MFTech Group Consultancy Project Documentation

Project Github Repository: <https://github.com/anster01/MFTech>

**Sources:**

Dataset Analysis:

|  |
| --- |
| Rossi, Alessio, et al. "Multilevel Monitoring of Activity and Sleep in Healthy People" (version 1.0.0). PhysioNet (2020), <https://doi.org/10.13026/cerq-fc86>. |
| <https://github.com/RossiAlessio/MMASH/blob/master/Code_MMASH.ipynb> |
| <https://oml.eular.org/sysModules/obxOML/docs/id_150/State-Trait-Anxiety-Inventory.pdf> |
| <https://github.com/RossiAlessio/MMASH> |

ML Techniques:

|  |
| --- |
| <https://www.sciencedirect.com/science/article/pii/S1532046420302380> |
| <https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html> |
| <https://dataespresso.com/en/2019/01/30/stress-detection-with-wearable-devices-and-machine-learning/> |
| <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> |

HRV Research:

|  |
| --- |
| <https://www.escardio.org/static-file/Escardio/Guidelines/Scientific-Statements/guidelines-Heart-Rate-Variability-FT-1996.pdf> |
| [https://www.firstbeat.com/en/science-and-physiology/heart-rate-variability/#:~:text=Heart%20rate%20variability%20increases%20during,have%20a%20generally%20inverse%20relationship](https://www.firstbeat.com/en/science-and-physiology/heart-rate-variability/%23:~:text=Heart%20rate%20variability%20increases%20during,have%20a%20generally%20inverse%20relationship). |
| <https://www.tandfonline.com/doi/full/10.1080/10615800802272251> |
| <https://imotions.com/blog/heart-rate-variability/> |
| <https://thesai.org/Downloads/Volume10No11/Paper_63-Time_and_Frequency_Analysis_of_Heart_Rate_Variability.pdf> |
| <https://www.health.harvard.edu/blog/heart-rate-variability-new-way-track-well-2017112212789> |
| <https://www.mdpi.com/1424-8220/21/12/3998/htm> |

Code Libraries:

|  |
| --- |
| <https://github.com/Aura-healthcare/hrv-analysis> |

Web Design:

|  |
| --- |
| <https://templatemo.com/> |
| <https://pyscript.net/> |
| <https://www.chartjs.org/> |
| <https://httpd.apache.org/> |

**Meeting notes with client:**

**Meeting 1:**

A close-up of a document

Description automatically generated with medium confidence

About the client:

* Miguel – major in computer & biomedical engineering
* Medical devices
* Wearable algorithms and systems à developing DSP and machine learning algorithms for medical devices - wellbeing wellness
* Wearable technologies to ensure people have good quality of life

About the project:

* Dataset (repository contained in physionet) containing data of 22 people well-monitored for 24 hours
* Two categories of data:
  + Physical activity data
  + Psychological data
* Develop a project to 1. Exploratory data analysis and characterise data à can we stratify different people into datasets? Same people within different time frames
* Explore correlation between physical activity and psychological parameters(labels)
* Need to decide what are the labels (anxiety status, emotions, etc)
* Beat-to-beat intervals (heart rates) à e.g. if it is strongly linked to stress
* If physical activity is related to stress
* Characteristics of all the data à produce Histograms, clusters to characterize
* Provide a prediction about people’s levels of stress

1. **Explore dataset (data analytics on this dataset) prepare clean sets of data, make sure no duplicates, clean the data, put the data in right form (prepare balanced set of data)**
2. **If appropriate to carry on à data science, select and propose models / algorithms (machine learning) (compromise between accuracy and overfitting), and test it.**
3. **Move model into cloud à something that can work in web app (maybe use AWS, gain weight API)**

**Meeting 2:**

* Do EDA and produce report  
  Heart Rate Variability (HRV)
* generate features, put in feature map/table
* "Feature selection"
* recursive feature elimination  
  maxim relevant
* Look for small deep learning models  
  or synthetic data but make sure the relationships are kept constant
* Try HRV 24 hour (long term) and short term as well
* create frequency spectrum --> then GAN to create the image of spectrum?
* time series analysis : exogenous input?
* feature space --> one column single value??
* time series: actigraph, RR-interval
* activity <-> HRV  
  consider many activities that affects your heart rate other than the psychological factor
* use SVM (small dataset)
* plot histograms, scatter plots, etc
* select a person and tell the person's score
* pandas profiling

**Timeline:**

|  |  |
| --- | --- |
| Week | Content |
| 1 | Meeting with the client and supervisor, brainstorming about the project. |
| 2 | Understanding the MMASH dataset and exploring correlations across various data by producing graphics (heatmap, etc). Doing research and understanding the biology behind stress and Heart Rate Variability. |
| 3 | Doing more research on STAI, Pittsburgh, BMI, HRV, and more data provided by the dataset. Reading papers to see potential biological relations across the data. Selecting features to use in developing machine learning algorithms, cleaning and preprocessing the data. Tried creating a linear regressor for stress detection, but the accuracy was too low due to small dataset. |
| 4 | Trying CT-GAN to synthesise more data as we encountered a problem of small dataset (22 people). Had another meeting with the client to discuss about this problem, and he recommended not to synthesise data artificially as the dataset takes in various data from one person and artificially synthesised data might not correspond to one another correctly. Read more papers and studies, then decided to change from regression to classification for Stress Detector. Created leaflet for the future customers. |
| 5 | Developing and training Activity Tracking machine learning algorithm to identify what kind of physical activity a user is engaged in currently, from accelerometer and heartbeat values from ‘Actigraph.csv’. Due to the large size of Actigraph.csv, we decided that the data is sufficient to develop an “Activity Tracker” algorithm. Analysing various HRV values for Stress-detecting algorithm, on 5-minute basis (short-term). |
| 6 | Developing a classifier for Stress detection using a machine learning algorithm. After trying different parameters as output (PANAS, STAI, Daily Stress), STAI was discovered to be the best parameter for use in the algorithm. Tried different classification algorithms from neural network, decision tree, random forest classifier, and SVM classifier. After comparing the accuracies and simplicities, SVM classifier was selected due to its suitability with small datasets.  Continued hyper-tuning and tried different classifiers for the Activity Tracking algorithm as well for best accuracy. Developed an AWS Web page which displays our data and will be accessed by our users. |
| 7 | Finished up our machine learning algorithms, and integrated them onto the AWS web page so that it can be put into use on the web page. Finalised our product and created poster. |

