CMPT 353 Computational Data Science

Summer 2023

Project: Walking Sensors

Mengdi Jin 301173198

Ruixuan Liu 301383988

Yuhao Xiao 301404905

Introduction

In this project, we will investigate whether a person's physical attributes such as gender, age, height, and BMI have any impacts on their step frequencies. Data were collected from cellphone's internal sensors while testers are walking on flat ground, upstairs, or downstairs. Lowpass Butterworth filter was applied to clean the raw data and Fast Fourier Transform was applied to transform signals into frequencies. At last, various regression and classification models were built to predict people's physical attributes from recording of their step frequencies.

Data Collection

Data used in this project were collected from a mobile application called Physics Toolbox Sensor Suite by Vieyra Software. This app uses internal smartphone sensors to collect, display, record, and export .csv data files.[1] To increase the data accuracy, a cellphone with the application installed was tied to the tester's leg between the knee and ankle, and all testers walk for approximately 30 seconds on flat ground or stairs. The mode we chose from the application is linear acceleration as it records the tester's linear acceleration data in x-, y-, and z– directions, and auto-calculates the Euclidean norm, i.e., $atotal = \sqrt{ax^2 + ay^2 + az^2}$.

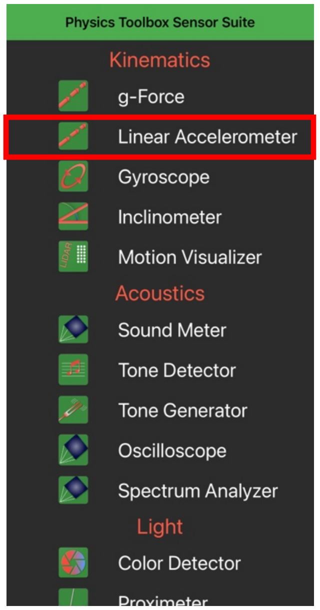


Figure 1 Homepage of the app

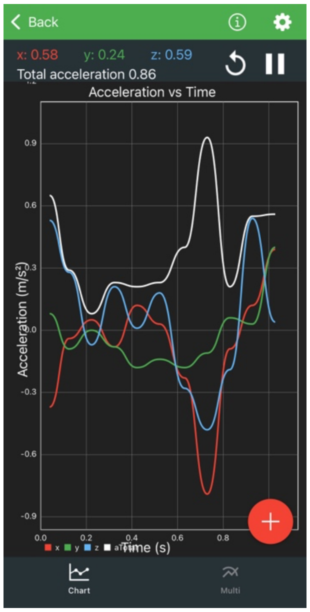
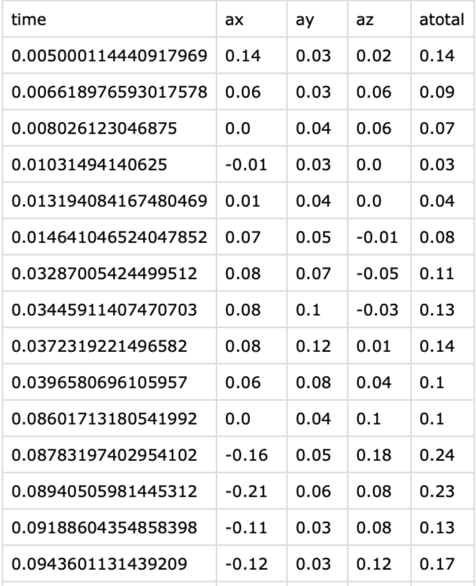


Figure 2 Recording page



time	ax	ay	az	atotal
0.005000114440917969	0.14	0.03	0.02	0.14
0.006618976593017578	0.06	0.03	0.06	0.09
0.008026123046875	0.0	0.04	0.06	0.07
0.01031494140625	-0.01	0.03	0.0	0.03
0.013194084167480469	0.01	0.04	0.0	0.04
0.014641046524047852	0.07	0.05	-0.01	0.08
0.03287005424499512	0.08	0.07	-0.05	0.11
0.03445911407470703	0.08	0.1	-0.03	0.13
0.0372319221496582	0.08	0.12	0.01	0.14
0.0396580696105957	0.06	0.08	0.04	0.1
0.08601713180541992	0.0	0.04	0.1	0.1
0.08783197402954102	-0.16	0.05	0.18	0.24
0.08940505981445312	-0.21	0.06	0.08	0.23
0.09188604354858398	-0.11	0.03	0.08	0.13
0.0943601131439209	-0.12	0.03	0.12	0.17

