# Autonomous Analytics Health Dashboard

Lu Wang

Supervisor: Attracta Brennan

November 2020

## Declaration

## Abstract

The purpose of this study is to showcase the importance of data visualization and how it can be useful in the healthcare domain. This study the applications of data visualization in supporting healthcare policy decisions as well as the development of personalized treatment plans. The initial sections of the study focus on the background of data visualization and going through available literature on this topic, the and the use of interactive visual analysis tools to help clinicians and medical managers gain insights from historical data to understand better which treatments have the best results.

<you will need to discuss more the gap – the application and the results also>

## Acknowledgements

## Table of Contents

## List of Figures

## List of Tables

## Glossary

| | |
|---|---|
| DV | Data Visualization |
| AL | Artificial Intelligence |
| ML | Machine Learning |
| IoT | Internet of Things |
| SaaS | Software-as-a-Service |
| PaaS | Platform-as-a-Service |
| IaaS | Infrastructure-as-a-Service |
| SVM | Support vector machine |
| SVC | Support Vector Classification |
| ROC | Receiver Operator Characteristic |
| AUC | Area Under ROC Curve |
| FPR | False positive rate |
| FNR | False Negative Rate |
| TPR | True Positive Rate |
| TNR | True Negative Rate |
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |

## Chapter 1: Introduction

Big data refers to complex data collections that cannot be accurately and efficiently completed with traditional data management systems in terms of data processing and application (ref). Due to the need to comply with regulatory requirements, historical and current health departments have generated a large amount of patient and practitioner data [7]. The effective use of medical big data can bring huge benefits to medical providers and patients. These benefits include: improving the quality and efficiency of healthcare planning and delivery [8], early disease detection (to facilitate intervention and treatment), customized health management, and effective detection of fraudulent healthcare behaviors [9]. The McKinsey Global Institute believes that the effective use of medical big data generated in the United States has the potential to increase the value of more than 300 billion US dollars each year, of which 66% of the value is related to the reduction of medical expenditure [10].

Although the development potential of big data is huge, it also brings many problems and new challenges. On the one hand, the storage and transmission of such a huge data set, analysis and processing are difficult to achieve. Furthermore, in big data, the meanings not only structured data, but also unstructured and semi-structured data. Therefore, many medical big data are underutilized and left unused in large digital repositories [12].

From the perspective of the development of the medical and health industry, it is foreseeable that as people's demand for medical and health services continues to increase, expenditures in the medical and health field will inevitably increase (reference). However, according to the report of the American Academy of Medical Sciences, in the current development of the global healthcare industry, at least one-third of the expenditure is wasted []. This obviously has no positive effect on improving medical health. Therefore, in order to ensure the healthy development of medical and health services, it is essential to use big data technology to evaluate the medical and health market and analyze the effects of medical and health diagnosis and treatment.

**Formatted:** Font: (Default) Times New Roman, Highlight

**Formatted:** Font: (Default) Times New Roman, Highlight

### 1.1 Primary Research Question

The research question for this study is "How can visualization and machine learning be used to create an automatous health analytic dashboard to 'best' present health status to users in different stages (eg. Patients , Doctors)? "

### 1.2 Research Motivation

Today, data plays an indispensable role in the healthcare industry[1]. Especially for patients with a medical history, doctors usually need to combine the patient's historical data for diagnosis and treatment (for example, diabetes is often accompanied by complications, and doctors must carefully consider the interaction between drugs before prescribing drugs (ref). Most importantly, patient data is usually very sensitive and is characterized by being complex and difficult to manage. The reality is that doctors spend a lot of energy attending to many patients every day. With the introduction of electronic medical records(EMRs), patient databases have gradually improved, but at the same time the data has also grown rapidly. As early as 2012, healthcare data accounted for 30% of global electronic data storage [1]. In 2013, the amount of healthcare data was 153 exabytes-expected to grow to 2,314 exabytes by 2020 [1]. Obviously, it is almost impossible to use human resources to process these data. How to effectively manage these data is critical.

Fortunately, the development of cloud computing, quantum computers, and other technologies has made the processing and management of these data 'easier' [ref]. As an example, graphic representation can help doctors understand the meaning behind complex data and support them in making 'more' accurate judgments with the fastest speed and risk control [ref].

The benefits of data visualization to the healthcare industry are obvious. For example, by combining artificial intelligence (AI) and dashboard technology, we can best display and help doctors understand patient data. Machine learning a subset of AI can help doctors get rid of the trouble of manually processing large and complex patient data. ML can be used to predict the direction of the disease based on historical and real-time data the more data the more accurate the prediction []. Therefore, the use of ML can save a lot of time and allow the doctor to focus on the main problem. Data visualization and algorithms can quickly extract useful information, organize data in an easy-to-understand way and present it to doctors to enhance treatment [].

All in all, from the perspective of medical and health development, research into AI applied to medical data the development of automatic medical and health big data analysis dashboards have become highly relevant to the health industry in the medical and health field [].

### 1.3 Secondary Research Questions

The main research question for this study "How can visualization and machine learning be used to create an automatous health analytic dashboard to 'best' present health status to users in different stages (eg. Patients, Doctors) has generated the following four research sub-questions that the author has set out to address for this thesis:

1. How to show a patient how compared to the average?
2. How to support doctors in selecting the 'best' statistical approach in analyzing medical data?
3. How to help doctors in selecting the 'best' Machine Learning approaches in analyzing medical data?
4. How to use best practices in data visualization to support doctors in the interpretation of the results from the Machine Learning and statistical approaches.

### 1.4 Thesis Structure

This thesis is structured as follows:

• Chapter 2 presents a general overview of the research background and selected technologies that can be used to fulfill the research questions.

• Chapter 3 summarizes the main and sub research question and presents the research objectives. It also outlines the research methodology that was used to address the research questions.

• Chapter 4 explores the functional and non-functional requirements, the system design, and finally, the technical requirements needed to develop the autonomous health analysis dashboard. It also explains the development of the personalized health analysis dashboard.

• Chapter 5 specifies the application of design principles in the dashboard design and health analytic dashboard.

- Chapter 6 describes the case study, how it was conducted and the accruing results and discussion. The study investigates all the four secondary research questions, how they were addressed and a discussion of the experiment results.

- Chapter 7 presents the conclusions of this research and describes further work that arose from the findings of both the primary and secondary research.

# Chapter 2: Data Visualization

## 2.1 Introduction

What is Visualization? Visualizations are external visual representations that are systematically related to the information that they represent[2]. The information represented may be related to the multiple abstract information of objects and events [2]. A large number of studies have shown that important information in visualization attracts the attention of the audience [2].

In this chapter, the history of data visualization is presented from the 17th century to the digital age. The big events of each age will be outlined and importance of data visualization and the psychology behind data visualization are discussed, in addition to best practice in data visualization.

Digital visualization means that the information to be conveyed by the data is presented to the users in a graphical manner such that professional users are supported in analyzing the data and non-professional users are supported in understanding the problem.

## 2.2: History of Data Visualization

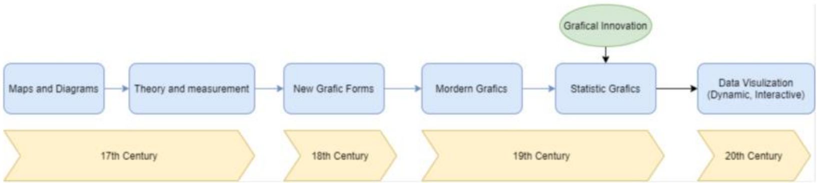


Figure 2.1: A brief overview of data visualization history

Visualization has a long history, it was first used to record the location of celestial bodies to help ancient people locate and make a map. People in this period were more focused on how to draw maps to better perform physical marking or map surveying. It is precisely because of the rise of navigation and mapping that the collection and representation of raw data and the proposal of some basic concepts have also laid the foundation for graphical representation.

In the 18th century, scientific research or statistical investigation produced a large amount of data, including qualitative and quantitative information [3]. With data collection in more and more fields, scientists urgently need to create some novel graphical representations to describe the attractiveness of data, which directly promoted the development of graphical representations in the early 18th century.

In the 19 century, based on previous design and technological innovation, statistical graphics and statistical data gradually showed explosive growth[3]. Most of the forms of modern data were basically produced during this period, such as the well-known pie charts, bar charts, line charts, scatter charts and histograms, etc [3](Figure 2.2). It was also during this period that people began to try to color graphics in order to highlight details for better visual distinction [4]. Therefore, this period is also called the golden period of graphic development.

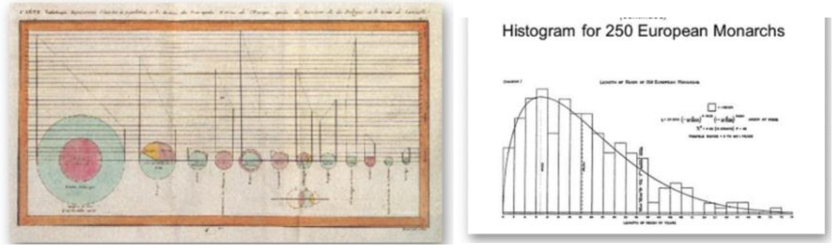


Figure 2.2: Pie charts and Histograms in 19th century [3], [5]

After the visualization of knowledge was popularized, a period of chart innovation was gradually ushered in, in the early 20th century. With the usefulness of graphical displays for understanding complex data and phenomena established, many new graphical forms were invented and extended to new areas of inquiry, particularly in the social, medical and business realm [6] e.g. ??????

It is also believed that the latest growth potential of data visualization comes from the development of interactive and dynamic graphics methods, real-time and direct manipulation of graphical objects and related statistical properties whereby classic linear statistical modeling is extended to a wider range of fields (general linear models, mixed models, spatial/geographic data models, etc.) [2].

### 2.3 The Importance of Data Visualization in the Medical Field

Through fields such as health data analysis and health informatics, medical organizations accrue large amounts of raw data, the patient's personal information, medical history, and attending doctors etc [7]. Dashboards developed by integrating this information can help medical staff quickly analyze large amounts of data to save time and even save lives. In one case study, the development of a dashboard to visualize electronic health record data resulted in a 65% reduction in time spent on data analysis in the first year alone, with further gains projected for the future [8].

There are many ways to build a visual dashboard. Medical organizations can synthesize raw data and convert it into graphics, charts and dashboards [7].Visualizing health data allows experts to present key trends and information via graphs, charts, and other visuals that show as well as tell [6]. Visually displaying health data points helps identify trends and significant "data" clusters [7]. What's more, it provides a way to share important findings with stakeholders who may not be data literate, such as hospital executives and administrators [7].

Patient-orientated visualization of medical information can improve their understanding and enhance the communication process, engaging the patient in an active information exchange and in the decision-making process[9]. This is essential as effective communication between providers and their patients has been shown to result in significantly improved medical outcomes[9].

Because the group of doctors far exceeds that of medical students and basic researchers [7]. Therefore, it is still important to consider the treatment plan from the perspective of the doctor. Research shows that digital graphics are easier to be recognized by the human brain than complex data. Knowledge of how to extract effective information from massive data will greatly help decision makers make accurate and rapid judgments. For example, if these target users are provided with better-to-use tools, this has a direct impact on patient health [10]. Electronic health records (EHRs) that process, organize, and visualize clinically meaningful information patterns were shown to significantly reduce physician cognitive workload in a study recently published in JAMA [11].

In summary, data Visualization can help us quickly locate and find the root cause of the problem. In the business field, saving time means saving a lot of expenses. This advantage is especially obvious in the medical industry-data visualization can help doctors quickly grasp the patient's real-time situation and analyse the cause so that they can prescribe the right medicine to save lives. Data visualization can greatly save resources, such as computer memory, and improve computer processing capabilities. Strong interaction. dynamic graphics can be a fixed unit in a certain period. Moreover, each team member can view and change the main data in real time, reducing the communication gap and cost.

### 2.4 The Psychology of Data Visualization and Communication

“The world you see is not necessarily the real world.”

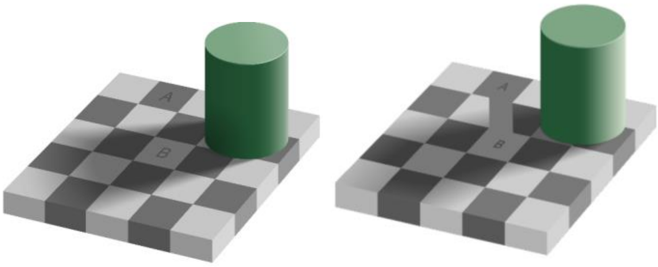


Figure 2.3 [12]: Checker shadow illusion

Figure [9] is the most famous Adelson’s same color illusion theory shows how our brain “Cheat” us. Your brain however perceives them differently, due to the surrounding colour and shadow details. More incredibly, is the fact that your brain continues to perceive square A as much darker than square B, even after learning that the two squares are the same colour [13].

The picture of conscious vision that emerges from these findings is that of limitation. We may think we see a full and detailed image of what is in front of our eyes, but we take in only a small subset of the information [14]. Cognitive psychology proves that the human brain can’t process a vast scope of information at the same time, our ability to have multiple consciousnesses at the same time is limited [15], [16]. This understanding makes us rely on the external form of information storage to enhance attention. One of the most powerful ways to do this is to encode information visually. Following encoding, viewers mentally search long-term memory for knowledge relevant for interpreting the visualization[2]. This knowledge is proposed to be in the form of a graph schema[2].

The visualization conventions associated with the graph schema can then help the viewer interpret the visualization(message assembly process) [2].

Data visualization is effective because it can change the balance between perception and cognition. It's psychological basis of data visualization is to use visualization to accurately convey information and allow users to understand the data with minimal effort [17]. Therefore, the graphical description used in data visualization should be subject to the following constraints [17]:

- Any information/data presented in visual form should be intuitive and accurate.
- Biases and traits subject to perceptual processing should be highlighted.
- Ensure that the message to be conveyed is not distorted.

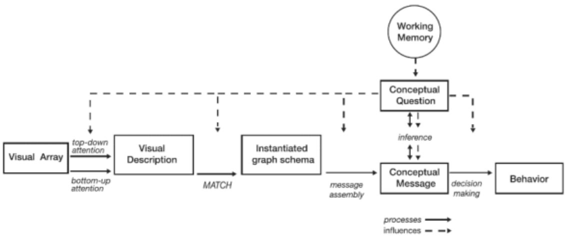
Data visualization technology also aims to make data manipulation and analysis possible. Therefore, the structure of the message needs to be compatible with the representation requirements and the preferences of human cognition processes [17]. This is one of the reasons why data modelling techniques used in data visualization should also be limited by the understanding of human memory and cognition.

### 2.4.1 Cognitive Data Visualization in decision making.

The extensive literature on human psychology demonstrates how the valence of an individual's emotional state (affect) appears to influence a number of cognitive functions including attention, memory and judgement[18]. Recent studies have used behavioural, physiological, and cognitive changes related to emotional responses as representatives of their valence1. Perception affects cognitive functions, such as attention and judgment[18].