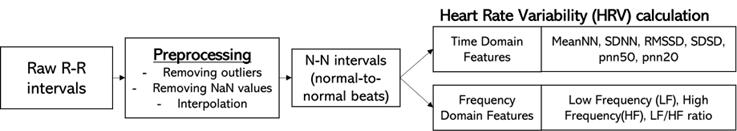
**Development process:**

**Stress Detector:**

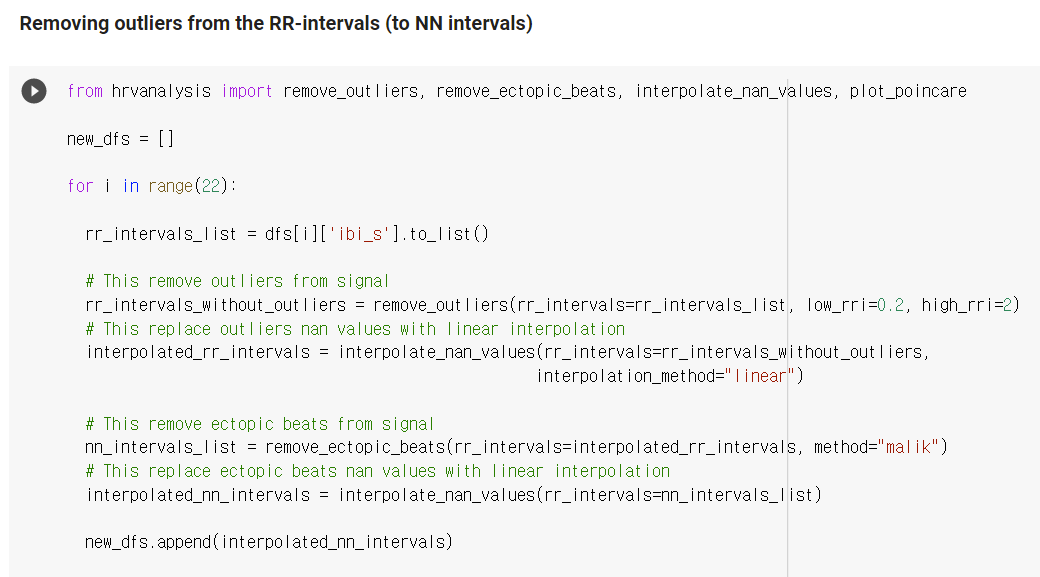
* Data analysis

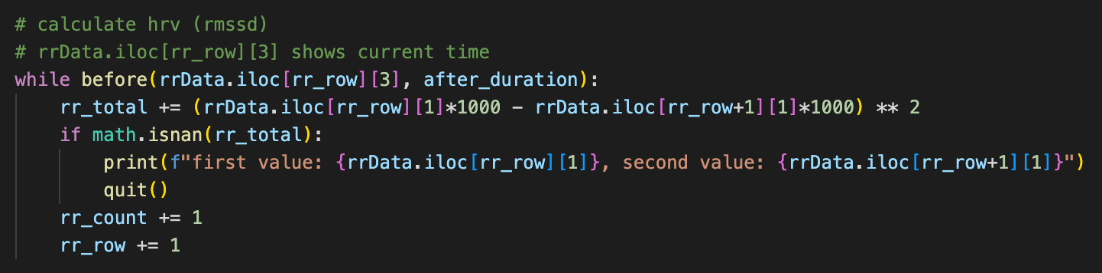
*Multilevel Monitoring of Activity and Sleep in Healthy People (MMASH)* dataset contains physiological data, such as beat-to-beat heart data, accelerometer data, sleep quality and physical activity, as well as the psychological markers such as anxiety, stress and emotional events. Various parameters have been explored to find any significant correlation between them.

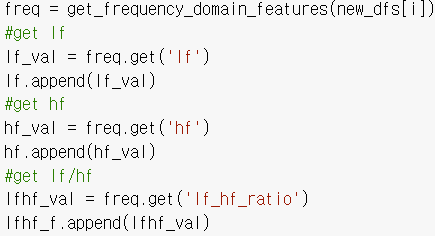
* + Heart Rate Variability (HRV) Analysis

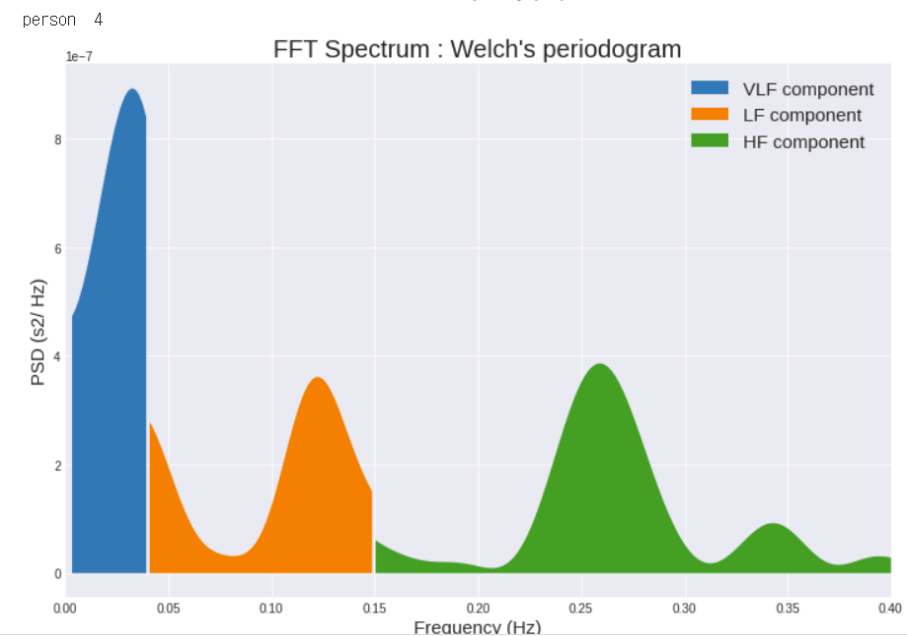


* + - Raw beat-to-beat intervals of 22 participants were analysed to calculate heart rate variability. From looking at the data, it was identified that there were some abnormalities in the data recorded due to potential errors in measurement. Anomalies were removed with ‘remove\_outliers’ functions. The removed values were then interpolated with linear interpolation. Ectopic beats were removed as well, in order to achieve normal beat-to-beat intervals, NN intervals, ready for HRV analysis.