Figure 3 Data collected by the app

During the process of data collection, 25 testers with different genders, ages, heights, and body shapes were invited to take the walking test, and 4 of them also took the test for walking upstairs and downstairs. We also collected the information of their physical attributes and calculated their BMI by the formula $BMI = \frac{weight\ (kg)}{height(m)^2}$.

[1] https://play.google.com/store/apps/details?id=com.chrystianvieyra.physicstoolboxsuite&hl=en

Data Cleaning

Even though the smartphones we used to collect the walking data have the high accuracy sensers, the data recorded by the Physics Toolbox Sensor Suite still contain large amount of noises. In order to mitigate the noises, we applied the lowpass Butterworth filter by `scipy.signal.butter()` and `scipy.signal.filtfilt()` on the raw data to remove the high frequency noise. A comparison between the raw data and the filtered data is showing below.

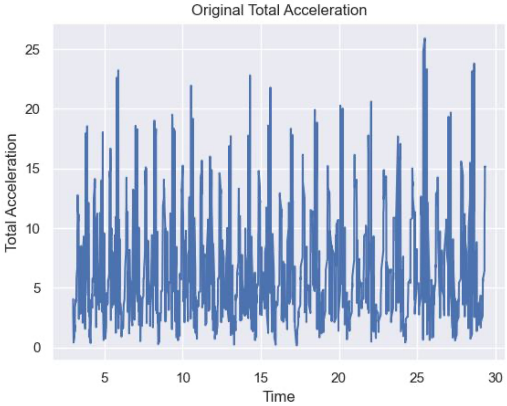


Figure 4 Raw data contains lots of noises

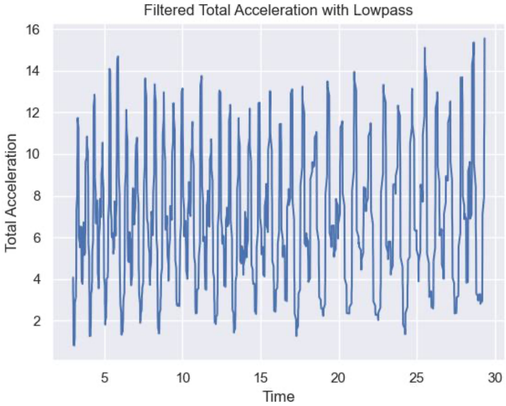


Figure 5 Filtered data with less noises

In addition, since there are few seconds between the testers pressed the record button and started walking, as well as they stopped walking and pressed the stop button, we will trim the data by discarding the data from the first 3 seconds and the last 3 seconds.

Data Processing

The raw data is in the form of total acceleration over time, and filtering process only mitigates the noises but not changes its data type. Therefore, we need to transform the acceleration signal into step frequency by applying Fast Fourier Transform with `scipy.fft.fft()`. It is a mathematical operation that changes the domain (x-axis) of a signal from time to frequency.[2] After doing that, we shifted the zero-frequency component to the center of the spectrum by calling the function `scipy.fft.fftshift()`. At last, in each dataset, we selected the largest value among all positive frequencies as the average frequency because the largest value represents the most commonly observed frequency within that dataset. That average frequency was stored in a .csv file along with all other physical attributes information.

