Assignment 3

Group 15: Xinyu Hu, Andrei Udrişte, Rui Pan

Vrije Universiteit Amsterdam

1 Introduction

A Modern life for people always means staying all the time with their computers, smartphones and in their chairs. Such a lifestyle makes it impossible to keep a healthy lifestyle, especially since every minute in our days is important and shouldn't be wasted. Hence, it is important for people, especially nowadays, to keep track of their daily exercises for managing a healthy balance between exercising and procrastinating. For reaching this goal, we plan to use sensor data, obtained using a piece of technology that everyone has with them, all the time, the smartphone. People can use their phones to self-tracked tasks in their daily lives, and used a formula to weight person's exercises and compare with Body mass index (BMI) to measure their health level and body shapes. The key idea behind is using a supervised learning approach to predict people's BMI with their own sensor data. Accelerometer and Gyroscope were used to quantified self for individual, and they could indirectly reflect whether a person regularly doing exercises, which guarantee a good body shape and a healthy lifestyle. Accordingly, in this project, we define the labels, which indicated personal height and weight, and collect data from the phone's sensors through *Phyphox*. The dataset contains personal activities which included running, walking, jumping, sitting, and laying over a period of time. Finally, we used the formula to weight each person's activity for evaluating their lifestyles and use the data to predict their BMI as a measurement for their healthy level [1].

2 Data

2.1 The data

The first step when dealing whit this kind of task is to determine what kind of task will the subject realize and what kind of measurements will be done. For this experiment we decided that we will ask our particiapant to perform 5 tasks: *Running*, *Walking*, *Jumping*, *Sitting* and *Laying*. We chose this tasks since they offer us different activity levels making it easier to classify if a person is active or not. After selecting the tasks the next step would be to select the sensor that will be used to measure each task. For a easier data processing we chose the same sensors for all activities, we chose the *Accelerometer* and the *Gyroscope*. We chose this sensor since our most important readings are correlated with the speed of the person and the *Accelerometer* offers a good measurement for this

task. The data was colected using the *Phyphox* app offered by RWTH Aachen University [3]. Furthermore, we also created two datasets called lazy and active, the lazy dataset contains activities like: sitting, laying and walking, while the active dataset contains activities like: running, jumping and walking. These two datasets would be used to predict each person's level of activity. Plus, there is another dataset for person's height and weight, which is used to predict BMI.

2.2 The formula

The formula we created to weight each activity is showing below:

$$score = (4 \times jump + 3 \times run + 2 \times walk + 1 \times sit + (-1) \times lay)$$

The weight is based on the difficult for each activity. We would like to give the most difficult activity more feedback. The score for each person could be used to measure how people engaged in these activities and also used to compare with BMI to see how the activities influenced the body shapes.

2.3 Body mass index (BMI)

The Body mass index (BMI) is a measurement of the body fat for each person. Its based on person's height and weight:

$$BMI = weight(kg)/([height(m)] \times 100)^2$$

It could be used to make a judgment about each person's healthy level and body shapes. If the output of a BMI were around 22, then we could consider this person is healthy.

2.4 The methods

One of the most important aspects of the project is the determination of the labels for each activity that will be used to determine the *activity score* based on the *formula* presented above. To determine the labels for each activity that our users may perform we decided to run some of the most well known classification algorithms: *Neural Networks*, *Random Forest*, *Support Vector Machine*, *K-Nearest Neighbours*, *Decision Tree* and *Naive Bayes*.

3 Result

3.1 Classification

Fig. 1: Boxpolts of Accelerometer

The Fig 1 summaries the feature—Accelerometer in our datasets, we can notice that all the data are evenly distributed and within the same range, which show the validity of datasets.