In addition, lots of evidence that shows the visualization can be used to draws viewers' attention and reduce some decision bias[2]. The most common methods for demonstrating that visualizations focus viewers' attention is by showing that viewers miss non-salient but task-relevant information[2]. As pointed out by Card, Mackinlay, and Shneiderman (1999), we can use vision to think, which means that when the visual deviation changed by the visualization is consistent with the correct interpretation, the visualization can use visual perception to easily explain the visualization, it can even improve decision-making performance[2].

Also, data visualization can help people reduce the workload of calculation and overcome barriers between disciplines. A study shows how designers use the first type of processing methods(See below figure) to create visualizations to help viewers make accurate decisions based on complex data even if they lack relevant knowledge[2]. This point particularly useful for the visual description of health data because health data usually takes the form of probabilities, which is not intuitive[2]. Visualization can express probability numbers in natural frequencies through precise computer processing, such as histograms which is more intuitive. That means by visually depicting natural frequencies, viewers can make perceptual comparisons instead of mathematical calculations[2]. This dual benefit may be the reason why visualization provides convenience for people with low health literacy, graphics literacy, and computing power[2].



~~Figure X[2]:~~ Model of visualization decision making, which emphasizes the influence of working memory

### 2.4.2 Bias and Traits

Bias is the difference between our actual and predicted values. Bias is the simple assumptions that our model makes about our data to be able to predict new data[19]. As previously discussed, unsupervised learning models can help us to figure out the potential structure from a dataset (Chapter 2.7.2). Our model after training learns these patterns and applies them to the test set to predict them[19]. But when the Bias is high that means the model can't capture the important features of our data[19]. This means that our model hasn't captured patterns in the training data and hence cannot perform well on the testing data too[19]. If this is the case, our model cannot perform on new data and cannot be sent into production[19]. This instance, where the model cannot find patterns in our training set and hence fails for both seen and unseen data, is called Underfitting[19].

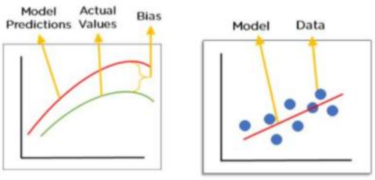


Figure 2.4 [19]: Bias and Underfitting

### 2.5 Data visualization Approaches

We have talked about the importance of data visualizations in Chapter 2.2, many advantages are outlined. Although visualization is a suitable tool, when visualizing big data, it is also important to choose the correct data representation [20],so we need to understand the advantages and disadvantages of each graph. Clear visual graphics can provide decision-makers with a scientific foundation for controlling the overall situation, quickly comprehending complicated and secret concepts or facts, and completely comprehending the actual situation, allowing them to make swift and targeted decisions.

#### 2.5.1 Visualization of Numerical Data

**Line Chart**

Line charts are often used for showcasing data patterns, they are clear and easy for the average user to analyse quickly[21]. Line charts can not only be used to observe the trend of data over time, but also compare the gap between different sets of data with the time changes, eg. comparison of the maximum and minimum values at a certain point in time. In addition, it can also estimate the trend of data with a scientific basis, and it often used to predict the market development trend.

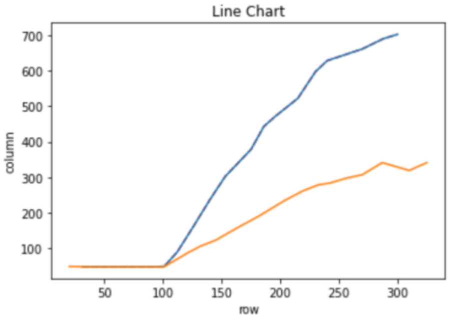


Figure 2.5: Line Chart

**Scatter Chart**

Scatter plots are useful for showing the association or correlation between two variables. A correlation can be quantified, such as a line of best fit, that too can be drawn as a line plot on the same chart, making the relationship clearer[22]. The scatter chart mainly explains the laws between the datasets. The scatter chart can not only show the relationship between two variables, but also reflect their distribution. If I want to see the distribution of variables, I can use either a scatter plot or a histogram.

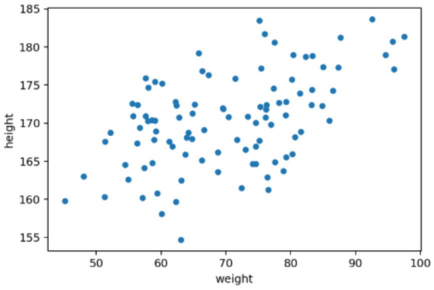


Figure 2.6: Scatter Plot

**Bar Chart**

Bar chart is more suitable for understanding the distribution and comparison between data.

Especially when we got data that has different categories, a bar graph is excellent for displaying the info[23]. We are able to easily compare several data sets and it's visually straightforward when someone reads it[23].

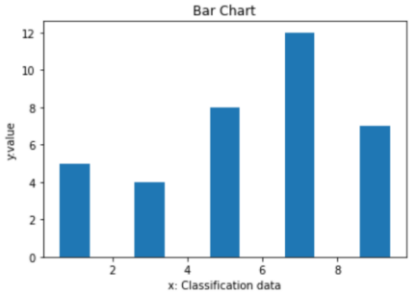


Figure 2.7: Bar Chart

**Histograms**

Another way to visualize the distribution of numeric variables is a histogram. A histogram is an accurate graphical representation of the distribution of numerical data[24]. The data distribution is represented by vertical stripes or line segments with different heights. Generally, the horizontal axis represents the data type, and the vertical axis represents the distribution. In other words, the histogram provides a view of the data density. It counts the number of times a value appears in each category and uses a bar graph to indicate its frequency. The more regular the values in the category, the larger the bars (Figure X). The benefit of histograms is we do not need to do mathematical calculation, our attention is naturally comparing the frequency.

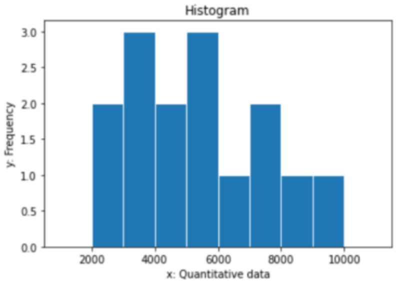


Figure 2.8: Histograms

**Pie Chart**

The pie chart can show the market share of various technologies well, but its disadvantage is that it cannot show the changes over time[25]. Qualitative variables can be represented using pie charts[25]. Table 1 shows various Machine Learning algorithms to compare the accuracy of them.

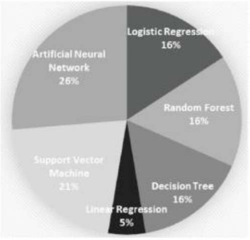


Figure 2.9[25]: Machine learning algorithms used in related works.

### Box Plot

Boxplots are useful to summarize the distribution of a data sample as an alternative to the histogram[22]. They can help to quickly get an idea of the range of common and sensible values in the box and in the whisker respectively[22]. Because we are not looking at the shape of the distribution explicitly, this method is often used when the data has an unknown or unusual distribution, such as non-Gaussian[22].

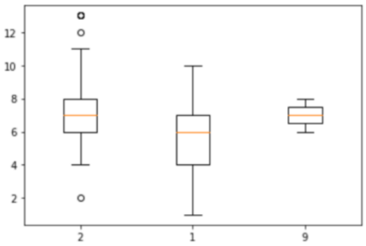


Figure 2.10: Box Plot

## 2.5.2 Visualization of Non-numerical Data

### Area Map

The area map can color the data by area to facilitate tracking and viewing of the area where the business is located.

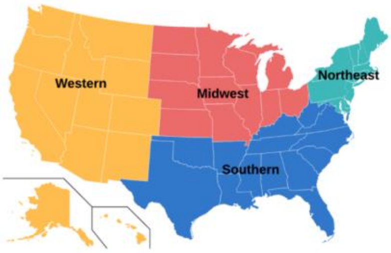
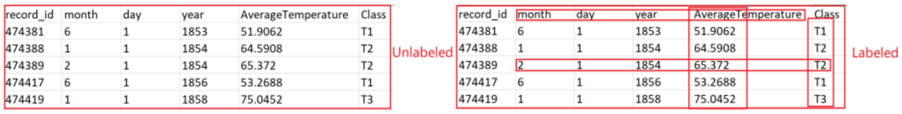


Figure 2.11: Area Map

### 2.5.3 Labeled vs Unlabeled data:



| record_id | month | day | year | AverageTemperature | Class |
|---|---|---|---|---|---|
| 474381 | 6 | 1 | 1853 | 51.9062 | T1 |
| 474388 | 1 | 1 | 1854 | 64.5908 | T2 |
| 474389 | 2 | 1 | 1854 | 65.372 | T2 |
| 474417 | 6 | 1 | 1856 | 53.2688 | T1 |
| 474419 | 1 | 1 | 1858 | 75.0452 | T3 |

Unlabeled

| record_id | month | day | year | AverageTemperature | Class |
|---|---|---|---|---|---|
| 474381 | 6 | 1 | 1853 | 51.9062 | T1 |
| 474388 | 1 | 1 | 1854 | 64.5908 | T2 |
| 474389 | 2 | 1 | 1854 | 65.372 | T2 |
| 474417 | 6 | 1 | 1856 | 53.2688 | T1 |
| 474419 | 1 | 1 | 1858 | 75.0452 | T3 |

Labeled

Table 2.1: Unlabelled data vs Labelled data

Before we understand what a dataset is, we must first understand what data is. Data is simply information, any time we have a table with information, we have data[26]. Normally, each row is a data point[26]. As figure shows the temperature dataset, each row represents the temperate that collected on specific date. Features are simply the columns of the table. As the figure shows, the features are day, month, year which uses to describe our data. In our figure, the last feature is special which is called labelled data.

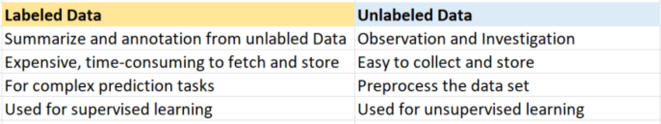
### Labelled Data

Data labelling typically starts by asking humans to make judgments about a given piece of unlabelled data[27]. That's because labelled data is much more valuable as it provides an accurate estimation of the conditions of our world[28]. It also shows understandable patterns and tells the machine what to look for[28]. This helps with advanced classification and building complex forecasting models[28]. After the ML algorithm is trained, it is able to find similar patterns in the new datasets that you feed into it[28]. Labeled data has high practical value, such as recommending purchases to customers, predicting stock market risks, weather forecasts, etc.

#### Unlabelled Data

Unlabelled data is also called raw data, and we usually refer to data that is easy to obtain but not easy to use[29]. Because it is available everywhere, it is relatively cheap and easy to store, and we don't need to invest time and resources on human annotators who will label the data.

Below is the summary of labelled data and unlabelled data.



| Labeled Data | Unlabeled Data |
|---|---|
| Summarize and annotation from unlabled Data | Observation and Investigation |
| Expensive, time-consuming to fetch and store | Easy to collect and store |
| For complex prediction tasks | Preprocess the data set |
| Used for supervised learning | Used for unsupervised learning |

Table 2.2: Comparison between labelled data and unlabelled data

### Labelled data and unlabelled data in Machine Learning

As we all know that machine learning has two learning models in general: supervised learning and unsupervised learning.

For supervised learning to work, you need a labelled set of data that the model can leam from to make correct decisions[27]. Data labelling typically starts by asking humans to make judgments about a given piece of unlabelled data[27].

While unsupervised learning (UL) is a machine learning algorithm that works with datasets without labelled responses[30]. It is most commonly used to find hidden patterns in large unlabelled datasets through cluster analysis that is to say dig out an learn the structures of the unlabelled data in order to simplify the unlabelled data or group it in accordance to the goals [30]. A common example is grouping the flower by their colour.

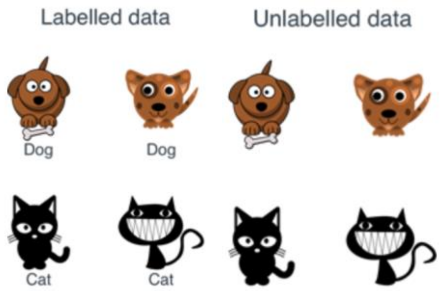


Figure 2.14[26]: Labelled vs Unlabelled data

We can also use supervised learning to figure out best prediction for the unlabeled data. The usual practice is divided the dataset into two parts, one part used for testing and the other part used for training purpose. Then combine the AI model to predict unsee data.

Take Figure dataset as an example, assume the dataset consists of “Dog” and “Cat”, machine learning will summarize experience from huge amount of history data, and then apply its learnings to do prediction. In this case, by using machine learning, we can summarize the characteristics of cats and dogs and classify and label these data. So when next time unlabelled data comes, based on its previous learning, we will quickly know the new guy is a “cat” or a “dog”.

### 2.5.4 numerical and categorical data

Generally, when examining an association, variables fit one of two types. The outcome variable, synonymous with the dependent variable, refers to the variable that we want to explain or predict as a result of the variation in the explanatory variable, or independent variable[31]. Both explanatory and outcome variables are further subcategorized by the distribution of the data as categorical or continuous variables[31].

When processing mathematical data, researchers need to consider and evaluate various forms of data, so analysts need to be familiar with the data forms used.

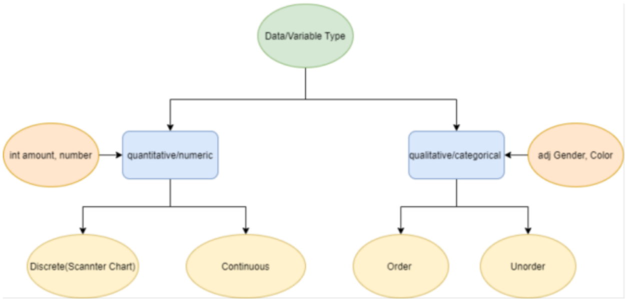


Figure 2.15: Data can be either quantitative or qualitative.

**Numerical data**

What is numerical data? Digital data is a data type represented by numbers, not natural language descriptions [12]. Numerical data, also called quantitative data, is usually collected in digital form and then expressed in a linear manner. For example, the warmest month of the year. The distribution of numerical data basically has two manifestations: discrete (representing countable objects) or continuous (representing data calculation).

For example, numerical statistics concerning the number of male and female students can be obtained, and then put together to obtain the overall number of students in the class. This function is one of the main ways to classify numerical results.

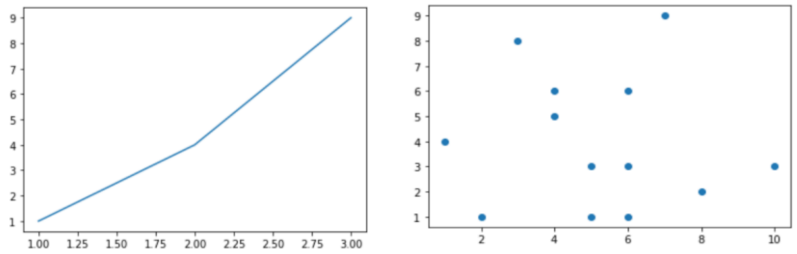


Figure 2.16: Continuous data vs Discrete data .