[2] https://towardsdatascience.com/fast-fourier-transform-937926e591cb

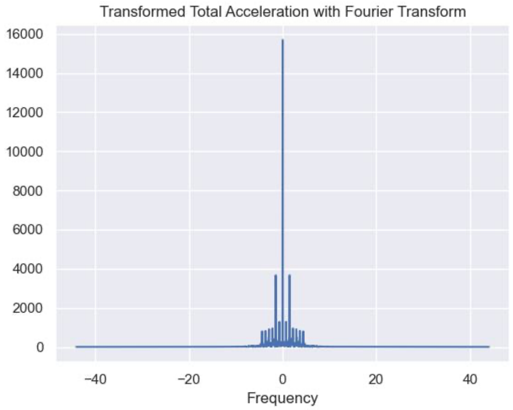
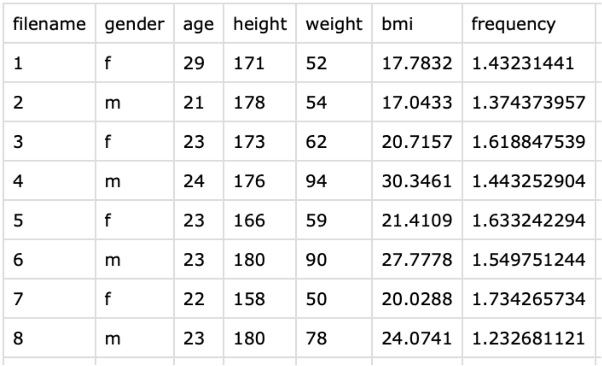


Figure 6 Transformed total acceleration



filename	gender	age	height	weight	bmi	frequency
1	f	29	171	52	17.7832	1.43231441
2	m	21	178	54	17.0433	1.374373957
3	f	23	173	62	20.7157	1.618847539
4	m	24	176	94	30.3461	1.443252904
5	f	23	166	59	21.4109	1.633242294
6	m	23	180	90	27.7778	1.549751244
7	f	22	158	50	20.0288	1.734265734
8	m	23	180	78	24.0741	1.232681121

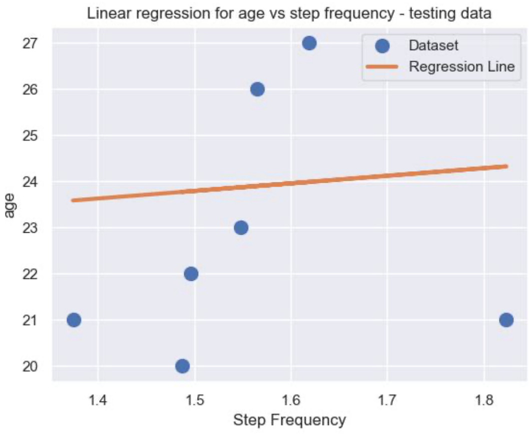
Figure 7 Summary of info and step frequency of testers

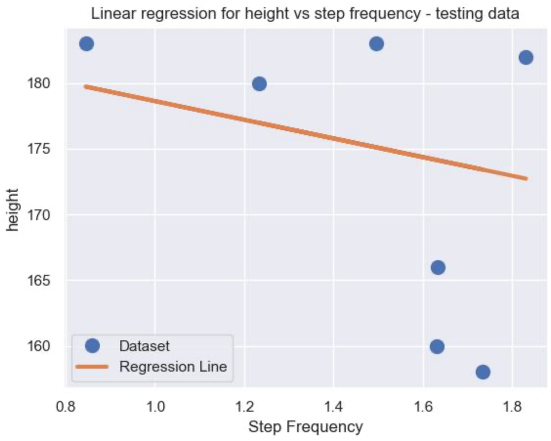
Regression

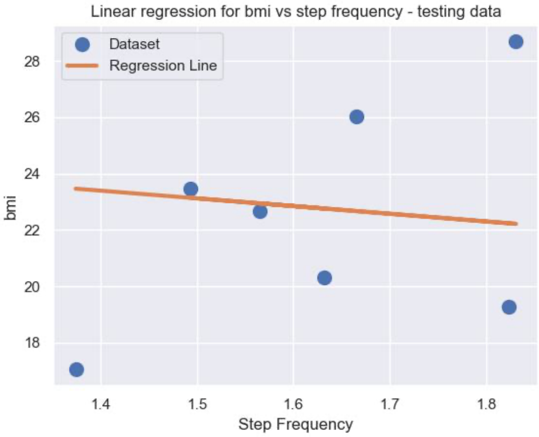
In order to find out the relationship between physical attributes (age, heights, and BMI) and step frequency, we applied regression models including `LinearRegression`, `KNeighborsRegressor`, `RandomForestRegressor`, and SVR from scikit-learn, as well as `stats.linregress()` from SciPy and `.poly1d()` from NumPy. For the four models in scikit-learn, data were split into training set to train the model and test set to test the accuracy of prediction on unseen data. The `stats.linregress()` and `.poly1d()` were only used to calculate the p-value and r^2.

