Dyson School of Design Engineering | MEng Design Engineering

Module Exam

Module code and Name	DE4-SIOT Sensing & IoT
Student CID	01228101
Assessment date	**10am, 13th Jan 2020**

Presentation URL (publicly accessible link): https://youtu.be/OdvbFIwX5To

Code & Data (publicly accessible link): https://github.com/rs2416/SensingandIoT

Coursework 1: Sensing

1. Introduction

In 2013, there were 8.2 million cases of anxiety in the UK [1]. The number of people with anxiety disorders in England is substantial and will grow over time. Despite high and rising prevalence of anxiety, current solutions are inadequate in providing treatment for everyone and in a timely manner [2].

In response to the inadequacy of current solutions and the greatly expanding commercial market for monitoring physiological data, researchers are increasingly trying to understand how to predict mental health through non-invasive data collection using wearables [3]. Nevertheless, there has been little effort to monitor and predict anxiety specifically.

This project is a preliminary investigation to see whether it is feasible to predict social-related anxiety using heart rate (HR) and sweat response (EDA) data and machine learning techniques (binary classification problem).

Physiological data was collected for 2 weeks, specifically during 5 anxiety inducing events which included 3 pitches, a job interview and a staff-student committee meeting. This physiological data was labelled through collection of contextual data such as calendar data (Google Calendar API), Global Positioning System (GPS) data and mobile phone speed (Followmee.com). The data was then cleaned and then split into 70% train and 30% test. Following this, various classification algorithm such as SVM, Logistic Regression and Decision Tree were used to train predictive models. These models were then evaluated for performance using 7-fold Cross Validation. A website platform was then created using React JavaScript to express the benefits of prediction of anxiety and how the contextual data can facilitate self-help activities.

2. Project objectives

This project had 7 main objectives:

1. Collect GPS and mobile speed data by implementing a web driver and crontab scheduler on Raspberry Pi to automatically collect the data from 'www.followmee.com' and then upload data to Google Sheets (for 15 days).
2. Collect Google Calendar data, using the Google Calendar API and crontab scheduler on Raspberry Pi and then upload data to Google Sheets (for 15 days).
3. Collect HR and EDA data by wearing E4 Empatica for 15 days, specifically during the 5 anxiety inducing events.
4. Clean and label data as anxious and non-anxious (0 or 1) using contextual data.
5. Train predictive models using various classification algorithms.
6. Evaluate the models for predictive capability.
7. Design and create a website platform using React JavaScript to indicate benefits of prediction and scope for self-help platforms.

3. Plan

The following plan was created to ensure timely delivery of the objectives, Figure 1. This involved considerations of various stages of the design, development and delivery. Effective planning during the project was particularly crucial as data needed to be collected during the anxiety induced events which took place from week beginning of 2nd December to 12th December.

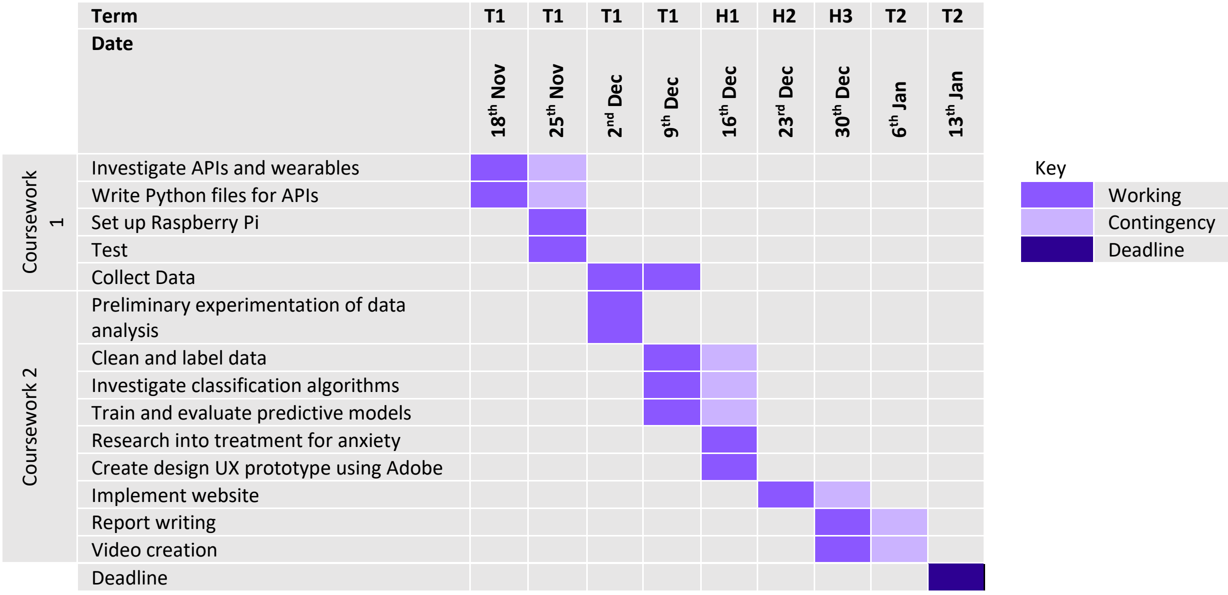

Figure 1. Gantt chart

4. Data sources and sensing set-up

Below are the justifications for the following selected data types and sources.

4.1. Physiological data – E4 Empatica

- **Type of data:** Elevated heart rate and sweat response are common responses during an episode of social anxiety [4]. In addition to this, these physiological responses are more common amongst other types of physiological responses experienced during social anxiety [5]. The E4 Empatica wearable offered simultaneous collection of both these types of data, contrasting to the majority of wearables on the market that don't offer an EDA sensor [6][7][8].
- **Ease of use and set-up:** As the device is designed for research purposes, downloading the CSV data is a frictionless process which involves using a USB; this was advantageous in comparison to using devices from companies such as Apple, who make it significantly difficult to download one's data.
- **Sensor sampling rate:** As the device was produced for use in research context, the device offered sophiscated sensors that sample at effective preset rates in order to ensure less information loss [9].

4.2. Calendar data - Google Calendar API

- **Type of data:** By collecting calendar data the activity of labelling the physiological data as anxious and non-anxious is more accurate. For example, I would be able to recall when I was pitching.
- **Ease of use:** The Google Calendar API was also selected as I had decided that I would store the data using Google Sheets. By using two types of Google APIs, the process of setting up the data collection was optimised due to familiarity with the process of setting up Google APIs.
- **Set-up:** The data was collected using a Python file that was scheduled to run at certain times using the crontab scheduler on Raspberry Pi, Figure 2.

Figure 2. My Raspberry Pi plugged into mains power, storing Python files to be run by crontab scheduler.

4.3. GPS data and mobile speed data – 'Followmee'

- **Type of data**: By collecting my GPS and mobile speed data in combination with calendar data, the activity of labelling the physiological data as anxious and non-anxious is further optimised. For example, using this data I can infer if I am late for events, in addition to this I can infer if elevated physiological activity is due to movement. This was assuming my mobile is on me constantly.
- **Ease of use:** I decided to use 'followmee.com' as it was one of the few platforms that allowed me to download my GPS and mobile speed data using one application. In addition to this, I wanted a platform that would allow me to download my data using a browser as I was automating the

process using a web driver. Alternative applications required a Mac or Android phone in order to download CSVs over the internet [10].

- **Set-up:** The data was collected using two Python files that were scheduled to run at certain times using the crontab scheduler on Raspberry Pi, Figure 2.

5. Data collection and storage process

5.1. Data collection and sampling rate

The EDA data was sampled at 4Hz and the HR (BPM) data was sampled at 1Hz. During the data analysis, in order to ensure that EDA information was not lost the HR data was resampled to 4Hz, since the more samples taken, the more accurate the digital representation. Post data collection and analysis, it was realised that the highest frequency for EDA (Fmax) is generally confined to 0.37Hz therefore a sampling rate of about 1Hz would have been appropriate based on Nyquist criteria [11][12].

Real-time data collection using the Empatica could have taken place using Bluetooth Low Energy (BLE). BLE involves low battery consumption and relatively low bandwidth compared to Wi-Fi. This data collection process would have involved the Empatica broadcasting a packet, which would have been read by other devices in range such as my mobile [13]. Using a prototyped app this data could have been sent to Google Sheets using Wi-Fi, Figure 3. Nevertheless, for reliable data collection, my phone needed to be within the range of Empatica constantly for 2 weeks, which I could not ensure. Therefore, a more reliable method was chosen, which involved transferring data by connecting a USB to the device and my laptop at the end of each day.

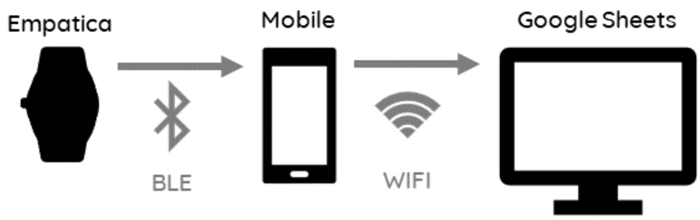

Figure 3. Real time data collection flow

The GPS and mobile speed data were sampled every 2 minutes (0.5Hz), this rate was preset by FollowMee. Although there was information loss, it can be argued that this is not a substantial information loss, as I would still be able to infer whether I am late for an event. A Python file called '*Web_Driver.py*' was used for the collection of this data. The file would activate a selenium web driver and log into the FollowMee platform and download my data from the last 24 hours in CSV format, Figure 4. This was scheduled to run at 6:01AM. A screen recording of the automated browser is featured on my GitHub.

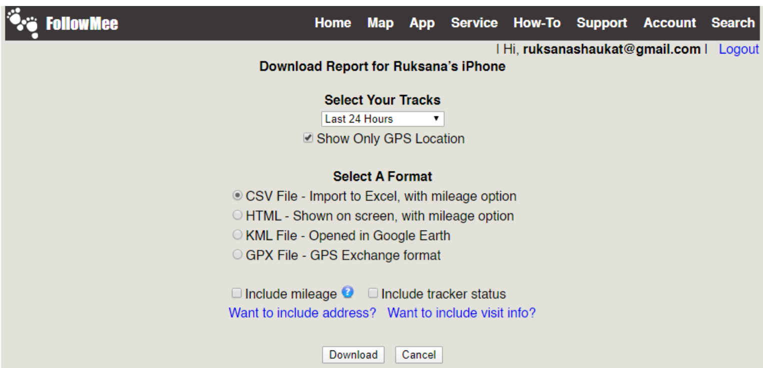

Figure 4. Screenshot taken from webdriver of download page

The calendar data was collected at the beginning of each day. A Python file called '*CalendarAPI.py*' would make an API request and return calendar events for the day (returning the events in the following 15 hours). Information such as event name, location, start time and end time was collected. This data was pushed to Google Sheets. The file was run at 6am. No information was lost, assuming that calendar event information was not adjusted on day of the event (after 6am).

5.2. Data storage

The GPS and mobile speed CSV were transferred to Google Sheets, using the Google Sheets API, the CSV was then deleted from Raspberry Pi. This was done using a file called '*Upload_GPS.py*' which was run 5 minutes after the file was downloaded. The deletion of the CSV ensured constant storage space on the Raspberry Pi. As the CSV could be repeatably downloaded from FollowMee, there were no concerns about information loss by deleting the CSV.

The calendar data was published directly to Google Sheets, using the Google Sheets API. This was done using a file called '*CalendarAPI.py*'. As mentioned previously, the physiological data was stored on my computer using a USB every day.

6. Basic characteristics of the end-to-end systems set up and data

Figure 5. summarises the end-to-end system set-up as well as the data collected and types of communication that took place. A Raspberry Pi 3 model B connected to Wi-Fi was used in order to automate the data collection. A Raspberry Pi was used instead of Arduino, due to its computational capacity and in-built Wireless LAN. The scripts used to collect and publish data as well as the data collected are published on my GitHub.

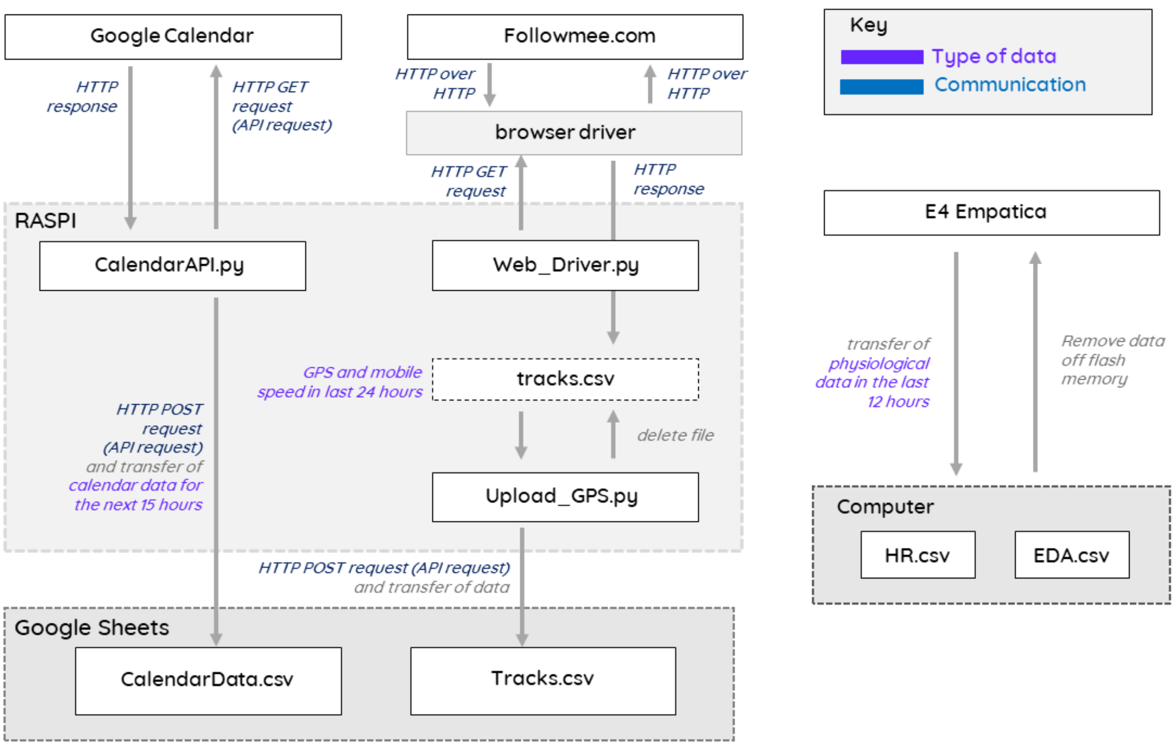

Figure 5. End-to-end system

7. Basic Time Series Data Analysis

Figure 6. indicates the continuous physiological time series with the anxiety labelled using the green background. This plot was formulated after the physiological data was cleaned and labelled; the data plotted is only the data collected during the days of the anxiety inducing events.

It is evident that there is great variation in the data collected, however it is somewhat clear that EDA (red) positively correlates with anxiety. From formulating a correlation matrix, Figure 15, it was evident that heart rate also correlates with anxiety. The high variance in data points and the limited data (only 5 anxiety inducing events analysed) suggests that the model that will be trained may experience overfitting despite applying the moving average filter.

From observing the plot, it is apparent that binary classification does not take into the account the severity of anxiety, as the plot indicates great variation and clear peaks of anxiety within the anxiety labelled data.

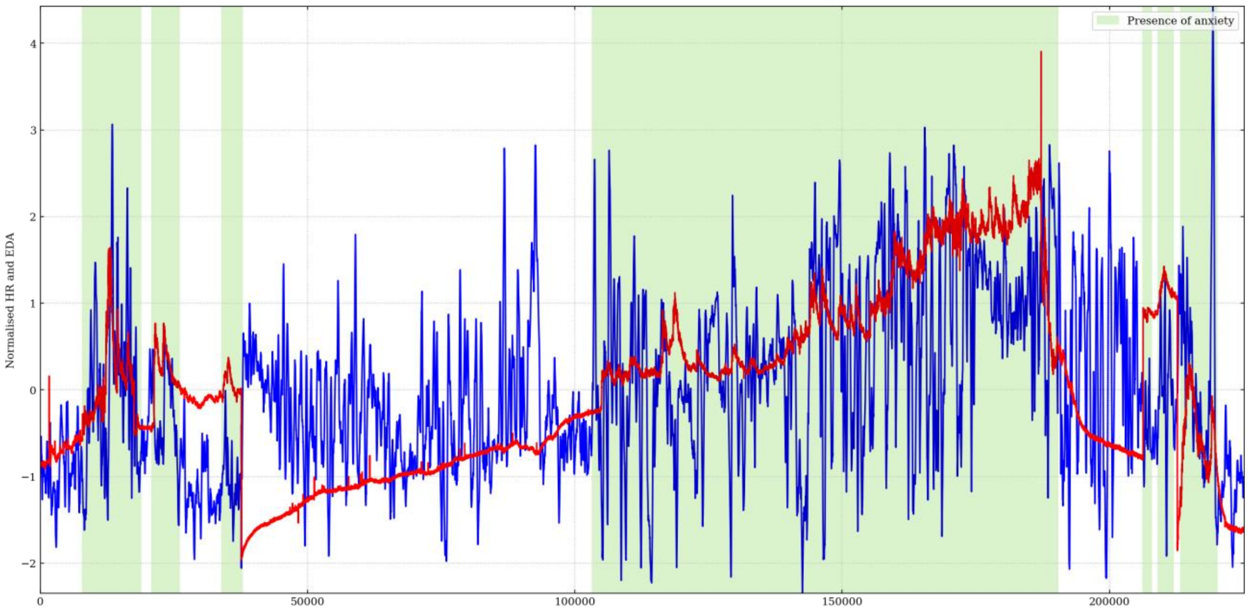

Figure 6. Plot of normalised physiological data and anxiety (in green).

Coursework 2: Internet of Things

1. Introduction

Within this section a website prototype was implemented using React JavaScript and CSS, in order to express the benefits of predicting anxiety and collecting contextual data. This section also expands on the data analysis that was performed, this involved pre-processing the data and training 7 predictive models using classification algorithms such as SVM, Random Tree and Logistic Regression. These models were then evaluated for accuracy using 7-fold cross validation.

2. Data interaction, visualisation and actuation platform

Following research into what aids social anxiety, I have designed a responsive self-help website platform that is also an access point to therapists. The website platform is tailored to the individual using it and the application becomes more personalised over time as the system begins to understand what makes the user anxious based on anxiety recognition combined with contextual information.

Figure 7. Sophie [14]

This prototype is designed for user persona called Sophie, Figure 7. Sophie is a working woman with social anxiety, she frequently presents pitches as part of her role as a businesswoman and finds it hard motivating herself to socialise, however, she is too embarrassed to seek help from a therapist.

2.1. User Centered Design, Visualisation, Interaction & Actuation

It is common during cognitive behavioural therapy (CBT) for social anxiety, to have individuals plan novel social activities. In response to this finding, an item on the dashboard (Figure 8) **visualises** Sophie's social interaction progress and the information in the blog section explains the need for social interactions for improving social anxiety [15]. The blog section can be **actuated** by clicking on the side bar. This section provides information to improve awareness of social anxiety as well as advice and therapy surrounding anxiety. As the system understands that Sophie is anxious about pitching it suggests blogs for improving anxiety (see video).

In addition to this, the dashboard allows Sophie to **visualise** her monthly anxiety progression (based on the model recognising her anxiety), which is reinforcement to incentivise Sophie to continue with the self-help platform, Figure 8. In addition to this, she can also **interact** with the graph and filter the days using the day selector. There is also a section which she can **actuate** by clicking on 'physiological data by event' presented on the dashboard page. This page (Figure 9) allows Sophie to view their data for each event in her calendar, it allows her to **visualise** the duration of anxiety during her pitching experience. The graph is available to **interact** with, allowing the viewer to see their heart rate and sweat values were per minute. When the 'view progress' button is clicked another graph is **actuated**, allowing her to **visualise** her anxiety since her last pitching experience, this plot compares the severity of each experience (see video).

The system also includes a feature where you can contact a therapist. Alongside this, there is a section which involves creating diary entries, this gives the platform further information such as sentiment and negative language patterns. This can also help Sophie realise her cognitive biases (see video).

2.2. Usability

Ease of use was considered while designing the platform. Familiar cues were used, to indicate to the user how to interact with the platform. For example, shadow was actuated as the user hovers over the side bar to indicate an interaction. The website is also responsive making the design accessible to many users. Anxinsights also featured a constant side bar to enable easy navigation. Other considerations such as font size and readability were taken into account. The handwriting style for the logo, also reminds the user that this is personal self-help platform, Figure 10.

Anxinsights

Figure 10. Logo font

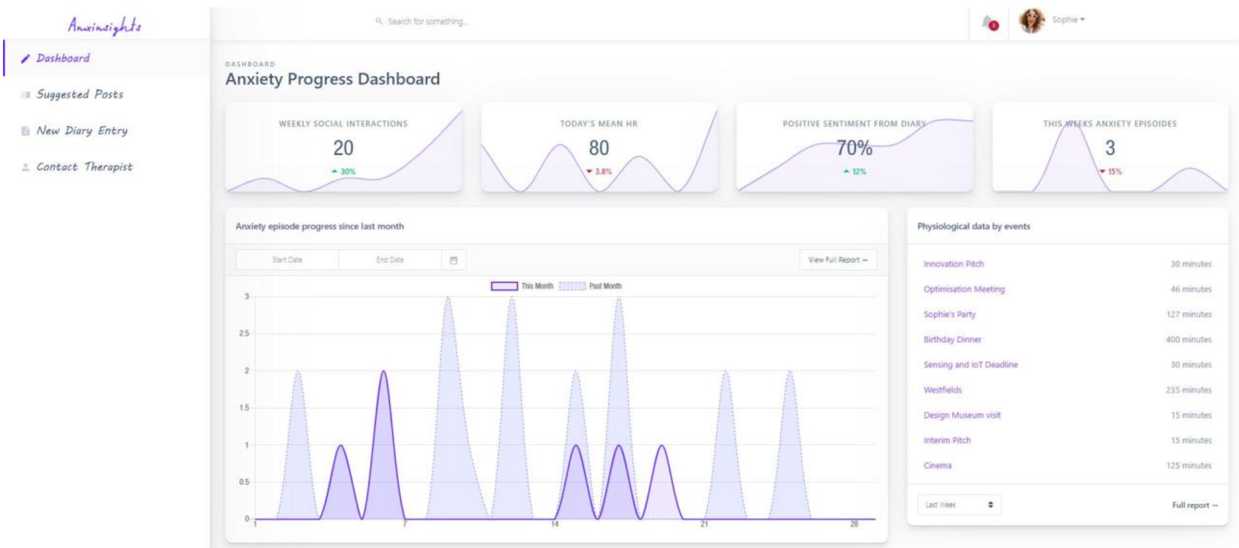

Figure 8. *Dashboard indicating anxiety progress, weekly social interaction progress and 'physiological data by events' section (clicking actuates Figure 9).*

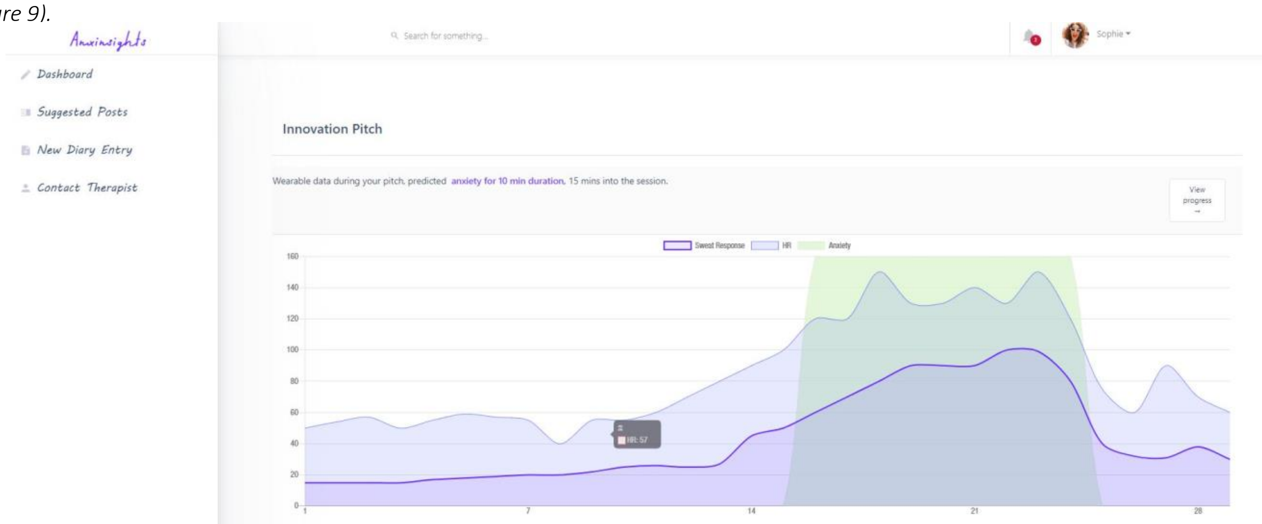

Figure 9. *In depth insight into physiological data per event and anxiety duration.*

2.3. Framework

Figure 11. *Left - small stats components (e.g. 'Weekly social interactions') from dashboard mapping data per component. Right – storing data as properties ('props').*

React is a popular JavaScript library for creating user interfaces. It is used for building fast and scalable web applications using component-based architecture. React has many benefits over other frameworks, one of the main benefits is that it involves reusing components (pieces of UI) [16]. React framework was used in this context as the initial UX/UI designs involved reusing components such as graphs, blog post formats etc. This platform was created by harnessing an open source template and bootstrap; the design was then improved by adding more components, routes and customising the CSS.

React components essentially make up a tree of components. Data will be passed down to the given UI in a structured manner, using properties and states, Figure 11. For the sake of the prototype the chart data was held in an array within the component rather a JSON file or through React properties [16].

3. Data analytics, inferences and insights

The data was cleaned, labelled and analysed. Data cleaning was done using a Python notebook called '*data_cleaning.ipynb'* and the analysis was executed using the '*data_analysis.ipynb'* notebook.

3.1. Data pre-processing

3.1.1. Resampling and smoothing

HR data was resampled to 4 Hz, instead of resampling EDA data in order to avoid information loss. Then a moving average filter was applied to both datasets, this decision was made in order to reduce overfitting, by reducing the noise associated with the collected data. Since overfitting can limit the generalisability of a predictive model.

```
anxietyduration1 = 86400
counter = 1

while counter < anxietyduration1:
        counter = counter + 1
        eda_1['Labelled'][start_anxiety] = 1
        start_anxiety = start_anxiety + 1
```

Figure 12. Using a for loop to assign anxiety labels. Start value was prelabelled, and duration of anxiety precalculated.

3.1.2. Labelling

The labelling process was informed by the collected calendar, GPS and mobile speed data. Another column was added to the EDA dataset, and considering the time started and sampling rate I was able to label the initial time of anxiety and then use a for loop to automate labelling for the duration of anxiety, Figure 12. Despite having the contextual data available, labelling the presence of anxiety was also subjective as the anxiety experienced was generally due to anticipation. The systematic approach that was taken when labelling the large amounts of data is further illustrated in Figure 13.

Date	Anxiety Events (Most anxiety first)	Start time of data collection	Duration of data collected that day (mins)	Total number of cells (mins x frequency x seconds)	Anxiety time (Hours, start - end)	Duration of anxiety period (mins, start - end)	Number of anxiety samples (mins x seconds x frequency)	time till start of starts (mins)	Cell start anxiety (without chopping)	Edited cells - clear anolymalies
10/12/19	Pitch (practicing 1:27, started presentation 7:00pm)	8:41	805	193200	1:16-7:16	360	86400	275	66000	Changed peak readings from 1.92 and 1.45 to 1.1. Cells 149608-149614
12/12/19	Interim Pitch	8:21	894	214560	10:17 - 10:42.5 (half sec)	25.5	6120	116	27840	Changed values due to peak at 26298-26303
25/11/19	Class pitch	8:43	159	38160	10:44 - 11:21	37	8880	121	29040	Changed EDA peaks rows 22429 - 22436, from 0.37 around to 0.23/0.24

Figure 13. Snippet of a spreadsheet that was used to optimise the process of labelling.

3.1.3. Combining and balancing data

The 3 types of data were then combined (HR, EDA and anxiety labels) for each day. The number of binary labels were then counted. Some 'non-anxious' data (data labelled 0), was then removed until the data was labelled about 50% non-anxious, 50% anxious, Figure 14. This improved the balance of the data and the bias of the model that would be created. The data from the 5 anxiety inducing events were then combined.

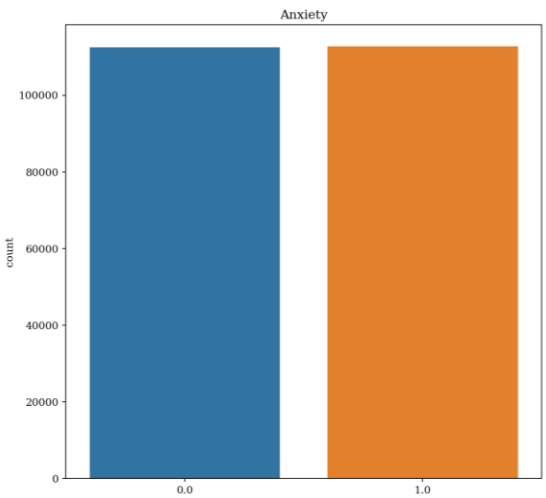

Figure 14. Anxiety label count

3.2. Data analysis

In order to investigate whether anxiety correlates with EDA and HR, a correlation matrix was formulated, Figure 15. The matrix indicated that there was a positive correlation between anxiety and both types physiological data. As well as indicating EDA as having a greater correlation with my anxiety (0.74) than HR (0.34). It was important to consider that these were correlations relating to my personal anxiety, rather than anxiety in general, reiterating that this is a predictive model for my personal anxiety response.

Data was then split into 70% training and 30% test data. Various binary classification algorithms were used to train the predictive models (Radial Support Vector Machines (SVM), Linear SVM, Logistic Regression, Decision Tree, K-nearest Neighbours, Gaussian Naïve Bayes and Random Forest). A majority of the models indicated accuracies above 90%.

These initial accuracy scores obtained weren't representative, due to the lack of data and high variance. Despite the removing of data to create a balanced data set, it would be more effective to train and test the algorithm on every instance of the data set. Cross validation is a resampling procedure used to evaluate models if there is limited data available [17]. A '7-fold Cross Validation' was performed, ideally, I would have performed a 10-Fold Cross Validation, but I was unable to do so due to a lack of computational power.

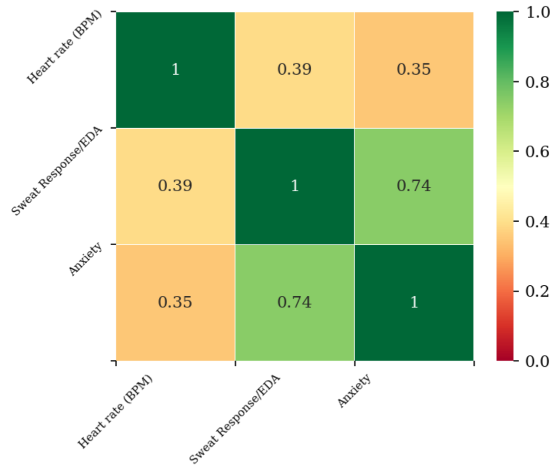

Figure 15. *Correlation matrix*

From the Cross Validation, it was evident that the Linear SVM and Logistic Regression were the most effective algorithms to train models to predict my anxiety, amongst the 7 algorithm, Figure 16. The predictive capability of these models were further illustrated by plotting confusion matrices, Figure 17. Confusion matrices are an also a common technique for evaluating the performance of a classification algorithm [18].

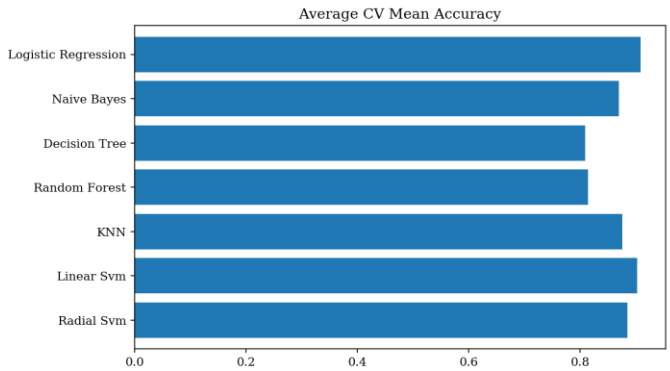

Figure 16. *Plot of average mean accuracies of each model using 7-fold Cross Validation method*

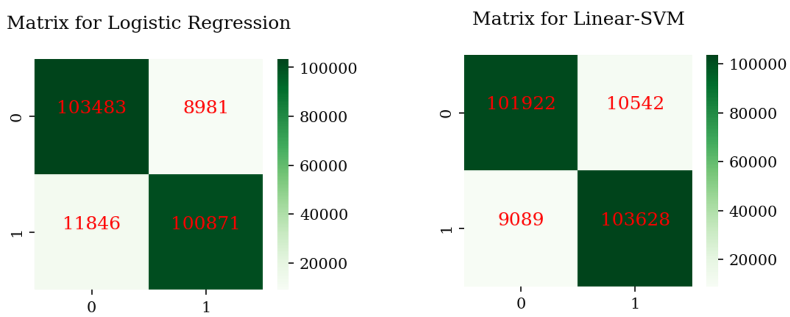

Figure 17. *Confusion matrices for Logistic Regression indicating 90.7% accuracy (left) and Linear SVM with 91.3% accuracy (right).*

4. Discussions on the important aspects of the project

The key objectives in this project were achieved. This premlimnary investigation seems to indicate that it is feasible to predict an individual's anxiety using physiological data. Algorithms such as Linear SVM and Logistic Regression seem effective for training models for prediction for social based anxiety (in a binary) way. Nevertheless, this premlimnary analysis only involved data collection from one participant, therefore it is still unclear whether it is feasible to create a more generalised model to predict social anxiety. In addition to this, the website platform clearly expressed some examples of the benefits of a system predicting social anxiety.

5. Avenues for future work and potential impact

5.1. Impact

5.1.1. Innovation, positive impact and scalability

This data-driven self-help platform is specifically innovative in a social anxiety context, as people with this anxiety often avoid seeking help from health services due to a fear of evaluation from healthcare professionals; therefore this platform offers an alternative and more comfortable method of treatment [15]. In addition to this, despite the potential benefits of a data-driven self-help platform and the scale of anxiety sufferers, no similar solution currently exists. There are also further benefits to the system, such as collection of data surrounding triggers of anxiety, which can be beneficial for psychology research.

5.1.2. Adverse impact

The project could also have a negative impact. There is scope for companies to harness these predictive models to identify vulnerable individuals; jeopardising the individual's 'emotional privacy'. In addition to this, the system handles sensitive data therefore it is important to consider the security of the underlying protocols (see '*avenues for future work'*). Furthermore, self-monitoring in this context may create an unhealthy obsession with one's personal health as well as revealing potential undiagnosed health issues and false positives.

5.2. Avenues for future work

Various steps need to be taken in order to improve the predictive model and the website dashboard, as well as creating an appropriate infrastructure for the system to exist.

5.2.1. Model Improvements

In order to improve the generalisability of the model, more data needs to be collected from a variation of participants. Moving forward, Neural Networks will also be explored as method for predicting anxiety, as they are seen to have a generally higher predictive accuracy. Feature extraction will be also be explored in order to improve the model's performance time and reduce dimensionality needed for the larger scale system.

5.2.2. Website Improvements

The website will be improved through testing from potential users. Expert interviews with healthcare professionals and therapists will also take place in order to ensure that the platform results in an improvement of anxiety.

5.2.3. Infrastructure

User's with social anxiety would be more likely to use this application if they were aware that the platform collects and interprets their data in a transparent way, specifically due to their fear of evaluation from others. A platform called 'Databox' could be used (once more developed) for managing the personal and sensitive data in a secure way. This would give the user greater control over who can access to their data, Figure 18. The Databox approach minimises the distribution of personal data and potential privacy threats compared to current cloud-based approaches, which often involves moving personal data to external locations for processing [19].

In addition to this, other areas need to be considered for the larger scale system. More advanced analysis such as the machine learning could take place in the cloud. Alongside this, a mobile app would accompany the website. This mobile app would collect GPS and calendar data and transfer this data wirelessly to the system, instead of using third party apps such as 'FollowMee'. If the user's wearable has a BLE component, the mobile app can also be used to collect the data packets to send wireless to the machine learning model in the cloud.

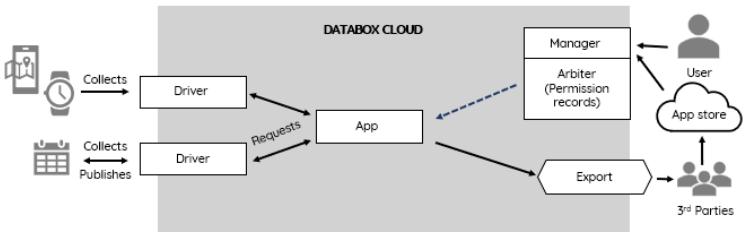

Figure 18. Databox example flow, allowing the user to manage how their personal data is distributed. Adapted from [19].

6. References

1. No Panic (2017). Anxiety Statistics - No Panic. [online] No Panic. Available at: https://nopanic.org.uk/anxiety-statistics/
2. The King's Fund. (2012). The King's Fund. [online] Available at: https://www.kingsfund.org.uk/projects/time-think-differently/trends-disease-and-disability-mental-physical-health
3. Peake, J.M., Kerr, G. and Sullivan, J.P. (2018). A Critical Review of Consumer Wearables, Mobile Applications, and Equipment for Providing Biofeedback, Monitoring Stress, and Sleep in Physically Active Populations. Frontiers in Physiology, [online] Available at: https://www.frontiersin.org/articles/10.3389/fphys.2018.00743/full.
4. Thompson, T., Kaminska, M., Marshall, C. and Van Zalk, N. (2019). Evaluation of the Social Phobia Scale and Social Interaction Anxiety Scale as assessments of performance... [online] ResearchGate. Available at: https://www.researchgate.net/publication/330796671_Evaluation_of_the_Social_Phobia_Scale_and_Social_Interaction_Anxiety_Scale_as_assessments_of_performance_and_interaction_anxiety.
5. Gramer, M. and Saria, K. (2007). Effects of social anxiety and evaluative threat on cardiovascular response to active performance situations. [online] ResearchGate. Available at: https://www.researchgate.net/publication/6838531_Effects_of_social_anxiety_and_evaluative_threat_on_cardiovascular_response_to_active_performance_situations
6. Apple (United Kingdom). (2020). Apple Watch Series 5. [online] Available at: https://www.apple.com/uk/apple-watch-series-5/health/
7. Sullivan, M. (2014). Fitbit announces new heartbeat-sensing fitness bracelet and watch. [online] VentureBeat. Available at: https://venturebeat.com/2014/10/27/fitbit-announces-new-fitness-bracelet-and-watch-with-heartbeat-sensors/ [Accessed 6 Jan. 2020].
8. Shimmersensing.com. (2020). Wearable GSR Sensor | Shimmer Galvanic Skin Response Sensor | EDA sensor. [online] Available at: https://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor [Accessed 6 Jan. 2020].
9. Empatica Support Center (2018). E4 wristband technical specifications. [online] Empatica Support. Available at: https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications
10. Qin, D. (2010). My Tracks. [online] Google.com. Available at: https://play.google.com/store/apps/details?id=com.zihua.android.mytracks&hl=en_GB
11. Posada-Quintero, H.F., Reljin, N., Mills, C., Mills, I., Florian, J.P., VanHeest, J.L. and Chon, K.H. (2018). Time-varying analysis of electrodermal activity during exercise. PLOS ONE, [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5983430/pdf/pone.0198328.pdf
12. Berkeley.edu. (2020). Nyquist sampling. [online] Available at: http://microscopy.berkeley.edu/courses/dib/sections/02Images/sampling.html
13. Wikipedia Contributors (2019). Bluetooth Low Energy. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Bluetooth_Low_Energy
14. Freepik. (2020c). Beautiful smiling model with afro curls hairstyle dressed in summer hipster white dress 6451399. [online] Available at: https://www.freepik.com/free-photo/beautiful-smiling-model-with-afro-curls-hairstyle-dressed-summer-hipster-white-dress_6451399.htm#page=1&query=female&position=10
15. Boukhechba, M., Chow, P., Fua, K., Teachman, B.A. and Barnes, L.E. (2018). Predicting Social Anxiety From Global Positioning System Traces of College Students: Feasibility Study. JMIR Mental Health, [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6053606/?report=classic
16. Reactjs.org. (2020). React – A JavaScript library for building user interfaces. [online] Available at: https://reactjs.org/
17. Sanjay.M (2018). Why and how to Cross Validate a Model? [online] Medium. Available at: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f
18. Sarang Narkhede (2018). Understanding Confusion Matrix. [online] Medium. Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
19. Bbc.co.uk. (2016). Databox - BBC R&D. [online] Available at: https://www.bbc.co.uk/rd/projects/databox