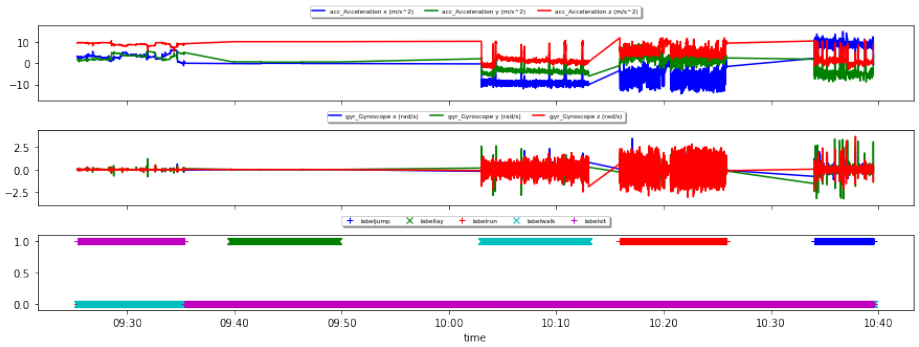

Fig. 2: All measured signals of one person

From the Fig 2, we can observe that all measured signals of one person and we use the following sensors: *Accelerometer, Gyoscope*. Since we have a break between each activities during the data collection phase, there exists some time intervals among the signals. First, we can look at what each sub-graph represents. The first sub-graph represents three accelerations of X, Y, and Z axis. The second sub-graph shows three rotations of X, Y, and Z axis. The 3rd sub-graph contains the label of all activities and the duration of time for each label. It can be observed that in the beginning there is little variation in the acceleration or in the gyroscope data, but as the experiments continues the variation in time start becoming more visible.

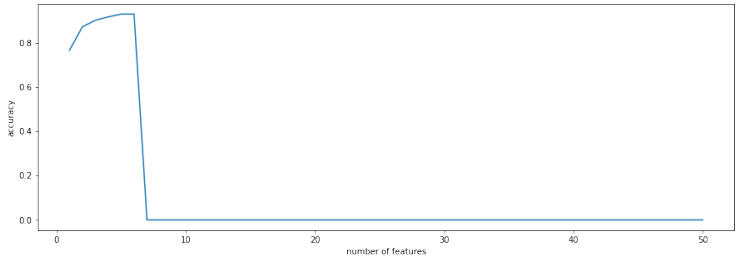

Fig. 3: Feature selection

The Fig 3 shows the result of feature selection, we find out that the three datasets have the same trend that the accuracy decreases with respect to the

number of features. The best feature number could approximate to 5, in this case we obtain an accuracy of approximately 0.7%. For this experiment we decided to use mainly the *Basic Features*, those consist of the three acceleration on X, Y and Z axis and the gyroscope data on X, Y and Z axis. The small number of features that have been used for the algorithm could represent the reason why the accuracy drops after approximately 6 features.

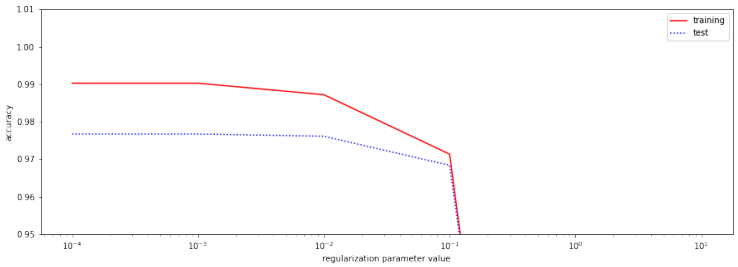

Fig. 4: The Influence of Regularization

The Fig 4 illustrates the influence of regularizer on the objective functions for both the training set and test set. As we can see, there is a trend that the accuracy goes down with the increasing of penalty on error functions. The best accuracy is obtained when we apply a regularization of 0.0001, with an accuracy of 0.99%. After the penalty is bigger than 0.1 we could see that the accuracy of the model drops under 0.95% and is no longer visible on the graph.