Models/Scores	Age	Height	BMI
Linear regression	-0.3999	0.1117	-1.8107
KNN	-1.2735	-0.0567	-1.8004
Random Forest	-0.1586	-0.3775	-2.3199
SVR	-0.0172	-0.0564	-1.1433

From the above table, we can tell that all models fitted the data very badly as almost all scores are negative. The only positive score is linear regression for height, but it is still considered as poorly since the accuracy is just roughly 11%. Here we plotted the regression lines for linear regression model to visualize the poor score.







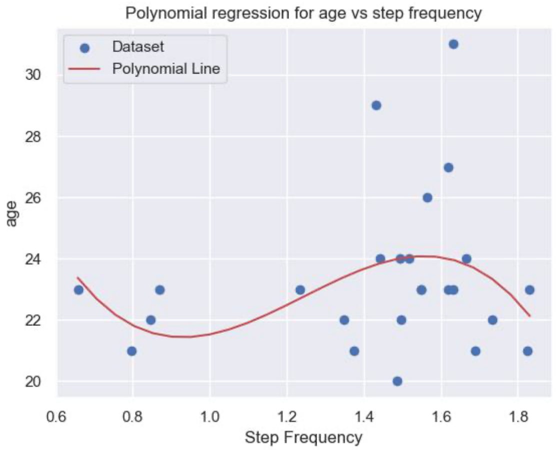
The poorly predictions can be explained by examining the p-value and r^2 from the linear regression and the 3-degree polynomial regression.

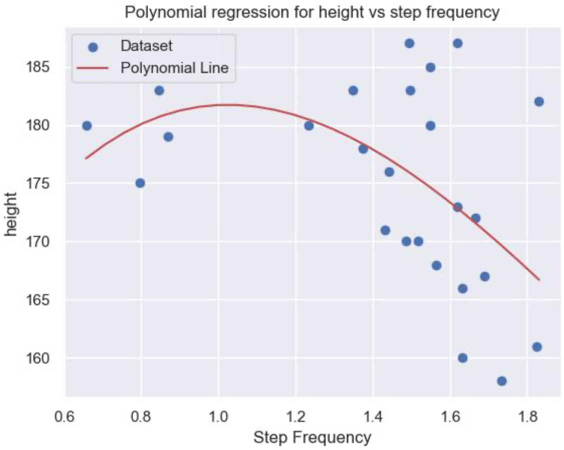
Linear regression	Age	Height	BMI
p-value	0.4135	0.0590	0.4292
r^2	0.0293	0.1464	0.0274

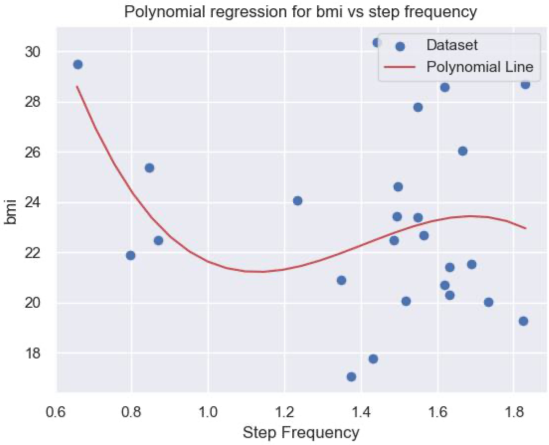
Polynomial regression	Age	Height	BMI
r^2	0.1142	0.2214	0.1218

All p-values of linear regression are greater than 0.05 which means that we will reject the null hypothesis and cannot conclude that age, height, or BMI affects step frequency. Since p-value of height is much smaller than age and BMI and close to 0.05, its r^2 is also much larger than other two, but still very close to 0. Since r^2 is the proportion of the variation in the dependent variable (age, height, and BMI) that is predictable from the independent variable (step frequency), closer to 0 means less correlation between 2 sets of data. Therefore, these measurements indicate that it is unlikely that there is a relationship between age, weight, BMI, and step frequency.