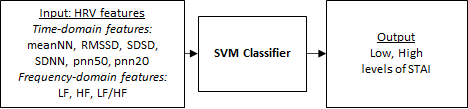
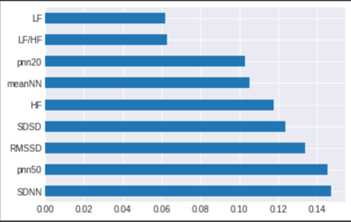
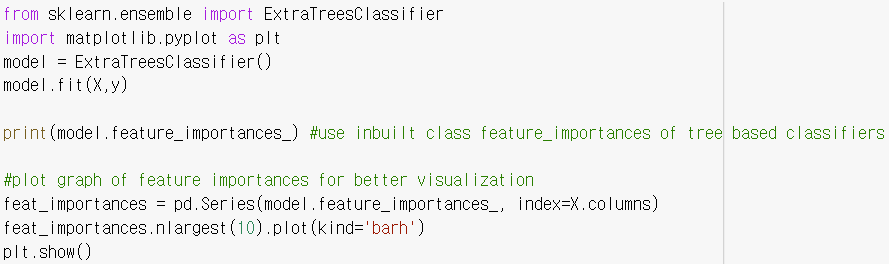


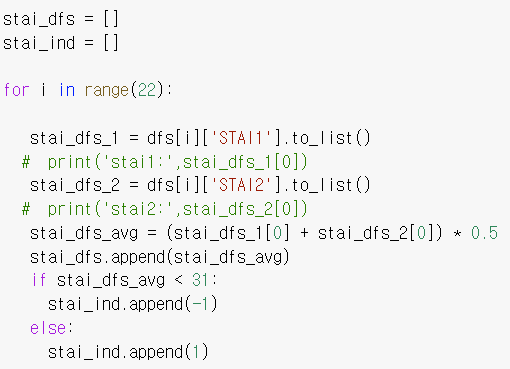
* + - HRV can be analysed with different parameters. Short-term HRV values for 5-minute time frame have been calculated for each subject. Then, mean HRV values have been calculated and have been compared between rest condition and mentally stressful condition. Differences between HRV values have been observed between these two conditions.
    - Time domain features such as MeanNN, SDNN, RMSSD, SDSD, pnn50, and pnn20 were extracted. For instance, it has been found that a greater stress level is related to a lower pNN50 (a time-domain HRV feature).
    - Frequency domain features have been extracted using get\_frequency\_domain\_features function in hrv-analysis library. Low Frequency (LF) and High Frequency (HF) have been calculated, as well as the ratio between them (LF/HF). Studies have found that LF is related to sympathetic (stressful) activities, while HF is related to parasympathetic (restoring, relaxing) activities. A correlation between high LF/HF ratio and high stress level has been observed.



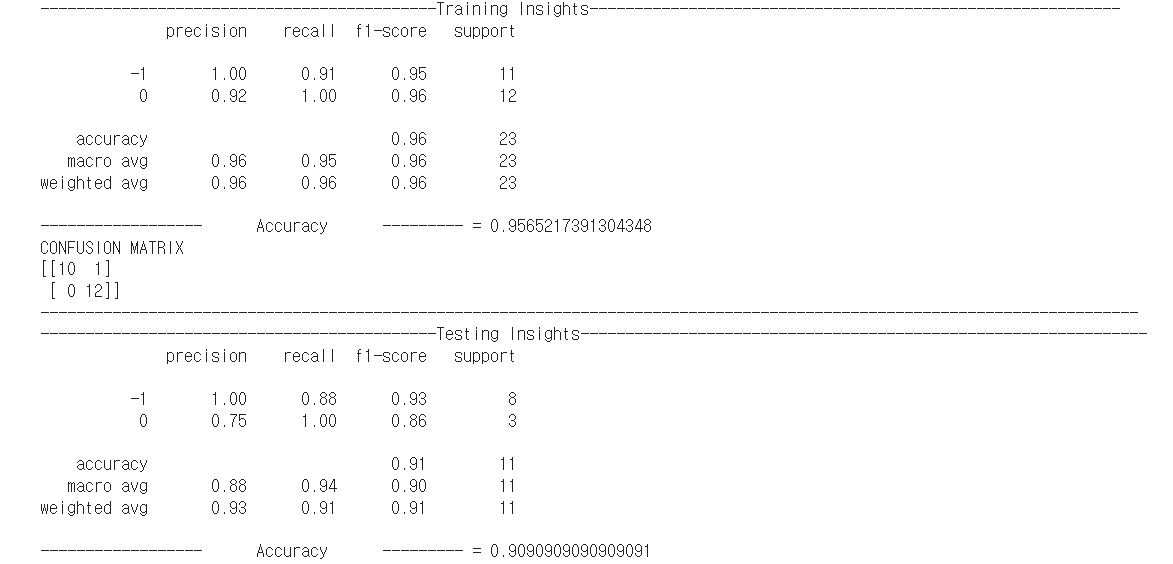


Example FFT spectrum graph of a subject

* + - Feature Importance map has been produced for feature selection, so that important features can be selected for the simplification of the Machine Learning model, while maintaining its accuracy.
* Machine Learning
  + Preparing X (input) and Y (output)
    - “Static Anxiety Index (STAI)”, “Daily Stress” and “Pittsburgh” recorded in a questionnaire done by the participants in MMASH dataset were selected as 3 potential psychological markers. After exploring the correlations, “Static Anxiety Index (STAI)” was selected as our psychological label for the machine learning algorithm. The threshold ‘31’ has been selected as recommended by the dataset. The Static Anxiety Index has been separated to two classes – low and high. This data was used as the label (Y) for the machine learning algorithm.