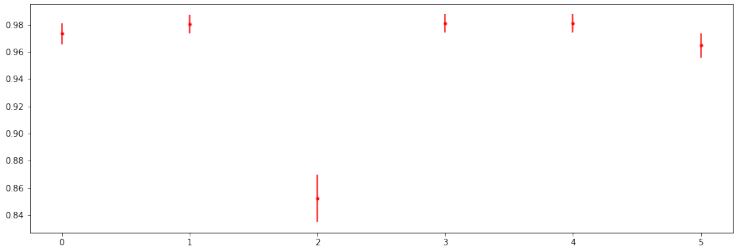

Fig. 5: Performance of algorithms

In Fig 5 we can observe the performance of each algorithm used to classify the labels. For the task at hand we used 5 classification algorithm *Neural Networks*, *Random Forest*, *Support Vector Machine*, *K-Nearest Neighbours*, *Decision Tree* and *Naive Bayes*, because we wanted to first test which of the algorithms will give us the best performance for the task at hand. The performance was measured using the Accuracy obtained on the test dataset. We can observe that almost all

the algorithms have the same performance over the dataset (97%), but there is an algorithm that had a lower accuracy compared to all other, *Support Vector Machine*, which had an accuracy of 85%. This means that we could use any of the other 4 algorithms to predict the labels for our dataset. If we are in the algorithm that offers the best performance, that would be the *K-Nearest Neighbours* with and accuracy of 98.2%.

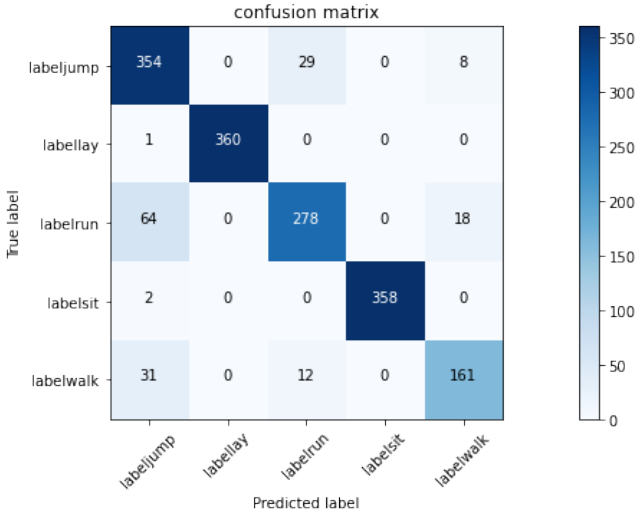

Fig. 6: Confusion matrix

Finally, Fig 6 gives us a confusion matrix. Here we could see the number of time the model was right about one of the prediction made or wrong. We can see that in the majority of the cases the algorithm makes correct predictions about all of the labels in the dataset. The main problems that could be observed are regarding the *Walking* and *Running* labels. The best predictions are realized for the label *Laying*, which has almost all of the labels predicted correctly, while the worst on is for the label *Running*, which has 82 labels predicted wrong. This would indicate that the model would require more training time and more running data, since this is one the the labels that are of interest for us.

3.2 Labels

After running the classification algorithms, we got the labels for each activity. The plot shows how many of them are in the merged dataset. For each activity, it is labeled by 10 minutes for each, and the total of them in time is 50 minutes. These labels shows how often each person worked in each exercise in 50 minutes in a day. One observation can be seen here is that which activity the person did

the most and least in the entire time, and these labels could be used to weight the scores. Using this method we ensure that even if a person does activities like running or jumping, but they only do the 2/3 minutes a day and the rest of the day they do activities like sitting or laying, they will still get a small score.