Furthermore, there are three graphs showing the whole dataset and a polynomial regression line.







Since data points are very scattered for all three graphs, it explains why it is hard to build a regression model with high accuracy.

Classification

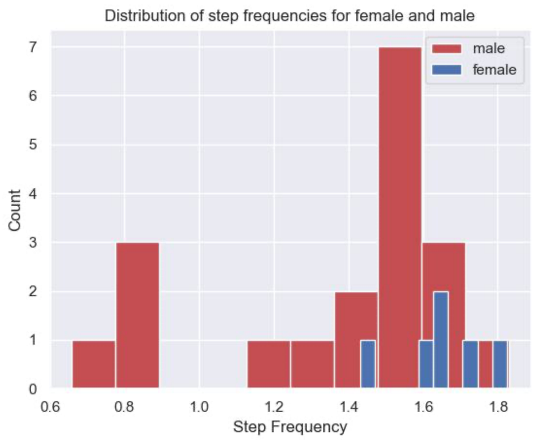
Since gender is a category data, we used classification to investigate its relationship to step frequency. We applied classification models including `GaussianNB`, `KNeighborsClassifier`, `RandomForestClassifier`, and SVC from scikit-learn. Same as regression, we also split data into training set to train the model and test set to test the accuracy of prediction on unseen data. To our surprise, all four models had much better scores compared with the regression models.

Models/Scores	Gender (training set)	Gender (testing set)
GaussianNB	0.8333	0.7143
KNN	0.8889	0.7143
Random Forest	0.9444	0.5714
SVR	0.7222	0.8571

Even running the tests multiple times to reduce the randomness, we still barely saw any scores below 0.5. Therefore, we can conclude that GaussianNB, KNN, and SVR can perform high accurate prediction for gender by given the step frequencies. Random Forest seems overfitting the training set as it generated the highest score for training set but lowest score for testing set, and these two scores differed a lot.

Since gender seems has the most impact on step frequency, we generated a summary table by taking the average of all step frequencies for two gender groups. From the following table, we can tell that female has a higher average step frequency than male, and it matches to the distribution of step frequencies for female and male. However, even the step frequencies seem different, the p-value from the ANOVA test is 0.064 and larger than 0.05. Therefore, we cannot reject the null hypothesis, so that cannot conclude that female and male has different average step frequency.

Gender	Step Frequency
Female	1.65
Male	1.37



General Data Analysis

Figure 8 shows the distribution of step frequencies for all testers walking on flat ground. As shown in the graph, the step frequencies varied widely from 0.6 steps per second to 1.9 steps per second, and the majority people have step frequency between 1.4 and 1.7.

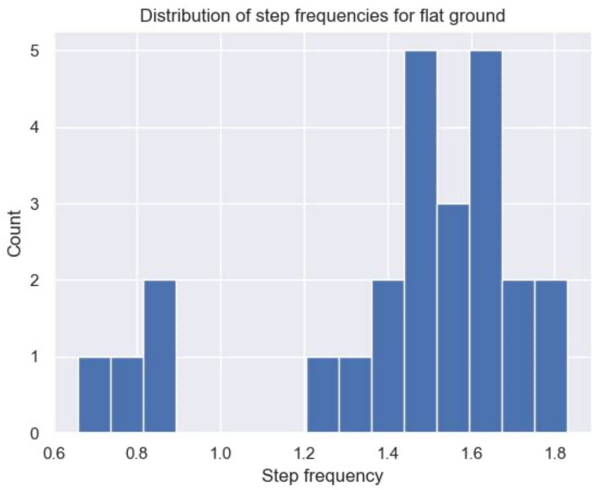


Figure 8 Distribution of step frequencies for flat ground.

