PPG-DaLiA Dataset Analysis

1.

Dataset

Introducing the data

Dataset

Description

PPG-DaLiA is a publicly available dataset for PPG-based heart rate estimation. This dataset features physiological and motion data, recorded from both a wrist- and a chest-worn device, of 15 subjects while performing a wide range of activities under close to real-life conditions.

Objective

From this data we were asked to predict the activity the patient is undergoing.

Dataset format

- The data of each patient is recorded in a .pkl file
- This file contains multiple information :
    - attributes of the patient (height, weight, age etc…)
    - the activity executed
    - signals from different captors
    - indexes of the identified R-peaks
    - heart rate ground truth extracted from the ECG-signal

However this data is not expressed uniformly, these features are stored in different frequencies (4 Hz, 32 Hz, 64 Hz …) thus the need to process the data.

2.

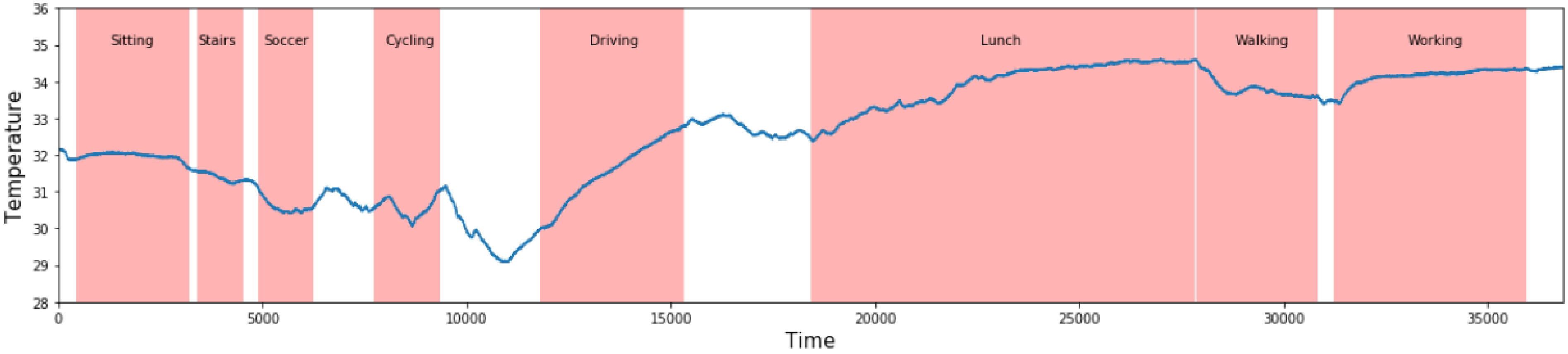
Study on one patient

Pre-processing, Visualization & Prediction

Pre-processing

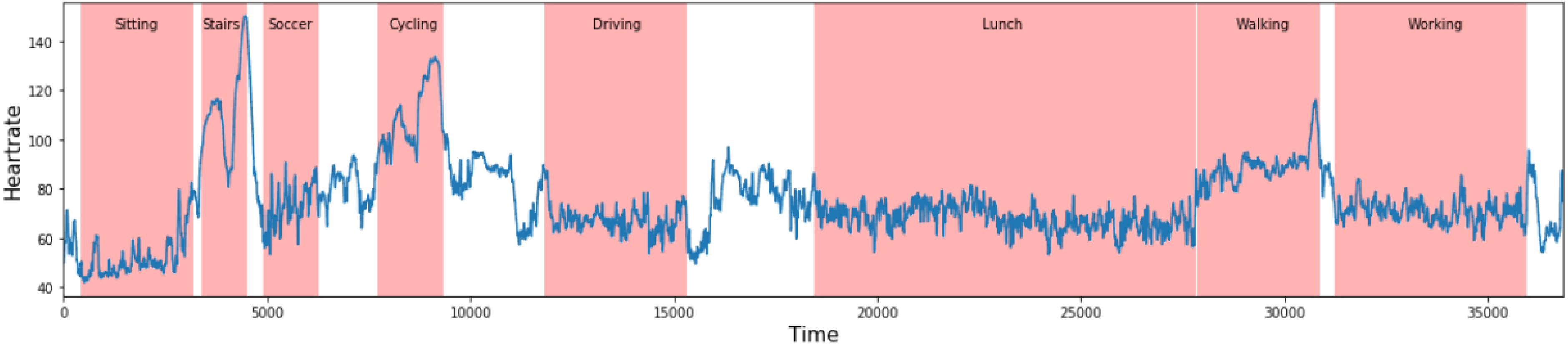
- As mentioned before the features are stored in different frequencies. The first step was to aggregate the data and put them in the same frequency. I chose to put it all to 4Hz as it is the frequency used to record the activity (our target variable).
- No feature engineering was needed except for the R-Peaks. Since the indexes were given, I chose to count the number of R-Peaks that occurred during our time period (4Hz so 0.25 seconds).