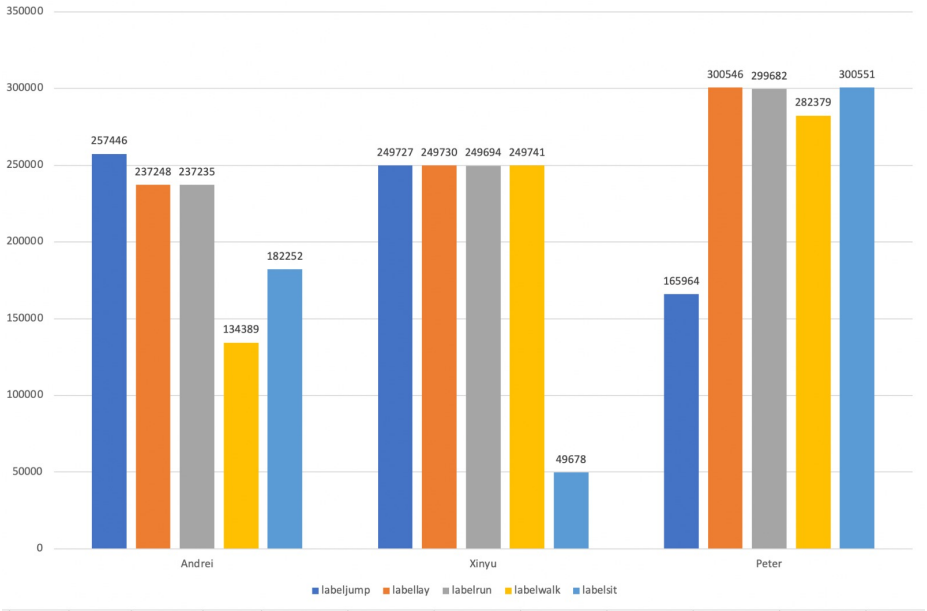

Fig. 7: Label count

3.3 Formula and BMI

The formula combined values from all activities and weighted them by the difficulty of each activity. The result here is based on the *Mean, Standard Deviation and Median* values of each activity. But different activities had different values to be considered. For example, walking and running would be use the data from y direction to compute mean, std and media, because these activities are people working from left to right in 3D case. However, the jump activity is based on z direction, because it is a top-down movement. In addition to that, the lay and sit are the movement on direction x. From the plot below we can see which person's score is higher or lower, which are useful for comparison between the scores and BMI. By comparison, the std usually has the closet values to the BMI. Moreover, the proper BMI range is between *18.5 and 24.9*. The three person are all in the proper range of BMI, which means all of them are with healthy body shapes.

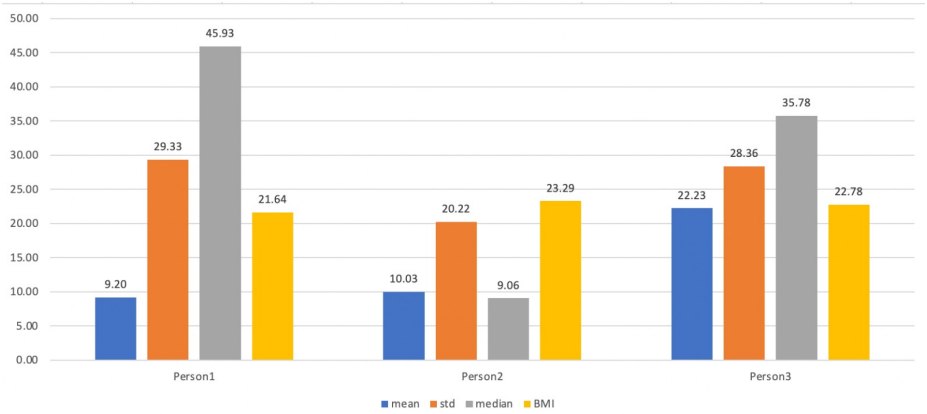

Fig. 8: Scores of the formula

3.4 Model

Because we need to obtain the relationship between BMI and running speed [2], the linear regression model is the best approach to find the correlation between these two features. The hypothesis is that people with high speed in running should have the BMI in the proper range which is between *18.5 and 24.9*. To do so, we need the speed first. Since the dataset contains acceleration and time, then the speed could be easy calculated by derivating the acceleration. Then, we used the average speed to fit the BMI. For the two plots, the speed is the same from the real person's dataset, but we generated 25 random BMI for testing the relationship for the regression. The results are showing below.