**Discrete data.** Discrete data is data that can only take certain values. These values do not have to be integers (for example, adult shoe size is 40 or 41), but these sizes are fixed values, that is, shoe size cannot be 40.2.

**Continuous Data.** Continuous data is data that may take any value. Weight, height, and age are examples of continuous data. Some constant data can adjust with time for example, the weight of a teenager during the first month of puberty, or the temperature and humidity of the room during the day. It is better to view this data on a line graph, so this type of graph will demonstrate how the data evolves over a given time span. Such constant statistics (e.g. freshman's height and weight) are typically clustered together to make it easier to read.

**Categorical Data.**

Categorical data is the data can not be measured and usually obtained from observation, it is also called absolute qualitative data. Take the product grade as an example: best, good, average, bad. Any string data (such as name, address, etc.) is also categorical data, it is usually an expression of characteristics. Bar charts, pie charts, and area map are often used to display results that contain non- categorical data.

### 2.6 Machine Learning in Data Visualization

#### 2.6.1 Machine Learning: Supervised Learning and Unsupervised Learning

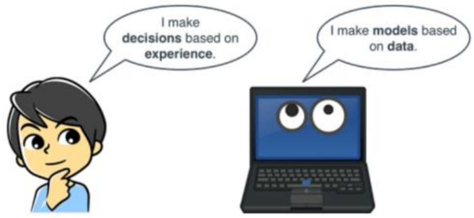

I make decisions based on experience.
I make models based on data.

Figure 2.17 [32]: Machine learning is when computers make decisions based on data.

Machine learning is a component of AI, which involves the training of algorithms or models that then give predictions about data it has yet to observe[33]. Machine learning techniques are used, for example, to automatically detect lesions in a follow-up exam that correspond to lesions in an earlier examination[10]. Thus, features are extracted and candidate regions are classified in order to determine corresponding pairs of lesions and to support their comparison[10]. Visualization techniques need to be considered that convey the results of such automatic processing[10]. This automatic processing cannot be 100% correct. Thus, users must also be enabled to verify the results or influence the parameters that guide the automatic process[10]. Moreover, datasets are analysed to identify probably relevant regions, such as certain organs, to focus the visual exploration to relevant subsets[10]. Again, it may be essential to enable the user to steer this process with appropriate parameters, including appropriate previews[10].

**Supervised Learning**

Supervised learning, similarly, as the name indicates, requires an operator to perform supervised corrections during the machine learning process. The operator provides a known data set containing the required inputs and outputs. Such machine learning algorithms must find an optimal way to determine how to derive these inputs and outputs[32]. Assuming that the operator knows the correct answer to the question, the algorithm can identify and analyse patterns in the data, learn from observations and make predictions accordingly[32]. This process will continue until the algorithm reaches a higher level of accuracy[32].

In general, using machine learning can solve three types of problems: Regression, Classification and Clustering. Most of the work done in Machine Learning has focused on supervised algorithms. Their main strength is that they produce models that we can incorporate in the decision-making process [34]. In order to choose the most suitable learning algorithm, a clear objective is required, and an analysis of previous data must be performing[34]. Thus, the feasibility of using a supervised algorithm over a not-supervised algorithm can be determined[34]. Supervised learning problems can be further grouped into regression and classification and prediction problems[32], [35].

**Supervised Learning: Regression**

Regression is a supervised learning problem that involves predicting a numerical label[36].At the same time, regression is a statistical method that estimates the strength of relationships between variables[26]. Regression-type problems are generally those where we attempt to predict the values of a continuous variable from one or more continuous and/or categorical predictor variables[37]. For example, the new health test data is compared with the existing known disease types to classify them, and the applications of this kind of program are numerous. If a variable is categorical it means that there is a finite/discrete number of groups or categories the variable can fit into[37].

There are mainly three major uses for regression analysis list as below:

**Determining the strength of predictors.** The regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable[38]. For example, how strong the relationship between eating habits and weight.

**Forecasting an effect.** It can be used to forecast effects or impact of changes[38]. That is, the regression analysis helps us to understand how much the dependent variable(Y) changes with a change in one or more independent variables(X)[38]. For example, the effect of having meat and vegetables on body weight.

**Trend forecasting.** regression analysis predicts trends and future values[38]. Example: To predict how much the patient would weigh after eating meat for a month.

### Supervised Learning: Classification

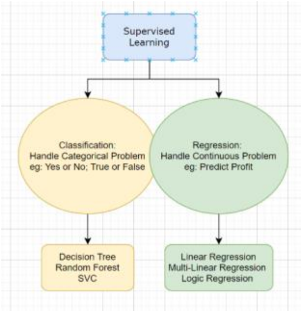

Supervised Learning
Classification: Handle Categorical Problem eg: Yes or No; True or False
Regression: Handle Continuous Problem eg: Predict Profit
Decision Tree Random Forest SVC
Linear Regression Multi-Linear Regression Logic Regression

Figure 2.18: Supervised Learning to solve Classification Problem.

In machine learning, there are two different types of supervised learning methods: classification and regression[37]. Both classification and regression problems may have one or more input variables and input variables may be any data type, such as numerical or categorical[36].

Classification is a supervised learning problem attempts to find relationships between variables and try to predict a state[26], [36], [37]. For example, trying to predict the type of animal (cat or dog). This supervised machine learning algorithm draws conclusions through observation and determines which type of new observations belong to and classifies them.

The main difference between classification and regression being the output of the model[37]. In a regression task, the output variable is numerical or continuous in nature, while for classification tasks the output variable is categorical or discrete in nature.

### Unsupervised Learning: Clustering

The author found that unsupervised learning can solve two types of problems: clustering and dimensionality reduction, in this section the author only illustrates the clustering part. The set of algorithms in which we use an unlabelled dataset, is called unsupervised learning[26]. In contrast, Unsupervised Learning focuses on finding data patterns[39].As there is no operator intervention and supervision, the program analyzes the data itself and summarizes a set of rules or summarizes the corresponding recognition mode. The algorithm attempts to organize data in a certain way to describe its structure and present certain data information. When the dataset is large enough, or the algorithm processes and evaluates more data sets, its decision-making ability will gradually improve and become more and more perfect. Obviously, this method is more rational and rigorous and is especially suitable for data collections with huge amounts of data. Unsupervised learning is a more difficult algorithm than supervised learning because we know little about the data or the expected results.

### 2.6.2 Machine learning models: Classification, Regression and Clustering

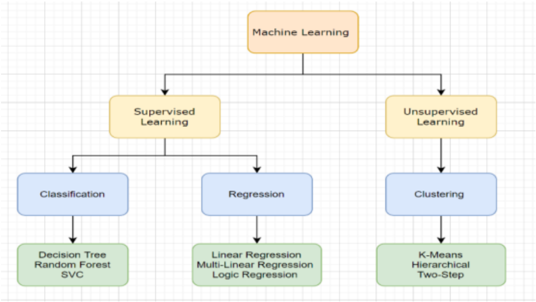


Figure 2.19: Machine Learning Models

As we already knew machine learning technique can be used to solve three types of problem: Regression, Classification and Clustering (See Chapter 2.5.2). Then which algorithms can be used to solve those problems? And how to choose the most suitable to solve specific questions? The author has summarized and outlined the machine learning models from two perspective (Figure above): Supervised and unsupervised learning methods. In the supervised learning methods, there are six most widely used algorithms are introduced: Decision Tree (DT), Random Forest (RF) and Support Vector Machine, Linear and multiple Linear Regression and Logic Regression. In supervised learning method, the author only simply introduced the clustering model to figure out the data pattern.

### Classification Model: Decision Tree

Before introducing random forests, we should firstly understand decision trees. A decision tree is a tree-based supervised learning method used to predict the output of a target variable[40]. Supervised learning uses labelled data (data with known output variables) to make predictions with the help of regression and classification algorithms[40]. Supervised learning algorithms act as a supervisor for training a model with a defined output variable. It learns from simple decision rules using the various data features[40].

Now let us see how decision trees affect decision making. Below figure clearly demonstrate how decision tree works. The decision tree distinguishes the animals based on its own features. The features as we talked before (Chapter 2.3.1), they are colour, height as the input. Here we also need to involve a concept of entropy. Entropy, it is a unit to describe the degree of confusion. In a word, The larger the value of entropy, the more confusion, and vice versa. By asking serious questions we can get two or more answers or output from each question. This process iterated until animals cannot be split. In this process, the entropy also decreases, which represents that the dispersion decreases and the target becomes orderly.

As the figure shows, we can make decisions using a tree structure. Each branch node represents a choice and leaf node represents a decision. From the graph we can easily figure out the biggest advantage of decision trees is it is simple to understand and to interpret and trees can be visualized[41]. It's disadvantage is decision makers may create too many complex trees, for example, irrelevant attributes can result in overfitting when training data set, they cannot superimpose the data well. This is called overfitting. In order to avoid this problem, some kind of top-end mechanism must be used to set the minimum number of samples required at the leaf routine or set the maximum depth of the tree.

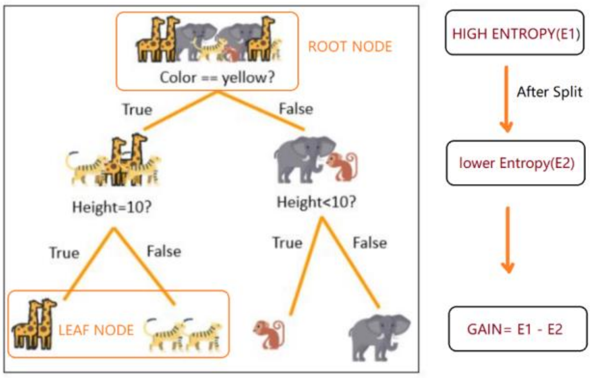


Figure 2.20[40]: Decision Tree

**Summary:**

Decision Trees are a supervised learning method used for classification and regression problems[41]. It is a highly used classifiers due to its simplicity for understanding and interpretation[34]. It requires little data preparation, handles numerical and categorical data, and performs very well with large data set in a short time[34]. Additionally, the hierarchical tree structure resembles a human way of decision-making, providing extending information about the sequence to classify and individually into a class, discovering rules in a more comprehensible manner[34].

**Classification Model: Random Forest**

As its name, random forest consists of a large collection of decision trees. Random forest is a supervised learning algorithm[42]. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method[42].
As decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures[43]. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees[43]. This process is known as bagging. The main idea is to try a combination of different models to determine the best effect[42].

One big advantage of random forest is that it can be used for both classification and regression problems[42]. Random forest adds additional randomness to the model, while growing the trees[42]. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features[42]. This results in a wide diversity that generally results in a better model[42]. Another advantage of random forest is it can significantly reduce the risk of overfitting as multiple trees are involved. Another advantage is it can run efficiently on large datasets and it needs less time in training. At the same time, the larger dataset, the more accuracy it predicts even though a large proportion of data is missing.

So in our random forest, we end up with trees that are not only trained on different sets of data (thanks to bagging) but also use different features to make decisions[43].

Random forest algorithm is used in many different fields, in the healthcare domain it is used to identify the correct combination of components in medicine and to analyse a patient’s medical history to identify diseases[42]. In e-commerce domain it used to determine customer preferences or predict profit.

**Summary:**

Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction[42]. It is a great algorithm to train early in the model development process, to see how it performs[42].

Taking the travelling example, if Peter asks more neighbour B, neighbour C and so on. Let’s assume most people suggest City D, then Peter finally decided the best city he would like to go. This is the general idea of Random Forest- it is similar to a voting mechanism, each tree voted for the solution that it thinks is the best.

**Classification Model: Support Vector Machines**

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems[44]. The core idea of SVM is to find a maximum marginal hyperplane(MMH) in multidimensional space that best divides the dataset into classes[44]. More accurately, SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error[44]. Compared with other classifiers (such as logistic regression and decision trees), SVM provides very high accuracy and it's known for its kernel skills for processing nonlinear input spaces. It can easily handle multiple continuous and categorical variables[44].

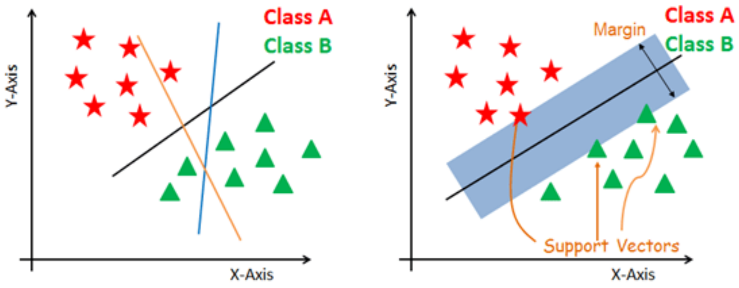


Figure 2.21 [44]: SVM

Now we know what is SVM, then how it works?

As previously stated, the main objective is to select a hyperplane with the maximum possible margin and segregate the given dataset in the best possible way. As the figure shows, the distance between the either nearest points is known as the margin[44]. SVM searches for the maximum marginal hyperplane in the following steps:

1. Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly[44].

2. Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure[44].

**Summary:**

SVM is a supervised machine learning algorithm which can be used for classification or regression problems[45]. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs[45]. Simply put, it does some extremely complex data transformations, then figures out how to separate your data based on the labels or outputs you've defined[45]. The purpose of svc is to classify the data set as much as possible.

**Regression Model: Linear Regression**

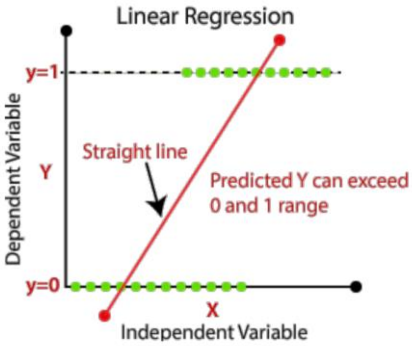


Figure 2.22 [19]: Regression Model-Linear Regression

The linear regression algorithm is one of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties[46]. In statistics, linear regression is a linear approach to modelling the strength of relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables, see Figure X)[46]. The Linear Regression model attempts to find the relationship between variables by finding the best fit line(the red line in the figure)[47]. Technically, in regression analysis, the independent variable is usually called the predictor variable and the dependent variable is called the criterion variable[48]. When implementing linear regression in a machine learning system, the variables must be continuous in nature, not categorical[37]. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data[49]. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest[49]. This does not necessarily imply that one variable causes the other, but that there is some significant association between the two variables[49].

Simple linear regression plots an independent variable X for a dependent variable Y [3]. If we increase the independent variable X from X0. . Xn, the process is called multiple linear regression[38], [48]. Regression analysis can result in linear or nonlinear graphs. A linear regression is where the relationships between your variables can be described with a straight line while non-linear regressions produce curved lines[48]

**Summary:**

In general, regression is a statistical method that estimates relationships between variables. Classification also attempts to find relationships between variables, with the main difference between classification and regression being the output of the model[37].In a regression task, the output variable is numerical or continuous in nature, while for classification tasks the output variable is categorical or discrete in nature. If a variable is categorical it means that there is a finite/discrete number of groups or categories the variable can fit into[37].