* + - Different time-domain and frequency-domain HRV features have been used as the input (X) for training for the machine learning algorithm.
    - With Activity Classifier to identify physical activities, mental stress could be differentiated from stress caused by physical movements, so that the changes in HRV features are only caused by mental stress. HRV values at times when physical stress was present were filtered.
    - Due to the imbalance in the dataset, sampling had to be done. Since the dataset is small, oversampling was done to achieve balanced dataset for training.
  + SVM classifier
    - Due to small dataset size of 22, a low-complexity algorithm should be adopted in order to avoid overfitting. Linear SVM classifier has been selected to avoid overfitting and maintain high accuracy.
    - Comparing with other classifiers such as Decision Tree Classifier, the SVM showed the highest testing accuracy.
    - The SVM classifier showed testing accuracy of 90.9%.



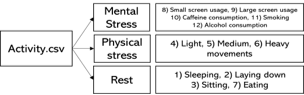
**Activity Tracker:**

* Data Analysis

In order to develop a classifier which predicts the current activity of the user, data that were continuously recorded throughout the day had to be analysed. From the MMASH dataset, Actigraph, which contains accelerometer data of the users, and RR, which contains the raw beat-to-beat intervals of the users, were the continuous dataset to be considered. Looking at the Actigraph dataset more thoroughly, it contains the following features:

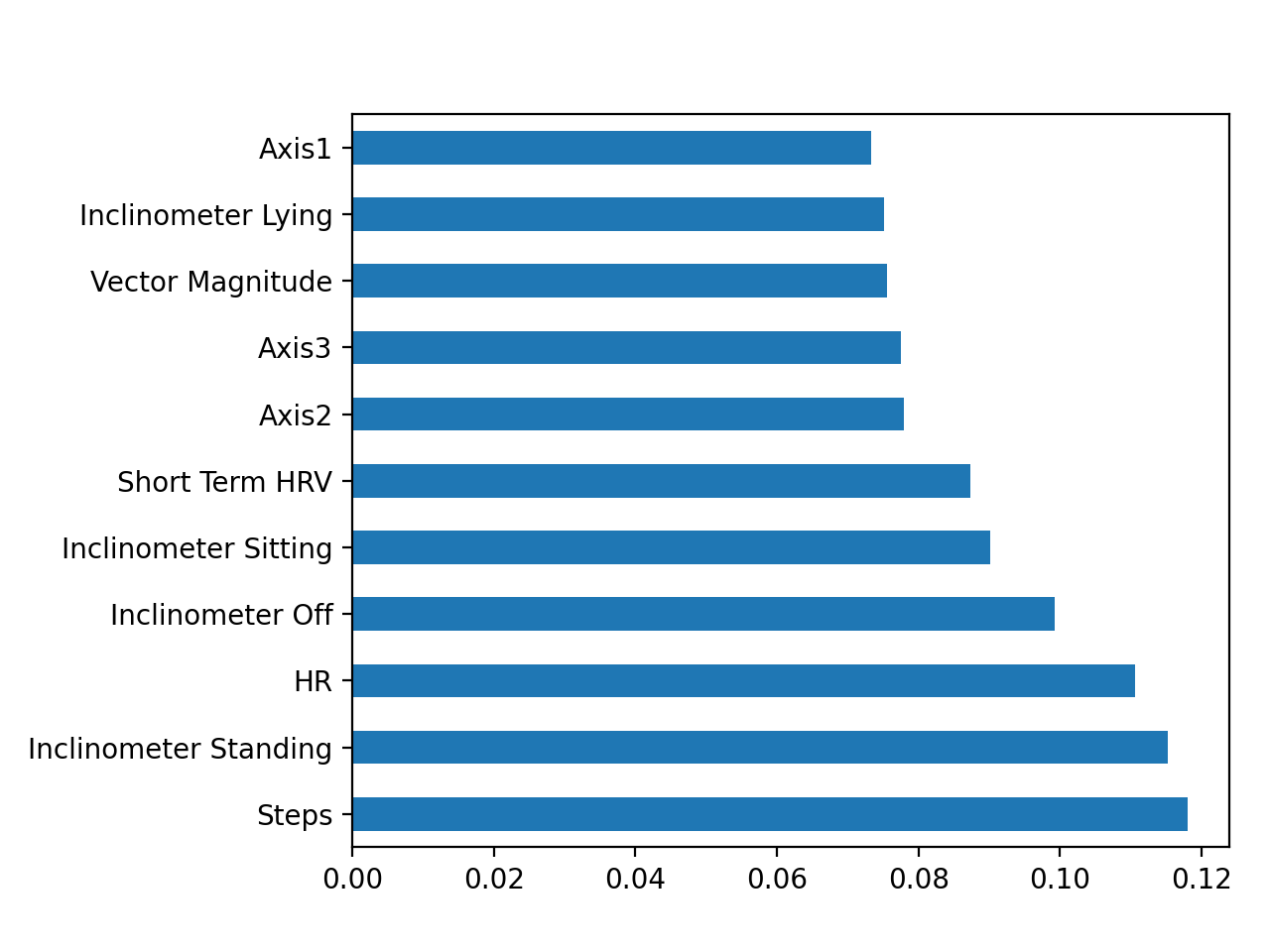
* Axis1/Axis2/Axis3: Raw acceleration data of the X/Y/Z-axis expressed in Newton-meter
* Steps: Number of steps per second
* HR: Heartbeats per minutes
* Inclinometer Off/Standing/Sitting/Lying: values of 0 or 1 indicating the current position of the user
* Vector magnitude: Vector movement derived from the raw acceleration data

The Activity questionnaire from the MMASH dataset shows which activity out of the 12 different categories each user was doing throughout the day at specific time intervals. The 12 categories are as follows and they were grouped into three classes to reduce the complexity and improve the accuracy of the classifier model:

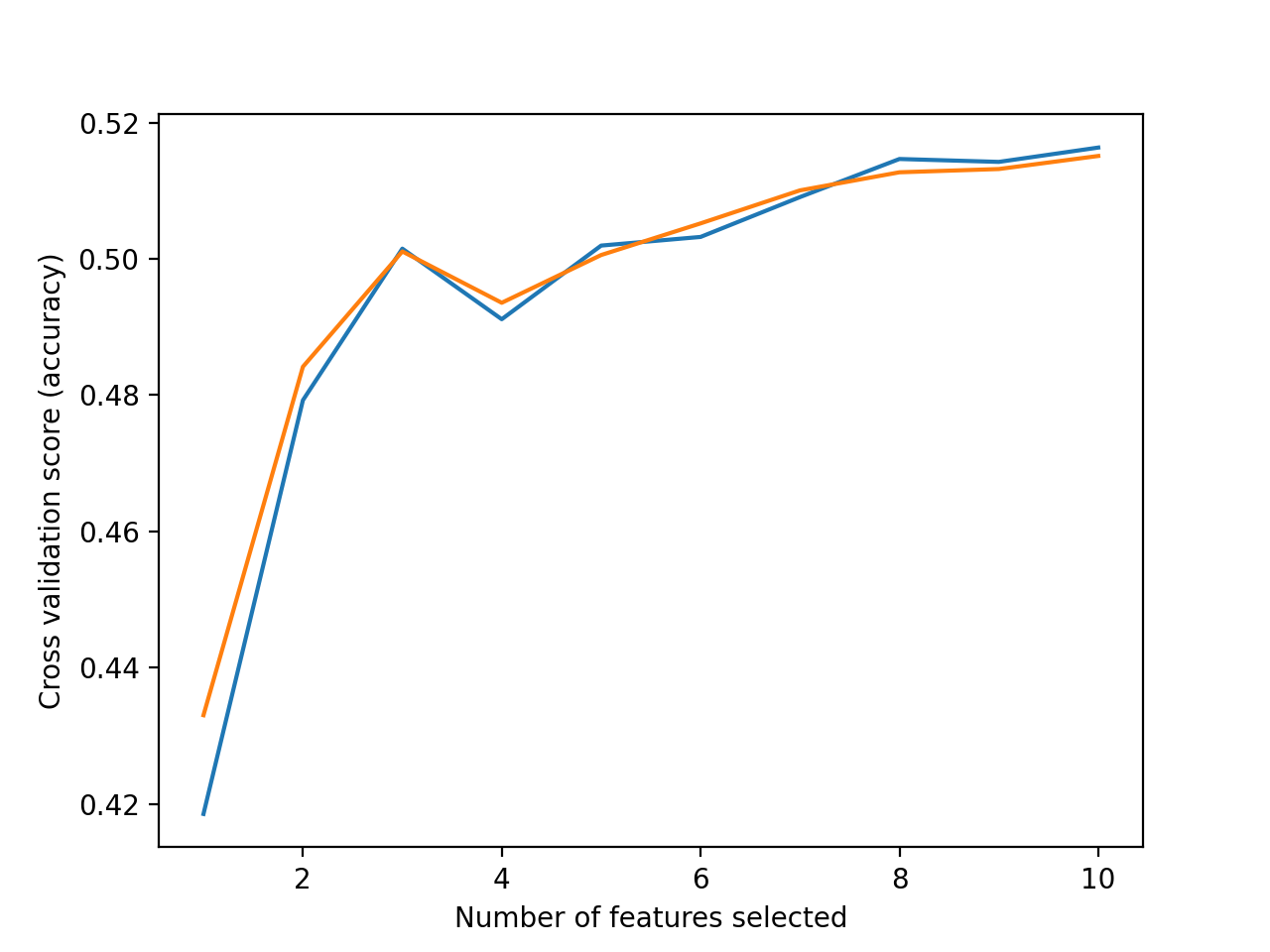


The activities have been categorised based on the effects each activity poses on one’s Central Nervous System, and the hormones they produce. The categorisation has made it easier to distinguish and identify different types of stress. These categorised activity classes were then used as the output labels of the classifier. However, it is important to note that as the Activity questionnaires were manually filled by the users, the time intervals may not be so accurate.

Initially, together with the accelerometer data, Short Term HRV values for 5-minute time frame were used as an input to the classifier. To match with the 5-minute time frame of the HRV values, the actigraph data which were recorded every second had been aggregated. However, the feature importance map below has shown that the short term HRV values were not essential in predicting activity classes.