Visualization



Plot of the temperature during the different activities

Visualization



Plot of the heart-rate during the different activities

Prediction

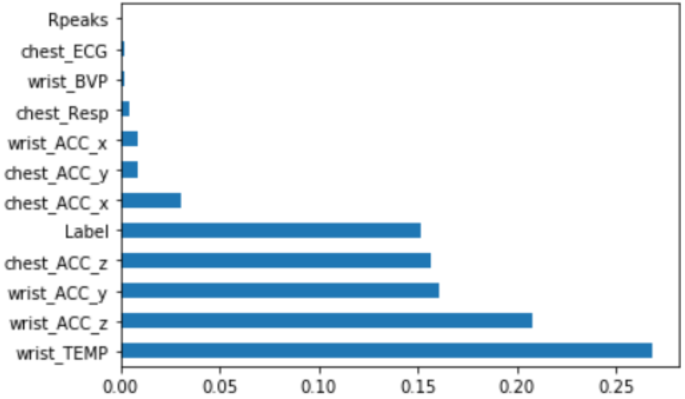
○ We then tried to build a model to predict the activities of the same patient. We didn’t use the attributes of our patient since we are working only on one subject.

○ We’ve essentially testes 2 models : Decision Tree & Random Forest. Since the results of those models were great, I was more interested in seeing how they would perform when faced with different data.

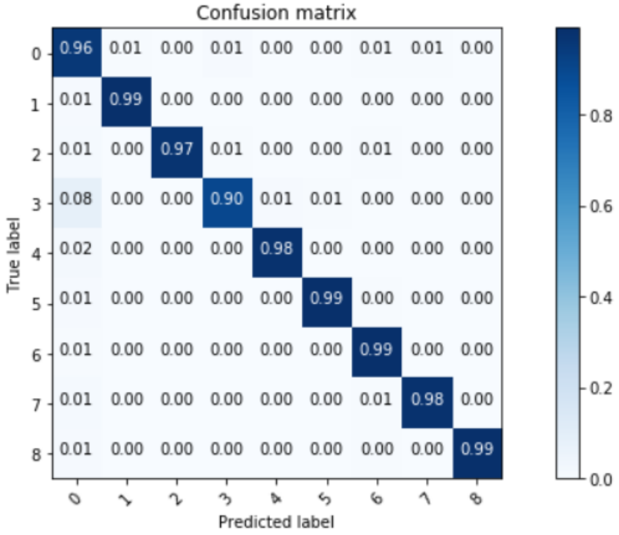
Prediction

Decision Tree

Score : 0.975



Feature importance

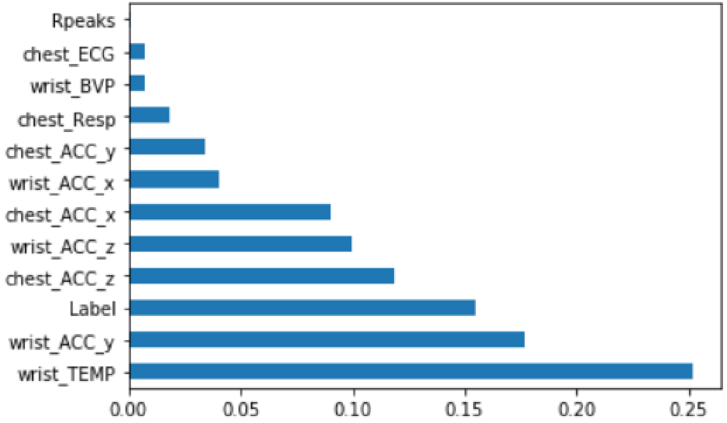


	0	1	2	3	4	5	6	7	8
0	0.96	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00
1	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.01	0.00	0.97	0.01	0.00	0.00	0.01	0.00	0.00
3	0.08	0.00	0.00	0.90	0.01	0.01	0.00	0.00	0.00
4	0.02	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
5	0.01	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00
6	0.01	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
7	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.98	0.00
8	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99