**Regression Model: Logic Regression**

Logistic regression is a statistical method that is used to describe data and the relationship between one dependent variable and one or more independent variables[50]. Logic Regression is a classification algorithm, used to predict binary outcomes for a given set of independent variables. The dependent variable's outcome is discrete. Application of logic regression is identified the different components that are present in the image or dataset and helps to categorize them. For example, determines the possibility of patient survival, taking the weight, age, blood fat into consideration.

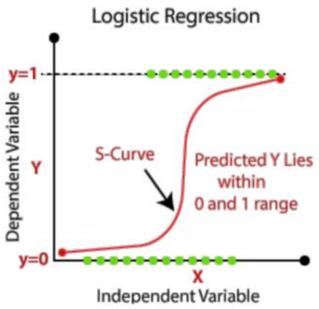


Figure 2.23[19] : Logic Regression

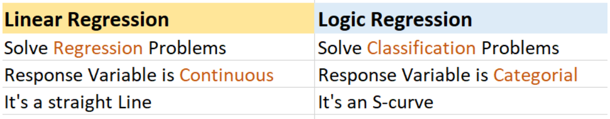
**Logic Regression formular**[51]**:**

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x)[51]. Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data[51]. In practical application, logic regression can help to predict the weather or help the doctor to see the trend of a disease based on the patient's history.

**Summary**:

Linear Regression used to solve Regression Problems while Logic Regression used to solve classification problems. The response variable of Linear Regression is continuous while Logic Regression is categorial. Logic Regression is used to predict the happening probability of an event.



| **Linear Regression** | **Logic Regression** |
|---|---|
| Solve Regression Problems | Solve Classification Problems |
| Response Variable is Continuous | Response Variable is Categorial |
| It's a straight Line | It's an S-curve |

Table 2.3: Linear Regression vs Logic Regresion

**Clustering Model**

The set of algorithms in which we use an unlabeled dataset, is called unsupervised learning[26]

As the name implies, there is no operator intervention and supervision. The program analyzes the data itself and summarizes a set of rules or summarizes the corresponding recognition mode. The algorithm attempts to organize data in a certain way to describe its structure and present certain data information. When the dataset is large enough, or the algorithm processes and evaluates more data sets, its decision-making ability will gradually improve and become more and more perfect. Obviously, this method is more rational and rigorous and is especially suitable for data collections with huge amounts of data. The author think Unsupervised learning is a more difficult algorithm than supervised learning because we don't know much about the data or the expected results.

### Clustering Analysis

In basic terms, the objective of clustering is to find different groups within the elements in the data[52]. To do so, clustering algorithms find the structure in the data so that elements of the same cluster (or group) are more similar to each other than to those from different clusters[52].

Clustering analysis contains several methods: K-means, Hierarchical Clustering and Two-Step. Due to space limitations, the author only introduces K-means method here.

#### K-means

K-Means algorithms are extremely easy to implement and very efficient computationally speaking[52]. This algorithm can deal with both numeric (continuous) and categorical data[53]. The K-Means algorithms aims to find and group in classes the data points that have high similarity between them[52]. In the terms of the algorithm, this similiarity is understood as the opposite of the distance between datapoints[52]. The closer the data points are, the more similar and more likely to belong to the same cluster they will be[52].

But they are not very good to identify classes when dealing with in groups that do not have a spherical distribution shape[52].

### Steps of K-means

1. Choosing the right number of clusters—that is the K value by using Elbow Method. A fundamental step for any unsupervised algorithm is to determine the optimal number of

clusters into which the data may be clustered56]. The Elbow Method is one of the most popular methods to determine this optimal value of k[54].

In short, as the value of K increases, that is, the more groups and the finer the number of groups, the smaller the distortion will be. Each cluster will have fewer constituent instances, and these instances will be closer to their respective centroids[55]. However, when the value of K increases to a certain extent, the improvement in average distortion will decrease as K increases. The value of K at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters[55]. Take below example, we can see the elbow point is 3 in this case.

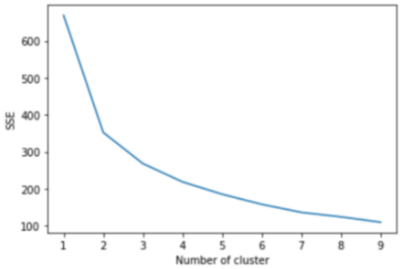


Figure 2.24: Elbow Method

### Hierarchical Clustering

Hierarchical clustering is an alternative to prototype-based clustering algorithms, is a bottom-up approach such that each data point begins in a separate cluster, and pairs of clusters at the bottom are merged together as we go up the hierarchy[52], [53]. The main advantage of Hierarchical clustering is that we do not need to specify the number of clusters, it will find it by itself[52]. In addition, it enables the plotting of dendrograms. Dendrograms are visualizations of a binary hierarchical clustering[52].

### 2.6.3 The workflow of Machine Learning

**Step1: Fetch Data.** This process depends on your project and data type [26].

We can fetch the data from various resources such as the real-time IoT system or static data from existing database or use online data such as google.

**Step2: Data Cleaning.** As we discussed before(in Chapter 2), the raw data is compiled from multiple sources, therefore, the information is displayed in various formats, for example, data may have errors or some columns may be empty [29], [56]. Data cleaning refers to the identification and correction of possible errors in the data set that have an impact on the predictive model. It is the process by which we address missing values and outliers, for Machine Learning success, after we chose our data, we need to clean, prepare, and manipulate the data[39], [56].

This process is a critical step, and people typically spend up to 80% of their time in this stage. Having a clean data set helps with your model's accuracy down the road[39]. After getting the data to a state you like, you need to convert the data sets into valid formats for your chosen ML platform[39]. Finally, you split your data into training and test data sets[39]. The training set is used to train the model in the next step, while the test data is used to validate the model in the fourth step[39].

**Step3: Model Selection.**

According to the knowledge summarized in Chapter 2.6.2, we need to make a rough screening of the model according to the data type of the variables in the data set that need to be judged and summarized. For example, for classification problems, we can choose SVC, decision tree, random forest; for the regression problem, we can choose linear or logic regression and so on.

Once we decide our problem type, we can use programming languages to build models.

**Step4: Train and Test Model.**

In this step, the data set connects to an algorithm, and the algorithm leverages sophisticated mathematical modelling to learn and develop predictions[39]. The dataset used to be divided into two parts, the larger dataset often used for training purpose and the smaller is used for test.

**Step5: Visualization.**

After going through all above four steps, we then used the analysis results to run data visualization. It can give the user a more intuitive feeling to observe the data, for example, the relationship between variables, the trend of the data etc.

~~Within the field of machine learning, there are two main types of tasks: supervised and unsupervised[33].~~

### 2.6.4 ETL Process

ETL process allows target data to be generated by combining data from multiple related source tables[57]. And the benefit of the ETL rules are composed in a human-readable format, allowing personnel with limited database programming expertise to compose, read, verify, and maintain them[57]. To simplify, ETL contains three steps: Extract, Load and Transform. In the data extraction step, required data elements from the source system are extracted to a temporary data storage from which they are transformed and loaded into the target database [57]. In the transform step, database programmers implement methods of data transformation and the schema mappings for loading data into the harmonized schema[57]. This process includes data "cleaning" (for example, deleting empty values or duplicate data) and establishing standardization (for example, standardizing the format of output variables, etc.), the process may iterate until the converted data is correct and complete, then it can be stored into the target database. This phase requires manually programming using database-programming languages such as structured query language (SQL) [57]. The loading step is much easier, just read the data from the target database. Below figure is the whole workflow of the ETL process:

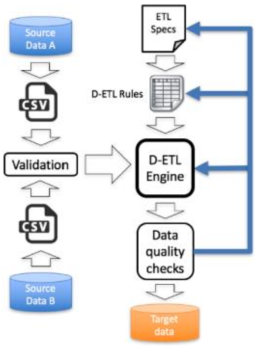


Figure 2.25[57]: Workflows of D-ETL approach to integration two source datasets

## 2.7 Statistical Application in Data Visualization

### 2.7.1 Introduction

In this part, the author discussed two statistical methods (parametric and non-parametric) and explained how to select appropriate statistical methods for analysis and interpretation of the data.

In biostatistics, for each of the specific situation, statistical methods are available for analysis and interpretation of the data[58]. To select the appropriate statistical method, one need to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected for data analysis[58].

Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean and median and another is inferential statistics, which draw conclusions from data using statistical tests such as student's *t*-test[58]. Selection of appropriate statistical method depends on the following three things: Aim and objective of the study, Type and distribution of the data used, and Nature of the observations (paired/unpaired) [58]. Other than knowledge of the statistical methods, another very important aspect is nature and type of the data collected and objective of the study because as per objective, corresponding statistical methods are selected which are suitable on given data[58].

Suppose our objective is to find out the predictors of the outcome variable, then regression analysis is used while to compare the means between two independent samples, unpaired samples t-test is used[58].

### 2.7.2 Statistical Concepts

The author after investigation found the purpose of the difference study is to compare the difference between data or multiple sets of data. It usually includes the following types of analysis methods, which are analysis of variance, T test and Chi-square test. Before introducing those statistical methods, it is mandatory to know the concept of correlation coefficient and p value.

**correlation coefficient**

The correlation coefficient, *r*, tells us about the strength and direction of the linear relationship between x and y[59]. A correlation is useful when you want to see the relationship between two (or more) normally distributed interval variables[60].

Relationships between variables. The correlation coefficient measures the degree of linear association between two variables, with a value in the range +1 to -1[61]. Positive values indicate that the two variables increase and decrease together[61]; negative values that one increases as the other decreases[61]. A correlation coefficient of zero indicates no linear relationship between the two variables[61].

The P value is used all over statistics, from t-tests to regression analysis, P values always used to determine statistical significance in a hypothesis test[62]. Using the p-value method, you could choose any appropriate significance level you want[59]. Assumes that we are using a significance level of 5%, $\alpha = 0.05$.

If p-value < ($\alpha$ =0.05), that means there is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero and we can reject the null hypothesis[59].

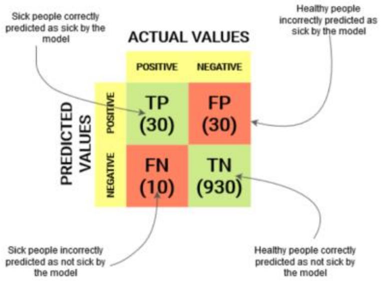
If p-value>= ($\alpha$ =0.05),that is to say there is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero and we cannot reject the null hypothesis[59].

Note that statistical significance does not mean that the results you have obtained actually have value in the context of your research[61]. If you have a large enough sample, a very small difference between groups can be identified as statistically significant, but such a small difference may be irrelevant in practice[61]. On the other hand, an apparently large difference may not be statistically significant in a small sample, due to the variation within the groups being compared[61].

**Confusion matrix**

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes[63]. The matrix compares the actual target values with those predicted by the machine learning model[63]. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making[63].

Below is picture clearly explained why and how the confusion matrix helps:



Sick people correctly predicted as sick by the model
ACTUAL VALUES
Healthy people incorrectly predicted as sick by the model
POSITIVE
NEGATIVE
PREDICTED VALUES
POSITIVE
NEGATIVE
TP (30)
FP (30)
FN (10)
TN (930)
Sick people incorrectly predicted as not sick by the model
Healthy people correctly predicted as not sick by the model

Figure 2.12 [63]: Confusion Matrix

**Sensitivity / True Positive Rate / Recall**

Sensitivity = $\frac{TP}{TP\ +\ FN}$

Sensitivity tells us what proportion of the positive class got correctly classified[63]. A simple example would be determining what percentage of the real sick people the model correctly identified.

**Specificity / True Negative Rate**

Specificity = $\frac{TN}{TN\ +\ FP}$

Specificity tells us what proportion of the negative class got correctly classified[63]..

Taking the same example as in Sensitivity, Specificity would mean determining the proportion of healthy people who were correctly identified by the model.

**False Negative Rate**

FNR = $\frac{FN}{TP\ +\ FN}$

False Negative Rate (FNR) tells us what proportion of the positive class got incorrectly classified by the classifier[63].A higher TPR and a lower FNR is desirable since we want to correctly classify the positive class.

**ROC-AUC**

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems[64]. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’[64]. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve[64]. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. It tells how much the model is capable of distinguishing between classes[63]. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease[63]. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis[64]. In a ROC curve, a higher X-axis value indicates a higher number of False positives than True negatives[64]s.

Take below figure as an example, according to the concept of confusion matrix A point has the highest Sensitivity and the lowest Specificity. That is to say all the negative points are classified incorrectly, and the positive points are classified correctly. B point compared with A point, it has same Sensitivity but higher Specificity.

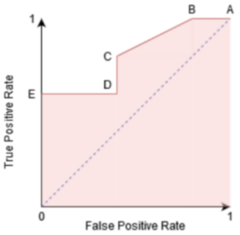


Figure 2.13 [64]:

### 2.7.3 Testing method

If the data we want to analyse is numerical data, then the corresponding statistical method should be selected. Statistical methods can be used to analyse averages, frequencies, patterns, and correlations between variables[65]. When creating the research design, one should clearly define the variables and formulate hypotheses about the relations between them[65]. Then we can choose appropriate statistical methods to test these hypotheses[65].

- **T test**

A t-test is a statistical test that is used to compare the means of two groups[66]. T-tests are a statistical way of testing a hypothesis when we do not know the population variance and our sample size is small, n < 30[67]. We perform a One-Sample t-test when we want to compare a sample mean with the population mean[67]. The difference from the Z Test is that we do not have the information on Population Variance here[67]. We use the sample standard deviation instead of population standard deviation in this case[67].It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another[66].

There are three types of T test: one sample t-test, two independent samples t-test and a paired t-test.

A one sample t-test allows us to test whether a sample mean (of a normally distributed interval variable) significantly differs from a hypothesized value[60]. Its purpose is to test whether the mean of the sample is equal to the mean of the known population. Example is when we want to test whether the average weight differs significantly than 6 pounds.

The independent t-test, also called the two sample t-test, independent-samples t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups[68]. Its purpose is to test whether the means of two independent samples are equal. Example is we want to check whether method of checking the boy and girl’s average weight is same.

A paired (samples) t-test is used when you have two related observations (i.e., two observations per subject) and you want to see if the means on these two normally distributed interval variables differ from one another[60]. Example is whether the boy and girl’s average weight is equal.

The null hypothesis for the independent t-test is that the population means from the two unrelated groups are equal[68]:

$$H0: u1 = u2$$

In most cases, we are looking to see if we can show that we can reject the null hypothesis and accept the alternative hypothesis, which is that the population means are not equal[68]:

$$HA: u1 \neq u2$$

Here, we need to involve the correlation coefficient concept to tell us the significance between variables that whether or not to accept or reject the alternative hypothesis.