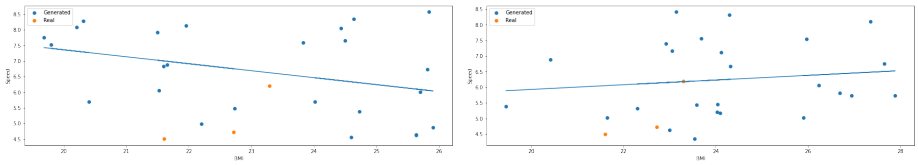

Fig. 9: 25 generation of BMI and the Linear Regression between them

This Fig. 9 left side shows that the correct relationship between speed and BMI. Since the higher BMI means carrying more weights and lower BMI means too little weights in the body, the speed goes down when the BMI goes high. However Fig 9 left plot shows that the regression line goes to oppose direction. It is because the 25 BMI was generated by random, the relationship between BMI and the real speed is based on the distribution of BMI.

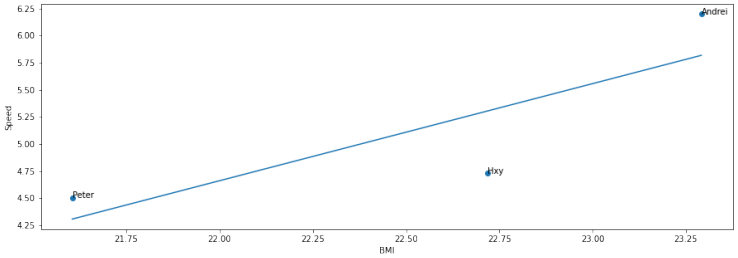

Fig. 10: 3 real person BMI and the Linear Regression between them

This Fig. 10 shows that the real person's BMI. It indicates that if the BMI were in a proper range, then the speed should increase as the BMI is increasing. Another observation about the plot would be that the regression line is in the opposite direction than what we are expecting. This could be because we have a very small dataset and the measurements may not be conclusive enough, but if we increase the number of participants we might obtain the desired result.

4 Conclusion

The main point of the project is to predict the healthy body shape of each participant by using sensor data. To do so, we did collect three person's data and used several classification algorithms to label them by the activity. The KNN made the best performance for this labeling process. In addition to measure the healthy body shape, we used two standards. The first one is an formula to weighted each activity and sum them together, and the second one is the BMI. The formula is based on ranking each activity by a certain weight, determined by its difficulty level. By comparing the result of the formula and BMI, we obtain that the formula has the best predictions. Additionally, one thing could be improved is increasing the number of participants, by doing this we should diminish the randomness induces by generating random points for the BMI. This project shows that the combination of activities and the BMI all predicted the healthy body shapes, and it also indicated that for getting a healthy body shapes, daily exercises are necessary. To do more on the jumping, running and walking and do less sitting and laying would guarantee a better and healthier lifestyle.

References

1. Suciati, Tri and Se, Ha Sakinah and others. : Body Mass Index as a Parameter of Running Speed. Bioscientia Medicina: Journal of Biomedicine and Translational Research **2**(5), 1–9 (2019)
2. Sedeaud, Adrien and Marc, Andy and Marck, Adrien and Dor, Frédéric and Schipman, Julien and Dorsey, Maya and Haida, Amal and Berthelot, Geoffroy and Toussaint, Jean-François: BMI, a performance parameter for speed improvement. PloS one **2**(5), (2014)
3. Staacks, Sebastian and Hütz, Simon and Heinke, Heidrun and Stampfer, Christoph: Advanced tools for smartphone-based experiments: phyphox. Physics education **4**(53), (2018)