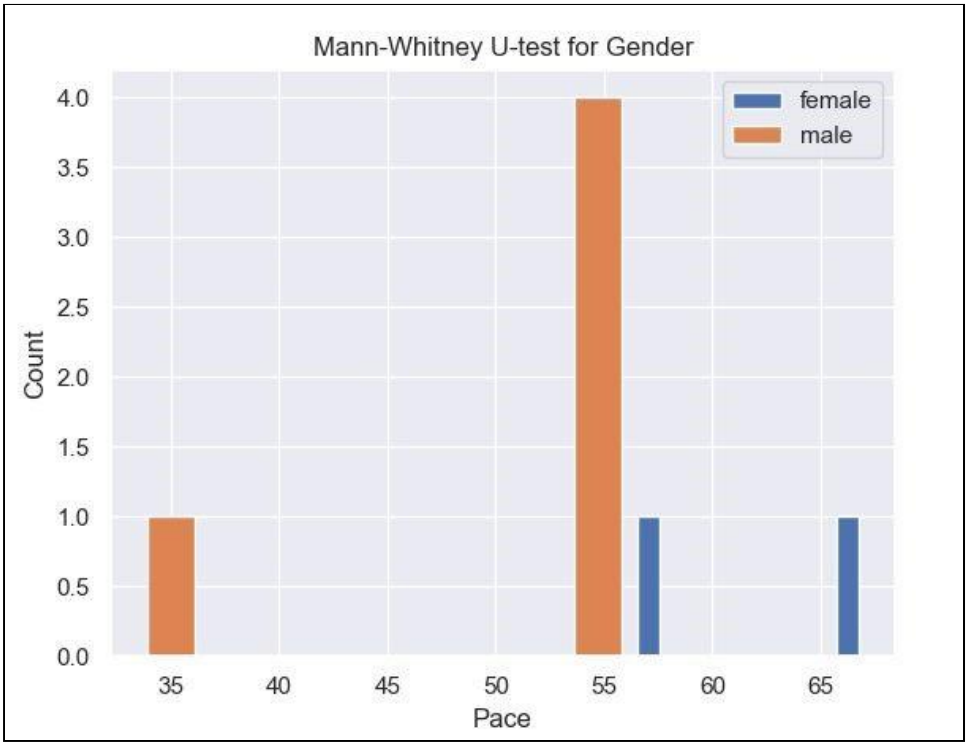

Figure 6, Mann-Whitney U-test for age

Finally, to further cement our findings, machine learning was done to validate the results from the Mann-Whitney U-test. We used the sample data in figure 5 to train the classification models to see if they could accurately predict the walking pace categories of slow, medium or fast. The function train_test_split was used to split the data into training and validating sets. Decision trees, K-nearest neighbors and Naïve Bayes were each trained on separate features age, gender, and height. The results of these tests aligned with the Mann-Whitney U-test as the accuracy score for gender for all three classifiers performed relatively well at 75%. We expected height to come as a close second as its p-value of 0.0558 on the Mann-Whitney test was very close to conclusive findings. The results can be found graphically in Figure 7.