Through investigation and research, the author concluded that whether it is a single-sample T-test, independent-sample T-test or paired-sample T-test, there are several basic premises:

(1) T test is a parametric test, which is used to test quantitative data (numbers have comparative significance). If the data are all qualitative data, non-parametric test is used.

(2) The sample data obeys normal or approximately normal distribution. If it does not meet the requirements, non-parametric testing can be considered.

(3) Check if the variances are equal. This can be done using the Levene’s test. If the variances of groups are equal, the p-value should be greater than 0.05[69].

- **Welch’s Test**

The independent samples t-test corrected for unequal variances is commonly known as Welch’s test, and is widely considered to be a robust alternative to the independent samples t-test[70].

When sample sizes are equal and variances are equal, both the independent samples t-test and Welch’s test perform similarly[70]. For unequal sample sizes and unequal variances, Welch’s test has superior Type I error robustness (Fagerland & Sandvik, 2009). Ruxton (2006) advocates the routine use of Welch’s test[70].

Here we need to use the Levene’s test. Levene’s test is an equal variance test, it can be used to check if our data sets fulfill the homogeneity of variance assumption before we perform the t-test or Analysis of Variance (ANOVA)[71].

- **Z Test**

z tests are a statistical way of testing a hypothesis when either we know the population variance, or we do not know the population variance but our sample size is large $n \geq 30$[67].

If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test[67]. We perform the One-Sample Z test when we want to compare a sample mean with the population mean[67].

- **Che-square Test**

Chi-Square test is used when we perform hypothesis testing on two categorical variables from a single population or we can say that to compare categorical variables from a single population[72]. By this we find is there any significant association between the two categorical variables[72].

An example is that if a researcher wants to understand the differences in users' preferences for electronic products, the researcher can select chi-square analysis and set options (single choice or multiple choice). Since chi-square analysis is used to compare the relationship between the classified data, it can be used to analyse the selection frequency or proportion of the classified data for comparison and difference analysis.

### 2.7.4 Summary

Here is the summary from the author:

| Data Type of X | Data Type of Y | Groups | Statistical Methods |
|---|---|---|---|
| Categorical | Numerical | 2 or more | Variance analysis |
| Categorical | Numerical | Only 2 | T test/Welch's test |
| Categorical | Categorical | 2 or more | Chi-square test |

Table 2.4: Summary of Statistical Approaches

As the table shows, the core difference between the three methods is the type of data is different. If X and Y axis data type are all categorical, then we need to use Chi-square test. If data type of X axis is categorial while Y is numerical, then we can use T test or Variance analysis. And the main difference between Variance analysis and T test is, T test can only compare two groups like compare the average weight between boys and girls. If there are more than 2 categories of independent variable: X like bad, good, excellent, then the variance analysis is a good choice.

According to above summary, the author generated an idea that we can based on the various type and number of groups to suggest the user a general statistical method (A logical judgement). As some test may have specific or more conditions, therefore, we probably need a second time logical judgement to give an accurate conclusion.

In a z-test, the sample is assumed to be normally distributed[73]. A z-score is calculated with population parameters such as “population mean” and “population standard deviation” and is used to validate a hypothesis that the sample drawn belongs to the same population[73].

Like a z-test, a t-test also assumes a normal distribution of the sample[73]. A t-test is used when the population parameters (mean and standard deviation) are not known[73].

ANOVA, also known as analysis of variance, is used to compare multiple (three or more) samples with a single test[73].

# Chapter 3. Research Methodology

## 3.1 Introduction

According to the American sociologist Earl Robert Babbie, “research is a systematic inquiry to describe, explain, predict, and control the observed phenomenon[74]”. A research purpose is met through forming hypotheses, collecting data, analysing results, forming conclusions, implementing findings into real-life applications and forming new research questions[75].

In this chapter, the author summarizes the main research problems and clearly puts forward the research goals, showing how to choose specific research methods to achieve the purpose of solving research problems. And introduced the research methods and limitations of this research.

## 3.2 Current Research Gaps

**Communication GAP.** What doctors think they are telling hospital patients, and what those patients actually hear, may be very different, a small study suggests[76].The findings, from a study of 89 patients at one U.S. hospital, add to research showing that doctors and patients are often not on the same page when discussing diagnoses and treatment[76]. The current study also shows that many doctors mistakenly believe their patients know more than they do[76].

**Recognition GAP**. The many burdens associated with the gap between patients and providers consist of the patient's anxiety and fears as well as the doctor's burden and stress of work[77]. Finding the bridges between these gaps and barriers can be beneficial in improving healthcare overall[77].Patients who understand their doctors are more likely to engage in healthier behaviours, understand their treatment options and follow their medication or check-up schedules[77]. In fact, research has proven that effective patient-provider communication can improve a patient's health as much as many drugs can[77]. Communication techniques are centralized in delivering any form of medical advice including medications[77].

**High Workload**. Quickly digesting a great deal of information about a patient, determining a diagnosis, and creating a treatment plan is a critical part of the training physicians’ complete medical school[78]. However, medical knowledge has exploded and the array of diagnostic tests and treatments has increased significantly[78]. It has become more challenging for physicians to decide on an optimal diagnosis and treatment plan. Interpreting new medical information can be difficult, especially when studies are sometimes contradictory[78]. Even a well-trained, intelligent physician simply cannot keep pace with all the latest research or have all of the knowledge available to them for every medical event[78].

**Lack of Personalized dashboard for the public.** Visualizations—visual representations of information, depicted in graphics—are studied by researchers in numerous ways, ranging from the study of the basic principles of creating visualizations, to the cognitive processes underlying their use, as well as how visualizations communicate complex information (such as in medical risk or spatial patterns)[2]. However, findings from different domains are rarely shared across domains though there may be domain-general principles underlying visualizations and their use[2]. The limited cross-domain communication may be due to a lack of a unifying cognitive framework[2]. For example, there are many statistical software nowadays like SPSS, R, Stata, and SAS are available and using these software, we can easily perform the statistical analysis but selection of appropriate statistical test is still a difficult task for the biomedical researchers especially those with nonstatistical background[58].

In summary, most dashboards on the market are aimed at professionals which are difficult to understand and operate, and most of them are static. Therefore, the author considers designing a personalized, interactive dashboard that can lead the users to dig out the relations of the disease data, try best to interpret the data in a more understandable way. This investigation could be enhanced further to below findings include:

1. Help them show the patient how they compare to the average.
2. Help the clinicians decide what model to use in machine learning.
3. Help the clinicians decide what approach to use in statistical.
4. Help them interpret the results.

### 3.3 Research Questions and Objectives

The primary research question of this thesis is “How to create an autonomous analysis health dashboard to the users in different stage?” This question has led to further research questions as follows:

**Research Question 1:** How to show a patient how they sit compared to the average.

**Objective:** Through extensive investigation and research, the author found that both doctors and patients like to compare with the average level as a criterion for judging their own health level, which seems to be a psychological effect. The most common example is that an adult man can quickly know whether he is overweight or underweight by comparing his weight with other people of the same age. In this article, the author will try to figure out how to show a patient can compare to the average.

**Research Question 2:** How to support doctors in selecting the ‘best’ statistical approaches?

**Objective:** Selection of appropriate statistical method is very important step in analysis of biomedical data[58]. A wrong selection of the statistical method not only creates some serious problem during the interpretation of the findings but also affects the conclusion of the study[58]. Therefore, the author will try to design a logic that can help the doctor automatically find the best statistical approach.

**Research Question 3:** How to help doctors in selecting the ‘best’ ML approaches?

**Objective:** As the doctors are quite busy in their daily work, they have no time and no such professional knowledge of how to select best ML approaches. For example, when to use Logic Regression, Clustering Analysis or variance analysis? What is the strength and limitations of them? Which model should the doctor use if they want to check the data relationship or predict the patient’s health status after taking the medicine? In addition, to the doctors it may be easier to understand those medical terminologies. While the ordinary users like patients they probably will find it difficult to understand what AUC is, specificity etc. Thus, the author will try to make a conclusion to suggest the ‘best’ ML model by comparing the AUC, sensitivity etc.

**Research Question 4:** How to use best practise in data visualisation to support users in the interpretation of the results from the ML/statistical approaches.

**Objective:** Although we have learned to collect and store large amounts of data, the data is so large that it is difficult for us to understand them on our own [79], [80].Artificial intelligence software can help humans create better computer vision models, and can obtain huge data sets, and show us various patterns and correlations, so as to help our intelligence play the best role [80], [81]. However, the output of the artificial intelligence development process is usually called the "black box" because it is not created by humans, so it is not easy to be interpreted or explained by humans[82]. As discussed in Chapter 2.4, data visualization enhances the interpretability of artificial intelligence. Therefore, the author tried to discuss how to interpret the results through a serious of graphs.

### 3.4 Sample Selection

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population[83].

Probability Sampling--this Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample[84]. It's alternatively known as random sampling and mainly used in quantitative research[84], [85]. While in non-probability sampling, the researcher chooses members for research at random[83]. Sample selection is a key factor in research design and can determine whether research questions will be answered before the study has even begun[86].

In our article, the author chose a non-probability sampling method because we need to randomly select the sample and be able to represent the public. Random sampling can ensure that the experimental results are objective and fair. Regardless of the educational or professional background of the participants, objective and fair random sampling can better test whether the public can benefit from the personalized dashboard.

Twelve people were invited to participate, and ten of them completed the interview. Before analysis, record and record the interview. Participants will be rewarded with a small cake. The author customized a guide based on the interview results, stipulating and explaining the design requirements of the analytical dashboard. A human-centered design team meets twice a week to evaluate and improve solutions to design deficiencies through discussions. The development front-end prototype used in the first three rounds of interviews was followed by feedback from people with medical backgrounds and dashboards in the fourth round of interviews, and the design was iteratively improved in the process.

### 3.5 Research Methodology

Now according to our research question, we need to choose the type of our research. If the objectives involve describing subjective experiences, interpreting meanings, and understanding concepts, we will need to do qualitative research[65]. If the objectives involve measuring variables, finding frequencies or correlations, and testing hypotheses, we will need to do quantitative research[65]. Quantitative research is expressed in numbers and graphs. It is used to test or confirm theories and assumptions[87]. Common quantitative methods is more datadriven research that uses surveys or questionnaires to derive numerical-based statistics or percentages[88].

In our article, the author has used hybrid the qualitative and quantitative methods in order to improve the practical use of the app. The methods include the focus group, observations, interview, case study and a questionnaire.

**Literature review:** This research is based on the author's extensive literature review to study how to solve the gap of the current topic. While completing the literature review, the author interacted with the current application.

This approach was required to address research questions 1 - 4.

**Focus Group:** The focus group method is a technique of group interview that generates data through the opinions expressed by participants[89]. A focus group is a group of deliberately selected people who participate in a facilitated discussion to obtain consumer perceptions about a particular topic or area of interest[90]. A focus group is qualitative research because it asks participants for open-ended responses conveying thoughts or feelings[88]. Some authors have recommended a minimum of three to four group meetings for simple research topics (Burrows & Kendall, 1997). The principle of theoretical saturation, where focus group discussion sessions are run until a clear pattern emerges and subsequent groups produce no new information (Krueger, 1994) .As the purpose of this research is to customize a personalized dashboard to visualize data through rigorous calculations and comparisons, which will help users make correct decisions and choices, and reduce cognitive pressure. Therefore, the author conducted the focus group. Focus group discussion usually yields both qualitative and observational data where analyses can be demanding[90]. The author insist on following the best practices when conducting the focus group[91] :

**Research design:**

- **Participants**: As per best practice, ten participants are considered large enough to gain a variety of perspectives and small enough not to become disorderly or fragmented[91]. And it needs to be carefully chosen by different methods, for example, by questionnaires or telephone.
- **Environment**: Researchers must choose a comfortable and safe environment for participants to minimize distractions.

**Data collection:**

- **Moderator**: The host always needs to create a relaxed and comfortable environment for participation and is also responsible for observing the behaviour of participants, taking notes or collecting data through recording. The length of the meeting must also be fully considered, because the longer the meeting time, the older the elderly are more likely to fatigue, and children will recover and lose the interest and attention of the project.

**Analysis**:

- Since focus group discussions will generate a lot of data (usually qualitative and observational data), the data needs to be analysed. Quantitative or qualitative analysis methods or a mixture of the two are usually used.

**Reporting of results:**

- After completing the above steps to analyze and process all the data, the researcher needs to make key decisions about the audience and tailor the target audience needs accordingly, and at the same time organize and merge the research results into a report for distribution. In addition to recording the key issues raised or emphasized by the participants, the report should also record the participants' information, such as gender, age, and education level.

Above method is needed to solve research questions 1-4.

**Observations:** Observational studies are non-experimental in nature, it's advantage is the investigator observes individuals without manipulation or intervention[92], [93].

**Interview:** An interview is generally a qualitative research technique[94]. which involves asking open-ended questions to converse with respondents and collect elicit data about a subject. Semi-structured interviews are used because the principles of human-centered design were used to translate these interviews into design requirements and to iteratively develop a front-end performance feedback dashboard[95]. Pre- and post- surveys were used to evaluate the effect of the dashboard on physicians' motivation and to measure their perception of the usefulness of the dashboard[95].

To ensure that the dashboard can be used normally, the author conducted one-on-one semi-structured interviews with each participant on some issues regarding the use of the program (see Appendix A for details) after each focus group meeting. The author uses a mixture of open and closed test questions, which allows participants to control their progress faster or participate more deeply.

**Case Study:** The case study approach allows in-depth, multi-faceted explorations of complex issues in their real-life settings[96]. In this project, the purpose of the use case study is to test whether the newly created application program satisfies the design requirements of different users well, whether it meets the practicability and ease of use at the same time; whether the functional requirements are fulfilled, and so on. For a complete case study analysis of this project, please refer to Chapter 6.

**Questionnaire**: A questionnaire is a research instrument consisting of a series of questions for the purpose of gathering information from respondents[97]. A research questionnaire is typically a mix of close-ended questions and open-ended questions[98]. For the design of closed questions, the author uses a scoring system (1-5 points, 1 point is strongly disagree ~ 5 points strongly agree). For development issues, the author encourages participants to express their views actively and boldly on the application and give certain material rewards. At the end of the first meeting, the author issued an initial questionnaire to each participant. The questionnaire consisted of one question (see Appendix B for details). After the meeting, the author provided the participants with a second questionnaire consisting of seven questions. Question composition (see Appendix C for details). The purpose of the questionnaire is to investigate the user experience, to make targeted improvements to the user's dissatisfaction.

This method is needed to solve the research questions Q1-Q4.

### 3.6 Limitations of the Study

Limitations of studies are always present. The limitation of this research lists as below:

**Subject limits.** This research involves many complex subject knowledges, and requires a certain understanding of statistics, medicine, computer, big data and other industries in order to complete it. Since the author's major is software related, not proficient in other subjects but only has a little understanding. Therefore, some insights in this article may be one-sided. In addition, the author has not conducted practical training on professional systems and has little experience in programming and designing big data visualization, so the dashboards he made are not as attractive as the professionals. However, every step of the author's design is supported by relevant theories and follows the principle of putting people first and starting from reality, so the project research has certain practical significance.