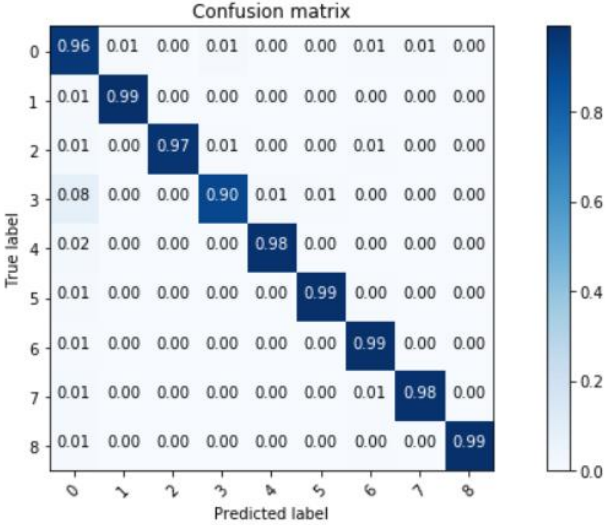
Prediction

Random Forest

Score : 0.978



Feature importance



Confusion matrix

True label \ Predicted label	0	1	2	3	4	5	6	7	8
0	0.96	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00
1	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.01	0.00	0.97	0.01	0.00	0.00	0.01	0.00	0.00
3	0.08	0.00	0.00	0.90	0.01	0.01	0.00	0.00	0.00
4	0.02	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
5	0.01	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00
6	0.01	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
7	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.98	0.00
8	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99

3.

Study on all patients

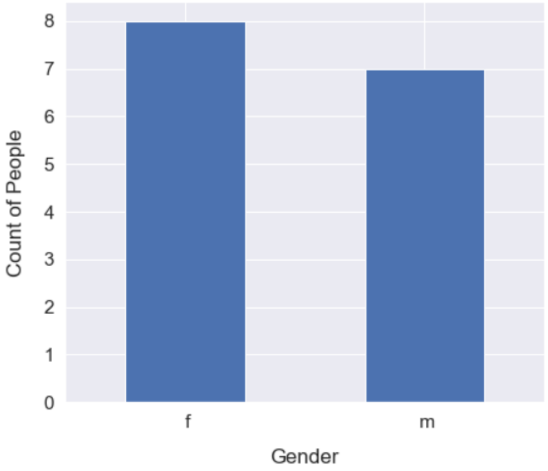
Pre-processing, Visualization & Prediction

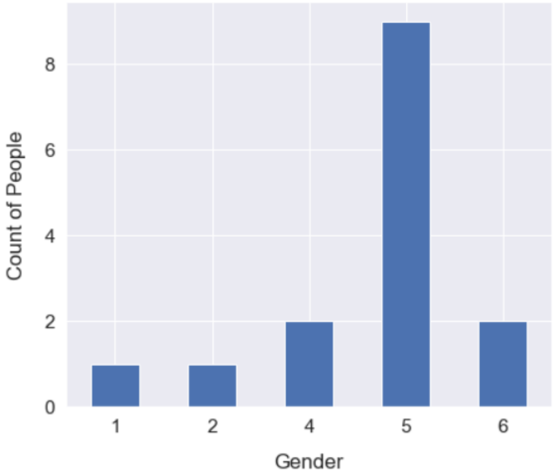
Pre-processing

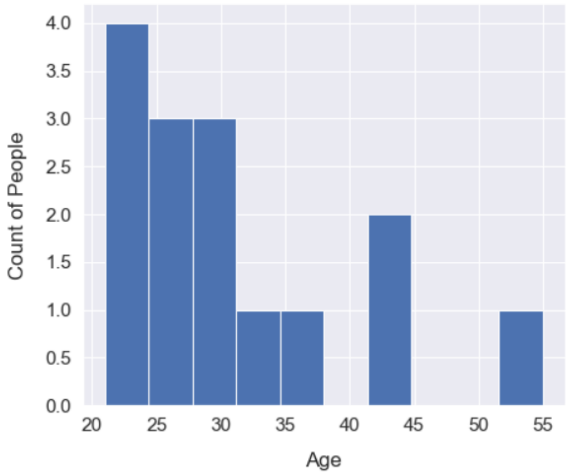
- The same steps of preprocessing will be done for each of the subjects. We will then aggregate the data of all subjects into one dataframe.