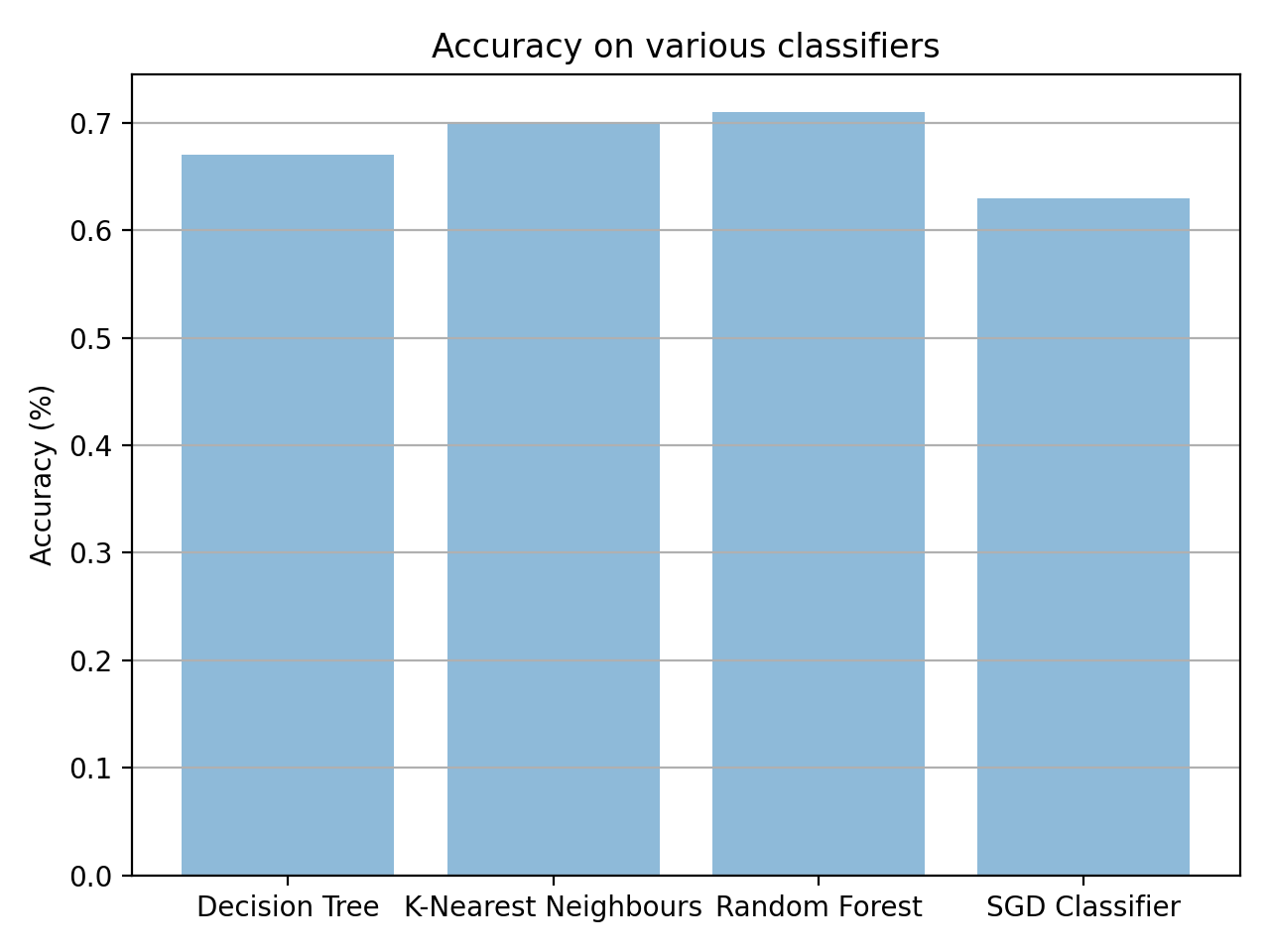
Therefore, the short term HRV values were removed from the inputs, and this allowed a greater number of data samples as the one-second time frame of the actigraph data could be maintained. To further analyse the input data and look for any possibility of reducing the number of input features, a feature selection algorithm called Recursive Feature Elimination has been applied. Recursive Feature Elimination recursively checks the accuracy of the model and removes the weakest feature in order to see how the accuracy of the model varies with the number of input features. The graph below shows the result of the algorithm.



The graph suggests that the accuracy of the model falls as the number of features decreases so it was decided to use all the actigraph features as the input to the classifier.

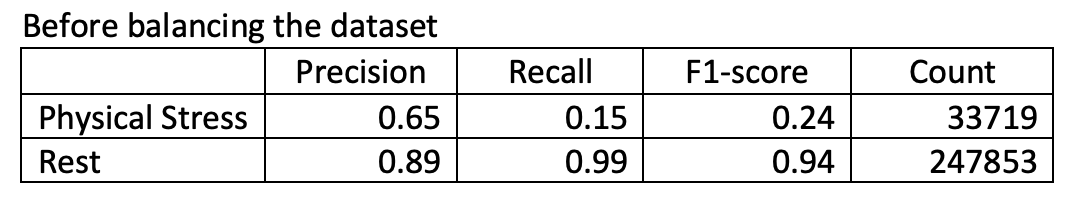
* Machine Learning

It was noted that the actigraph data used are not suitable to predict “mentally stressful activity”. However, it could effectively distinguish between “physically stressful activity” and “rest”. Therefore, the activity tracker was developed as a binary classifier. The data samples with “mentally stressful activity” as output label were removed due to that reason. Various classifier models were compared to find the best classification algorithm for the developing activity classifier and as shown in the following bar graph, Random Forest Classifier showed the highest accuracy hence it has been chosen.

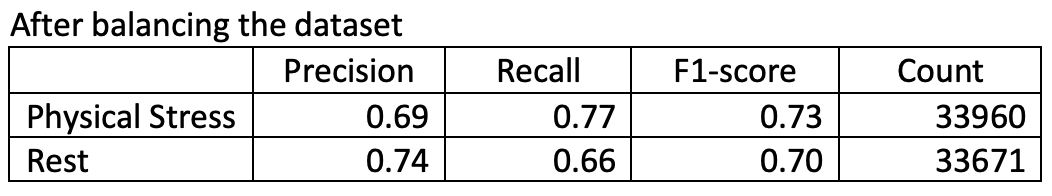


Random Forest Classifier is a classification algorithm which consists of multiple decision trees. Each independently trained decision tree makes a prediction of an output class for the given input and the output class which receives the most votes from the trees becomes the prediction of the entire model. As multiple independently trained trees are combined, Random Forest Classifier can make robust predictions without being biased.

It was observed that most of the samples had “rest” as output label, which could cause various problems while training a machine learning model as an imbalanced dataset. Most classification algorithms are designed to deal with a balanced dataset where the distribution of the classes among the samples is equal. Thus, if an imbalanced dataset is given the algorithm may treat it as a balanced dataset. This involves the risk of the classifier model only learning characteristics of the majority class and neglecting the minority class. This could be shown when a model was trained with the imbalanced dataset.