**Data.** Due to the new released General Data Protection Regulation (GDPR), the author had to use the NHANES data instead. Although the data of NHANES is professionally collected and carefully designed, the conclusions drawn may be different compared with the actual data of patients. Therefore, the results of this paper need further verification and verification. In addition, the NHANES dataset selected by the author has certain advantages, it is also true that the data the author used are raw data and has not undergone professional and complex processing. On the other hand, the sample size is quite small with lots of empty and invalid data, after data cleaning step the sample size become smaller. This may cause a certain amount of data distortion and result in insufficient accuracy of test results more or less. The author has tried to integrate NHANES data for years to compensate for the small sample size, but still not sure that the experimental results are completely accurate.

**Participants.** Finally, the author also identified the limitations of the project: due to the coronavirus government issued travel restrictions within 20 kilometers, people are also advised to reduce unnecessary gatherings, so the sample size is limited, and the author cannot collect more user data to test the applicability and practicality of the app.

### 3.7 Conclusions

After discussing the current research gaps, such as the lack of dynamic interactive web applications. How to help the clinicians decide what statistical approach to use and how to select the best fit models. How to help the doctors to interpret those analysis results.
How to show the patient compared to the average etc. The author designed and developed an autonomous analytic dashboard to overcome above those gaps.

In order to improve the use of the dashboard, we adopted a hybrid approach, through extensive literature reviews, group interviews, observation of the use of experimental participants, and questionnaires before and after the experiment to collect feedback to complete the case study.

Finally, the author also summarized the shortcomings of the project. Due to the global pandemic of the new coronavirus, the author was unable to gather more participants, so the sample size was limited. And due to professional limitations, the author may not have a deeper understanding of the real situation of some disciplines, but only meets the functional requirements of the dashboard.

# Chapter 4. Design and Architecture

### 4.1 Introduction

In this chapter, the author adopts a human-centered design method, mainly for doctors and patients to carefully design an analytical and personalized visual dashboard, so as to help people in different professional fields to make decisions quickly and accurately. To efficiently support tasks in clinical practice, it is essential that not only efficient and reliable visual computing algorithms are developed, but that they are integrated in a carefully designed user interface[10]. The goal of regular clinical adopts a user-centered design approach, this comprises an in-depth analysis of the tasks to be solved and the target user group[10]. The design and development of prototypes should be carried out only based on verified assumptions about the essential usage scenarios[10]. This process usually starts with some kind of representation of the current solution, e.g., a workflow diagram, then it continues with a representation of user needs and their priorities as well as a representation of the envisioned solution[10].Thus, user interface design is much more than a nice visual wrap-up for some algorithms. It is instead a complex and highly iterative process, which requires early and continuous feedback from the target user group and other relevant stakeholders[10].

Evidence shows the involvement of end-users in the design of the intervention is likely to improve its effectiveness[95]. Human-centered design is an approach that systematically incorporates end-user feedback throughout the design process [18]. This process helps to ensure that the design is functional, supports the end-user’s goals, and fits the organizational context [19]. Additionally, it ensures ease-of-use and usefulness of the technology, thereby engendering more positive attitudes towards the technology [20].

The purpose of this article is to correctly guide doctors in the busy medical procedures to quickly grasp the patient’s condition and make decisions (such as quickly and accurately selecting the best statistical method or AI model), and to help ordinary patients understand the severity of their condition (through Compared with average).

### 4.2 Design thinking

Figure 4.1 illustrates the design thinking that can be applied to software development.

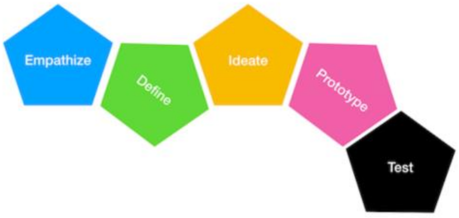


Figure 4.1[99]: How to Use Design Thinking Principles to Develop a Project

The above framework was implemented when designing the autonomous analytic health dashboard. The framework was applied as follows:

**Empathize**. Learn about that audience through interviews and/or observation and find out what truly matters to them[100].

**Define user's requirements.** The key concept of design thinking is to understand need and insights of users[101]. Besides an analysis of tasks, it is also essential to understand users, their habits and preferences, their educational background and problem-solving strategies[10]. This stage in a user interface lifecycle aims at understanding users' qualifications, preferences, needs and attitudes in order to create solutions which are acceptable and appropriate for them[10]. User analysis usually follows task analysis and adds information[10].

In real life, doctors in different subjects have different workflows and operating methods. Take doctor and radiologists as examples. The former relies more on empirical cognition and manual technology for diagnosis, while the latter cannot do without a computer. All in all, most people prefer simple easy-to-use interfaces. Intuitive and streamlined task interfaces to help minimize cognitive load and context switching for human labellers[27].This distinction has far-reaching implications for what constitutes an effective visualization and interaction technique.

**Ideate**: Measures and dimensions are content we wish to display for your users to spot a problem, identify a cause of a problem, and take actions to solve a problem[101].Here the author employ a human-centered design approach to identify the most suitable measures to include on the dashboard and to ensure essential design requirements are met[95]. The author designed a prototype in the beginning and invited several people to interact and make suggestions. This would improve physicians’ likelihood of using the performance feedback dashboard and increase their motivation to change practice patterns[95].

**Test**: Substantial feedback on your latest medical visualization or image analysis problem usually requires that users can solve real problems with it[10].Graphs should be selected based on types of data[101]. The dashboard then contains the relevant information to help support decision-making[102]. Once the iterative process is over, it’s time to create the final product and deliver -- just make sure all issues are addressed [100].

**同情**：在设计思维的核心是同情。这一步是关于了解客户，理解他们的问题，观察，倾听和观察他们所苦苦挣扎的事情。对于软件开发，这意味着在着手解决方案之前真正理解问题。例如，在与潜在客户的访谈中可能会发生这种情况。

**定义**：下一步是确定问题。这个想法是，当您寻找正确的解决方案时，构架至关重要。如何解决用户问题？您实际上是为谁设计的？您的用户是谁？

**Ideate:** 构思阶段是要产生最广泛的可能性。构思技术包括集思广益，素描，思维导图，以及原型设计，以学习和探索有关问题和上下文的更多信息。

**原型**：在“设计思维”中，您可以一次构建多个原型，以帮助您更好地了解用户。这个想法不是在每个原型上花费太多时间，而是探索各种选择。

**测试**：最后一步是测试阶段。您正在与用户一起测试原型，并获得有关他们的反馈。这里的关键是在用户浏览原型时进行聆听和观看。

The user experience covers perceived attractiveness, joy of use and other aspects that influence how engaged users actually are[10]. Among others, a distinct visual design, typography, the careful use of colours, shapes, and animations contribute to the user experience[10]. Over and above visual design, user experience encompasses the complete life-cycle of user interaction with a product, including marketing, purchasing, service, packaging, training. Among many others, color schemes, typography, well-balanced layouts and the use of certain shapes play an essential role[10].In research settings, usability issues are often considered an optional or minor aspect[10]. This attitude is not effective, since

**Scenarios**. Scenarios are now widely used in HCI, in particular to characterize and envision radically new software systems[10]. From a user’s perspective in natural language, user stories include explicit statements of expectations and preferences[10]. After discussion and refinement, the user stories are refined to conceptual scenarios that abstract from expectations and preferences and may summarize user stories[10]. Concrete scenarios are derived to precisely describe how the interaction should be performed and how the system responds[10].

**Group related metrics.** Positioning the information on the dashboard logically is essential. Grouping related metrics next to each other makes them easy to find — and makes the dashboard's design more attractive[103]. As the author aims to develop a dashboard that can automatically analysis or calculate the corresponding index and find out the 'best' approach to the users in the different stage. Therefore, the author considered to make three groups, one is the machine learning part to figure out the best model, the other is to suggest the most suitable statistical approach and the last is comparing the average value by data visualization.

**Task analysis.** Task analysis is a key element for research and development that is targeted at supporting clinical work[10]. The failure of many attempts to create useful systems for clinicians is often largely due to an incomplete task analysis, where major requirements were not identified or their priority was underestimated[10]. Typical users have no idea what could be done with adequate technical support, they are accustomed to certain kind of technology and try to cope with it[10]. Thus, user needs have to be very carefully elicited[10]. In order to determine task requirements, the traditional question-and-answer format may be a safe and simple way, but it needs more careful planning.
Method: Modern task analysis combines a variety of methods. Here we focus on observations and interviews, which are particularly important in practice[10]. That is to say confirm the input and output of a project.

**Workflow analysis and design.** Workflow represents a process as a graph representations which contain actions or events(nodes in the graph) and their logical sequence (edges in the graph)[10].Workflows may contain variants and may emphasize typical sequences of actions. The design of medical visualization may borrow from these experiences, notations and tools to identify such workflows and thus to characterize diagnosis, treatment planning, interventional procedures and outcome control[10].

Also as previously talked, at different sites or even among different doctors at one site, there might be huge differences in their specific workflows (unlike in manufacturing and administrative procedures, medical treatment is and must be more individualized with respect to the patient and the medical doctor)[10]. Serving various personas is one of the most difficult issues of dashboard design. Once each user role is defined, it becomes critical to understand where their needs overlap and where they diverge[104].

Effective communication is the underlying principle of every successful dashboard design[104]. Predicting various scenarios in which users may find themselves probably can help to aid in gaining a better grasp of the user's situation. Consistent sizes and clear relationships between elements will help create patterns and visual flow[103].

**Design Principles for better user experience.**

As user interface design and development involves a large number of decisions, principles at different levels provide a basis for discussions and decisions[10]. At the highest level, general rules give an orientation how certain desirable properties of a user interface may be achieved[10]. Since these high-level principles are quite strive for consistency, dialog icons and other graphics and interface elements should be designed in a consistent manner and provide users with the same operations and feedback[10]. The layout also needs to be consistent, for example, interface elements (such as "Stop" and "Continue" buttons) should be placed in similar locations.

**Aim for high performance.** Pursue high performance, there are various experiments showing that a response time of less than 0.1 second is perfect because the user does not notice any delay and can interact with the system smoothly[10].

**Colour and Bias**. Colour is widely used as a way of representing and distinguishing information[10]. According to a recent study, it is also a key factor in user decision-making[10]. Therefore, we can use high-contrast colours (black and which etc) combine with patterns to show the differences of the information. ~~Good data visualization should tell the story clearly and avoid distortion[10].~~ Avoid using visual representations that cannot accurately represent the data set, such as 3D pie charts[10]. A good design of visualization can ensure no distortion and the users could make a conclusion from it.

The dashboard developed by these principles is not necessarily very beautiful and eye-catching user interface, but it can help designers develop better and more practical dashboards[10].

### 4.3 Requirements

Functional requirements are product features or functions that developers must implement to enable users to accomplish their tasks, it mainly focus on the user’s requirements[105], [106].

Non-functional requirements define system behavior, features, and general characteristics that affect the user experience[106]. It focuses on the ease of use and performance of the system. In other words, functional requirements tell the system what to do and what not to do while non-functional requirements tell the system how to do it.

#### 4.3.1 Functional Requirements

The author has followed the best practice principles, the functional requirements of the autonomous health analysis dashboard are:

•Allow users to operate with the mouse to achieve click input, such as selecting buttons and select list option and so on.

•The app allows users to store objects in a secure environment.

• The app allows users to fully control their experience in the system.

• The app can guide users through the selection of the best statistical method and ML model.

• The app can guide users to compare with the average.

•The app can guide users to check the data distribution.

#### 4.3.2 Non-Functional Requirements

The non-functional requirements of the personalized dashboard are:

• The app must be simple to operate and easy to understand.

• The app must have a consistent layout and a consistent design such as fonts, formatting, and so on.

• For people of different classes, regardless of professional background or cognitive ability, they must be able to use the app.

• For privacy protection, the user's data authorization permission must be obtained when the app is started, because under normal circumstances, the user will directly operate and analyse the patient's personal data.

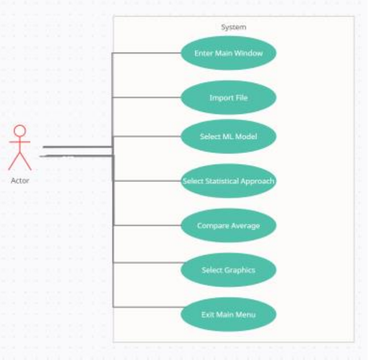
- The system must not crash during start-up, and the delay must not be greater than 10s.

- The system must be easy to maintain and malleable to allow adaptation to future update needs. Since we are a personalized analytical dashboard, if we need to update the app in the future, such as adding new technologies or improving system performance, solving some bugs, etc. New developers do not need to cut over with the author, and can directly add new user needs.

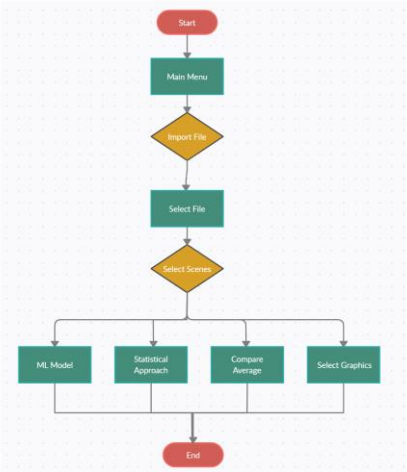
- The system is as simple and beautiful as possible.

### 4.4 Use Case and Activity Diagram

A use case is a methodology used in system analysis to identify, clarify and organize system requirements[107]. The best app strategy is one that uses not more than two use cases[108]. Because too many use cases can overwhelm users. Activity diagram is basically a flowchart to represent the flow from one activity to another activity, it can be described as an operation of the system[109].



### 4.5 System's Logic



### 4.6 System's Architecture and Tools

In this chapter, the author elaborated how to best present the data to the users considering from many aspects: the language selection, tool selection, which graphic can be used to best present the data in both main windows and sub-window of the dashboard.

#### 4.6.1 Language,Tool and Frame

**Language**.To realize the purpose of the research, coding is a mandatory part. Both R, Python, Java language can be used here. In fact, most medical personnel are more inclined to use R language as it is a statistical programming language that can implement various statistical and graphical techniques etc. But here the author will use Python as the author thinks it is easier to understand and it needs less coding which looks more concise. But the most important, Python has lots of statistical and AI models to do data analytics and the seaborn (A data visualization library) can provide high level interface and draw attractive graphics.

**Tools**: After investigation, the author found Jupiter Notebook is a powerful tool. It is an open-source web application that allows the user to do machine learning, data visualization, statistical modelling, data cleaning and transformation etc.

**Frame**. To develop a beautiful and interactive web-based dashboard, the author has used dash as it is a wonderful library framework in pure python for creating analytical web applications. We even do not need to know HTML, CSS, JavaScript as it is all in Python. In general, there are three steps can be taken to do data visualization.