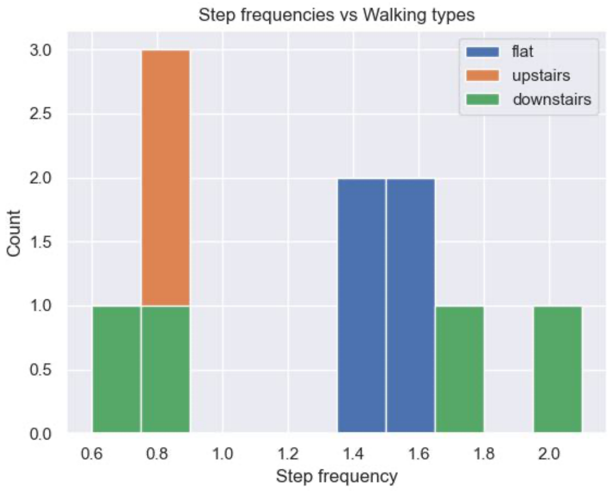


Figure 9 Step frequencies vs Walking types

Other than investigating the relationships between physical attributes and step frequency, we also noticed some interesting facts from analyzing the collected data. Aside from the flat ground, we also collected some dataset from walking upstairs and downstairs. From the figure 9, we can tell that walking upstairs has a very consistent step frequency around 0.8 for all testers and is much lower than walking on flat ground which is almost double the downstairs frequency. The step frequencies of walking downstairs depends heavily on testers. It could be either higher than flat ground or lower than upstairs.

Conclusion

In conclusion, among all physical attributes, only classifications on gender performed well with testing scores above 0.7. Almost all regression models performed extremely bad on predicting age, height, and BMI as they could not even get positive scores. The high p-value and low r^2 value of the step frequency data also proved that the scattered data is the reason why regression models cannot perform well. We also found out some interesting facts through the project such as females walk faster than males, walking upstairs is always slower than walking on flat ground, and walking downstairs really depends on the walking habit of the testers.

Limitations

The most significant limitation for our project is the lack of variety of data, especially the attribute of age. Although all group members tried so hard to get as much data as we can, but since we are all international students living on campus, the data we collected were almost all from university students between 20 and 24. Therefore, our prediction on age can only cover a very narrow age range which affects the performance of the regression models. Another limitation we faced is the unbalanced data from male and female testers. Among all 25 testers, only 6 are females and only accounts for 24%, so that male testers are three times larger than female testers. Hence, the gathered data from the male testers are dominated during the tests and may cause bias in the conclusions of our project. The last limitation is that compared with 25 sample data from walking on flat ground, we only have 4 testers who walked upstairs and downstairs. Therefore, we can barely use classification models or try Mann–Whitney U test on different walking type groups.

Reference

https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.filtfilt.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.fft.fft.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.fft.fftshift.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

https://www.w3schools.com/python/python_ml_polynomial_regression.asp

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.pie.html

https://towardsdatascience.com/fast-fourier-transform-937926e591cb

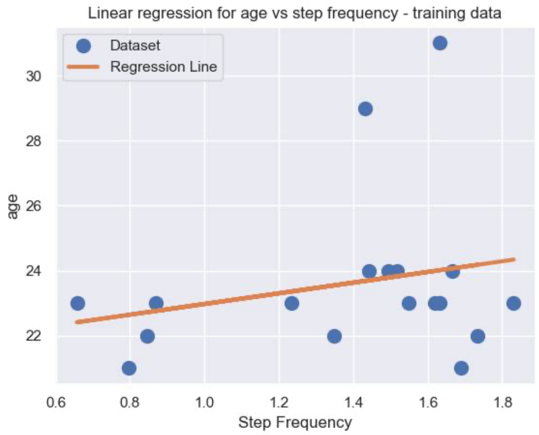
https://play.google.com/store/apps/details?id=com.chrystianvieyra.physicstoolboxsuite&hl=en