Visualization

Plotting the attributes of the subjects.



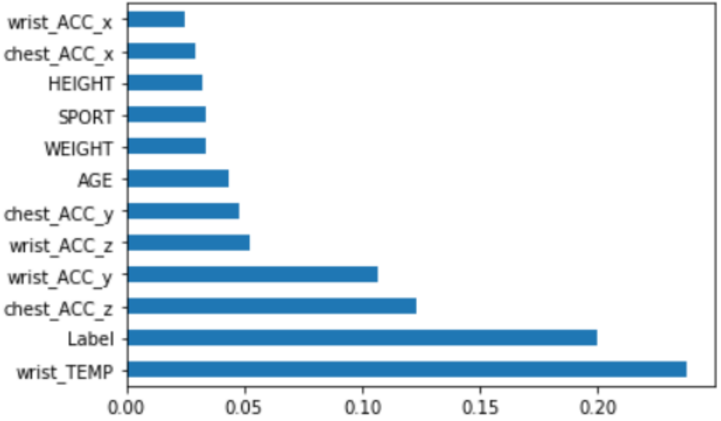




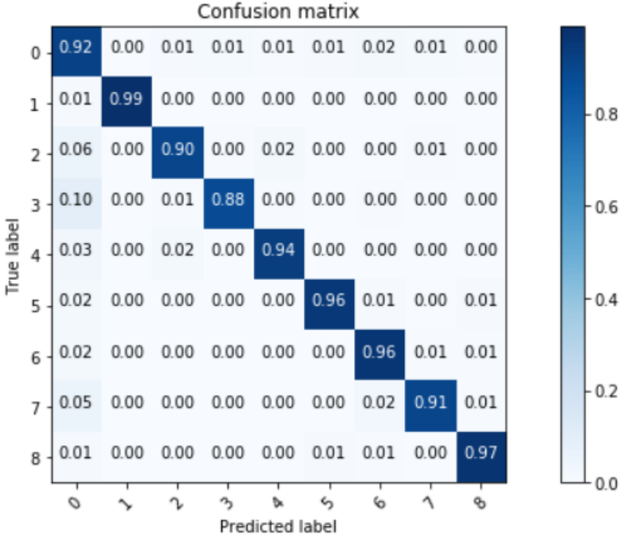
Prediction

Decision Tree

Score : 0.94



Feature importance

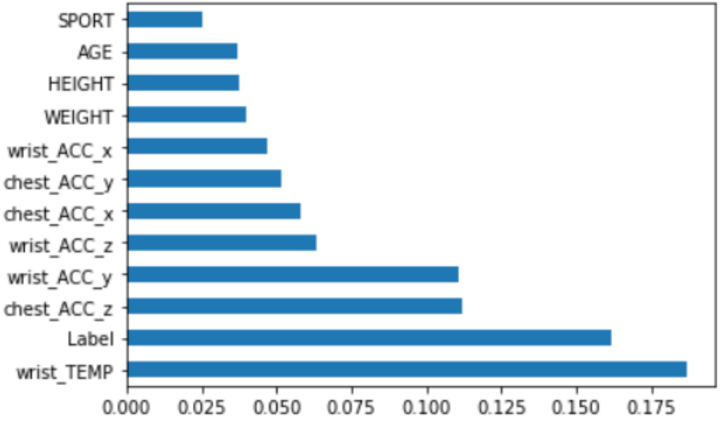


	0	1	2	3	4	5	6	7	8
0	0.92	0.00	0.01	0.01	0.01	0.01	0.02	0.01	0.00
1	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.06	0.00	0.90	0.00	0.02	0.00	0.00	0.01	0.00
3	0.10	0.00	0.01	0.88	0.00	0.00	0.00	0.00	0.00
4	0.03	0.00	0.02	0.00	0.94	0.00	0.00	0.00	0.00
5	0.02	0.00	0.00	0.00	0.00	0.96	0.01	0.00	0.01
6	0.02	0.00	0.00	0.00	0.00	0.00	0.96	0.01	0.01
7	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.91	0.01
8	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.97