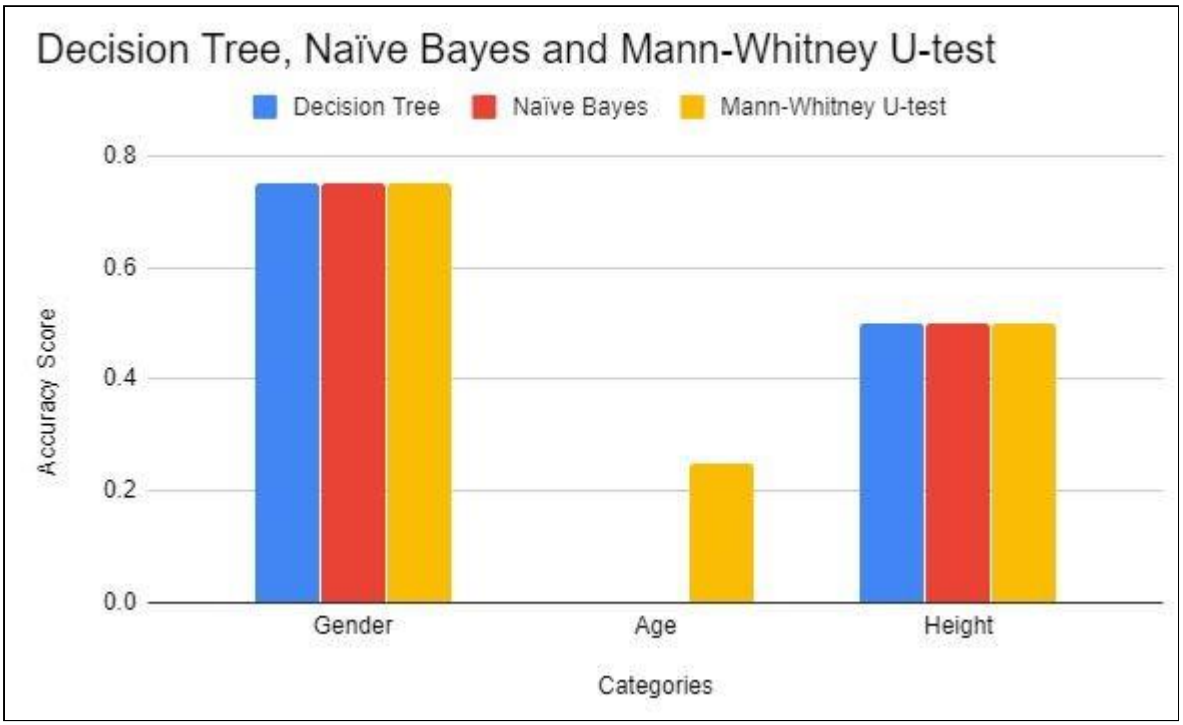

Figure 7, Machine learning tests using Decision tree, K-nearest neighbors and Naïve Bayes

Limitations

Throughout the project, there were many unexpected problems encountered. The app used to collect the data only allows up to around 10,000 data points before crashing. This resulted in many walks being redone which consumes a lot of time. Another problem with collecting data was the placement of the phone. If the phone was placed too high, the y value for acceleration would be inaccurate or if the phone wasn't securely taped, it would wobble and start giving off unnecessary noise. So when it came to unbiasing the data, this made it so no walking pattern could be found within any ranges of indices.

A larger scale limitation was gathering enough data since there is a small sample size with mostly the same age group, and therefore not representing the population. If there were more time and resources, more data would be collected from a wider range of ages. Also, the credibility of the data should be doubted as the collection is done on a cheap phone sensor. The researchers also participated in this experiment which may result in biased data.

A problem we had with the Naive Bayes was sometimes a *RuntimeWarning: divide by zero encountered in log* occurs on the command line. We suspect that it may be because of similar data points being assigned different walking pace labels, which is unavoidable due to a limited data set. Also in the Machine Learning section, the categories of slow, medium, and fast were arbitrarily chosen based on the limited data

Conclusion

The data analysis process was followed starting with choosing the topic and question which was to explore what affects walking pace the most (gender, age, height). Using a phone sensor to record accelerometer information, data was gathered to be analyzed using data analysis techniques. The data was first cleaned through a Butterworth filter. Then a fourier transformation was used to obtain the frequency of steps. After this step of the data analysis process, statistical tests were performed. Applying the central limit theorem to the ANOVA test resulted in a normal distribution. Following this step, a Mann-Whitney U-test determined that age and height are inconclusive towards affecting walking pace, but the p-value for gender was 0.0407 which means gender affects walking pace. To further our understanding of these results, machine learning classifications such as Decision trees, K-nearest neighbors and Naïve Bayes were trained on the features age, gender and height. The results from this test matched the Mann-Whitney U-test. This firmly answers our original question. From this experiment, it can be concluded that gender does affect walking pace; females, on average, tend to walk faster than males.

Project Experience Summary

Angus Fan

- Gathered multiple test subjects to record walking data by attaching the phone to the leg for raw data which could then be filtered and analyzed for meaningful results
- Cleaned data and transformed it into meaningful information using Butterworth filter and fourier transformation for data analysis
- Reviewed and debugged code by testing resulting in efficient and effective code

- Created graphs using excel, google sheets and seaborn to present the data and results
- Plotted acceleration vs time and Mann Whitney U test graph of gender, creating a visual consisting of the main results for the report

Phong Le

- Applied the Mann-Whitney U statistical test to determine if feature groups could be meaningfully sorted, resulting in a conclusive p-value for gender
- Collected walking data for TJ and Hana by taping the phone to the leg, resulting in raw accelerometer data ready to be cleaned
- Wrote code to clean, extract, and input step frequency data into a CSV file, preparing to do analysis on it
- Trained and tested Decision Tree, Naive Bayes and KNN classifier models on separate features to see if their accuracies supported the statistical conclusions, resulting in gender models performing the best.

References

[unutbu]. (September 12, 2010). *how to extract frequency associated with fft values in python* [Online forum post]. Stackoverflow. https://stackoverflow.com/questions/3694918/how-to-extract-frequency-associated-with-fft-values-in-python

Baker, G. (May 12, 2020). *Statistical Tests* [Lecture Notes] https://ggbaker.ca/data-science/content/stats-tests.html