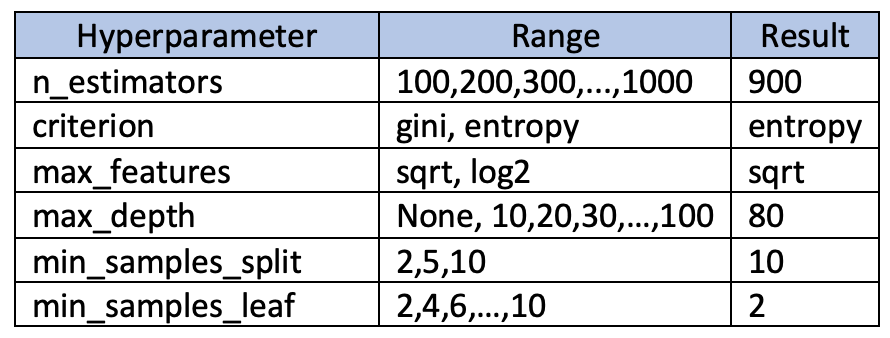


The above table is the classification report of a random forest classifier which was trained on the imbalanced dataset. Even though a high accuracy of 89% was recorded, the recall and F1-score for “Physical Stress” were only 0.15 and 0.24 respectively. This means that lots of samples from “Physical Stress” were predicted as “Rest” showing that the trained model is extremely biased towards the majority class, “Rest”. So accuracy is not a suitable evaluator in this case with the imbalanced dataset. There were roughly 8 times more samples from “Rest” than those from “Physical Stress”. To fix this problem, under-sampling was done by reducing the number of samples from the majority class while keeping the number of samples from the minority class. The following table shows the classification report of a random forest classifier which was trained on the balanced dataset after under-sampling.



The number of samples from the two classes are very similar now and the recall and F1-score of “Physical Stress” greatly improved.

To further improve the classification model, hyperparameter tuning had been attempted. The Randomised Search approach was used to save time because constructing thousands of random forest classifiers would take a huge amount of time. The following ranges of parameters were used for tuning and the result was as shown.



However, only a slight increase in the accuracy was observed.

The developed activity tracker can be used to predict the current activity of the user between “Rest” and “Physically stressful activity”. This gets rid of the inconvenience that users may experience while filling in forms manually to indicate when they have done physical activities. With the website showing the result of the tracker, users would be able to easily view how long they have done physical activities throughout the day.

**AWS Webpage:**

<http://13.40.221.14/>

The web app is hosted and stored on AWS and can be accessed via the link above. The website was based on a free-to-use and open-source template from <https://templatemo.com/>.

The web app is used to display and access all the work we have done in this project in a clear way. The web app shows three key statistics:

* Live Analysis- predicts stress levels of the user based on live data from wearable biosensors. If stress levels are detected to be too high, then advice and interventions based on the recorded physiological data is given
* Activity Tracking- uses live data from biosensors to predict whether the user is being active or at rest. Data from this is also used to keep track of exercise and daily activity of the user
* Dashboard- results of data analytics of stress and anxiety, from the MMASH dataset, of the 22 subjects are shown here where you can view the day of the cohort as a whole and individual subjects, with comparisons to the rest of the participants. If a subject is found to have more stress and anxiety than the rest of the cohort, then possible interventions for that subject, based on their own data, are given.

Linking the two:

Source code:

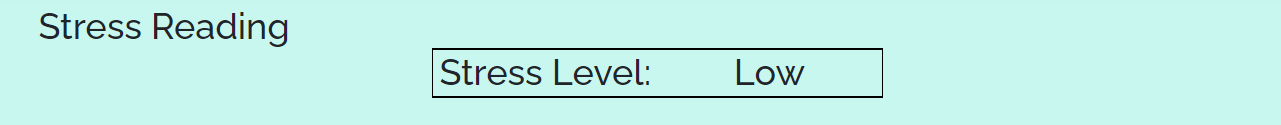
<https://github.com/anster01/MFTech>

The GitHub above displays all the code including the website, data analytics and machine learning. The website is coded in HTML, CSS and JavaScript with some PyScript to run the machine learning programs.

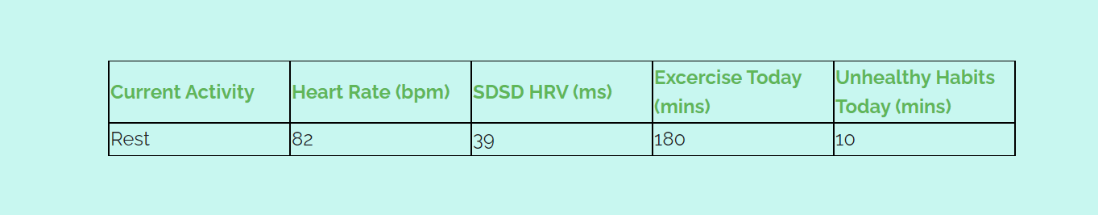
On AWS, a web server is set-up and then the website can be accessed via the link above. The web server is set-up using Apache2, which is an open-source provider for HTTP servers. On the server, by inputting *‘start service apache2’* in Terminalwithin the directory of the *index.html* file, the web server runs and so the website is accessible publicly via the link above. A web server is needed so that the website can access local files, such as databases and Python scripts, that are stored on the server within the same directory.

Live Analysis:

1. On page load, the website uses PyScript to call the script *./assets/SVM\_model.py* which runs the ML model for stress prediction
2. A value representing either ‘Low’ or ‘High’ is returned and this is displayed clearly on the website. A stressmeter graphic (coded using Chart.js) is also shown on the website.



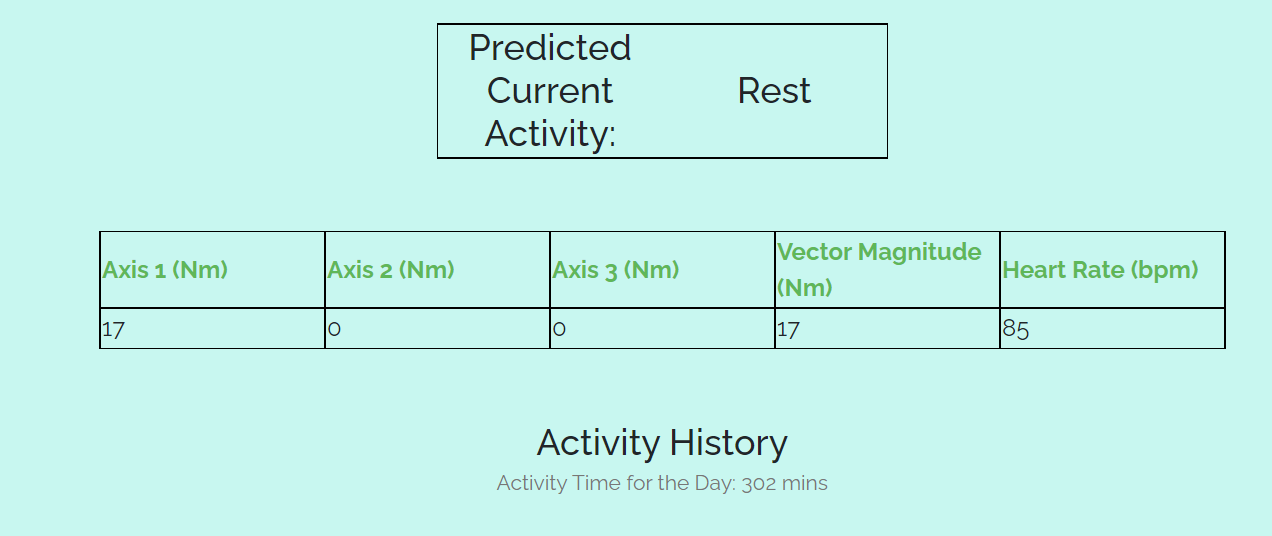
1. All the live bio-sensor data, such as HRV and heartrate, which the SVM model uses to predict stress levels are displayed in a table.