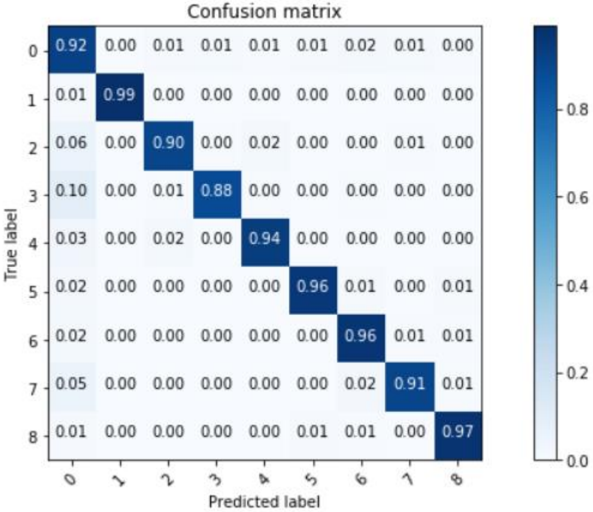
Prediction

Random Forest

Score : 0.962



Feature importance



Prediction

- Our model are still doing very well. My guess is that those results come from the fact that since the activities are done during a certain period, the “almost” same rows appear both in the train set and the test set. The model probably only predicts things it has already seen which means it is very likely our model is overfitting.

- To challenge that guess, I will train the model on only 13 subjects, leaving the 2 other subjects to score our model.

4.