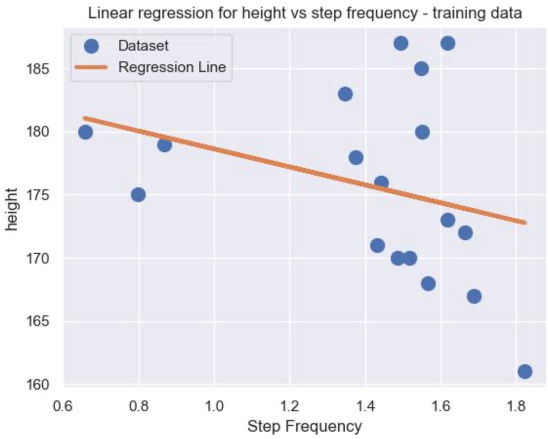
Appendix

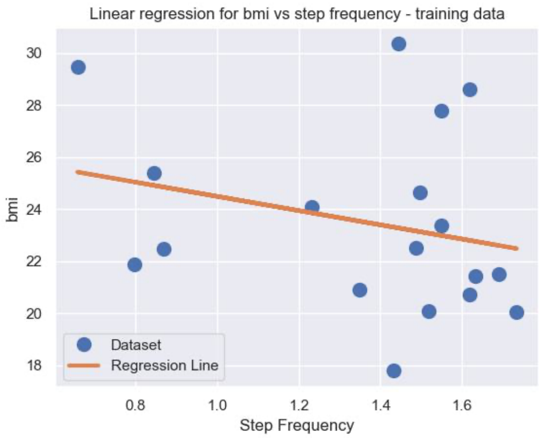
Summary of all testers’ physical attributes and their step frequencies for flat ground

no.	gender	age	Height (cm)	Weight (kg)	BMI	frequency
1	f	29	171	52	17.7832	1.43231441
2	m	21	178	54	17.0433	1.37437396
3	f	23	173	62	20.7157	1.61884754
4	m	24	176	94	30.3461	1.4432529
5	f	23	166	59	21.4109	1.63324229
6	m	23	180	90	27.7778	1.54975124
7	f	22	158	50	20.0288	1.73426573
8	m	23	180	78	24.0741	1.23268112
9	m	23	179	72	22.4712	0.86862106
10	m	24	187	82	23.4493	1.4937688
11	m	23	182	95	28.6801	1.83116883
12	m	24	172	77	26.0276	1.66541916
13	m	22	183	82.5	24.635	1.49608893
14	m	23	185	80	23.3747	1.54846727
15	m	20	170	65	22.4913	1.48725607
16	m	27	187	100	28.5968	1.6194905
17	m	23	180	95.5	29.4753	0.65738679
18	m	22	183	70	20.9024	1.34753521
19	f	21	161	50	19.2894	1.82375479
20	m	26	168	64	22.6757	1.56521739
21	m	21	167	60	21.5139	1.68932039
22	m	21	175	67	21.8776	0.79762401
23	m	22	183	85	25.3815	0.84480764
24	m	24	170	58	20.0692	1.51805226
25	f	31	160	52	20.3125	1.63176265

Linear regression graphs for training data of age, height, and BMI


Linear regression for age vs step frequency - training data


Linear regression for height vs step frequency - training data


Linear regression for bmi vs step frequency - training data

Accomplishment Statements

Mengdi Jin

- Brainstormed with group members to select the topic
- Discussed with group members to determine the methods to collect the data
- Collected walking data
- Implemented data cleaning and data processing methods
- Wrote data collection, cleaning, and processing parts of the report
- Edit the coding format of all files

Ruixuan Liu

- Brainstormed with group members to select the topic
- Discussed with group members to determine the methods to collect the data
- Collected walking data
- Implemented classification models and general analysis
- Wrote README.md
- Wrote classification and general analysis parts of the report
- Added appendix and reference to the report

Yuhao Xiao

- Brainstormed with group members to select the topic
- Discussed with group members to determine the methods to collect the data
- Collected walking data
- Implemented regression models
- Wrote regression, introduction, and conclusion part of the report