1. If stress levels are predicted to be high, then from the data given, possible lifestyle interventions are provided to help lower the stress levels

Activity Tracking:

1. On page load, the website uses PyScript to call the script *./assets/RFC\_model.py* which runs the ML model for activity detection
2. A value representing either ‘Rest’ or ‘Physically Active’ is returned and this is displayed on the website along with the live the accelerometer and axial data measured by the biosensors
3. Data from this is also used to calculate physical activity throughout the day and is this is also displayed on the website



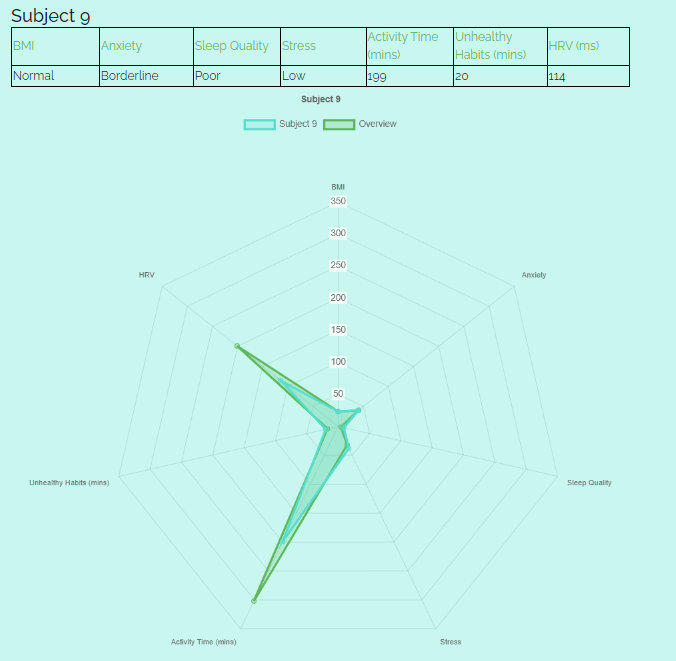
Dashboard:

This section is a small case-study on the 22 people from the MMASH dataset given to view their, both individually and the group, to find correlations between their physiological data and their mental health and wellbeing markers, in particular stress, anxiety, and sleep quality. If a subject is found to have any of these markers particularly high or bad, then possible interventions are given for this subject to improve their mental wellness.

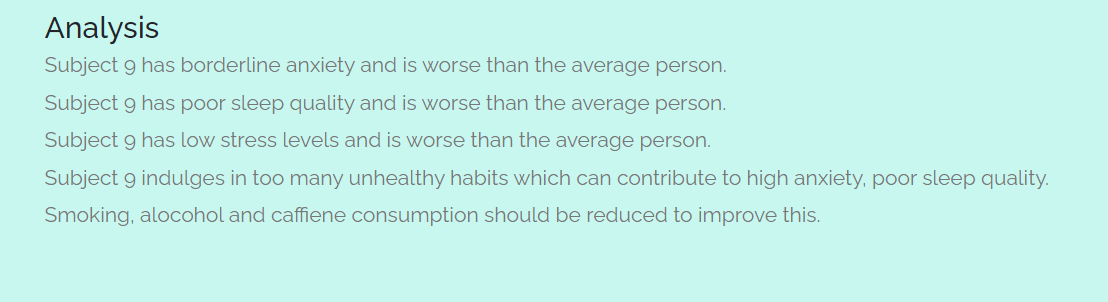
1. Data is loaded from a local database stored on the server.
2. The user can select whether to view data about the whole group or about individual subjects
   1. For the group, Chart.js displays a stacked bar chart showing the number of subjects in each level of each marker e.g., low anxiety, borderline anxiety, high anxiety of which the thresholds are defined by MMASH.



* 1. For individual subjects, Chart.js display a radar graph comparing the selected subject to the average of the group.



1. If levels of the markers are found to be too high, then possible lifestyle interventions for that subject are provided based on their physiological data.



**Ethical Considerations:**

All personal data used in the development of this project was taken from the MMASH dataset. The MMASH dataset provides the following ethical disclaimer: “22 healthy young adult males were recruited. Before starting, the participants signed an informed consent to take part in this study. This provided information about the research protocol, possible risks and data usage, in accordance with the General Data Protection Regulation: Regulation - EU 2016/679 of the European Parliament and of the Council 27/04/2016 - on the protection of private persons with regard to the processing of personal data and on the free movement of such data. In accordance with the Helsinki Declaration as revised in 2013, the study was approved by the Ethical Committee of the University of Pisa (#0077455/2018).”

For future uses, the data shown on the web page will be kept anonymous, to protect individual privacy. The data will strictly be used for the purpose of machine learning training and analysis for the dashboard only. The individuals using the Health Tracker will only have access to their specific health details.

**Sustainability report:**

The project does not require the use of additional hardware to function since it is primarily based in software. Although an external recording device, such as a smart watch or health band, must be used, the software does not require proprietary hardware, so can take advantage of existing hardware. Since smart watches are very common and inexpensive, they eliminate the need to produce more hardware to support the product, thus minimising its environmental impact.