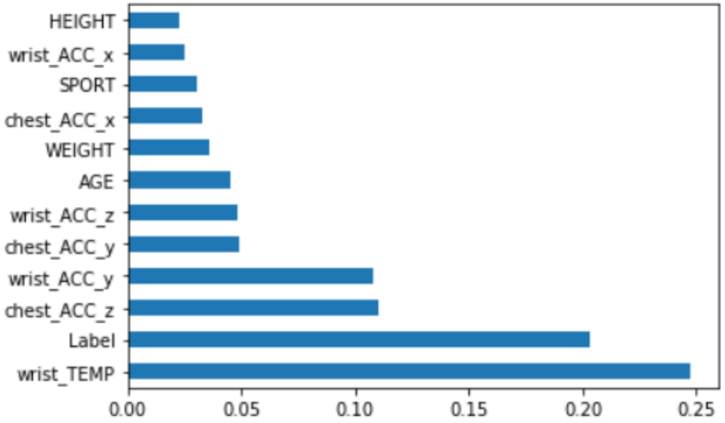
Study on 13/2 patients

Prediction

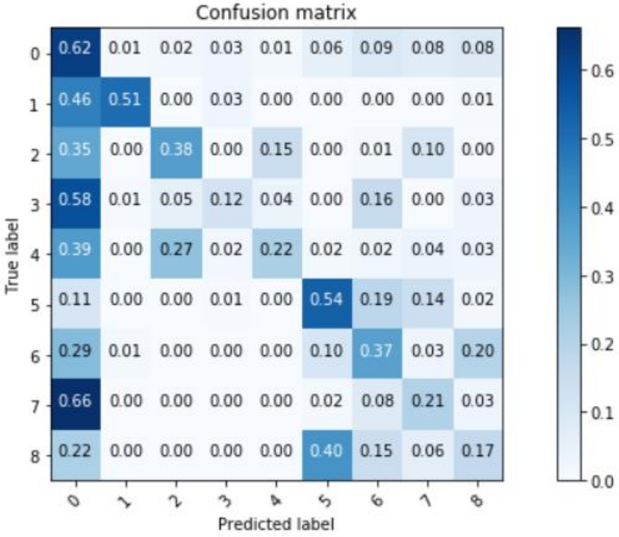
Prediction

Decision Tree

Score : 0.407



Feature importance

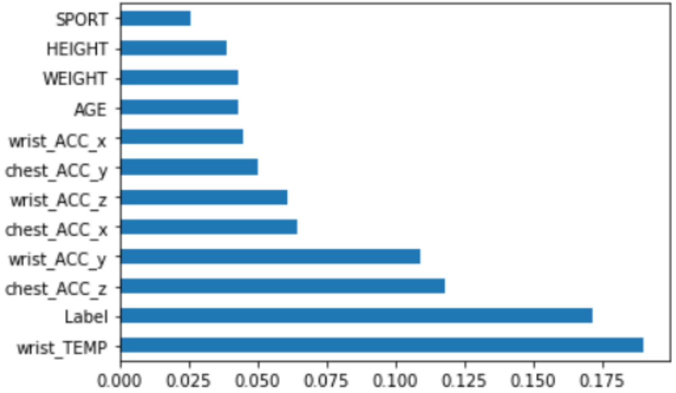


	0	1	2	3	4	5	6	7	8
0	0.62	0.01	0.02	0.03	0.01	0.06	0.09	0.08	0.08
1	0.46	0.51	0.00	0.03	0.00	0.00	0.00	0.00	0.01
2	0.35	0.00	0.38	0.00	0.15	0.00	0.01	0.10	0.00
3	0.58	0.01	0.05	0.12	0.04	0.00	0.16	0.00	0.03
4	0.39	0.00	0.27	0.02	0.22	0.02	0.02	0.04	0.03
5	0.11	0.00	0.00	0.01	0.00	0.54	0.19	0.14	0.02
6	0.29	0.01	0.00	0.00	0.00	0.10	0.37	0.03	0.20
7	0.66	0.00	0.00	0.00	0.00	0.02	0.08	0.21	0.03
8	0.22	0.00	0.00	0.00	0.00	0.40	0.15	0.06	0.17

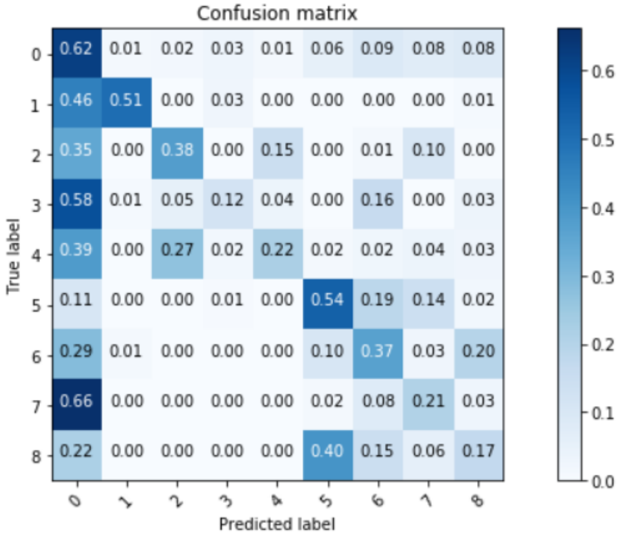
Prediction

Random Forest

Score : 0.552



Feature importance



	0	1	2	3	4	5	6	7	8
0	0.62	0.01	0.02	0.03	0.01	0.06	0.09	0.08	0.08
1	0.46	0.51	0.00	0.03	0.00	0.00	0.00	0.00	0.01
2	0.35	0.00	0.38	0.00	0.15	0.00	0.01	0.10	0.00
3	0.58	0.01	0.05	0.12	0.04	0.00	0.16	0.00	0.03
4	0.39	0.00	0.27	0.02	0.22	0.02	0.02	0.04	0.03
5	0.11	0.00	0.00	0.01	0.00	0.54	0.19	0.14	0.02
6	0.29	0.01	0.00	0.00	0.00	0.10	0.37	0.03	0.20
7	0.66	0.00	0.00	0.00	0.00	0.02	0.08	0.21	0.03
8	0.22	0.00	0.00	0.00	0.00	0.40	0.15	0.06	0.17

Prediction

○ We can clearly see now that the model trained does very poorly when we face it to new subjects. The previous models were overfitting because they were only trained on 15 sets for each activities.

○ To have better results we would need data from more subjects.

END

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash