

UNIVERSITY OF VAASA

SCHOOL OF TECHNOLOGY AND INNOVATIONS

COMMUNICATION AND SYSTEMS ENGINEERING

Hilda Amo Henaku

**MOBILE SOLUTION FOR DATA MINING AND DECISION SUPPORT:
WEIGHT MONITORING AND EARLY PREDICTION OF CARDIAC ARREST**

Master's thesis for the degree of Master of Science in Technology submitted for assessment.

Vaasa, April 27, 2018.

Supervisor

Professor Mohammed Elmusrati

Instructor

Shaima Abdelmageed

ACKNOWLEDGEMENTS

I am very grateful to the Lord God Almighty for his blessings and grace upon me and for His guidance through my studies.

My profound appreciation goes to my supervisor, Professor Mohammed Elmusrati for his guidance and support throughout my masters programme and this thesis work. I would also like to express my deepest gratitude to my instructor Shaima Abdelmageed for her guidance, assistance and support throughout this thesis.

Special thanks go to my mum, Miss Patience Addo and my siblings; Priscilla Amo Henaku and Ernest Amo Henaku for their love and support throughout this period. I am very grateful for all the love and care you have shown me. I would also like to thank Yayra Asare, Emmanuel Tettey, Godwin Ashong, Edward Pie and all those who contributed to making my studies a success. May God bless you all and guide you in all you do.

TABLE OF CONTENT

ACKNOWLEDGEMENTS	2
TABLE OF CONTENT	3
LIST OF FIGURES	6
LIST OF TABLES	10
ABBREVIATIONS	11
ABSTRACT	12
1 INTRODUCTION	13
1.1 Background	13
1.2 Thesis Statement	15
1.3 Motivation	15
1.4 Methodology	16
1.5 Expectation	16
2 PRINCLIPLES OF WEIGHT MANAGEMENT AND DATA MINING	17
2.1 The significance of calorie counting in weight management	17
2.2 Interpreting Nutrition Fact Labels	19
2.3 Wearables for Activity tracking	20
2.4 Diabetes and Physical Activity	22
2.5 A brief description of the android mobile operating system	24
2.6 System design and architecture	25
2.7 Data mining and machine learning	27
2.7.1 Introduction to Logistic Regression	30
3 MOBILE APP DEVELOPMENT	33

3.1	Graphic User Interface and functionalities of the mobile app	35
3.1.1	User authentication (Firebase Authentication)	36
3.1.2	Database for app (Firebase Realtime database)	38
3.2	Calorie Counting	39
3.3	Per hundred-gram labelling against per-serving	40
3.4	Daily required calories calculation	43
3.5	Blood glucose interpretations	46
3.5.1	Wireless syncing of glucometers	48
3.6	Physical activity monitoring with Fitbit	49
4	LOGISTIC REGRESSION MODEL AND IMPLEMENTATION	51
4.1	What is classification in machine learning?	52
4.2	Building the predictive model with Python	52
4.2.1	Data preparation	53
4.2.2	Data exploration	56
4.3	Model implementation	58
4.3.1	Determining coefficients	59
4.3.2	Predicting test results and calculating accuracy	60
4.3.3	Performing cross validation	60
4.3.4	Confusion matrix	61
4.3.5	Precision, recall, F-measure and Support	61
4.3.6	Receiver Operating Characteristic (ROC) curve	62
4.4	Python – Firebase Connection	63
4.4.1	Conversion of JSON into Dataframe.	65
4.4.2	Prediction of cardiac arrest	67
4.5	Building the machine learning model with MATLAB	68

4.5.1	Description of the MATLAB toolbox.	68
4.6	Challenges with data acquisition	72
5	DISCUSSION AND FUTURE WORKS	74
6	CONCLUSION	76
	REFERENCES	77

LIST OF FIGURES

Figure 1. A sample NFL (National Health Service UK website, 2015).....	20
Figure 2. An illustration of the network architecture of wearable devices (Ching & Singh, 2016).....	22
Figure 3. A simple representation of the system architecture of this project	26
Figure 4. An example of a wearable data collection architecture.	27
Figure 5. The S-shaped logistic curve (Simonoff ,2018).	31
Figure 6. A flowchart of the algorithm of the mobile app - <i>my Calbuddy</i>	34
Figure 7. Splash screen for NFLs and fitness tracking.....	35
Figure 8. Splash screen for daily calorie needs.	36
Figure 9. A screenshot of the Firebase authentication platform.....	37
Figure 10. User interface for login options and email entry.....	37
Figure 11. User interface for creating an account.	38
Figure 12. A screenshot of my Calbuddy food label interpretation.	40
Figure 13. NFL with labelling showing amount of ingredients per 100 g and per serving.	41
Figure 14. A screenshot showing calorie calculations per serving.....	42
Figure 15. A screenshot showing calorie calculations per 100g.	43
Figure 16. Screenshots showing user interfaces for selecting option for calorie calculations.	44
Figure 17. Image showing results for Scenario 3.	45
Figure 18. Other images showing interfaces for other weight goals which includes datepickers for the selection of the time duration for goal.....	46
Figure 19. Display of a case of hypoglycaemia form Scenario 4.....	47

Figure 20. This also shows an incident of hyperglycemia where the user’s blood sugar is too high.	48
Figure 21. Login screens for Fitbit account and selection of needed information.	49
Figure 22. Display of physical activity information and account settings.	50
Figure 23. Commands for importing excel file and converting into a dataframe.	55
Figure 24. Command window showing model building and testing.	56
Figure 25. Bar chat of ages versus frequency of cardiac arrest.	56
Figure 26. Bar chat of gender versus frequency of cardiac arrest.	57
Figure 27. Bar Chat showing diabetes status against frequency of cardiac arrests.	58
Figure 28. Command window for implementation of logistic regression model.	58
Figure 29. Syntax for obtaining logistic regression coefficients.	59
Figure 30. Display of coefficients in console.	59
Figure 31. Display of model accuracy.	60
Figure 32. Implementation of cross validation and display of accuracy of cross validation.	60
Figure 33. Confusion matrix of the model.	61
Figure 34. Console showing precision, recall, F-measure and support values.	62
Figure 35. ROC curve implementation syntax and display.	63
Figure 36. Python codes for firebase database connection and imported JSON file from database after connection.	64
Figure 37. Syntax of retrieving user data and display of results.	64
Figure 38. This shows details of the user in scenario IV in the Firebase database.	65
Figure 39. Converting data from JSON format into dataframes.	65
Figure 40. Creating new dataframes with needed information.	66
Figure 41. Creating new columns and replacing missing data.	67

Figure 42. Syntax for display prediction.	67
Figure 43. Output and display of the dummy user's chances of getting a cardiac arrest.	68
Figure 44. MATLAB codes for assigning predictors and the target of the model.....	68
Figure 45. Model training MATLAB codes.....	69
Figure 46. Codes for implementing cross-validation.	69
Figure 47. Output showing the estimated coefficients obtained from the model.....	70
Figure 48. Prediction using dummy user with information in Table 4.	70
Figure 49. Confusion matrix of the model.	71
Figure 50. The ROC curve of the model using the MATLAB classifier app.	72
Figure 51. Screenshot showing accuracy of model using the machine learning classifier app.	72

LIST OF TABLES

Table 1 Activity levels and their corresponding activity multipliers	18
Table 2. Recommended BG levels pre prandial and post prandial	47
Table 3. Dataset for training and testing of the model.	54
Table 4. Dummy user information for testing of logistic regression model in MATLAB.	71

ABBREVIATIONS

ADL	Activities of Daily Living
API	Application Program Interface
AUC	Area Under Curve
BG	Blood Glucose
BMR	Body Mass Ratio
CSS3	Cascading Style Sheets 3
EHR	Electronic Health Record
HTML	Hyper Text Mark Up Language
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
KDD	Knowlegde Discovery in Database
NFC	Near Field Communication
NFL	Nutrition Fact Label
OCR	Optical Character Recognition
PA	Physical activity
PAL	Physical activity level
PAN	Public Area Network
ROC	Receiving Operating Characteristic
SCA	Sudden Cardiac Arrest
SDK	Software Development Kit
SQL	Structured Query Language
UI	User Interface

Faculty of technology

Author: Hilda Amo Henaku
Topic of the Thesis: Mobile solution for data mining and decision support: weight monitoring and early prediction of cardiac arrest
Supervisor: Professor Mohammed Elmusrati
Instructor: Shaima Abdelmageed
Degree: Master of Science in Technology
Major of Subject: Communication and Systems Engineering
Year of Entering the University: 2015
Year of Completing the Thesis: 2018

Pages: 83

ABSTRACT

The daily accumulation of data through various means has led to the popularity of data mining in recent times. Through the use of the data mining techniques, data that are collected are used for problem-solving and other purposes. In data mining, patterns and trends of large datasets are studied using computer-based techniques. This thesis is using an Android mobile application as a data sampling tool for data mining purposes. Using this application, a predictive machine learning model, was developed to predict the probability of occurrence of cardiac of arrest in users of a mobile app over a ten- year span. The designed mobile application also functions as a support tool for weight management and fitness. The mobile application was connected to a real-time database and a machine learning tool using a Python program to perform prediction. The machine learning was based on Logistic Regression that is one of the predominant models used in the healthcare sector for predictions. The system used the user's age, height, weight, activity level and diabetes status to predict the user's chances of getting a Sudden Cardiac Arrest (SCA) over a ten-year period. A detailed account of the implementation processes and principles are discussed throughout this work.

KEYWORDS: Android, BG, calories, data mining, Logistic Regression, machine learning, my Calbuddy, Python, SCA.

1 INTRODUCTION

1.1 Background

Due to the rapid growth of information technology and the daily use of systems, applications and software, data is being collected on daily basis across various fields and institutions. Besides data being collected and stored, novel techniques have been introduced to help make data collected into databases more useful for human development, and thus, the concept of data mining. Data mining involves the use of techniques to study large amount of sampled data to decide patterns and trends that yield the outcomes that are observed from the data (Morias et al., 2017). The idea behind this project is to use an android application as a datamining tool and a decision support tool for promoting fitness and wellbeing. A simple mobile application was developed for this project to purposely demonstrate how effectively android applications can be used for data mining purposes and was also developed with the aim of future enhancement and modifications for public use. This mobile application is named “*my Calbuddy*” and would be referred to as that throughout this report. *My Calbuddy* is an android application that provides a platform for its users to determine the total amount of calories on the food label of an item they are either about to consume or purchase, to calculate the daily required calories for losing, gaining or maintaining weight and for diabetes patients to interpret their blood sugar(glucose) level readings from their glucometers. A real-time database (Firebase Database) is used to store the information that is entered by the users of the application for data mining purposes.

Losing, maintaining or gaining weight may not just involve cutting down or increasing one’s food intake but may involve physical activity as well. Dieting however, to achieve any of these goals should be done healthily so that one does not get complications afterwards. As keeping blood glucose (BG) concentrations in check when exercising could be very strenuous for people with diabetes who are on insulin therapy (Zaharieva & Riddell, 2017), *my Calbuddy* uses its glucose interpretation interface to help reduce the stress involved in doing this.

Smart watches and wristbands have in recent times become trendy for weight management and healthy living purposes by tracking the physical activities of people who wear them. Some of these smartwatch brands include, Fitbit, Polar, Jawbone, Garmin, Samsung, and many others. For activity level measurements and accurate classification of individuals' activity levels, *my Calbuddy* provides a feature that allows its users to retrieve information on their daily physical activities by connecting their Fitbit watches or wristbands via Bluetooth to their mobile devices. With this function, users will be able to tell exactly how physically active they are which will also help them choose their correct activity levels for accurate answers of their daily required calories for weight gain, weight loss or weight maintenance.

Another important aspect of weight loss that is taken into consideration in this work is the calculation of calories on food labels as mentioned earlier. One of the methods that have been introduced to help prevent obesity is the labelling of calories on food items (Nikalauo, et.al., 2017). The reason for this to help users make the right choices when shopping and planning their meals.

As an engineer, one should have the attitude of problem solving and innovation and consider everything as an area of interest and a potential solution or a step to solving a problem. It is for this reason that this project focuses extensively on healthcare and wellbeing support by not just calculating calories and tracking physical activity but using an aspect of data mining and artificial intelligence known as machine learning to predict the likeliness of the occurrence of cardiac arrest in individual over a ten (10) year period using data of *my Calbuddy* users. Information of weight, height, gender, age, and diabetes states are variables for this prediction. The machine level programming is explored with two programming languages which are **Python** and **MATLAB** which are later discussed into details.

The continuing chapters of this work include the literature review in Chapter two which introduces the ideas, theories and the background of various aspects and factors considered for this project. It would include topics such as, the significance of calorie counting in weight management, Interpreting Nutrition Fact Labels, A brief description of the android software, logistic regression and others. Chapter three will provide information on the details and procedures involved in achieve the goals of this project. Chapter four will discuss the experimentation using Python and MATLAB, challenges

that were faced during the project and alternative approaches that could help make things easier. Chapter five follows with the discussion and future works and Chapter six will be the conclusion.

1.2 Thesis Statement

Using an android mobile application named *my Calbuddy* for data mining purposes. The main task is to predict the chance of getting a cardiac arrest over a ten-year span using machine learning for prediction and decision support and, also to calculate daily required calories, calories on food labels and interpret blood glucose (**BG**) level of diabetes patients.

1.3 Motivation

The quest of engineers to make life better and easier has led to the invention of varying forms of technology for various areas of life. These technologies tackle issues related to health, wellbeing, energy, power, finances and all relevant areas of our day to day lives. Problem solving, and innovation are traits that every inventor possesses which makes them the unique and creative people we see them to be. With the dream of being an inventor of a novel technology someday, I have realised that gaining these traits would go a long way in helping me to achieve my dream. During my bachelor's degree (2007 – 2011), I had a roommate who had diabetes and I saw how she battled with some issues especially those related to weight management and sustaining her blood glucose level. This was what led to the concept of my '*my Calbuddy*', an android application which focuses on not only on regular people but people with diabetes as well. As machine learning has become a very important subject of interest in recent times, the focus broadened to not just making an application, but also to use the data sampled from it to help solve another problem as well, and to serve as a decision support system for Healthcare facilities in the near future.

1.4 Methodology

My Calbuddy is installed on an android smartphone. A user enters information such as his or her weight, height, age, gender, activity level, diabetes status (this information is stored in a Firebase Realtime Database) and the Harris Benedict formula is used to calculate the his or her Body Mass Ratio (B.M.R.). The BMR value is then multiplied by the Physical Activity level (P.A.L.) factor of the user is to obtain the total daily required calories he or she needs. Daily calories for weight loss and gain are also calculated based on standard recommended guidelines. Calorie calculations on food labels are also done using the Atwater general system formula (Atwater & Benedict, 1902). A diabetes patient also uses a portal to enter his or her blood glucose readings from his glucometer which is interpreted based on the type of diabetes he or she has, and prandial button he chooses from the options available. The application tells the user if his blood sugar level is too high (hyperglycemia) at that moment, too low (hypoglycemia) or normal. Activity level is also tracked with a Fitbit wristband which is connected via Bluetooth to the user's mobile device. A model for predicting a user's likeliness of getting a cardiac arrest in the next 10 years is also done using Python and is demonstrated as well using MATLAB's machine learning toolbox.

1.5 Expectation

What we seek to achieve by the end of this project is to predict the probability of getting a cardiac arrest in the next ten years using a logistic regression (a machine learning technique). Predictors that will be used for building this model will be based on data that *my Calbuddy* users enter when using the app. This further discussed in explained in Chapter 4.

2 PRINCLIPLES OF WEIGHT MANAGEMENT AND DATA MINING

This chapters gives insight into the various principles and theories that are used for developing this project and discusses previous and current works that have been done in relation to this topic as well. It is also to help the reader get a better understanding of the all the principles and criteria used for implementing this work.

2.1 The significance of calorie counting in weight management

Keeping one's weight in check is very essential for a healthy lifestyle and for confidence building as well. In doing so, one may either want to lose, gain, or maintain his or her weight, which must however, be done cautiously to get the right results. The ideology of '*calories in calories out*' indicates that weight is either lost or gained based on how much kilocalories (also referred to as calories) is consumed and how much energy is dispensed (Riera-Crichton & Tefft, 2014). If calories consumed is more than energy dispensed, weight is lost and if it is less, weight is gained. There are various formulas that are used to calculate the daily caloric needs of individuals but, in this study, the Harris Benedict Equation is used to find the Basal Metabolic rate and multiplied by the Physical Activity Level multiplier to get the total daily energy expenditure or calories. The various Physical Activity levels and their corresponding multipliers are showed in Table 1 below. The Harris Benedict formula for both Men and women is given as:

Women: BMR

$$= 655 + (9.6 \times \text{weight in kg}) + (1.8 \times \text{height in cm}) \\ - (4.7 \times \text{age in years})$$

$$\mathbf{Men:} \text{ BMR} = 66 + (13.7 \times \text{weight in kg}) + (5 \times \text{height in cm}) \\ - (6.8 \times \text{age in years})$$

(Hsu., et al, 2018)

Where B.M.R. represents Body Mass Ratio.

Table 1. Activity levels and their corresponding activity multipliers

Activity Multipliers	
1.2	Sedentary; little or no exercise
1.375	Lightly Active; light exercise 1-3 days per week
1.55	Moderately Active; moderate exercise 3-5 days per week
1.725	Very Active; hard exercise 6-7 days per week
1.9	Extremely Active; hard daily exercise

(getphysical.co.uk, 2015).

In a research done by Johnson et.al in 2006, it was proposed that by undertaking restriction in calories through reducing consumed food quantities on some days and eating more than required on certain days increased the length of life and the quality of health in several species. Results from this experiment was observed as early as two weeks from the beginning of the experiment. They observed significant health benefits in cases of asthma, bacteria, fungal and viral infection, Insulin Resistance (I.R). There are several factors which influence the way we control and monitor our diet such as our emotions. Studies have showed that cognitive restraints are associated with a higher desire for food with low calories than food with higher calories. It is also said that there is a higher probability that people with binge-eating disorder and people on a strict diet will rely on food when they are unhappy (Racine, 2018). Several other factors that influence dietary control such memory on food quality, the repercussions of concentrating on the sensory properties of food, among others have also been investigated by Seguias & Tapper (2018).

Nutrition Fact Labels (NFL) furnish consumers with details about nutrients and calories in food items they purchase from grocery stores. NFLs have been made compulsory by various food regulatory institutions and in most countries and some continents. NFLs have benefits for consumers which include helping them to select healthy foods to help them eat healthily (Wolfson, et al., 2017). The visibility of labels is hence very important in aiding with decision making. In a study done in 2017 by Nikolaou, et.al., it was proposed that since about 60 to 70% of adults in Europe and the United States were overweight, increasing the font size of food labels by about 10 folds larger than the current

sizes of fonts on food labels could be helpful in combating obesity. According to a research done in Canada (McCrory, et al., 2016), it was observed that though most participants of the study used calorie information on NFLs when purchasing food items, very few knew the exact amount of calories they needed in order to help maintain their weight. There are several calorie counting apps currently available on the market that calculate calories and help with weight management and fitness. Some of these include, My fitnesspal, Lose it, FatSecret, Cron-o-meter and others which are also based on similar formulas and concepts as discussed earlier. There are also apps which have been made purposely for diabetes patients to monitor their Blood Glucose levels and manage their weight as well. Some of which are, Glucosio, BG Monitor, Glooko, Fooducate and others.

2.2 Interpreting Nutrition Fact Labels

The soft tissue framework of the body is made up mainly of protein, carbohydrates and fats which are referred to as Macronutrients (Campbell, 2017). Interpreting Nutrition Fact Labels (NFL) can be very stressful and confusing but there are key items to look out for on these labels which can help with proper decision making. These include the serving size, amount of calories, the Percent (%) Daily Value.

1. **Serving Size:** This provides information on the number of servings the package contains and is sometimes referred to as cups or sizes. It is important to note that nutrition on labels are provided per one serving of food.
2. **Amount of calories:** Calories are stated in one serving of food.
3. **Percent (%) Daily Value:** This gives information on how nutrients in one serving of food add up to one's entire daily diet. Based on this, one can select items based on which nutrients he or she needs more and which he or she need less (US Food and Drug Administration, 2015). Figure 1 is an example of an NFL which shows how nutrients and all other details are displayed of food items.

The Atwater general factor system created by W. Atwater, an American Chemist is used to assign values of energy to food items. It uses a factor for the major energy nutrients which are protein, carbohydrates, and fat (Kent, 2006). It is based on this system that most

food label calorie calculations are done. The formula used for the project to calculate macronutrients is stated below.

Total Calories(**Macronutrients**)

$$= (\text{grams of fat} \times 9) + (\text{grams of carbohydrates} \times 4) \\ + (\text{grams of protein} \times 4)$$

Nutrition				
Typical values	100g contains	Each slice (typically 44g) contains	% RI*	RI* for an average adult
Energy	985kJ 235kcal	435kJ 105kcal	5%	8400kJ 2000kcal
Fat	1.5g	0.7g	1%	70g
of which saturates	0.3g	0.1g	1%	20g
Carbohydrate	45.5g	20.0g		
of which sugars	3.8g	1.7g	2%	90g
Fibre	2.8g	1.2g		
Protein	7.7g	3.4g		
Salt	1.0g	0.4g	7%	6g

This pack contains 16 servings

*Reference intake of an average adult (8400kJ / 2000kcal)

Figure 1. A sample NFL (National Health Service UK website, 2015).

2.3 Wearables for Activity tracking

Engaging in more physical activity (P.A.) helps promote quality health and reduces the risk of one acquiring diabetes, cancer, heart diseases. It also improves mental health and reduces cognitive dysfunction as well (Philips, et al., 2018). This implies that, quality of life is improved when more physical activity is done. Usually, factors such as intensity, regularity of activity and the time taken to perform exercise or physical activity cause us to estimate our physical activity above what it usually is. Modern technological improvements that have been made to wearable activity monitors are helping us to estimate our physical activities correctly through readings of the number of steps we take per day, our heart rate and other parameters (Gresham, et al., 2018). Wrist band wearables

such as smart watches and fitness trackers are being used to for physical activity measurements in various research fields and are currently of a high demand on the market. It was recorded that, the annual sale of wearables in the United States in 2013, was 84 million which is implied high sales and demand ((Philips, et al., 2018). Wearables have been recognised by the European Parliament Scientific and Technology Options Assessment Panel (S.T.O.A) as one of the life changing inventions of recent times. Aside health, well-being and fitness purposes, there are future prospect of diagnosis based on data collected with these wearables to help reduce cost of accessing health facilities for treatment (European Commission, Directorate-General for Communications Networks, Content and Technology, 2016). There are currently several brands of wristband wearables and mobile applications for tracking heart rate, sleep, movement, and other factors associated with health and well-being. Some of these wristbands include, Fitbit, Apple health, Google fit, S-health, Jawbone, Polar, Microsoft health, etc. De Arriba-Pérez et al (2016), investigated how data can be gathered and integrated by all manufacturers of these wearables where they identified key issues related to harmonising all data models and transfer modes of these wearables and proposed that all vendors should standardize their data models and transfer modes. Hwang & Lee (2017), conducted a research to determine the physical demands of construction workers using wristband wearable devices. Measures of the Heart rate reserves (%HRR) were taken for this research and it was observed that physical demands changed due to the work patterns of these construction workers and other factors such as age and heart stress. Pevnick et al (2018) also discussed the future and current status of wearables in the field of cardiology. They highlighted the significance of wearables in improving healthcare and as a good source of data for various forms of health studies. Glyn et al (2014) also determined how efficiently a smartphone app could be use in primary care to encourage physical activity. Pedometers have been known as tools for measuring and encouraging indulgence in physical activity. (Åkerberg, et al., 2012). Aside pedometers, other forms of devices for activity monitoring and tracking are accelerometers and Integrated multisensory systems. Pedometers evaluate how many steps are taken along the vertical plane via either digital or mechanical readings. Accelerometers discover acceleration in either three directions, one direction or even two directions so that the regularity, amount of movement and magnitude of motion can be obtained. Integrated multisensory systems, however use

accelerometry alongside other sensors to measure the reaction or feedback of the body such as heart rate during exercise to help improve the evaluation of physical activity (Gresham, et al., 2018). These activity tracking tools are what are incorporated into the wearables and physical activity tracking mobile application that are currently available. The network connectivity and framework of wearables is illustrated in Figure 2 below.

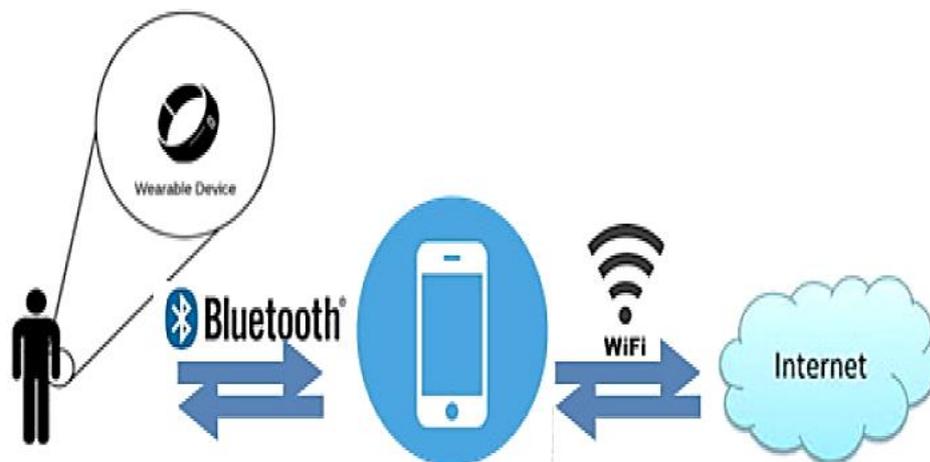


Figure 2. An illustration of the network architecture of wearable devices (Ching & Singh, 2016).

2.4 Diabetes and Physical Activity

Diagnosis for diabetes is done when BG is greater than or equal to 126 mg\ dL (7 mmol/L) when fasting or when BG is greater than 200 mg\ dL (11 mmol/L) post prandial or any at time instant or when haemoglobin A1C (HbA1c) is greater than 6.5 percent (Jain, et al., 2017). When the blood sugar level (referred to as **BG** in this work) is too low for the body to produce sufficient energy for its functions, **hypoglycaemia** occurs. People with BG levels below 70 mg\ dL (3.9 mmol/L) are usually said to have low blood sugar. **Hypoglycemia** is a very critical condition which can affect the standard of life of patient and hence requires regular monitoring of BG (Khunti, et.al., 2017). Symptoms of hypoglycaemia include, difficulty in speaking, dizziness, nervousness, sweating,

shakiness, confusion and can even lead to mortality in adverse cases (NorthWest Memorial Hospital, 2013). Measures taken to control this are insulin dose reduction, performing less physical activity and increasing calorie consumption. Doctors should however be seen for advice before taking these steps (Khunti, et.al., 2017). On the contrary, **Hyperglycemia** refers to the situation where BG is high and is mostly predominant in people with diabetes Type 2. Measures to control **hyperglycemia** include increasing physical activity and regulating calories (Northwestern Memorial Healthcare, 2016).

Though exercising regularly comes with several health benefits, people with diabetes face challenges with managing their blood glucose levels during exercise which becomes even more challenging for patients on insulin therapy or on some sort of hypoglycemic agents. Hypoglycemia is the most threatening effect of exercising for diabetes patients. Despite that, it can be controlled with the right insulin dosage and alterations and with carbohydrate supplements in some scenarios. Hyperglycemia may however be experienced as well, after vigorous exercising since insulin sensitivity could be heightened due to severe reaction to stress. (Zaharieva & Ridell, 2017).

People with diabetes must put into consideration the above-mentioned conditions (which are Hyperglycemia and Hypoclycemia) and other factors before engaging in physical activity, and as such, diabetes patients who take regular injections and insulin infusions should be very careful when exercising (Zaharieva & Ridell, 2017). These other factors to be considered may include, reasons for exercise, timing of physical activity, insulin sensitivity, how long the exercise may take and others (Siomos, et al., 2017). Due to these challenges, people with diabetes are usually not motivated to exercise. Jenkins and Jenks (2017), proposed that medical practitioners should play a part in motivating diabetes patients to engage in physical activity. They also suggested that exercise prescription and other related subject to diabetes, be added to the medical school curriculum. Some physical activity guidance on how physical activity could be prescribed for diabetes patients was also discussed by Siomos, et al. (2017) who also suggested that aside taking notes and observing tailored prescriptions for patients, nurse practitioners should also engage patients in motivational interviews to help motivate them and give them enough audience when having sessions with them. They are suggested that nurse practitioners

should try to understand what the limitations and fears of the patients individually, and how willing they are to physical activity into improve with their health.

2.5 A brief description of the android mobile operating system

Android is a Linux based open source operating system for mobile devices and other platforms (Android TV and android Auto). Android is fast evolving as; novel functionalities and abilities are introduced to its users and developers from time to time. It equips developers with what is needed for creating applications and utilizes hardware abilities of all devices. Android user interfaces adjust to suit each device and can be managed in various devices.

An institution for developing software called Android Inc. was formed in October 2003 in California in the United States of America. The founders were Andy Rubin, Nick Sears, Rich Miner and Chris White. Their initial plan was to develop a Linux based operating system which could work on digital cameras that are able to connect to computers. Two years later, Android Inc. was bought by Google (August 2005) and in November 2007, Google unveiled an alliance for mobile technology companies and investors called the Open Handset Alliance Consortium (OHAC). There was however, no handset release under OHAC until HTC released its first smartphone in October 2008. That was when the first version of android which had just 35 apps was disclosed. Android has introduced many versions over time, some of which are cupcake, lollipop, donut, froyo, Éclair, Ice cream sandwich and others (Perera, et al., 2017).

The Android Software Development Kit (SDK) contains libraries that are used for the creating android applications and has tools for developing applications, executing them and testing them as well. Integrated development environments (IDEs) are user interfaces used to make software development easier. Android studio which is the official IDE for the development of android applications can be downloaded and installed for android software development (Barry, 2015). Android studio was the platform used for coding and developing *my Calbuddy*.

Android has some very important features and components that are discussed below:

User interface (UI): Android has already built UI components which are managed by the UI to develop GUIs (Graphic User Interfaces) for apps.

Connectivity: Aside supporting WANs (Wide Area Networks) like 3G, CDMA, 4G and GSM, Ethernet, Wifi, LAN technologies, PAN (Bluetooth), current versions of android also support Near Field Communication(NFC).

Storage: There vast alternatives for saving data can be selected based on whether data should be accessed by other applications or not and the space needed for storage. These storage options include SQLite database, Network connection, cloud storage, shared preferences, Internal and external storage.

Media Support: Some supported media include MPS, MDI for audio, GIF, PNG, BMP, MPEG-4 SP for audio and JPEG for images.

Messaging: GCM (Google Cloud Messaging) is also form of Push messaging for Android aside the most common ones which are, SMS (Short Message Service), Multimedia Messaging Service (MMS).

Web browser: The web browser for android supports Cascading Style Sheets (CSS3) and Hyper Text Mark up Languages (HTMLs).

Android beam: It is used to transfer photos, URLs and other forms of data form one device to another via Bluetooth or NFC.

Other features of android include multi-tasking, multi touch and multi-languages (Perera, et al., 2017).

2.6 System design and architecture

The system is made of connections that form a complete circle from the app to the prediction. The app is connected to the real-time database and then to the machine learning system. Connections include, user device connection to the internet (via a cellular network), connection of *my Calbuddy* to the Firebase Realtime Database and the Firebase Authentication console (via internet), connection of the Fitbit wristband to the android device through Bluetooth (a Public Area Network, PAN) and connection of the Realtime Database to machine level prediction algorithm (Python Environment). Figure 3 is a diagrammatic representation of the entire system architecture described above.



Figure 3. A simple representation of the system architecture of this project

Telecommunication companies and Internet Service Providers (ISPs) are usually the main suppliers of mobile internet services for devices. There are however factors that can interrupt the smooth transfer of data packets between mobile devices and transmission nodes of these mobile networks. Some of the problems that can occur during transmission may be due to poor network coverage, network congestion as well as security issues. These problems can cause delay in the synchronisation of the user's information to Firebase and may cause issues with authentication, and delay in information transfer onto the Firebase Realtime Database.

Data transfer from wearables to databases is either done through direct or indirect access. In indirect access, third parties access data by using an app an intermediary to either collect data from a wearable or send data to a data warehouse while in indirect access, a data is retrieved using a warehouse REST API to a third-party server. A warehouse REST API provided by Fitbit for third party systems allows data from Fitbit wristbands to be obtained through the REST API. In spite of this, developers must be authorized before they can get access to data. Successful transfer of Data from wearables from the devices to the data warehouses are also dependent on the quality of the mobile network and may face the same challenges as pointed out. Figure 4 shows an example of a data collection framework of a wearable device.

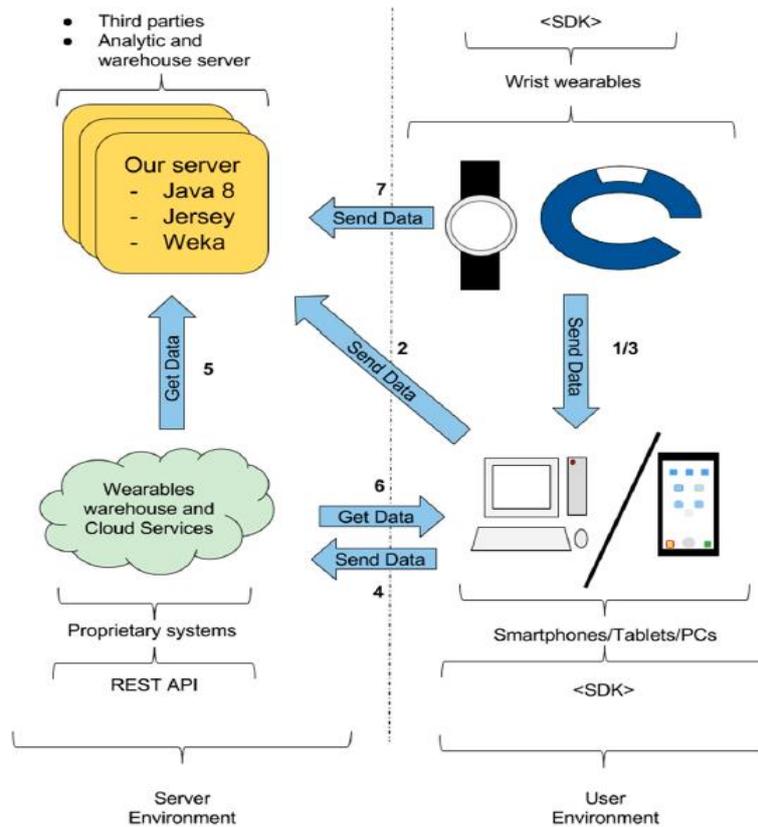


Figure 4. An example of a wearable data collection architecture.

2.7 Data mining and machine learning

Data mining is a very key aspect of Knowledge discovery in database (KDD) which is being greatly explored lately. In recent times, there have been the introduction of several data mining techniques for extraction and discovery of patterns since there is a large amount of information for use in various sectors and industries (Neesha, et al., 2015). There is therefore a need to use these collected data to solve problems to help make life better. With the use of computer-based Information Systems, hidden patterns and trends can be identified from bulk amount of data using techniques in a process called Data mining (Morias, et al., 2017).

Data mining consists of several areas which are, statistics, probability, artificial intelligence and machine learning. Data mining models can be categorised into two main groups; predictive and descriptive models. In the health sector, predictive models are

usually utilised for research and decision support (Neesha, et al., 2015). Amutha et al. (2018), used a mobile application to predict outcomes of a Treadmill test using data mining algorithms. Age, sex, BMI, diabetes, Dyslipidemia, and Systematic hypertension were used as the predictors for the machine learning algorithms that were used. During the research they realised that their model yielded low accuracy when the ages were more than 60, which implied that there were other parameters that should be taken into consideration when future works are done on the subject. A model for predicting the chances of death in females with ST- elevation myocardial infarction was also designed by Mansoor et al. In the work of Mansoor et al. (2017), logistic regression, a data mining model was compared with another model called the Random rain forest where, the former outperformed the later. Kavakiotis et al. (2017) also did a study on methods used for machine learning and data mining for people with diabetes. They discussed major papers and findings in relation to the subject of research which included healthcare, diagnosis, environmental factors, amongst others. An appropriate data mining model for predicting how essential resuscitation of new born babies was, was investigated through the trial of several models by Morais et al. (2017). who finally got the best model from a case study that used the K-Nearest Neighbour technique, cross validation and data with oversampling. All the examples prove that a lot of work is being done with data mining in the health sector.

The discipline of data mining that is of much interest as far as this project is concerned is machine learning. Machine learning is a field of study that enables machines to “learn” from historic data that they are fed with to improve experience. Machine learning algorithms can be classified into three main groups namely; supervised learning, unsupervised learning, and reinforcement learning. Machine learning is an aspect of artificial intelligence but is sometimes used interchangeably (Kavakiotis, et al, 2017).

In supervised learning, predictors are used to estimate an outcome. Examples are Decision tree, Random forest, Logistic regression, KNN (K-nearest neighbour). On the contrary, in unsupervised learning there is no outcome to be predicted. Unsupervised learning is used for grouping population into clusters. Examples of unsupervised learning are; K-means and Apriori algorithm. However, in reinforcement learning, the machine learns from experience through trials to make explicit decisions based on previous experiences (Ray, 2017). Machine learning is being used for prediction in the healthcare sector,

business sector and even in criminal investigations. Amrit et al., identified that it was feasible to use machine learning classification methods to detect child abuse using free-text or structured data and therefore proposed a decision support system API and its implementation for this purpose. Support vector machine and Random forest models were both used for this study. Forsyth et al, used machine learning to investigate symptoms that occur in breast cancer patients that undergo chemotherapy. Data used for this research was retrieved from an electronic health record and trained with a model to identify words that indicated the presence of a symptom, absence of a symptom or no identified symptoms at all. 10,000 sentences were labelled manually and divided into training, validation and test datasets. The performance of the model was finally tested using data that had about two percent of data that did not exist in the training dataset. The most prevalent symptoms were, nausea, fatigue and pain. Bzdog & Meyer-Lindenberg (2017), in their article, discussed the pros and cons of incorporating machine learning into precision psychiatry. The conventional way for making decisions in psychiatry for a while now has been through classical psychiatry which is based on few datasets, Bzdod & Meyer- Lindenberg (2017), suggested that machine learning could be used in the area of psychiatry since they successfully used a large amount of data was to identify trends and patterns in machine learning. Lynch et al. (2017), also investigated how prediction could be made on the survival of people suffering from lung cancer using supervised machine learning classification algorithms. According to their findings, they stated that predicting the survival of lung cancer patients was possible with the various types of supervised classification methods except for decision tree which was not feasible. Based on data collected when stroke patients were being admitted, the Bathel index status was correctly classified using Logistic regression, Support Vector Machine (SMV) and Random Forest. The Bathel scores of these patients were also correctly classified using, Logistic regression, Support Vector Machine (SMV). This machine learning built model also correctly predicted the Activities of daily living (ADL) for patients recovering from stroke. It was therefore concluded that this approach of using Machine learning for ADL prediction was very pragmatic and applicable (Lin, et al., 2018). Psychological ways of suppressing pain in the treatment of sickle cell disease was explored by Yang, et al. (2018) too. Classification methods were used in this approach which included Multinomial logistic regression, K-Nearest Neighbors, Support Vector Machine and Random Forest.

2.7.1 Introduction to Logistic Regression

Logistic regression is one of the major multivariate analysis methods used in the healthcare sector. Multivariate analysis involves the prediction of just one outcome variable with a number of variables (more than one). It also investigates the correlation between the outcome variable which is the dependent variable and the predictors which are the independent variables. This correlation is described by a model which uses an expression in which the outcome is expressed as a product of the coefficient of the independent variables and their values. Each coefficient represents how each independent variable affects the outcome variable, and hence regulating the effect of the other independent variable. The outcome of logistic regression is usually binary. Aside logistic regression, other multivariate methods used in the healthcare sector are linear regression, discriminant analysis and proportional hazard regression (Park, 2013).

Logistic regression is one of the generalized linear models where $y_i \sim \text{Binomial}(1, p_i)$, with the model satisfying p_i . Compared to other methods, logistic regression has more benefits and due to its efficiency especially on small datasets, it is also referred to as the canonical link function (Simonoff, 2018). Studies and research have been done using logistic regression for prediction in various fields of study. Huang et al. (2017), used logistic regression to examine the loss of grain during harvest. Jabeur (2017), also used Partial Least Squares Logistic Regression to create a model to forecast bankruptcy in a French company. Elio et al., (2017) did a study on areas liable to be affected by Radon in Ireland. They used indoor Radon readings sampled by the Environmental Protection Agency and four additional dataset which yielded a better spatial resolution for the Indoor Radon map that already existed. Wu et al (2018), used data mining techniques which included logistic regression to build a model to predict Diabetes Mellitus Type 2 using 8 predictors from a dataset and an outcome, which was the diabetes status of the patient. The categories were classified as positive and negative classes. Using logistic regression, Eström, et al. (2018), utilised data from a program that detected certain relevant environmental issues to investigate issues like climate change and factors that caused climate change. Tripoliti, et al. (2017) also discussed machine learning techniques which

included logistic regression, for detecting heart failure, the intensity of heart failure and the likely occurrence of catastrophic situations.

It is very important to comprehend the coefficients that are evaluated in any fitted model. This can be achieved by identifying the correlation between dependent and independent variables and suitably determining the dimension with which the independent variable changes. The extent to which the logit is altered in relation to the how much alteration is done in one unit of the independent variable is represented by the coefficient of the slope in logistic regression (Hosmer, Lemeshow & Sturdivant, 2013). The logistic curve is used for fitting data to predict the probability that an event will occur. Logistic model is more known due to the s-shaped curve (shown in Figure 5) and the logistic function. Logistic regression can however be subcategorised into binomial and multinomial logistic regression. Binomial logistic regression is used when the independent variable is either categorical or continuous and the dependent variable has two categories while multinomial logistic regression is used when the dependent variable has more than one category (Park, 2013).

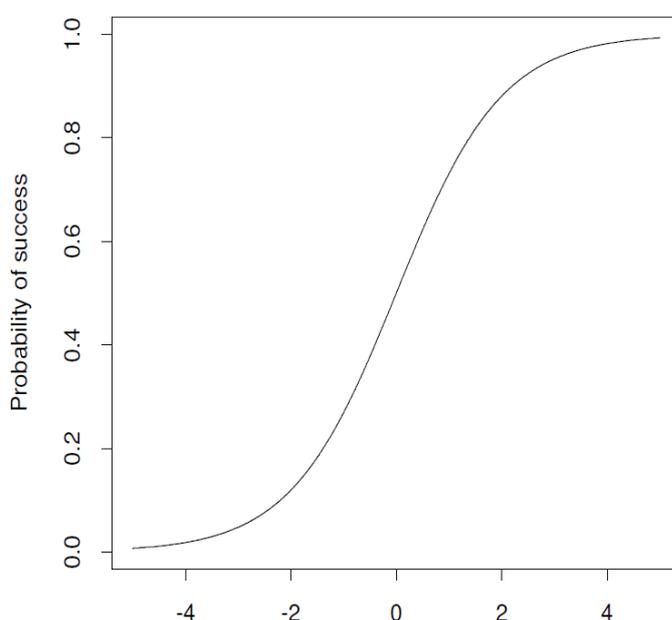


Figure 5.The S-shaped logistic curve (Simonoff ,2018).

In logistic regression, the response variable has an average which can be expressed in terms of x as, $p = a + \beta x$. This yields values that can fall out of the 0 to 1 range and for this reason, the natural log of the odds is found.

$$\text{Logit}(y) = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = a + \beta x$$

Where p is the probability of the outcome which is of interest, x is the interpretive variable and α, β , the variables of the logistic regression.

When the antilog of the predictive equation it gives:

$$\begin{aligned} P &= P(Y = \text{interested outcome}/X = x, \text{ a specific value}) = \frac{e^{a+\beta x}}{1 + e^{a+\beta x}} \\ &= \frac{1}{1 + e^{-(a+\beta x)}} \end{aligned}$$

In cases where the predictors are more, we get

$$\text{Logit}(y) = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = a + \beta_1 X_1 + \dots + \beta_n X_n$$

Hence

$$\begin{aligned} P &= P(Y = \text{interested outcome}/X = x_1, x_2, \dots, x_n, \text{ a specific value}) = \\ &= \frac{e^{a+\beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{a+\beta_1 X_1 + \dots + \beta_n X_n}} = \frac{1}{1 + e^{-(a+\beta_1 X_1 + \dots + \beta_n X_n)}} \end{aligned}$$

(Park, 2013).

The odd ratio (OR) is the ratio of one odd to the another odd, the OR of an increase in x can be expressed as e^{β_1} . If $OR = 1$, the outcome is not affected by the other odd, that is, the exposure and if the $OR > 1$, the exposure is associated with a greater odd outcome and vice versa when $OR < 1$.

In this project, the prediction of cardiac arrest is done which yield a categorical outcome of 'Yes' as 1 and 'No' as 0. Details of this is discussed in Chapter 4.

3 MOBILE APP DEVELOPMENT

This chapter gives a detailed description of *my Calbuddy*, the functions it performs and the features it presents to its users. It can be said that, how well a software product does, depends on the people that use it (Blackwell, A., 2010). It is based on this notion that the GUI (Graphic User Interface) of *my Calbuddy* is created is hence very user friendly and simple for all its users. The main functions of *my Calbuddy* are calculating and determining the right amount of calories needed for individuals to either lose, gain or maintain weight, calculating the total amount of calories on the labels of food items user's want to either buy or consume, receiving information on the daily physical activities through a Fitbit wearable device, interpreting of BG readings for diabetes patients so they can take the necessary steps to keep their BG at the right levels during physical activity sessions or at any random time. A flow chart of the sequence of functions of *my Calbuddy* is shown in detail in the Figure 6.



Figure 6. A flowchart of the algorithm of the mobile app - *my Calbuddy*.

3.1 Graphic User Interface and functionalities of the mobile app

My Calbuddy first introduces its users to the its functionalities using a set of splash screens. After these introductory screens (which can be skipped), the user creates an account either by using his or her email address, or phone number which is authenticated using Firebase. *My Calbuddy's* introductory screens (splash screens) are shown in Figures 7 and 8.

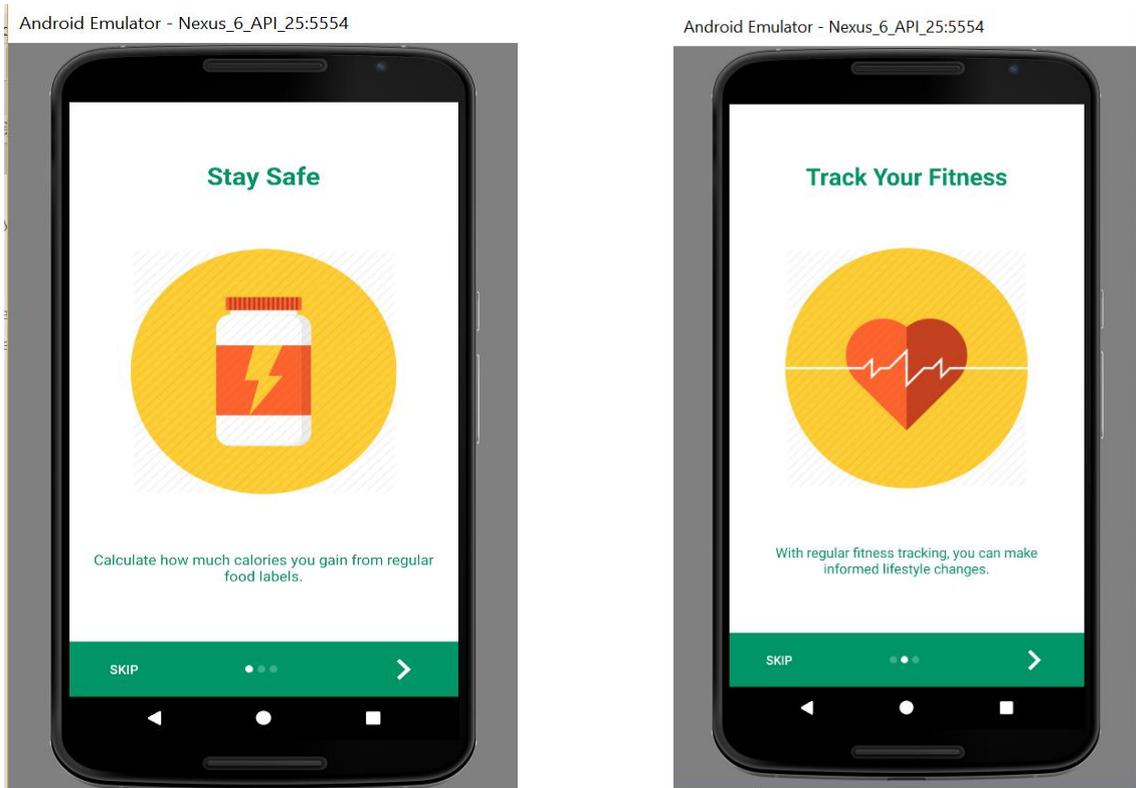


Figure 7. Splash screen for NFLs and fitness tracking.

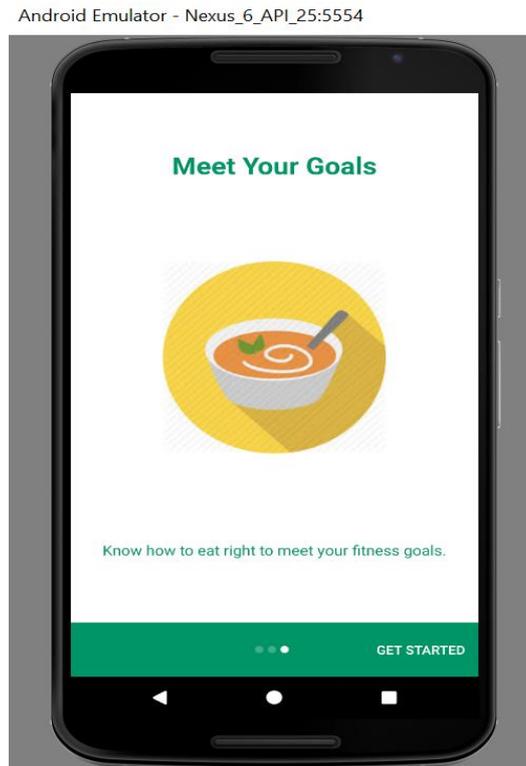


Figure 8. Splash screen for daily calorie needs.

3.1.1 User authentication (Firebase Authentication)

Firestore provides SDKs and libraries that enable android app developers to identify their users and also save their data in a cloud. Authentication with Firestore can be done using one's email, phone number, social media accounts such as twitter, Facebook, and others (Firestore website, 2018). Figure 9 shows what the Firestore authentication platform looks like. Preferred authentication methods the app developer wants can be chosen from the Firestore console and enabled using the using the required procedures. The app authenticates its users using email, and phone. These are demonstrated in Figures 10 and 11.

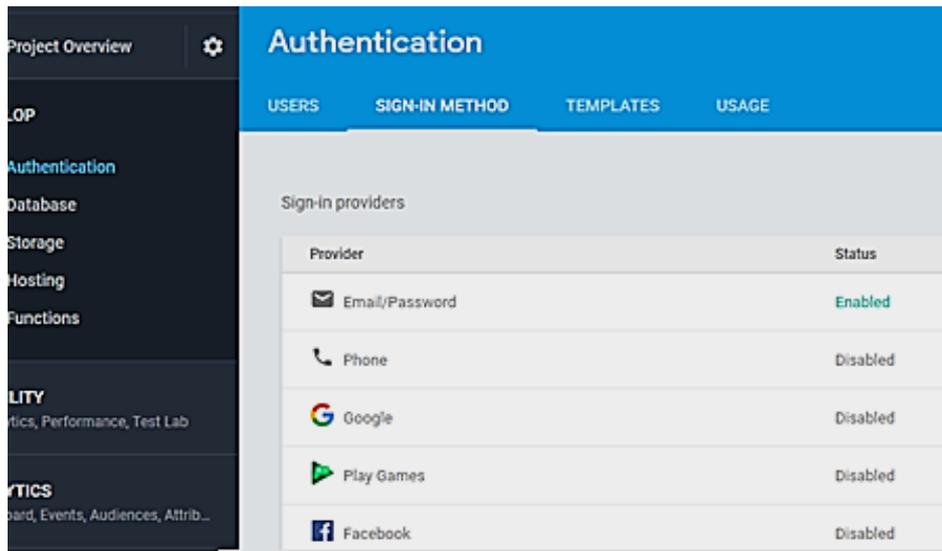


Figure 9. A screenshot of the Firebase authentication platform.

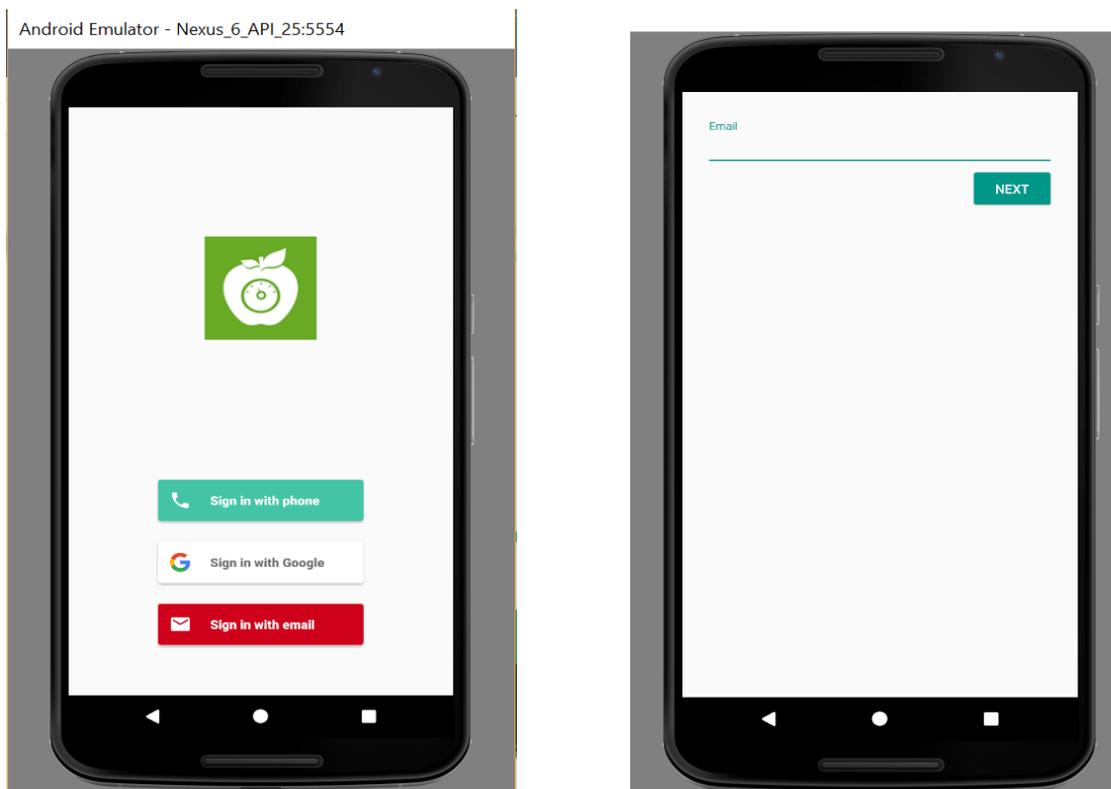


Figure 10. User interface for login options and email entry.

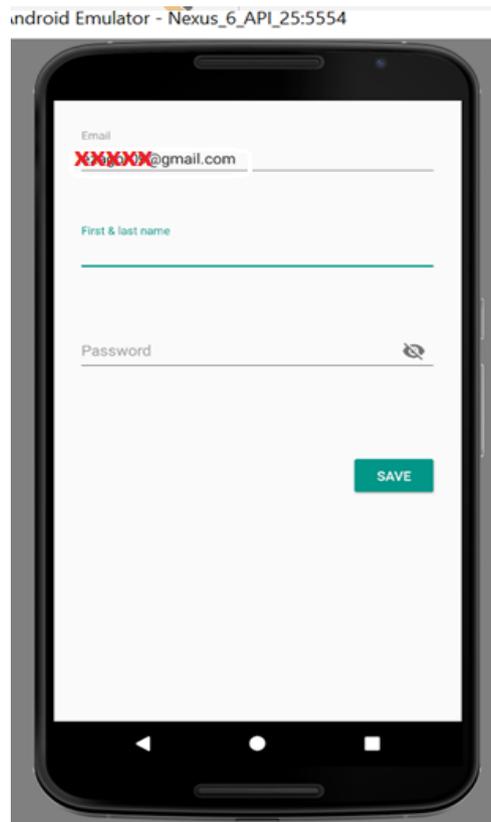


Figure 11. User interface for creating an account.

3.1.2 Database for app (Firebase Realtime database)

Firebase Realtime database is a database hosted on a cloud. It does not require SQL queries or any complicated set up process like building a database from scratch. Data of users are synchronised in real-time via internet and is stored as JSON. Data can be accessed at all times, not only when the user is online. Security policies however apply to the use of Firebase which must be adhered by the database administrators who in this context is the app owner or developer. Databases are very important for all applications that store bulk data (Lee, 2012). For this reason and for the purposes of data mining, integrating this application with a database was very important and critical in achieving our goal.

3.2 Calorie Counting

The application calculates the calories of macronutrients on the food label using the Atwater formula quoted above. This to help users make choices on which products to pick in the grocery store based on their preferences and to enable them to know how much they may be consuming. The initial aim of this was to create a way for non-Finnish speakers to interpret food labels. This is a functionality that may be developed properly for users in the near future. Reducing the intake of carbohydrates has been known to control diabetes, although there have been several debates on low carbohydrate diets (Fienman, et al., 2015). Since total carbohydrates printed on food labels include all types of carbohydrates and not just sugars, it is very important for diabetes patients to take note of the amount of carbohydrates they are consuming. Colours are known to affect our emotions, intentions, and behavioural patterns (Bonnardel, et al., 2011). It is for this reason that *my Calbuddy* uses a different text colour for the inputted text of the total amount of carbohydrate on food labels so that it will catch the attention of diabetes patients (this is illustrated in Figure12). In order to assist non-Finnish speaker to identify ingredients in Finnish, the Finnish names of ingredients are also given by the app.

Android Emulator - Nexus_6_API_25:5554

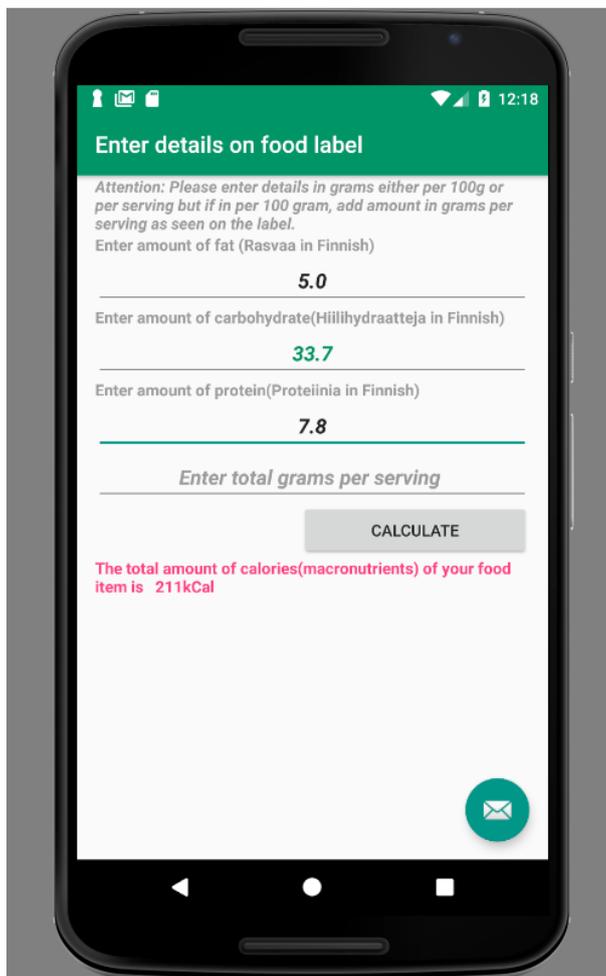


Figure 12. A screenshot of my Calbuddy food label interpretation.

3.3 Per hundred-gram labelling against per-serving

Provision is made the app for the two labelling types that are found on food items. Items labelled in per hundred gram give the grams of ingredients in every hundred gram of the food item whereas items labelling per serving straight away give the exact amount of ingredients per serving. Figure 13 shows a Nutrition Fact Label with labelling done per 100g of nutrients and per serving size of each nutrient. What should however, be noted by consumers is that if one eats more that the portion or serving size, he or she is to multiply the size per serving on the food label by the number of portions he or she consumed. For instance, if 5 pieces or biscuit make up one serving size of the food item

and an individual eats 10 pieces, he or she has taken 2 portions at a serving and should therefore multiple the total number of calories per serving by 2.

Typical values	100g contains	each slice (typically 44g) contains	% RI*	RI* for an average adult
Energy	985kJ 235kcal	435kJ 105kcal	5%	8400kJ 2000kcal
Fat	1.5g	0.7g	1%	70g
of which saturates	0.3g	0.1g	1%	20g
Carbohydrate	45.5g	20.0g	2%	90g
of which sugars	3.8g	1.7g		
Fibre	2.8g	1.2g		
Protein	7.7g	3.4g		
Salt	1.0g	0.4g	7%	6g

This pack contains 16 servings
*Reference intake of an average adult (8400kJ / 2000kcal)

Figure 13. NFL with labelling showing amount of ingredients per 100 g and per serving.

I. Scenario 1

A straight forward calculation of calories in food items is done in this scenario. The total calories (macronutrients) in an item which has **5g** of fat, **33.7g** of carbohydrates and **7.8g** of protein per serving is done using the app. The total calories is obtained as **211 kCal** which is showed in Figure 14 below.

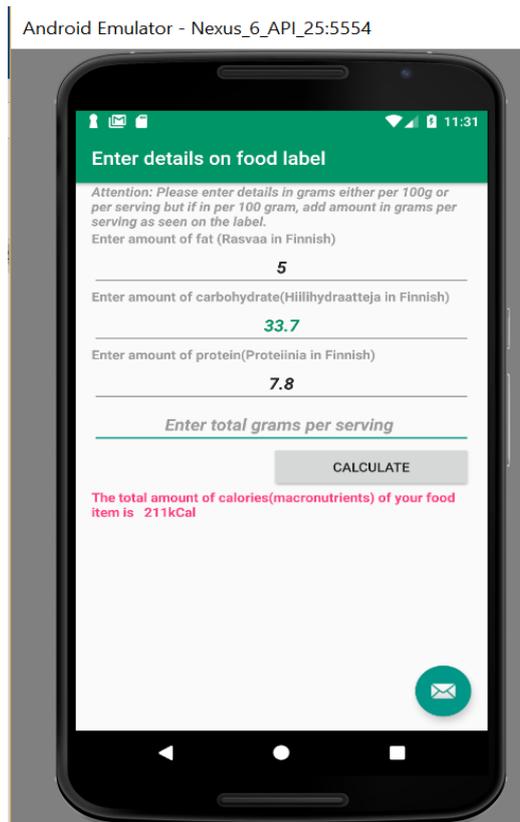


Figure 14. A screenshot showing calorie calculations per serving.

II. Scenario 2

This aim of this scenario is to show accuracy of food labelling calculations done by *my Calbuddy*. The information used are from the same food label used in scenario 2. To get the total number of calories when labelling is done per 100g, total calories of per 100g is divided by 100 and multiplied by the total grams of the food item per serving/ portion. The food label contained **10.0 g** of fat, **67.3g** of carbohydrates and **15.5g** protein, all per 100g and its the grams per portion was **50g**. The total number of calories calculated is **211 kCal** establishing the fact that both labelling types yield the same results. Figure 15 show the app's user interface and results of the above scenario.

Enter details on food label

Attention: Please enter details in grams either per 100g or per serving but if in per 100 gram, add amount in grams per serving as seen on the label.

Enter amount of fat (Rasvaa in Finnish)

10.0

Enter amount of carbohydrate(Hiilihydraatteja in Finnish)

67.3

Enter amount of protein(Proteiinia in Finnish)

15.5

50

CALCULATE

The total amount of calories(macronutrients) of your food item is 211kCal

Figure 15. A screenshot showing calorie calculations per 100g.

3.4 Daily required calories calculation

The daily required calories for an individual is obtain using the BMR and activity multiplier as explained in Chapter 2. Estimations of calories needed to gain and maintain weight are done using guidelines from the National Institute of Health and Care Excellence (NICE) in the UK which are stated on the Nation Health Service website of the UK. A total amount of 2500 kCal is estimated per day for average men and 2000k Cal for average women. These however, vary based on the height, weight, age and activity level individuals. The ideal calorie allowance for weight gain per day is 250 calories for men and 125 calories for women. Medical advice should sort first before one gains weight, that is when he or she is underweight and has a BMI (Body Mass Index less done 1.8). The recommended safe range weight loss is also between 0.5 kg to 1 kg per week. It based on all these principles that calculations are made for both weight maintenance, loss and gain by the app.

III. Scenario 3

In this scenario, a **male** user who is **19** years old, **172** cm tall and **moderately active** tries find out how much to consume to on a daily basis to keep his weight as it is. The user interfaces for performing this task are shown in Figures 16, 17, 18.

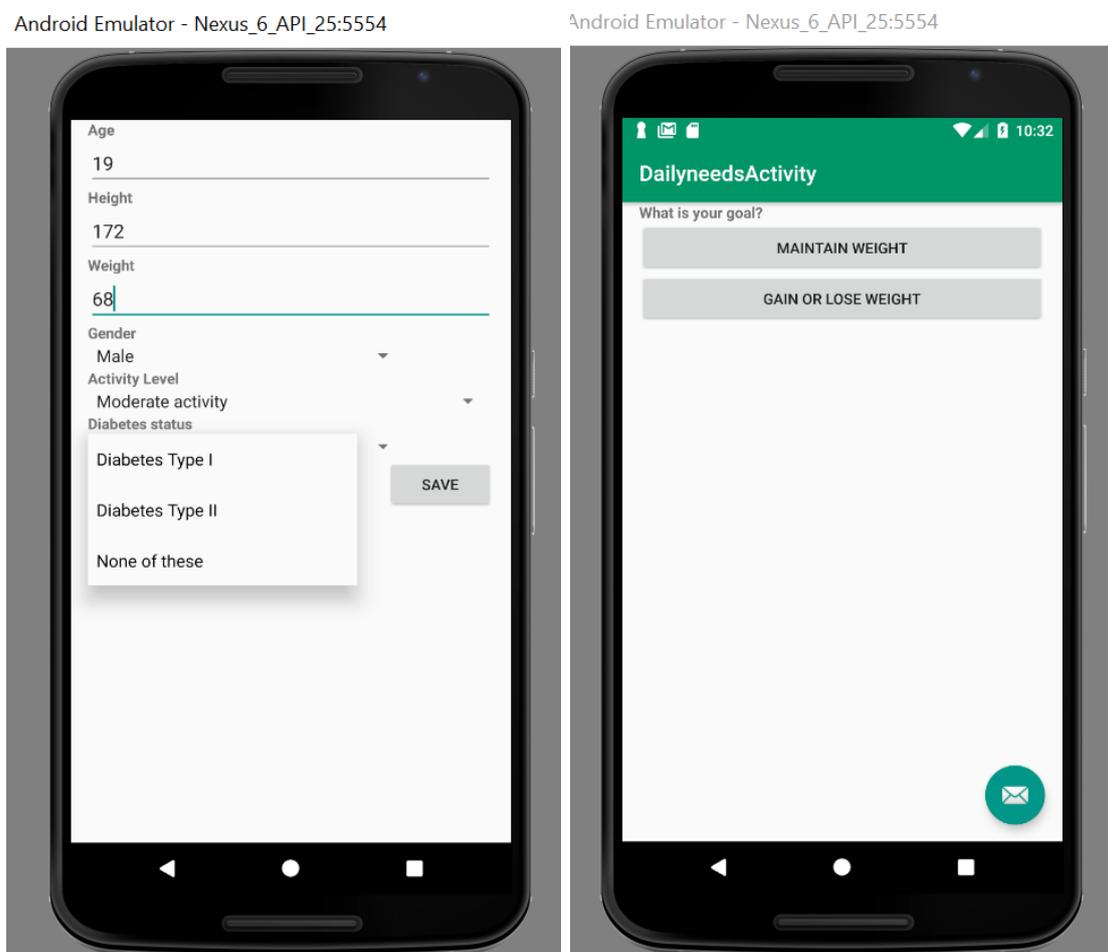


Figure 16. Screenshots showing user interfaces for selecting option for calorie calculations.

Android Emulator - Nexus_6_API_25:5554

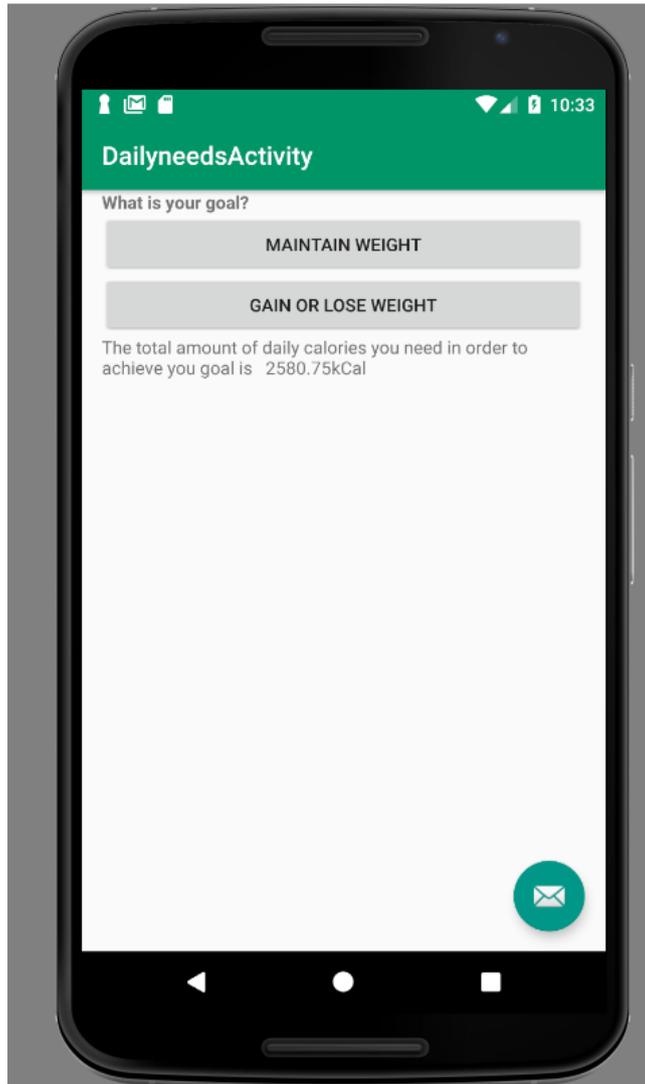


Figure 17. Image showing results for Scenario 3.

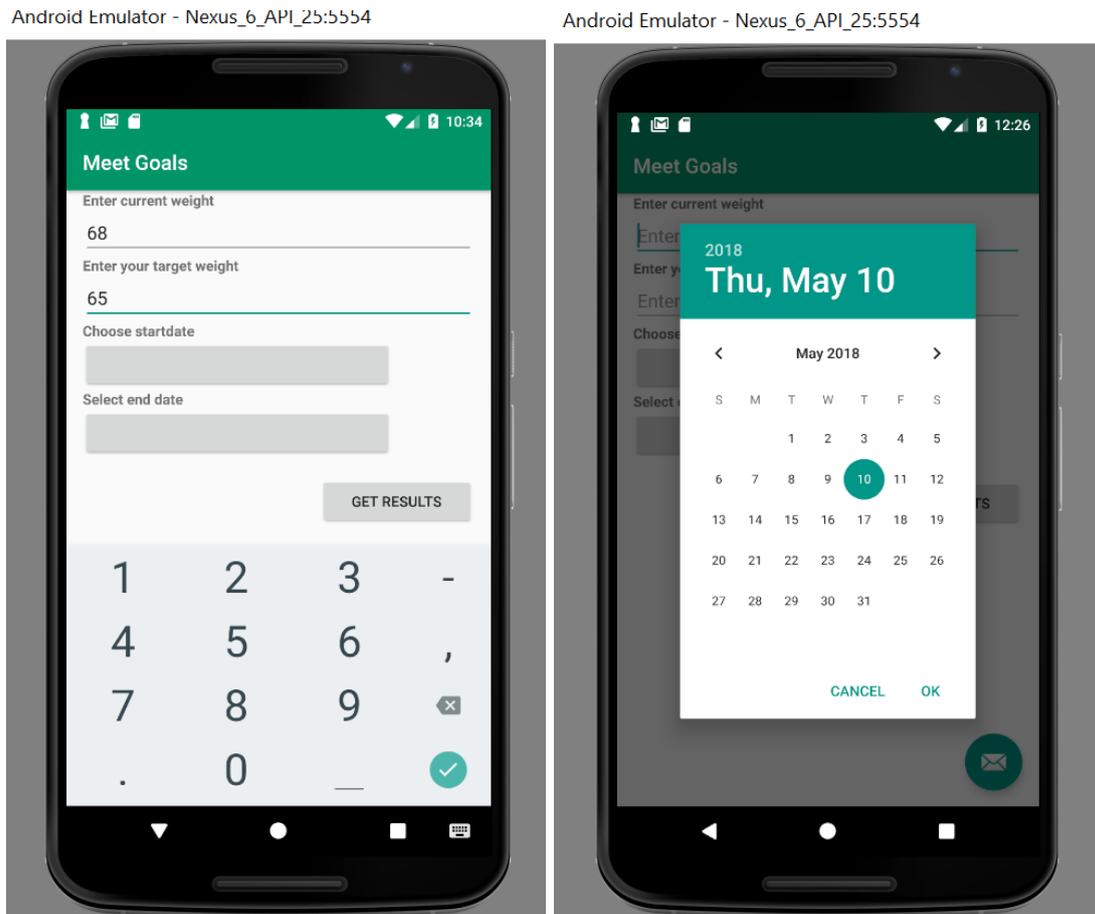


Figure 18. Other images showing interfaces for other weight goals which includes date-pickers for the selection of the time duration for goal.

3.5 Blood glucose interpretations

This function which is built for users with diabetes was also done using guidelines of the **NICE** UK published on the diabetes.co.uk website. Ranges for glucose reading for people with diabetes are provided on this site which is seen in table 2. It is to help people with diabetes keep their blood glucose in check during their daily activities and while during exercise. This informs users of conditions of *hypoglycemia* and *hyperglycemia* and normal BG levels as well.

Table 2. Recommended BG levels pre prandial and post prandial

Target Levels by Type	Upon walking	Before meals (pre prandial)	At least 90 minutes after meals (post prandial)
Non-diabetic*		4.0 to 5.9 mmol/L	Under 7.8 mmol/L
Type 2 diabetes		4 to 7 mmol/L	Under 8.5 mmol/L
Type 1 diabetes	5 to 7 mmol/L	4 to 7 mmol/L	5 to 9 mmol/L
Children w/ type 1 diabetes	4 to 7 mmol/L	4 to 7 mmol/L	5 to 9 mmol/L

IV. Scenario 4

This demonstrates how BG level interpretations are displayed to users. A person with **diabetes Type II** whose glucometer reading is **2 mmol/L** and has not eaten in the **past 90 minutes** (post prandial) gets a feedback that his blood sugar level is low and thus would have to take the needed steps in normalising it. Figures 19 and 20 depict the scenario above and a case of hyperglycemia respectively.

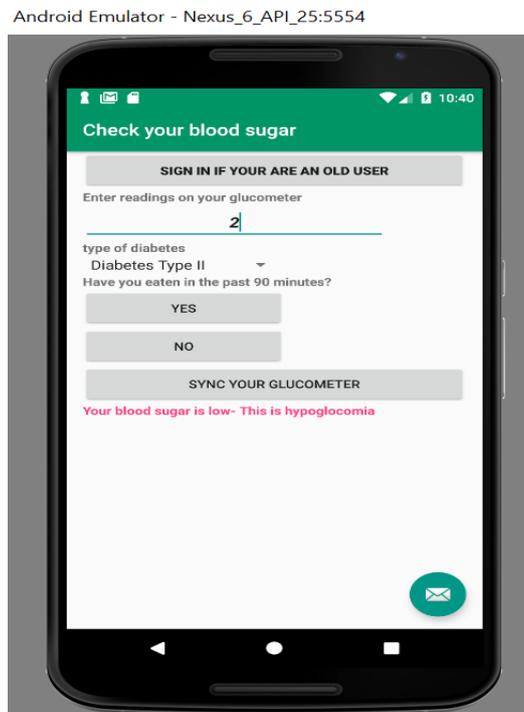


Figure 19. Display of a case of hypoglycaemia form Scenario 4.

Android Emulator - Nexus_6_API_25:5554

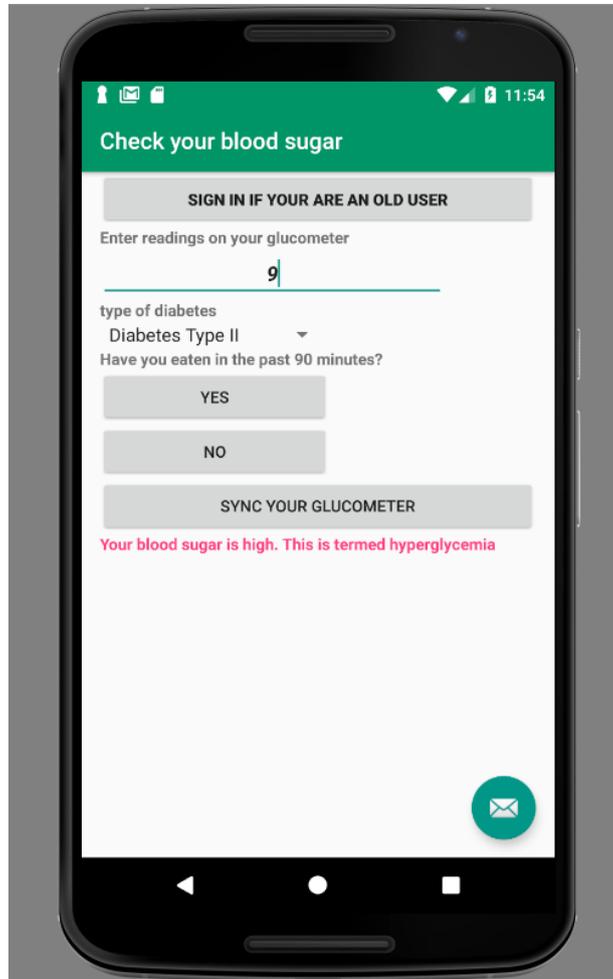


Figure 20. This also shows an incident of hyperglycemia where the user's blood sugar is too high.

3.5.1 Wireless syncing of glucometers

Syncing of one's glucometer is however disabled and left for future development so that users can wirelessly sync their glucometers for automatic interpretation rather than done the manual entry that is currently being done.

3.6 Physical activity monitoring with Fitbit

Fitbit is one of the best known smart band brands worldwide. Fitbit provides APIs for developer want to integrate their apps with Fitbit. The developer however must register his app before he can be able to integrate his application. *My Calbuddy* is integrated with Fitbit to retrieve information on the physical activities and heart rate of it users who have Fitbit wearables. The user must have a Fitbit account before he can access her or her Fitbit information. The user logs in to his or her Fitbit account first, where information that has been retrieved from the Fitbit server and stored in the database is fetched for him or her. In implementation and testing of this functionality, the Fitbit charge wristband was used. This feature was added to help encourage fitness and to help users know their physical activities levels so that they can choose the right level of physical activity for their daily needed calorie consumption calculations since most often than not, people estimate how active they are wrongly. *My Calbuddy's* Fitbit interfaces are shown in Figures 21 and 22.

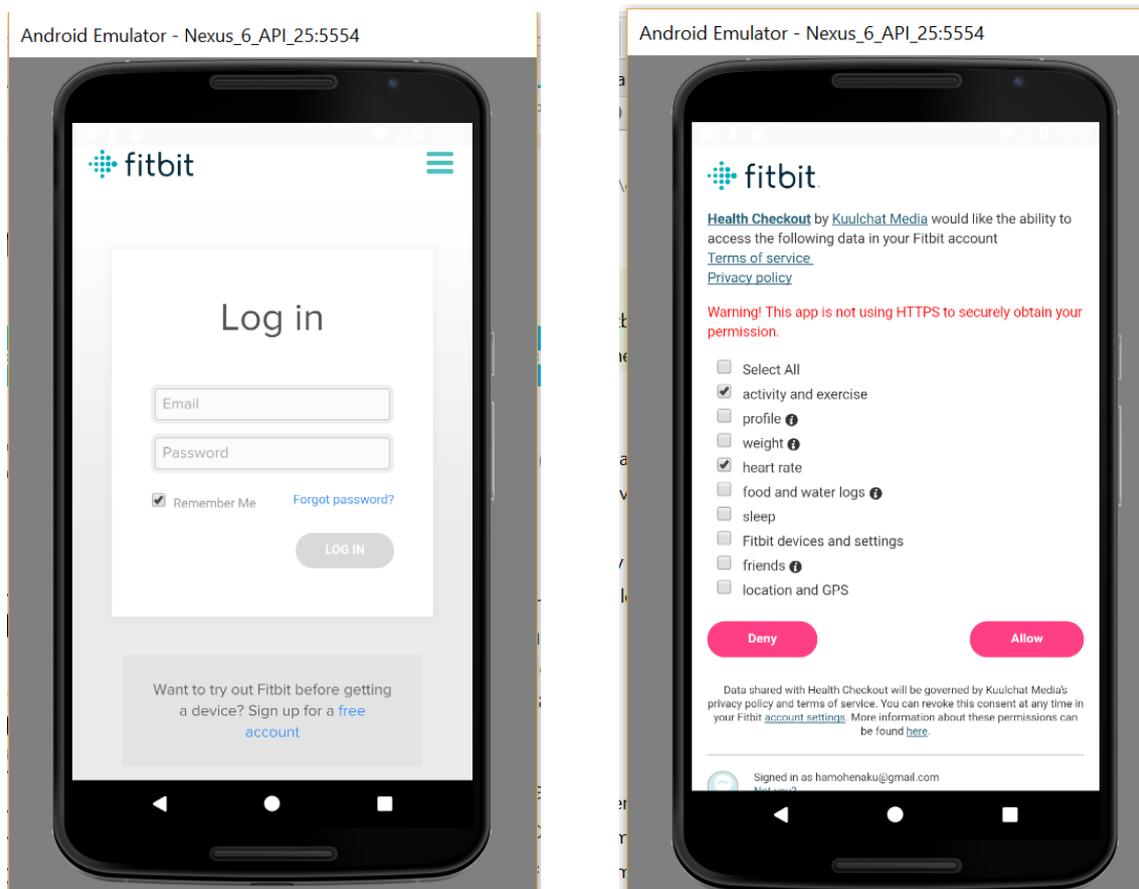


Figure 21. Login screens for Fitbit account and selection of needed information.

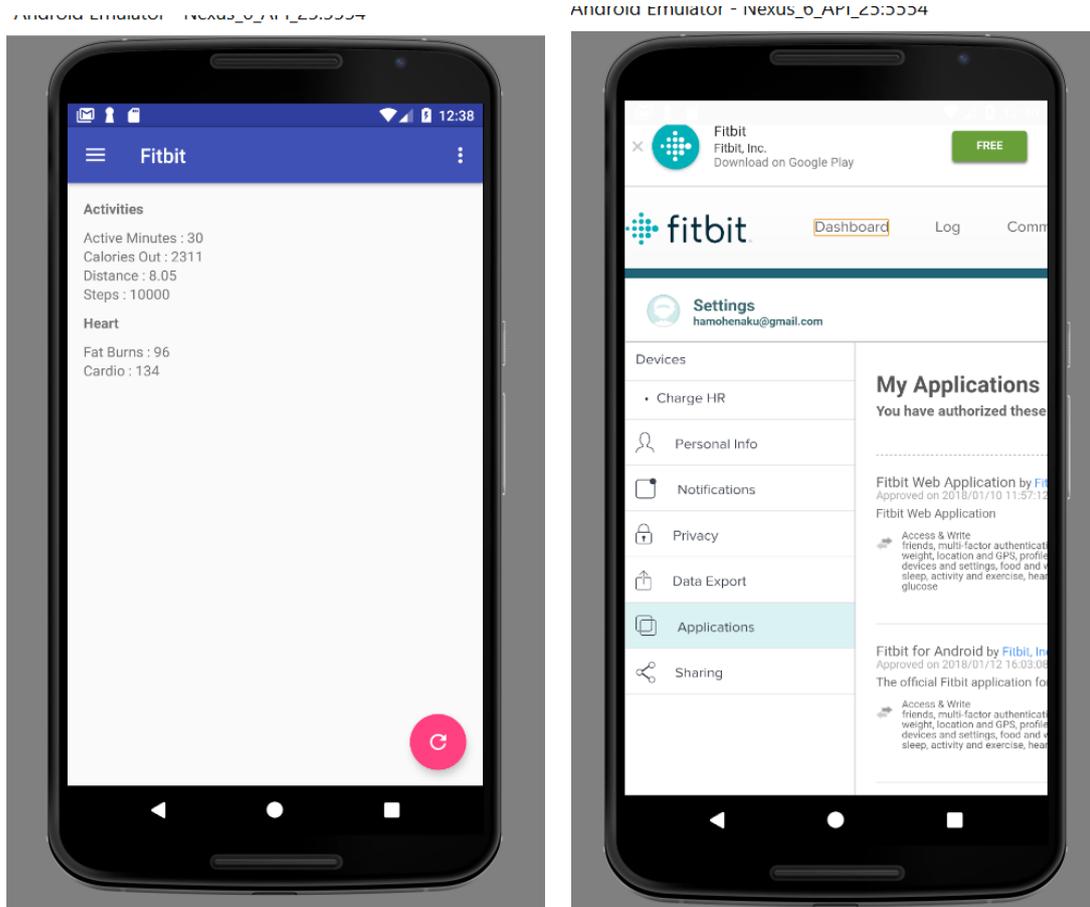


Figure 22. Display of physical activity information and account settings.

4 LOGISTIC REGRESSION MODEL AND IMPLEMENTATION

Sudden cardiac arrest (SCA) is a very critical condition that leads to mortality and has been an issue of concern in the health sector. SCA occurs when someone collapses because the heart has stopped pumping blood across his or her body, mostly due to an issue of the heart's electrical signals. It has a high risk of mortality and occurs without the foreknowledge of the patients (British heart foundation). Recently, a European Union funded project (ESCAPE- NET) was done to determine the risks and occurrence of SCA through the study of 857,790 SCA cases (Empana et al, 2018). Nanayakkara, et al. (2018) also did a study of machine learning techniques that could help make better a forecast of mortality of patients that had already had cardiac arrest. Though cardiac arrest usually occurs abruptly, there are risk factors associated with it, which are; diabetes, obesity, sedentary lifestyle, age, gender, previous history of heart attack and others. The occurrence of heart attack in individuals increases as a person grows and also, males more likely of getting cardiac arrest than women (Mayo clinic website).

As data is being gathered from day to day from various fields, there is the need for novel concepts and implementations to retrieve very important information from all these huge data storehouses for mankind to use to make life better (Neesha, et al., 2015). Mobile devices such as mobile phones have been come a part of our daily lives and can be very good tools for collecting data for analysis. That is why the mobile app is used as a data collection tool in this work. As discussed in chapter 2, machine learning techniques have been used in recent times to make predictions on health- related issues and diseases such as lung cancer, stroke, diabetes mellitus and many others.

This project aims at pre-diagnosing cardiac arrest before its occurrence and is to use this as decision support system for cardiac arrest predictions over a 10-year period. When individuals get a fair idea of their chances of getting cardiac arrests, they can be more careful and take some steps in order to help prevent its occurrence. Since *my Calbuddy* collected data on age, gender, diabetes status and physical activity of its users, these parameters are used predictors for the logistic regression model built for this project. A 10- year span is chosen based on the assumption that since age, weight and diabetes are variable and can change with time (age increases annually),10 years could be a logical duration for a person to change his or her lifestyle to create a permanent change to his

diabetes status and weight. The machine learning algorithm was first explored and implemented using MATLAB but, in order to build complete chain from the app to the database and to the machine learning algorithm, a much flexible programming language (Python) was use used later for implementation. Both methods are described below with more insight on the implementation process using Python.

4.1 What is classification in machine learning?

Classification is the process of finding a function that learns to group data items into classes that have already been designated (Neesha, et al., 2015). The model used for the classification process is called the classifier (Miskuf, et al. 2017). In this work. Logistic Regression is used for this purpose to classification were predictors are used to classify the outcome of cardiac arrest into an occurrence as 1 and a non-occurrence as 0.

4.2 Building the predictive model with Python

Python is a simple high-level, object-oriented programming language which can be used as a scripting language for binding components together in application development. It was developed by a man named, Guido van Rossum and is quite similar in nature to Fortran which is one of the oldest programming languages. Unlike other programming languages, variables can be used in python without declaring them and classes are also used only when needed. Python employs the use of indentation as a way of its structural check and management (Doty, 2008). This language was chosen for the implementation of the machine learning algorithm and the for the database connection due to its simplicity and flexibility.

4.2.1 Data preparation

In order to build a logistic regression model to make predictions, a dataset is needed. Getting information from Health facilities around was not possible within the timeframe of this project hence, data used was not real data gathered from a system but was data that was conjured. The dataset has 6 independent variables (predictors) which were age, gender, height, weight, activity level, and the diabetes status of user. The outcome (dependent variable) was categorised as to whether records of hospital patients indicate an occurrence of cardiac arrest or not. The data which was in excel format was imported into the python IDE and used. Table 3 is the dataset used for training and testing of the model.

Table 3. Dataset for training and testing of the model.

Age	Height	Weight	Gender (Male =1, Female 0)	Activity_level	Diabetes_Status(Diabetic = 1, Non-diabetic= 0)	Cardiac_Arrest
28	173	55	1	Sedentary	0	Yes
55	165	80	0	Lightly_Active	0	Yes
56	162	65	0	Moderate_activity	1	Yes
67	163.5	68	0	Moderate_activity	0	No
75	164	66	0	Extremely_Active	0	No
66	174	75	1	Sedentary	1	Yes
80	175	78	1	Sedentary	0	Yes
55	164	67	0	Moderate_activity	0	Yes
35	162	60	0	Very_Active	0	No
26	170	65	1	Extremely_Active	1	No
78	166	70	0	Moderate_activity	1	No
53	172	64	1	Moderate_activity	0	No
42	180	80	1	Extremely_Active	1	Yes
72	163.5	69	0	Sedentary	0	Yes
56	165	78	0	Sedentary	0	Yes
33	163	62	0	Very_Active	0	No
25	178	75	1	Very_Active	1	No
46	181	86	1	Moderate_activity	0	Yes
56	172	83	1	Sedentary	1	Yes
76	175	67	0	Very_Active	0	No
55	174	80	1	Lightly_Active	1	Yes
65	170	70	1	Lightly_Active	0	Yes
70	166	56	0	Moderate_activity	1	No
60	165	54	0	Very_Active	1	No
62	168	58	0	Very_Active	0	No
61	169	90	1	Moderate_activity	1	Yes
62	174	76	1	Moderate_activity	0	Yes
72	173	72	1	Lightly_Active	1	Yes
35	162	67	0	Lightly_Active	0	No
67	160	63	0	Moderate_activity	0	No
87	161	62	0	Very_Active	0	No
65	171	85	1	Very_Active	1	Yes
78	158	67	0	Moderate_activity	0	No
56	158	56	0	Moderate_activity	1	No
77	176	72	1	Very_Active	1	Yes
89	159	64	0	Very_Active	1	No
99	163.5	60	0	Very_Active	0	No
67	164	67	0	Moderate_activity	0	No
87	180	77	1	Very_Active	1	Yes
67	162	82	0	Lightly_Active	1	Yes
66	161	75	0	Moderate_activity	0	No
78	175	68	1	Sedentary	1	No
88	165	69	0	Sedentary	1	Yes
70	167	65	0	Moderate_activity	0	No

I. Independent Variables (Predictors)

The predictors (X) used are age (numeric), weight (numeric), height (numeric), Activity level (categorical: ‘Extremely active’, ‘Lightly active’, ‘Moderate activity’, ‘Sedentary’, ‘Very Active’), Diabetes status (categorical: Non-diabetes :0, diabetes :1).

II. Dependent Variable (Desired target)

The predicted variable y is classified based on question there is an incident of cardiac arrest in the patient’s hospital records. This classification was done with “1” as Yes and “0” as No.

The python implementation of these concepts is shown in Figure 23. Dummy variables were created for the various categories of Activity level to indicate their presence as 1 and absence as 0.

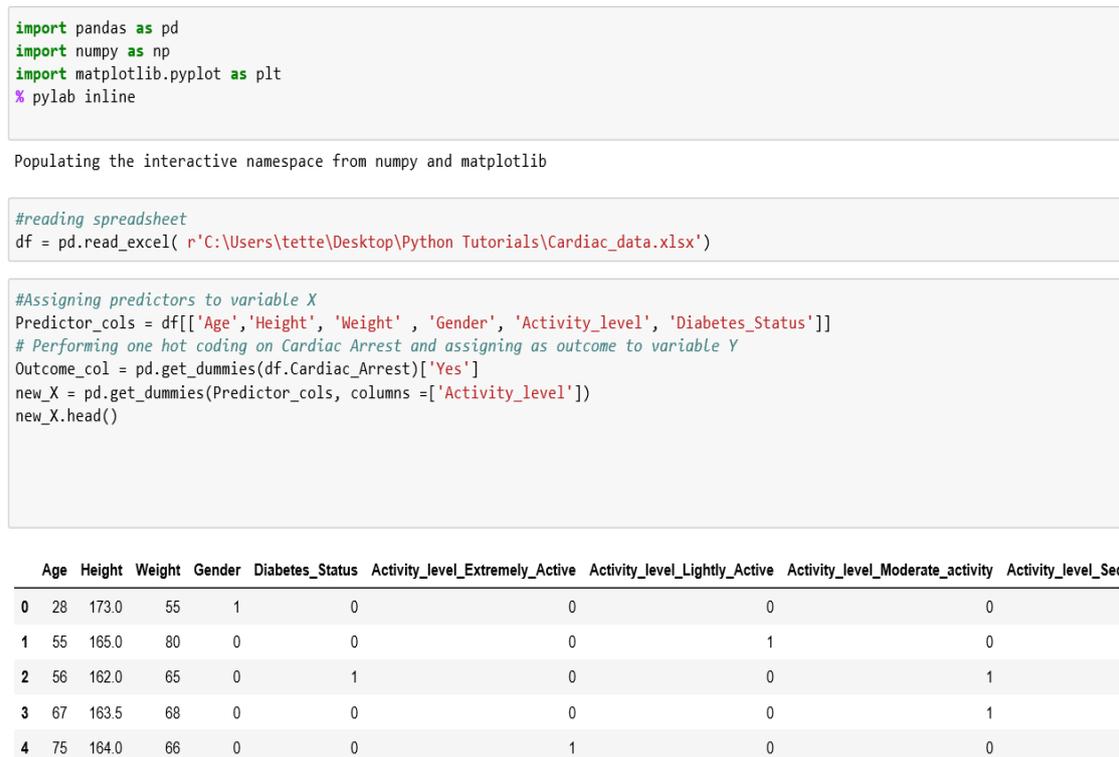


Figure 23. Commands for importing excel file and converting into a dataframe.

A test size of 33% (0.33) of the original data was used to testing the model as indicated below in Figure 24.

	Age	Height	Weight	Gender	Diabetes_Status	Activity_level_Extremely_Active	Activity_level_Lightly_Active	Activity_level_Moderate_activity	Activity_level_Sec
0	28	173.0	55	1	0	0	0	0	0
1	55	165.0	80	0	0	0	1	0	0
2	56	162.0	65	0	1	0	0	0	1
3	67	163.5	68	0	0	0	0	0	1
4	75	164.0	66	0	0	1	0	0	0

```

X = new_X
y = Outcome_col

#build and test training_set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

```

Figure 24. Command window showing model building and testing.

4.2.2 Data exploration

- I. **Age:** Though age is one of the contribution factors to cardiac arrest, the dataset did not show a significant relationship between cardiac arrest and age. This shown in the bar chart of age against frequency of cardiac arrest in Figure 25. There is no particular pattern shown.

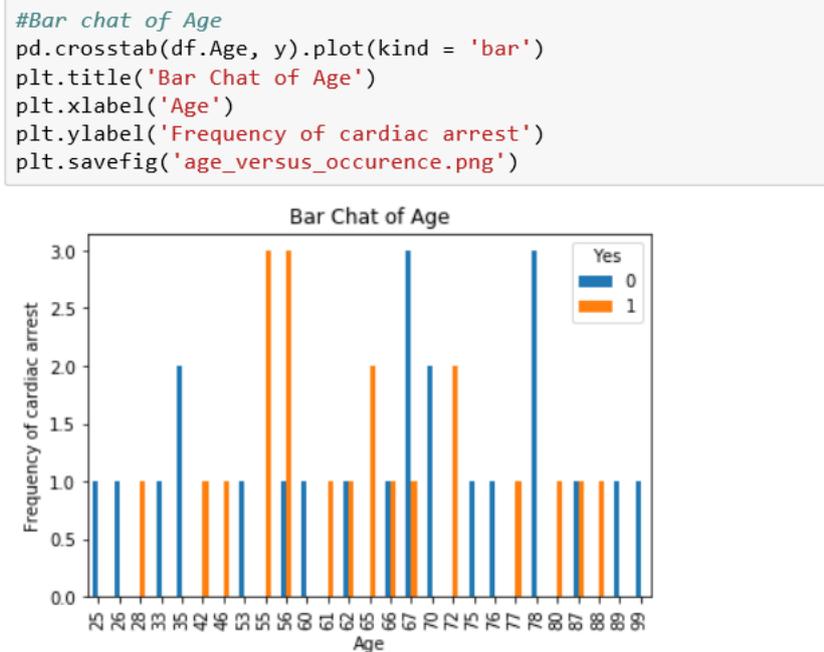


Figure 25. Bar chat of ages versus frequency of cardiac arrest.

II. Gender: Our dataset shows that more males had cardiac arrests than females which supports existing literature. This is illustrated by a bar chart in Figure 26. From this, it can be reliably said that gender has a significant correlation with the occurrence of cardiac arrest.

```
: #Bar chat of Gender
pd.crosstab(df.Gender, y).plot(kind = 'bar')
plt.title('Bar Chat of Gender')
plt.xlabel('gender')
plt.ylabel('Frequency of cardiac arrest')
plt.savefig('gender_versus_occurence.png')
```

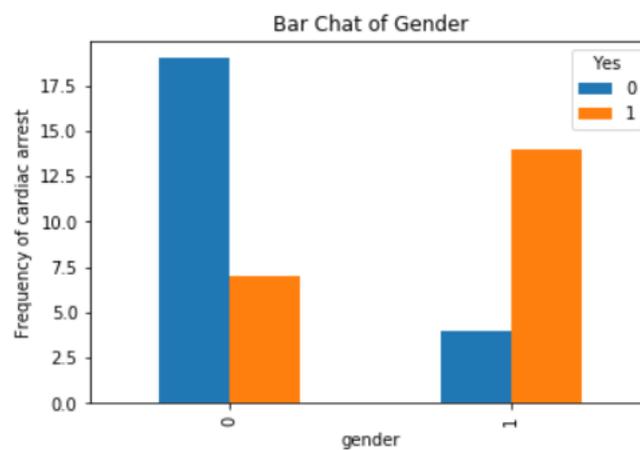


Figure 26. Bar chat of gender versus frequency of cardiac arrest.

III. Diabetes: Since diabetes is one of the contributing factors of cardiac arrest, diabetes was also plotted against the frequency of cardiac arrest. Diabetes patients had a higher frequency of cardiac arrest than non-diabetic patients. Though the difference in non-diabetic and diabetic patients was not so vast when observed from the plot in Figure 27, we can however say that diabetes has a significant effect on the occurrence of cardiac arrest.

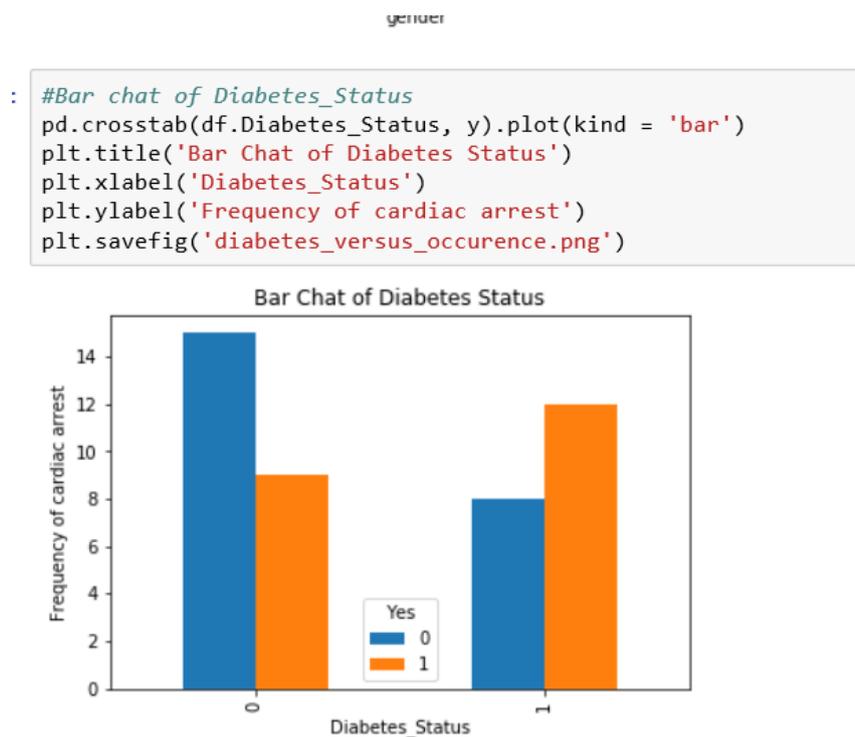


Figure 27. Bar Chat showing diabetes status against frequency of cardiac arrests.

4.3 Model implementation

Implementation of the Logistic Regression using Python commands is shown below in Figure 28.

```

: #Logist Regression model
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
model = LogisticRegression(penalty = 'l2', C=1)

: model.fit(X_train, y_train)

: LogisticRegression(C=1, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
verbose=0, warm_start=False)

```

Figure 28. Command window for implementation of logistic regression model.

4.3.1 Determining coefficients

The individual coefficients of the logistic regression model help in determining the probability of the occurrence of an event as elaborated in chapter 2. When working with logistic regression, the null hypothesis is that X variables are not associated with the Y variables which means that the Y variables predicted are far from the actual Y values that can be obtained unexpectedly (McDonald, 2014). As previously mentioned, these coefficients can be used to determine the likelihood of an outcome. The logistic regression coefficients for the model and the python syntax for getting these coefficients are shown in Figures 29 and 30.

```
]: import statsmodels.api as sm
   from scipy import stats
   model = LogisticRegression(fit_intercept = False, C = 1e9)
   mdl = model.fit(X_train, y_train)
   model.coef_

#sm
logit = sm.Logit(y, X)
logit.fit().params
```

Figure 29. Syntax for obtaining logistic regression coefficients.

```
Age                0.000912
Height             0.035481
Weight            0.195030
Gender            1.076056
Diabetes_Status   0.356852
Activity_level_Extremely_Active -21.957570
Activity_level_Lightly_Active   -19.221830
Activity_level_Moderate_activity -20.553362
Activity_level_Sedentary        -18.023496
Activity_level_Very_Active      -21.810796
dtype: float64
```

Figure 30. Display of coefficients in console.

4.3.2 Predicting test results and calculating accuracy

The prediction accuracy shown in of the model was 87% as shown in Figure 31. The aim of every model is to achieve an accuracy close to 100%, hence obtaining an accuracy of 87% is good.

```

: #print accuracy of model
y_pred = model.fit(X_train, y_train).predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(model.score(X_test, y_test)))

Accuracy of logistic regression classifier on test set: 0.87

```

Figure 31. Display of model accuracy.

4.3.3 Performing cross validation

Cross validation which gives better results than hold out sets is used for performance evaluation in this project. K-folds cross validation is use do determine how well the model performs (Wu, et al, 2018). K-folds cross validation is done when the model puts the dataset into k number of blocks handles each as a place holder set and trains these blocks k number of times. Hold out sets are small sample sets kept for the purpose of tuning the model (Thorton, C., 2011). In this project, 5-Folds cross validation is used which is shown in Figure 32. The importance of cross validation is to prevent over fitting. The cross validation accuracy was obtained as 65%.

```

#Cross validation
from sklearn import model_selection
from sklearn.model_selection import cross_val_score
kfold = model_selection.KFold(n_splits=5, random_state=7)
modelCV = LogisticRegression()
scoring = 'accuracy'
results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold, scoring=scoring)
print("5-fold cross validation average accuracy: {:.3f}" % (results.mean()))

5-fold cross validation average accuracy: 0.647

```

Figure 32. Implementation of cross validation and display of accuracy of cross validation.

4.3.4 Confusion matrix

The counts of the correct and incorrect classifications done by a classifier are usually shown by the confusion matrix. The confusion matrix of the model is as seen in Figure 33 below.

```

: #Confusion matrix
  from sklearn.metrics import confusion_matrix
  confusion_matrix = confusion_matrix(y_test, y_pred)
  print(confusion_matrix)

[[7 1]
 [1 6]]

```

Figure 33. Confusion matrix of the model.

This result tells us that, we have 7+6 correct predictions and 1 +1 incorrect predictions.

4.3.5 Precision, recall, F-measure and Support

$$Precision = \frac{T_p}{(T_p + F_p)}$$

$$Recall = \frac{T_p}{(T_p + T_n)}$$

Where T_p refers to True positives, F_p refers to False positives and T_n refers to True negatives.

F-beta score uses $\beta = 1.0$ to weight recall higher than precision and hence making recall as essential as precision. Figure 34 shows the precision, recall and F-beta score of the model.

Support represents the number of times each class occurs in y-test. (Li , 2017)

```
#compute precision, recall, F-measure and support
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.88	0.88	0.88	8
1	0.86	0.86	0.86	7
avg / total	0.87	0.87	0.87	15

Figure 34. Console showing precision, recall, F-measure and support values.

Interpretation: This means that the model is able to label the occurrence of cardiac arrest correctly and not otherwise by 87% (precision) and is also able to identify all occurrences of cardiac arrest by 87%.

4.3.6 Receiver Operating Characteristic (ROC) curve

ROC curves illustrate how distinguishing between two classes is done properly by a logistic regression model. It is very good to use ROC when much precision is needed in the classification being done. If the ROC region of the 1x1 square is on the top left corner in a manner that, the Area Under the Curve (AUC) becomes 100%, the model is said to distinguish perfectly, but, if the ROC curve falls on the diagonal line and has an are less than 50%, then the model does no good work (Sainani, K. L. ,2014). In this model, the AUC was 87% as indicated in Figure 35, which implies a good ability to discriminate.

```

#ROC
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, model.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, model.predict_proba(X_test)[:,-1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC.png')
plt.show()

```

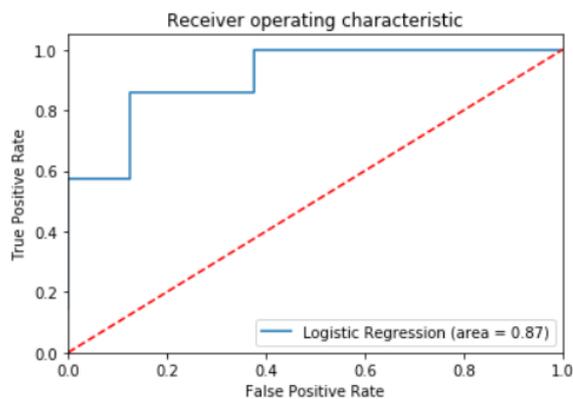


Figure 35. ROC curve implementation syntax and display.

4.4 Python – Firebase Connection

This connection is done so that user data of Calorie Buddy users can be fetched for predictions. When the real-time database is connected to python, the user information is obtained in JSON format which is then converted into Dataframes for predictions with the model. The implementation and this Firebase-python connection and the fetching of data from Firebase is shown in Figure 36. The information of a particular user is obtained using the User Id which is unique for every user.

```
In [33]: #Importing user data from Firebase Realtime database
from firebase import firebase
firebase = firebase.FirebaseApplication('https://my-calorie-buddy.firebaseio.com/')
result = firebase.get('/Userinfo', None)
print(result)

{'-L9L5p7iXUpsGacmREXr': {'actlevel': 'Moderate activity', 'userId': '-L9L5p7iXUpsGacmREXr', 'userage': '28', 'userdisease': 'None of these', 'usergender': 'Female', 'userheight': '170', 'userweight': '68'}, '-L9L7Zeza5H_Dev2S45Z': {'actlevel': 'Moderate activity', 'userId': '-L9L7Zeza5H_Dev2S45Z', 'userage': '28', 'userdisease': 'None of these', 'usergender': 'Female', 'userheight': '168', 'userweight': '68'}, '-L9LDnOntgdFuIfdK3_1': {'actlevel': 'Moderate activity', 'userId': '-L9LDnOntgdFuIfdK3_1', 'userage': '28', 'userdisease': 'Diabetes Type II', 'usergender': 'Female', 'userheight': '168', 'userweight': '68'}, '-L9ea1PxPcTXp758iwHR': {'actlevel': 'Moderate activity', 'userId': '-L9ea1PxPcTXp758iwHR', 'userage': '18', 'userdisease': 'Diabetes Type I', 'usergender': 'Female', 'userheight': '168', 'userweight': '68'}, '-L9j6578CCf3oviRaHT4': {'actlevel': 'Moderate activity', 'userId': '-L9j6578CCf3oviRaHT4', 'userage': '19', 'userdisease': 'Diabetes Type II', 'usergender': 'Male', 'userheight': '172', 'userweight': '68'}, '-L9jWHz8g8f4yggZ2uWg': {'actlevel': 'Moderate activity', 'userId': '-L9jWHz8g8f4yggZ2uWg', 'userage': '19', 'userdisease': 'None of these', 'usergender': 'Male', 'userheight': '172', 'userweight': '68'}}
```

Figure 36. Python codes for firebase database connection and imported JSON file from database after connection.

I. Scenario 5

In the scenario being used, a dummy user was created with age 28, height 170 cm, weight 68 kg, gender *Female*, and a *non-diabetes* status. The user's id was '**L9L5P7iXUpsGacmREXr**'. The details of this particular user was retrieved using the User id as illustrated in Figure 37. Figure 38 also shows this user's details as saved in the database.

```
#Getting data of a particular user for prediction
result1 = result['-L9L5p7iXUpsGacmREXr'] #the user ID is unique for every user.
result1

{'actlevel': 'Moderate activity',
 'userId': '-L9L5p7iXUpsGacmREXr',
 'userage': '28',
 'userdisease': 'None of these',
 'usergender': 'Female',
 'userheight': '170',
 'userweight': '68'}
```

Figure 37. Syntax of retrieving user data and display of results.

A new dataframe is created afterwards with only needed information. This included all the columns except the column for userID. The commands used are shown in Figure 40.

```

: #Creating new Dataframe with only needed columns
newdata = pd.DataFrame()
newdata['Age'] = W['userage']
newdata['Height'] = W['userheight']
newdata['Weight'] = W['userweight']
newdata['Gender'] = W['usergender']
newdata['Diabetes_Status'] = W['userdisease']
newdata['Activity_level_Moderate_activity'] = W['actlevel']

newdata['Diabetes_Status'] = newdata['Diabetes_Status'].map({'None of these': 0, 'Diabetes Type I': 1, 'Diabetes Type II': 1 })

: #output
newdata

:
   Age  Height  Weight  Gender  Diabetes_Status  Activity_level_Moderate_activity
0    28    170     68  Female                0                Moderate activity

:
Activity_df = pd.DataFrame(columns = ['Activity_level_Extremely_active', 'Activity_level_Lightly_active', 'Activity_level_Sedentary', 'Activity_level_Very_active'])
Activity_df

:
   Activity_level_Extremely_active  Activity_level_Lightly_active  Activity_level_Sedentary  Activity_level_Very_active

:
fdata= newdata.join(Activity_df)

```

Figure 40. Creating new dataframes with needed information.

Relevant information were kept and missing numbers were replace for the final dataframe as demonstrated in Figure 41.

```
fdata['Gender'] = fdata['Gender'].map({'Female': 0, 'Male': 1})

#pd.get_dummies(new_cols, columns =['Activity_Level'])

fdata
```

	Age	Height	Weight	Gender	Diabetes_Status	Activity_level_Moderate_activity	Activity_level_Extremely_active	Activity_level_Lightly_active	Activity_level_Sed
0	28	170	68	0	0	Moderate activity	NaN	NaN	

```
#Replacing missing values
fdata.Activity_level_Extremely_active = fdata.Activity_level_Extremely_active.fillna(value=0)
fdata.Activity_level_Lightly_active = fdata.Activity_level_Lightly_active.fillna(value=0)
fdata.Activity_level_Sedentary= fdata.Activity_level_Sedentary.fillna(value=0)
fdata.Activity_level_Very_active = fdata.Activity_level_Very_active.fillna(value=0)
fdata['Activity_level_Moderate_activity'] = fdata['Activity_level_Moderate_activity'].map({'Moderate activity': 1})
fdata
#prob1 = model.predict_proba(fdata)
```

	Age	Height	Weight	Gender	Diabetes_Status	Activity_level_Moderate_activity	Activity_level_Extremely_active	Activity_level_Lightly_active	Activity_level_Sed
0	28	170	68	0	0	1	0	0	

Figure 41. Creating new columns and replacing missing data.

4.4.2 Prediction of cardiac arrest

The syntax for the prediction of the dummy user using the model in shown in Figure 42 below.

```
: #predicition on user
prob1 = model.predict_proba(fdata)

: prob_s = np.around(prob1, decimals=5)
prob_s = prob_s* 100
model.predict(fdata)

#print ("")
#print ("Prediction: ")
print ("Probability: ")
#print( prob_s[0,0])
print ( "You are", prob_s[0,1], "% likely to get a cardiac arrest in the next 10 years" )
```

Figure 42. Syntax for display prediction.

Final answer to the question: “How likely are you (the user) to get a cardiac arrest in the next 10 years?” is given below in Figure 43.

```
Probability:
You are 0.058 % likely to get a cardiac arrest in the next 10 years
```

Figure 43. Output and display of the dummy user’s chances of getting a cardiac arrest.

4.5 Building the machine learning model with MATLAB

This section gives a brief description of how the model was initially implemented with MATLAB.

4.5.1 Description of the MATLAB toolbox.

There are many modern software packages available either freely or on sale for analysing data which includes MATLAB. MATLAB has been designed in such a way that it executes task quickly. It is user friendly and has libraries and functions that are pre-written and can be used for executing various tasks easily. MATLAB uses matrices as its main way of representing variables for data mining and other purposes. The Statics and Machine learning toolboxes or curve fitting toolbox are the commonly used data mining tool boxes among the toolboxes MATLAB has. It can be installed on computers with Linux, Windows and other operating systems (Mikuf & Zolotova, 2017).

a. Data preparation

The train dataset had 6 independent variable (predictors) which were age, gender, height, weight, activity level, and the diabetes status of patients. The outcome (dependent variable) was categorised into ‘yes’ and ‘no’ as to whether patients had had cardiac arrest or not. The data which was in excel format (Table 3) was imported into MATLAB using the import tool. Syntax for assigning the predictors and outcome variable is in Figure 44.

```
inputTable = Cardiacrecs;
predictorNames = {'Age', 'Height', 'Weight', 'Gender', 'Activitylevel', 'Diabetes'};
predictors = inputTable(:, predictorNames);
response = inputTable.HeartAttack;
isCategoricalPredictor = [false, false, false, true, true, true];
```

Figure 44. MATLAB codes for assigning predictors and the target of the model.

b. Training the model

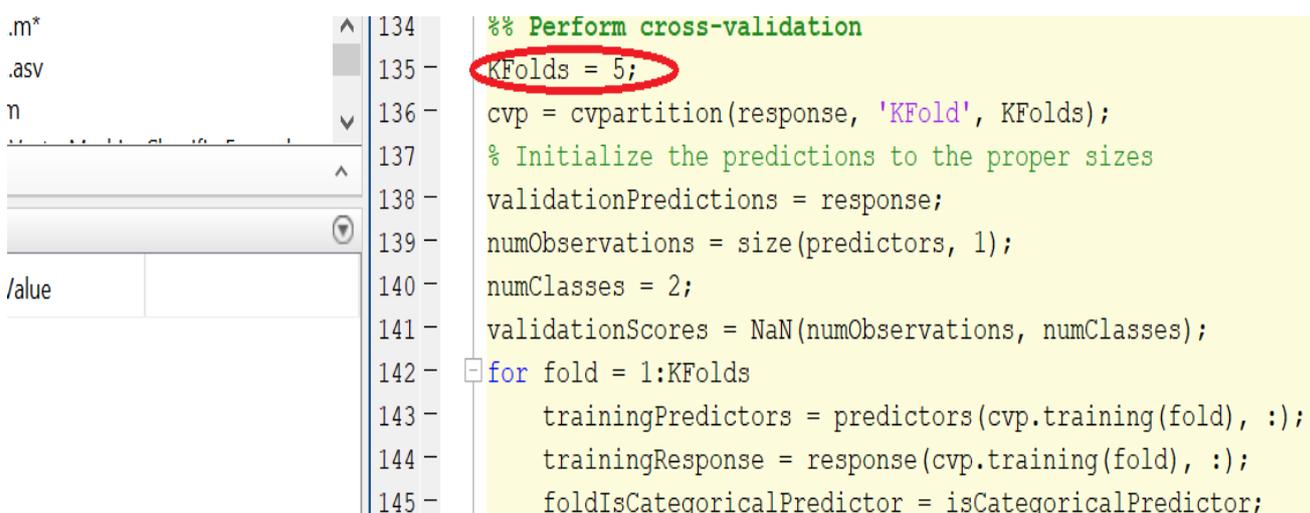
The syntax for training the model is shown in the codes in Figure 45.

```
% Train using fitglm.
GeneralizedLinearModel = fitglm(...
    concatenatedPredictorsAndResponse, ...
    'Distribution', 'binomial', ...
    'link', 'logit');
```

Figure 45. Model training MATLAB codes.

c. Cross validation

5-Folds cross validation was used just as done with Python. The syntax in Figure 46 shows how it was implemented.



```
.m* 134 %% Perform cross-validation
.asv 135 - KFolds = 5;
n    136 - cvp = cvpartition(response, 'KFold', KFolds);
137 % Initialize the predictions to the proper sizes
138 validationPredictions = response;
139 numObservations = size(predictors, 1);
140 numClasses = 2;
141 validationScores = NaN(numObservations, numClasses);
142 - for fold = 1:KFolds
143     trainingPredictors = predictors(cvp.training(fold), :);
144     trainingResponse = response(cvp.training(fold), :);
145     foldIsCategoricalPredictor = isCategoricalPredictor;
```

Figure 46. Codes for implementing cross-validation.

d. Interpreting logistic regression coefficients

The P values determine the relation between two variables, that is if there is a relationship between them or not. If the p value is small, the null hypothesis is rejected and vice versa. However, the p values of multiple logistic regression are not to be applied null hypothesis that are biological (McDonald, 2014). The regression coefficients of the model using MATLAB is shown in Figure 47.

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-102.57	4.9428e+08	-2.0751e-07	1
Age	-4.3191e-14	6.3946e+05	-6.7544e-20	1
Height	-8.959e-15	3.0782e+06	-2.9105e-21	1
Weight	-1.5384e-13	1.4266e+06	-1.0783e-19	1
Gender_1	1.4377e-12	4.2343e+07	3.3953e-20	1
Activitylevel_Lightly active'	1.7358e-12	4.3014e+07	4.0355e-20	1
Activitylevel_Moderately active'	1.7918e-12	4.1116e+07	4.3578e-20	1
Activitylevel_Sedentary'	3.3246e-12	4.3752e+07	7.5988e-20	1
Activitylevel_Very Active'	3.1884e-13	4.2453e+07	7.5104e-21	1
Diabetes_1	-7.046e-13	2.3107e+07	-3.0493e-20	1

44 observations, 34 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 0, p-value = 1

Figure 47. Output showing the estimated coefficients obtained from the model.

e. Predicting outcome for a new user

After modelling was done, testing was done with a dummy user with the information in table 4. The model predicted that there was a chance of the person getting a cardiac arrest which is indicated with a 'Yes' as shown in Figure 48 below.

```

> yfit = trainedClassifier.predictFcn(Newinput4)

yfit =
    categorical
    Yes

```

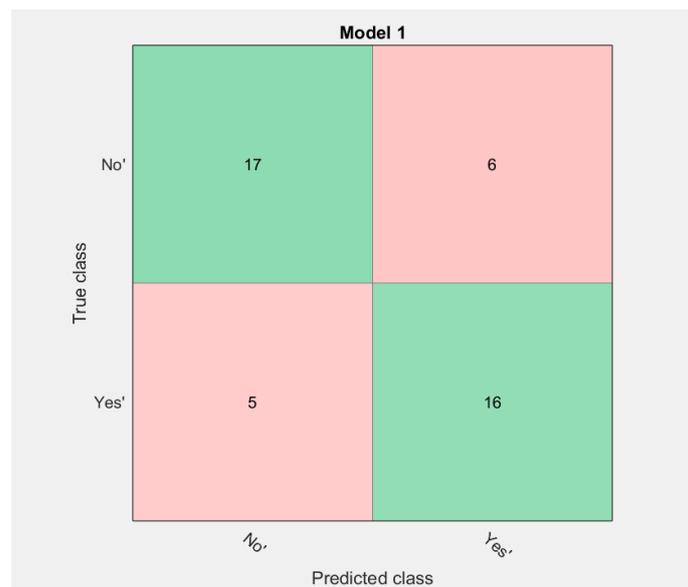
Figure 48. Prediction using dummy user with information in Table 4.

Table 4. Dummy user information for testing of logistic regression model in MATLAB.

Age	Height	Weight	Gender(Male =1, Female 0)	Activity level	Diabetes Status(Diabetic = 1, Non- diabetic= 0)
30	173	65	0	Sedentary	1

f. Confusion matrix and ROC (Receiving Operating Characteristic) curve

Below in Figure 49 is the confusion matrix of model obtained using the MATLAB Machine learning toolbox. It shows 17 correctly predicted 'No' responses and 16 correctly predicted 'Yes' responses.

**Figure 49.** Confusion matrix of the model.

g. ROC curve for the model

The ROC was 75% as indicated in Figure 50, which implies a moderate power to discriminate.

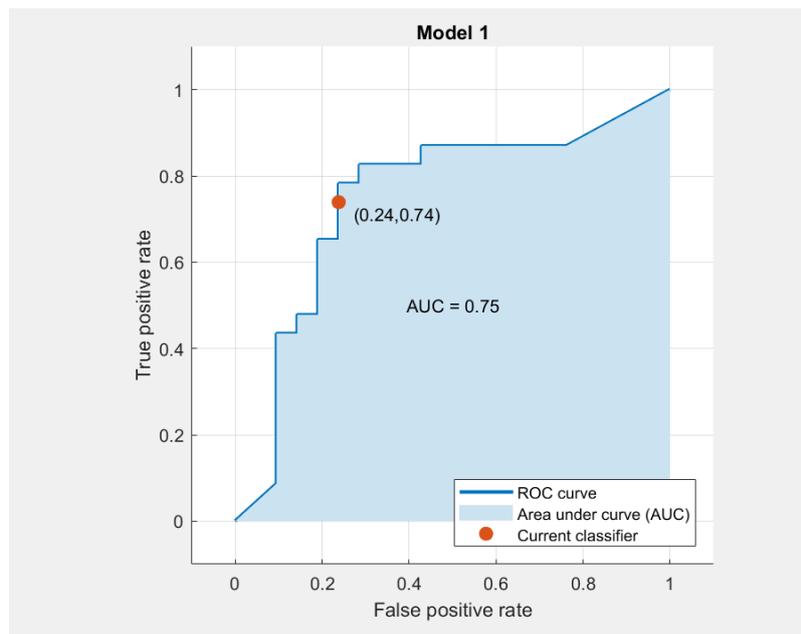


Figure 50. The ROC curve of the model using the MATLAB classifier app.

h. Accuracy of the model.

The validation accuracy of the model is estimated as 75.0 % which is shown in Figure 51 below. The percentage implies a fair model accuracy as well.

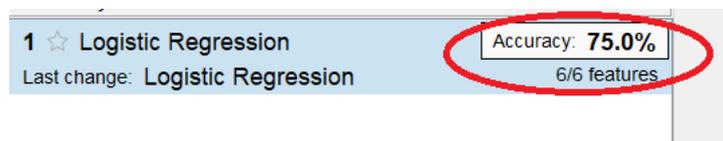


Figure 51. Screenshot showing accuracy of model using the machine learning classifier app.

4.6 Challenges with data acquisition

The main challenge with building the model was with data acquisition. Acquiring data for training and testing was not possible as stated earlier. Hence the dataset used was a dummy created just for this purpose. Due to privacy concerns, health care facilities were very sceptical about given out data of their patients. Several procedures and approvals had to be followed and done which could not be completed due to the limited time frame

of this project. In order to get a practical system for prediction, procedures for acquiring data should be started early in future works so that realistic results can be obtained.

5 DISCUSSION AND FUTURE WORKS

The reason for this project was to determine how efficient it is to solve health issues with software applications like mobile apps. Since android is a very simple mobile software for application development, it was chosen for the app development for this work. Collecting data for databases using smartphones and wearables can go a long way in solving health related issues and would help save lives in the near future. These devices could be used to send alerts to emergency numbers when individuals are faced with sudden tragic incidences such as heart attacks for immediate attention by ambulances. These alerts can be in the form of a text messages to emergency numbers set by default by the users. These numbers could be numbers of carers, emergency units or close relations. Through prompt and immediate responses from these emergency number owners, mortality can be avoided. This emergency alert feature can be incorporate in the *my Calbuddy* app when future work is done on this project.

With the dream of making *my Calbuddy* very easy to use and a well- known application extra functions other than the alert function above such are bar code scanning of food labels for calorie interpretation, synchronisation of app with glucose meters for diabetic patients for automated glucose level alerts and interpretations and a customisation portal for customers on insulin dosage for alerts on when to take their insulin doses, could be future adds on to it.

There are existing barcode databases that could be used to help make calorie checks more efficient and less stressful. Though this feature is already available in some apps, adding to the app of great help for its users. This would save users the trouble and hustle of having to look for information on food labels and manually entering them into the system.

Another feature that could be incorporated in this mobile app is the interpretation of food labels from Finnish to English for non-Finnish speakers. This would help make grocery shopping less stressful for people who do not speak and understand Finnish but can speak English. This implementation could be done using **Optical Character Recognition (OCR)** whereby texts are scanned and interpreted into English for users or through

scanners which would fetch information from a database with English interpretations for users.

Some food allergies can cause swelling of cells in the body which can lead to insulin resistance in diabetes patients (Derrell, 2011). Due to this proven connection between Diabetes and food allergies, adding allergy checks that would provide information on likely allergies of food ingredients on food labels could be a very healthy tool for diabetes patients. In doing this, available allergen databases can be worked with so that when barcodes are scanned for information of risks of allergies, likely allergies can be fetched for users. Gaining prior knowledge of allergies can help people chose grocery product without the fear of the aftermaths of consuming them.

Due to the inability to get real data for building the predictive model, I would recommend that the process of data acquisition be commenced early when future works are being done. This can would help build a more accurate system for accurate real-life predictions. Records can be obtained from EHRs on cardiac arrest in-patients and out-patients of health facilities.

6 CONCLUSION

Mobile devices are not just fancy tools that are used for calls and social media purposes but can be used for a wide range of activities. By installing a simple mobile app on a mobile device, certain goals and targets can be achieved, some of which are depicted in this thesis work. The work did not just focus on the mobile software decision support, but also addressed the issue of SCA (Sudden Cardiac Arrest) which is an area of concern in the healthcare sector. The Android platform was chosen for the app development due to its wide usage and flexibility.

The use of wearables for health care purposes has a great potential of solving bigger problems in the health sector in the future. With the aim of promoting fitness and exercise for a healthy life, the Fitbit SDK was incorporated into the mobile application so that users would be motivated to achieve their daily weight goals by not just counting calories but by exercising as well, since exercising can help solve the problems of obesity, heart diseases and other forms of health conditions.

At the end of this thesis, we were able to perform prediction of SCA for users of a mobile software application. Though the dataset used was (simulated) not from a real-life system or database, the steps and processes involved can be implemented using accurate data from real systems for predictive purposes in the health care sector in the future. Achieving this result has made me glad because, at the end of the day, different issues were tackled and addressed at various stages in this very project, i.e. from weight management to physical activity tracking to cardiac arrest prediction thus implying that used for problem solving in diverse ways.

REFERENCES

- Äkerberg, A., Lindén, M., & Folke, M. (2012). “How accurate are pedometer cell phone applications?.” *Procedia Technology*, 5, 787-792.
- Atwater, W.O & Benedict, F.G. 1902. Experiments on the metabolism of matter and energy in human body, 1898-1900. *US Office of Experiment Stations Bulletin* No. 109, Government Printing Office, Washington, DC.
- Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert systems with applications*, 88, 402-418.
- Amutha, A., J., Padmajavalli, R., & Prabhakar, D. (2018). A Novel Approach for the Prediction of Treadmill Test in Cardiology using Data Mining Algorithms Implemented as a Mobile Application. *Indian Hear Journal*.
- Backwell A. (2009). Humanc Computer Interaction- Lecture Notes, *Cambridge Computer Science Tripos, Part II, 2*. Available: <https://www.cl.cam.ac.uk/teaching/1011/HCI/HCI2010.pdf>
- Barry, B. (2015). *Android application development all-in-one for dummies*. Second Edition, 31. New Jersey, NJ: John Wiley & Sons, Inc.
- Bonnardel, N., Piolat, A., & Le Bigot, L. (2011). The impact of colour on Website appeal and users’ cognitive processes. *Displays*, 32(2), 69-80.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Patten recognition*, 30(7), 1145-1159.
- British Heart Foundation. Cardiac Arrest. Available: <https://www.bhf.org.uk/heart-health/conditions/cardiac-arrest>
- Bzdok, D., & Meyer-Linderberg, A. (2017). Machine learning for precision psychiatry. Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

- Campbell, I., (2017). Macronutrients, minerals, vitamins and energy. *Anaesthesia & Intensive Care Medicine*, 18(3), 141-146.
- Ching, K. W., & Singh, M. M. (2016). Wearable technology devices security and privacy vulnerability analysis. *Int. J. Netw. Secur. Appl*, 8 (3), 19-30. https://www.researchgate.net/publication/303870892_Wearable_Technology_Devices_Security_and_Privacy_Vulnerability_Analysis
- De Arriba-Pérez, F., Caeiro-Rodríguez, M., & Santos-Gago, J. M. (2016). Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. *Sensors*, 16(9), 1558.
- Derrell J. (2011). Understanding the connection between food allergies and diabetes. *Natural News*. Available: https://www.naturalnews.com/033125_food_allergies_diabetes.html
- Diabetes.co.uk. Blood sugar level ranges. *The global diabetes community*. Available: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
- Doty, S.R. (2008). Lecture notes- Python Basics. Loyola University of Chicago. Available: <https://anh.cs.luc.edu/331/notes/PythonBasics.pdf>
- Ekström, M., Esseen, P.A., Westerlund, B., Grafström, A., Jonsson, B. G., & Ståhl, G. (2018). Logistic regression for clustered data from environmental monitoring programs. *Ecological Informatics*, 43, 165- 173.
- Elío, J., Crowley, Q., Scanlon, R., Hodgson, J., & Long, S. (2017). Logistic regression model for detecting radon prone areas in Ireland. *Science of The Total Environment*, 599, 1317-1329.
- Empana, J. P., Blom, M. T., Böttiger, B. W., Dargès, N., Dekker, J. M., Gislason, G., ... & Jonsson, M. (2018). Determinants of occurrence and survival after sudden cardiac arrest—A European perspective: The ESCAPE-NET project. *Resuscitation*, 124, 7-13.
- Fienman, R. D., Pogozelski, W. K., Astrup, A., Bernstein, R. K., Fine, E. J., Westman, E. C., ... & Nielsen, J. V. (2015). Dietary carbohydrate restriction as the first approach in diabetes management: critical review and evidence base. *Nutrition*, 31(1), 1-13.

Get physical. Calculating your daily calorie expenditure. Available: <http://getphysical.co.uk/index.php/training-resources/calculating-your-calorie-expenditure/>

Glynn, L. G., Hayes, P. s., Cassey, M., Glynn, F., Alvarez-Iglesias, A., Newell, J., ... & Murphy, A. W. (2014). Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial. *Br J Gen Pract*, *64*(624), e384-e391.

Gresham, G., Schrack, J., Gresham, L. M., Shinde, A. M., Hendifar, A. E., Tuli, R., ... & Piantadosi, S. (2018). Wearable activity monitors in oncology trials: Current use of an emerging technology. *Contemporary clinical trials*, *64*, 13-21.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Huang, T., Li, B., Shen, D., Cao, J., & Mao, B. (2017). Analysis of the grain loss in harvest based logistic regression. *Procedia Computer Science*, *122*, 698-705.

Hwang, S., & Lee, S. (2017). Wristband-type wearable health devices to measure construction workers' physical demands. *Automation in Construction*, *83*, 330-340.

Jabeur, S. B. (2017). Bankruptcy prediction using Partial Least Squares Logistic Regression. *Journal of Retailing and Consumer Services*, *36*, 197-202.

Jain, V., Patel, R. K., Kapadia, Z., Galiveeti, S., Banerji, M., & Hope, L. (2017). Drugs and hyperglycemia: A practical guide. *Maturitas*, *104*, 80-83.

Jenkins, D. W., & Jenks, A. (2017). Exercise and diabetes: A narrative review. *The Journal of Foot and Ankle Surgery*, *56*(5), 968-974.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

- Kent M. (2006). Atwater General factor System. *The Oxford Dictionary of Sports and Science Medicine*. Third Edition. Published in print January 2006 | ISBN: 9780198568506. Published online January 2007 | e-ISBN: 9780191727788. <http://oxfordindex.oup.com/view/10.1093/acref/9780198568506.013.0679>
- Khunti, K., Alsifri, S., Aronson, R., Berković, M. C., Enters-Weijnen, C., Forsén, T., ... & Kapur, R. (2017). Impact of hypoglycaemia on patient-reported outcomes from a global, 24-country study of 27,585 people with type 1 and insulin-treated type 2 diabetes. *diabetes research and clinical practice*, 130, 121-129.
- Lee, S. (2012). Creating and using databases for android applications. *International Journal of Database Theory and Application*, 5(2), 99-106.
- Li, S (2017). Building a logistic regression model with python step-by-step. Towards Data Science. Available: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>
- Lin, W. Y., Chen, C. H., Tseng, Y. J., Tsai, Y. T., Chang, C. Y., Wang, H. Y., & Chen, C. K. (2018). Predicting Post-Stroke Activities of Daily Living Through Machine Learning-Based Approach on Initiating Rehabilitation. *International Journal of Medical Informatics*.
- Lynch, C., M., Abdollai, B., Fuqua, J. D., Alexandra, R., Bartholomai, J. A., Blgemann, R. N., ... & Frieboes, H. B. (2017). Prediction of lung cancer for patient survival via supervise machine learning classification techniques. *International journal of medical informatics*, 108, 1-8.
- Mansoor, H., Elgendy, I. Y., Segal, R., Bavry, A. A., & Bian, J. (2017). Risk prediction model for in-hospital mortality in women with ST -elevation myocardial infarction: A machine learning approach. *Heart & Lung: The Journal of Acute and Critical Care*, 46(6), 405-411.
- Mayo Clinic. Sudden Cardiac Arrest. Available: <https://www.mayoclinic.org/diseases-conditions/sudden-cardiac-arrest/symptoms-causes/syc-20350634>

- McCrorry, C., Vanderlee, L., White, C., M., Reid, J.L., & Hammond, D. (2016). Knowledge of recommended calorie intake and influence of calories on food selection among Canadians. *Journal of nutrition and education and behaviour*, 48(3), 199-207.
- McDonald, J.H. 2014. Handbook of Biological Statistics, Third Edition. Sparky House Publishing, Baltimore, Maryland.
<http://www.biostathandbook.com/multiplelogistic.html>
- Morias, A., Pexoto, Hö, Coimbra, C., Abelha, A., & Machado, J. (2017). Predicting the need of Neonatal Resuscitation using Data Mining. *Procedia Computer Science*, 113, 571-576.
- Nanayakkara, S., Fogarty, S., Ross, K., Milosevic, Z., Richards, B., Liew, D., ... & Kaye, D. (2018). Machine learning models significantly improve outcome prediction after cardiac arrest. *Journal of the American College of Cardiology*, 71(11 Supplement), A775.
- Neesha, J., Nur'Aini, A. R., Wahidah, H. (2015). Data mining in healthcare- A review. *Procedia Computer Science*, 72, 306-313.
- Nikolaou, C. K., McPartland, M., Demkova, L., & Lean, M. E. (2017). Supersize the label: The effect of prominent calorie labelling on sales. *Nutrition*, 25, 112-113.
- Park H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43(2), 154-164.
- Perera. W. A. S. N., Nananyakkara, J., Werapitiya, B. K., Rajasingham, S., Premaratne, U.S., Erandaka, W.U., ... & Ratnayake, H. U. W. (2017). Introduction to Android.
- Pevnick, J. M., Birkeland, K., Zimmer, R., Elad, Y., & Kedan, I. (2018). Wearable technology for cardiology: An update and framework for the future. *Trends in cardiovascular medicine*, 28(2), 144-150.
- Phillips, S. M., Cadmus-Bertram, L., Rosenberg, D., Buman, M. P., & Lynch, B. M. (2018). Wearable Technology and Physical Activity in Chronic Disease: Opportunities and Challenges. *American journal of preventive medicine*, 54(1), 144-150.

- Racine, S. E. (2018). Emotional ratings of high-and low-calorie food are differentially associated with cognitive restraint and dietary restriction. *Appetite*, 121, 302-308
- Riera-Crichton, D., Tefft, N. (2014). Macronutrients and obesity: revisiting the calories in, calories out framework. *Economics and Human Biology*, 14, 33-49.
- Sainani, K.L. (2014). Logistic regression. *PM&R*, 6(12), 1157-1162.
- Seguias, L., & Tapper, K. (2018). The effect of mindful eating on subsequent intake of a high calorie snack. *Appetite*, 121, 93-100.
- Simonoff, J. S. (2012). Logistic regression- modelling the probability of success. *Teaching paper of Leonard N. Stern School of Business, New York University*.
- Siomos, M., Z., Andreoni, M., Buchholz, S. W., & Dickins, K. (2017). A guide to physical activity for individuals with diabetes. *The Journal for Nurse Practitioners*, 13(1), 82-88.
- Thornton, C. (2011). Machine Learning: Lecture 5- Cross validation. *University of Sussex*, 15-18. Available: <http://users.sussex.ac.uk/~christ/crs/ml/lec03a.pdf>
- Tripepi, G., Jager, K. J., Dekker, F.W., & Zoccali, C. (2008). Linear and logistic regression analysis. *Kidney international*, 73(7), 806-810.
- Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Hear failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Computation and structural journal*, 15, 26-47.
- US Food and Drug Administration. (2015). Using nutrition fact labels: A how -to guide for older adults. Available: <https://www.fda.gov/Food/ResourcesForYou/Consumers/ucm267499.htm>
- Wolfson, J. A., Graham, D. J., & Bleich, S. N. (2017). Attention to physical activity-equivalent calorie information on nutrition facts labels: An eye-tracking investigation. *Journal of nutrition education and behaviour*, 49(1), 35-42.

- Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked, 10*, 100-107.
- Yang, F., Banerjee, T., Narine, K., & Shah, N. (2018). Improving pain management in patients with sickle cell disease form physiological measures using machine learning techniques. *Smart Health*.
- Zaharieva, D. P., & Ridell, M. C. (2017). Insulin Management Strategies for Exercise in Diabetes. *Canadian journal of diabetes, 41(5)*, 507-516.