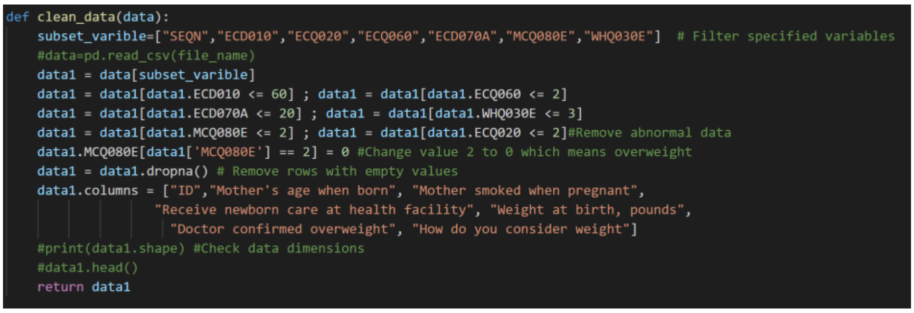
First, components. Components like dropdown list, the check box, the button etc all these components interactive capability inside our data can help us to create a data visualization dashboard.
Second, Plotly graphs. By using plotly graphs or charts like scatter graphs, bar charts we can make data visualization.

Third, the callback. This is the most important element because the callback connects between the dash components and plotly graphs in order to create an interactive dashboard.

Please be aware that the components and the plotly graphs are all going into the inside of the app layout while the callback is outside the app layout.

### 4.6.2 ETL Process

As Chapter 2.5.3 discussed, to remove the bias and noise we must do data cleaning to remove the invalid data. Below is the code to do data cleaning.



```python
def clean_data(data):
    subset_varible=["SEQN","ECD010","ECQ020","ECQ060","ECD070A","MCQ080E","WHQ030E"]  # Filter specified variables
    #data=pd.read_csv(file_name)
    data1 = data[subset_varible]
    data1 = data1[data1.ECD010 <= 60] ; data1 = data1[data1.ECQ060 <= 2]
    data1 = data1[data1.ECD070A <= 20] ; data1 = data1[data1.WHQ030E <= 3]
    data1 = data1[data1.MCQ080E <= 2] ; data1 = data1[data1.ECQ020 <= 2]#Remove abnormal data
    data1.MCQ080E[data1['MCQ080E'] == 2] = 0 #Change value 2 to 0 which means overweight
    data1 = data1.dropna() # Remove rows with empty values
    data1.columns = ["ID","Mother's age when born", "Mother smoked when pregnant",
                    "Receive newborn care at health facility", "Weight at birth, pounds",
                      "Doctor confirmed overweight", "How do you consider weight"]
    #print(data1.shape) #Check data dimensions
    #data1.head()
    return data1
```

As the Chapter 2.6.4 outlined, ETL process also involved to format the variables. As we plan to design a personalized dashboard, the ETL can help to ensure only the allow the selected variables to come into the database. Below shows how the ETL process involves in the whole system.

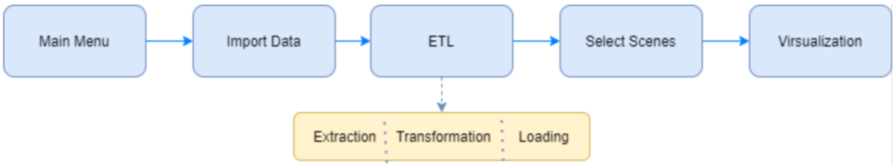
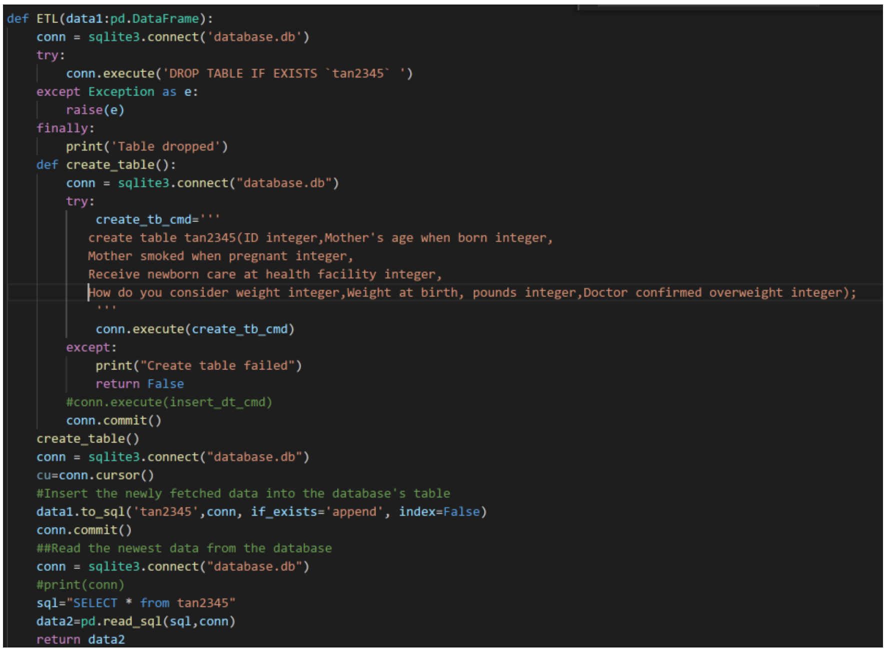


Figure 1: The ETL process in the system overview.

Figure 2 shows the code realization of the ETL process.



```
def ETL(data1:pd.DataFrame):
    conn = sqlite3.connect('database.db')
    try:
        conn.execute('DROP TABLE IF EXISTS `tan2345` ')
    except Exception as e:
        raise(e)
    finally:
        print('Table dropped')
    def create_table():
        conn = sqlite3.connect("database.db")
        try:
            create_tb_cmd='''
           create table tan2345(ID integer,Mother's age when born integer,
           Mother smoked when pregnant integer,
           Receive newborn care at health facility integer,
           How do you consider weight integer,Weight at birth, pounds integer,Doctor confirmed overweight integer);
            '''
            conn.execute(create_tb_cmd)
        except:
            print("Create table failed")
            return False
        #conn.execute(insert_dt_cmd)
        conn.commit()
    create_table()
    conn = sqlite3.connect("database.db")
    cu=conn.cursor()
    #Insert the newly fetched data into the database's table
    data1.to_sql('tan2345',conn, if_exists='append', index=False)
    conn.commit()
    ##Read the newest data from the database
    conn = sqlite3.connect("database.db")
    #print(conn)
    sql="SELECT * from tan2345"
    data2=pd.read_sql(sql,conn)
    return data2
```

Figure:

### 4.6.3 Scene1: Graph Selection for Statistical approach.

**Defining the Purpose of the Analysis**

It is important to determine the purpose of the analysis to choose the appropriate statistical test to support the research question[31]. It is important to ensure that the statistical analysis is appropriate for the way that the study was designed and the data were collected[31].

In this study, we choose the following items as independent variables: whether the mother smoked during pregnancy (divided into smoking and non-smoking), and the mother's age at delivery (according to relevant data, we classified the mothers at delivery as older women who were older than 35 Years old, the age of middle-aged women is between 20 and 35 years old, and the age of younger women is less than 20 years old). Explore the relationship between these independent variables and baby weight.

Through the questionnaire survey, we have obtained data about the baby's weight at birth, the mother's age at delivery, whether the mother smoked during pregnancy, and whether the doctor's diagnosis was overweight. In order to ensure confidentiality, each sample data is identified with a different number code (SEQN) to match the obtained data.

In the research we use a cross-sectional method, it is a type of research design in which we can collect data from many different individuals at a single point in time and we can observe variables without influencing them[110].Therefore, it can be used for the description and comparative analysis of various types of research objects. Time-saving and low-cost, the researcher does not have to wait for the subjects to grow up.

**Select Statistical Method**

If the purpose is to determine if two continuous variables in the study population are correlated, a Pearson correlation should be used if both variables are normally distributed or if the relationship between the two variables is linear, and a Spearman correlation should be used if at least one variable is not normally distributed[31]. For example, we can check whether a person's smoking times are related to cancer.

In this study, we used the following items as independent variables: whether the mother smoked during pregnancy (divided into smoking and non-smoking), the mother's age at delivery, and explored the relationship between these independent variables and the baby's weight.

A total of 3603 mothers were surveyed in this study. After excluding invalid samples, there were 2726 valid sample data.

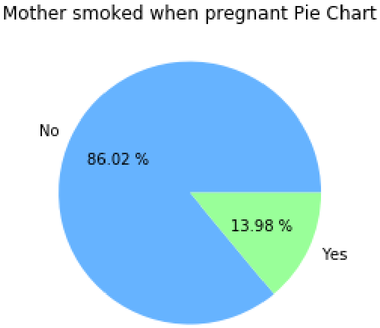


Figure 4.1: The percentage of mothers smoking during pregnancy.

Among the 2726 mothers, 13.98% (381) of the mothers smoked during pregnancy, and 86.02% (2345) of the mothers did not smoke during pregnancy.

Using the mother's age at childbirth as the abscissa and the newborn baby's weight as the ordinate, a scatter diagram is drawn as shown below:

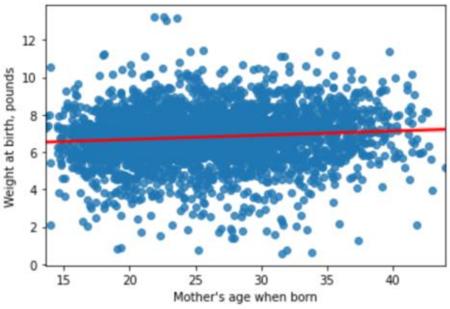


Figure 4.2: The linear regression line of the baby's weight with the mother's age.

It can be seen from the figure that most newborn babies weigh between 5 and 8 pounds, the lightest newborn baby is 1 pound, the heaviest newborn baby is 13 pounds, and the average newborn baby's weight floating between 6~8 pounds. The age of the mother at the time of giving birth is between 20 and 30 years old, the youngest mother at the time of delivery is 14 years old while the oldest is 45 years old.

A Spearman correlation coefficient was used to describe relations between continuous data[95]. Rho correlations coefficients of >0.7 were considered strong correlations, coefficients between 0.5–0.7 were considered moderate, and coefficients between 0.3 and 0.5

were considered weak[95]. A p-value of < 0.05 was considered significant[95]. Analyses were performed with Python in Jupiter Notebook. Now we start to analyze the correlation coefficients of the mother's age at childbirth, whether the mother smoked during pregnancy, and the weight of the newborn baby.

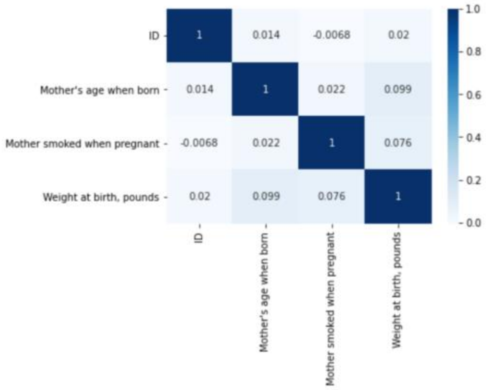


Figure 4.3: Confusion Matrix to show the correlation coefficient between variables.

It can be seen from the figure that the correlation coefficient values of the independent variable (the mother's age at delivery, whether the mother smoked during pregnancy) and the dependent variable (newborn baby weight) are all less than 0.3, which seems to indicate the degree of correlation between the independent variable and the dependent variable Extremely weak.

By consulting related literature, it is found that in actual problems, the correlation coefficient is generally calculated using sample data, so it has a certain degree of randomness, especially if the sample size is relatively small, this randomness is greater. In this case, use the sample the reliability of the correlation coefficient estimation of the overall correlation coefficient will be greatly questioned, that is, the sample correlation coefficient does not indicate whether the two populations from which the sample comes from have a significant linear relationship. Therefore, it needs to be statistically inferred to determine whether there is a correlation between variables through testing methods, and we need to conduct further testing and analysis between variables.

**Statistical Approach Selection**

In order to explore whether the mother’s smoking during pregnancy affects the weight of the newborn baby, we divide the data in the questionnaire into the pregnant mother’s smoking group and the mother’s non-smoker group. In order to study the effect of smoking on the weight of newborn babies, we also set up a group of mothers who did not smoke during pregnancy as a control group.

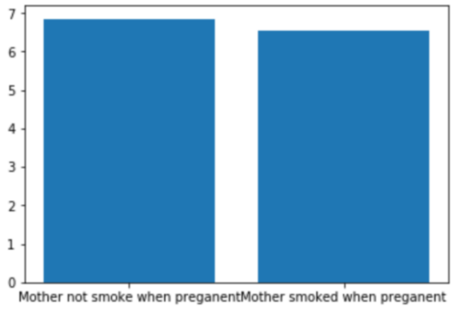


Figure 4.4:

From the above table, we find that the average weight of newborn babies in the mother's non-smoking group during pregnancy is 6.85 pounds, and the average weight of newborn babies in the mother's smoking group during pregnancy is 6.54 pounds. The weight of 50% of newborn babies in the mothers who did not smoke during pregnancy was 6 to 8 pounds, while the data of 50% of babies whose mothers smoked during pregnancy was 6 to 7 pounds. Indeed, there is a difference between the two sets of data, but whether this difference comes from whether the mother smoked during pregnancy or because of sample errors, we conducted a T test analysis on the two sets of data to find out the reason. As there are three types of T test, we will look into which type of T test is best fit in the case.

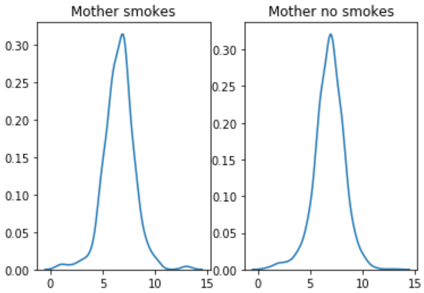


Figure 4.5: Infant weight’s plot of standard normal probability density

It can be seen from the standard normal probability density estimation graph that the two sample populations approximately obey a normal distribution. Since we are analyzing whether the mother's smoking during pregnancy and the mother's non-smoking during pregnancy (quantitative data) have an effect on the weight of the newborn baby (quantitative data), according to the analysis in Chapter 2.4.5, our study is suitable for independent T-test.

Now we establish the hypothesis of the independent sample T test:

H0: μ1-μ2 ＝ 0

H1: μ1-μ2 ≠ 0

We performed Levene's homogeneity test of variance (α=0.05) on the data of the weight of the newborn baby when the mother smoked during pregnancy and the weight of the newborn baby when the mother did not smoke during pregnancy. The Student t test was used to judge whether the variances of the two populations were equal (Student's t test or Welch T test (Welch's t test). If the p-value is inferior or equal to the significance level 0.05, we can reject the null hypothesis and accept the alternative hypothesis[111]. In other words, we can conclude that the mean values of group A and B are significantly different[111].

The P value of Levene's homogeneity of variance test was calculated to be 0.258. In our example, the tested p-value is less than the threshold 0.05, so the variances of the sample groups are not equal. The P value (0.258) is greater than the significance level α (0.05), so the null hypothesis cannot be rejected. The variance of the two populations is not significantly different, and the Student's t test should be used.

Based on the above analysis, we found that the average weight (6.55 pounds) of newborn babies in the mother's smoking group during pregnancy is lower than the weight of newborn babies in the mother's non-smoking group (6.84 pounds). Smoking during pregnancy does affect the weight of the newborn baby, leading to light weight.

**Conclusion:**

In order to explore the impact of mothers' smoking during pregnancy on the weight of newborn babies, we conducted a questionnaire survey of 3603 mothers to investigate whether they smoked during pregnancy and the weight of the newborn baby. We obtained 3603 relevant data.

For those mothers who answered "clear/forgot", we deleted them and retained 2,726 valid data. Among them, 13.98% (381 people) of mothers smoked during pregnancy, and 86.02% (2345 people) of mothers did not smoke during pregnancy.

After analyzing these data, we found that the average weight (6.55 pounds) of newborn babies in the mother's smoking group during pregnancy was lower than the weight of newborn babies in the mother's non-smoking group during pregnancy (6.84 pounds). In order to further understand whether this difference comes from the overall difference or Due to sampling errors, we conducted independent sample T-test analysis on the data of the smoking group and the non-smoker group, and the results showed that smoking during pregnancy does affect the weight of the newborn baby, causing the baby to be lighter.
Therefore, smoking cessation during pregnancy is particularly important for the health of newborn babies.

### 4.6.4 Scene2: Graph Selection for ML models

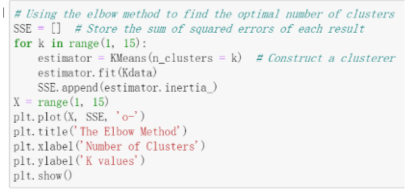
As we already knew that the unsupervised learning can be used to dig out the data structure of an unknown dataset (See Chapter 2). Here the author will firstly do clustering analysis to observe how many categories our data can be made. And build several machine learning models (SVC, Random Forest, and Logic Regression) and compare their medical related KPIs (accuracy, specificity and sensitivity etc). Then the conclusions can be made based on all the analysis.

After the ETL process the data are in standard format, the author will show how to build and select the correct AI models based on machine learning knowledges in Chapter 2.5.1.

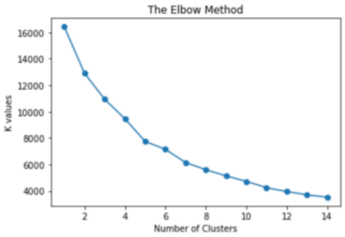
**Clustering Analysis**

In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale[112].

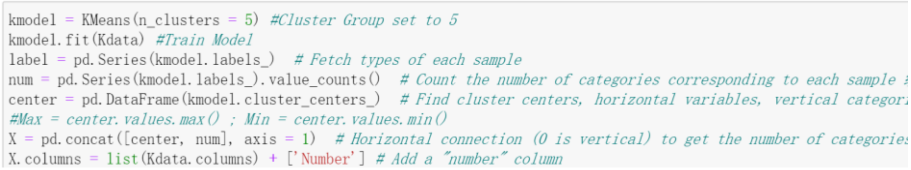
Here we need to use elbow method to find the best clustering point based on Chapter 2. As the figure shows, the best elbow point is 5.



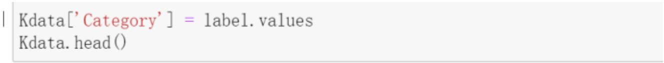
```
# Using the elbow method to find the optimal number of clusters
SSE = []  # Store the sum of squared errors of each result
for k in range(1, 15):
    estimator = KMeans(n_clusters = k)  # Construct a clusterer
    estimator.fit(Kdata)
    SSE.append(estimator.inertia_)
X = range(1, 15)
plt.plot(X, SSE, 'o-')
plt.title('The Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('K values')
plt.show()
```



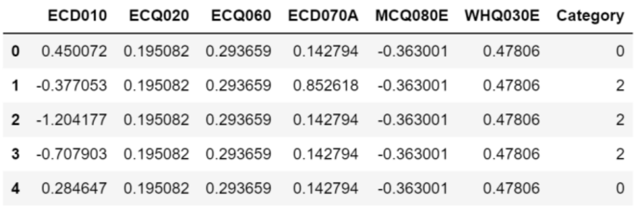
We can see the point 5 is the turning point which is decreasing sharply.That means clustering the dataset into 5 categories should be the best choice。



```
kmodel = KMeans(n_clusters = 5) #Cluster Group set to 5
kmodel.fit(Kdata) #Train Model
label = pd.Series(kmodel.labels_)  # Fetch types of each sample
num = pd.Series(kmodel.labels_).value_counts()  # Count the number of categories corresponding to each sample i
center = pd.DataFrame(kmodel.cluster_centers_)  # Find cluster centers, horizontal variables, vertical categori
#Max = center.values.max() ; Min = center.values.min()
X = pd.concat([center, num], axis = 1)  # Horizontal connection (0 is vertical) to get the number of categories
X.columns = list(Kdata.columns) + ['Number'] # Add a "number" column
```

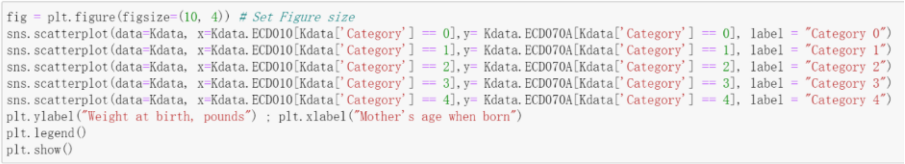


```
Kdata['Category'] = label.values
Kdata.head()
```

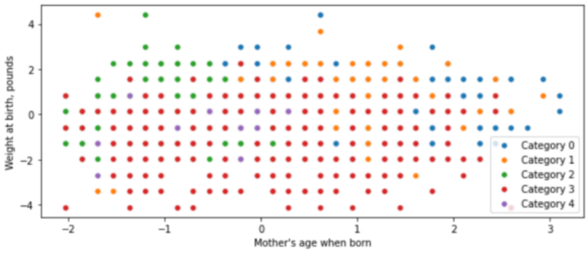


| | ECD010 | ECQ020 | ECQ060 | ECD070A | MCQ080E | WHQ030E | Category |
|---|---|---|---|---|---|---|---|
| **0** | 0.450072 | 0.195082 | 0.293659 | 0.142794 | -0.363001 | 0.47806 | 0 |
| **1** | -0.377053 | 0.195082 | 0.293659 | 0.852618 | -0.363001 | 0.47806 | 2 |
| **2** | -1.204177 | 0.195082 | 0.293659 | 0.142794 | -0.363001 | 0.47806 | 2 |
| **3** | -0.707903 | 0.195082 | 0.293659 | 0.142794 | -0.363001 | 0.47806 | 2 |
| **4** | 0.284647 | 0.195082 | 0.293659 | 0.142794 | -0.363001 | 0.47806 | 0 |

As we know, the dataset can be clustered into five groups/categories。



```
fig = plt.figure(figsize=(10, 4)) # Set Figure size
sns.scatterplot(data=Kdata, x=Kdata.ECD010[Kdata['Category'] == 0],y= Kdata.ECD070A[Kdata['Category'] == 0], label = "Category 0")
sns.scatterplot(data=Kdata, x=Kdata.ECD010[Kdata['Category'] == 1],y= Kdata.ECD070A[Kdata['Category'] == 1], label = "Category 1")
sns.scatterplot(data=Kdata, x=Kdata.ECD010[Kdata['Category'] == 2],y= Kdata.ECD070A[Kdata['Category'] == 2], label = "Category 2")
sns.scatterplot(data=Kdata, x=Kdata.ECD010[Kdata['Category'] == 3],y= Kdata.ECD070A[Kdata['Category'] == 3], label = "Category 3")
sns.scatterplot(data=Kdata, x=Kdata.ECD010[Kdata['Category'] == 4],y= Kdata.ECD070A[Kdata['Category'] == 4], label = "Category 4")
plt.ylabel("Weight at birth, pounds") ; plt.xlabel("Mother's age when born")
plt.legend()
plt.show()
```

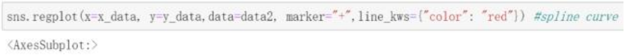


Conclusion: From the Scatter we can see the Category 3 is the majority, after checking the table, we found the younger the mother's age, the higher baby's weight.

**Machine Learning Model Selection**

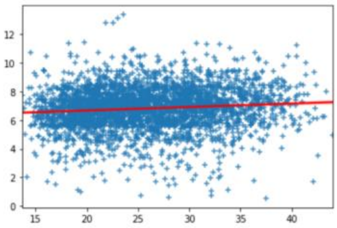
**Simple Linear Regression**

The function in Seaborn to find the linear regression relationship is regplot[113]. Now using the mother's age at childbirth as the abscissa and the newborn baby's weight as the ordinate, a scatter diagram is drawn as shown below:



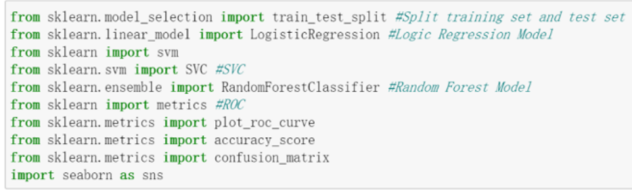
```
sns.regplot(x=x_data, y=y_data,data=data2, marker="+",line_kws={"color": "red"}) #spline curve
```

```
<AxesSubplot:>
```

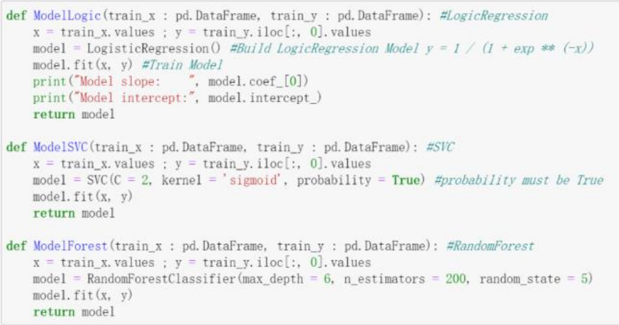


It can be seen from the figure that most new-born babies weigh between 5 and 8 pounds. The lightest new-born baby is 1 pound, the heaviest new-born baby is 13 pounds, and the average new-born baby's weight is around 6.8 pounds. The age of the mother at the time of childbirth is between 20 and 30 years old, the youngest mother at the time of childbirth is 14 years old, and the oldest mother at the time of childbirth is 45 years old.

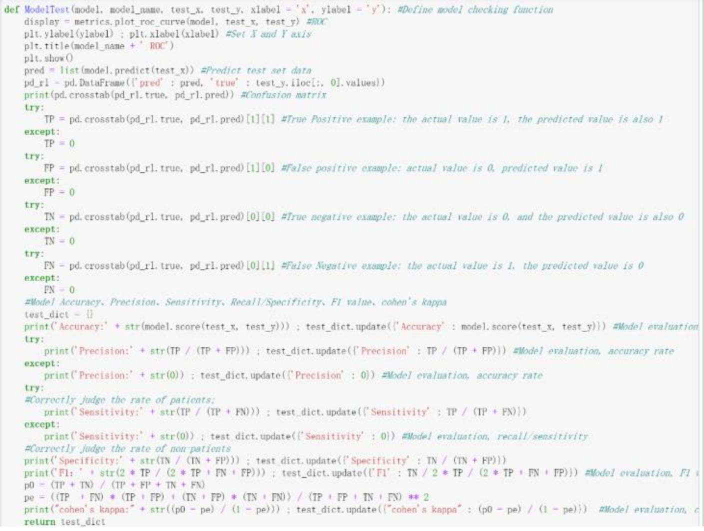
**Build Models:**



```
from sklearn.model_selection import train_test_split #Split training set and test set
from sklearn.linear_model import LogisticRegression #Logic Regression Model
from sklearn import svm
from sklearn.svm import SVC #SVC
from sklearn.ensemble import RandomForestClassifier #Random Forest Model
from sklearn import metrics #ROC
from sklearn.metrics import plot_roc_curve
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
import seaborn as sns
```

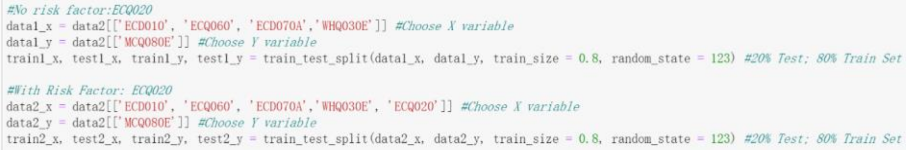


```
def ModelLogic(train_x : pd.DataFrame, train_y : pd.DataFrame): #LogicRegression
    x = train_x.values ; y = train_y.iloc[:, 0].values
    model = LogisticRegression() #Build LogicRegression Model y = 1 / (1 + exp ** (-x))
    model.fit(x, y) #Train Model
    print("Model slope:    ", model.coef_[0])
    print("Model intercept:", model.intercept_)
    return model

def ModelSVC(train_x : pd.DataFrame, train_y : pd.DataFrame): #SVC
    x = train_x.values ; y = train_y.iloc[:, 0].values
    model = SVC(C = 2, kernel = 'sigmoid', probability = True) #probability must be True
    model.fit(x, y)
    return model

def ModelForest(train_x : pd.DataFrame, train_y : pd.DataFrame): #RandomForest
    x = train_x.values ; y = train_y.iloc[:, 0].values
    model = RandomForestClassifier(max_depth = 6, n_estimators = 200, random_state = 5)
    model.fit(x, y)
    return model
```



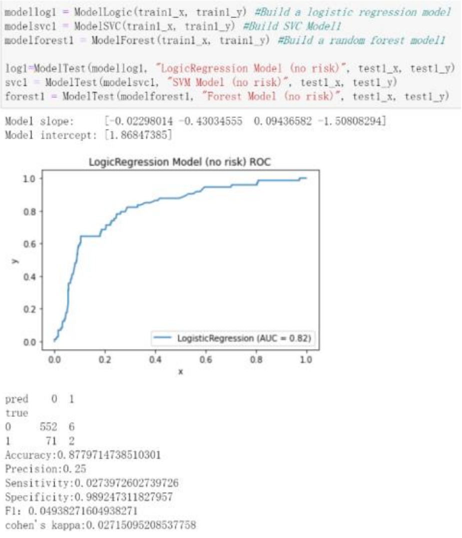
```
def ModelTest(model, model_name, test_x, test_y, xlabel = 'x', ylabel = 'y'): #Define model checking function
    display = metrics.plot_roc_curve(model, test_x, test_y) #ROC
    plt.ylabel(ylabel) ; plt.xlabel(xlabel) #Set X and Y axis
    plt.title(model_name + ' ROC')
    plt.show()
    pred = list(model.predict(test_x)) #Predict test set data
    pd_rl = pd.DataFrame({'pred' : pred, 'true' : test_y.iloc[:, 0].values})
    print(pd.crosstab(pd_rl.true, pd_rl.pred)) #Confusion matrix
    try:
        TP = pd.crosstab(pd_rl.true, pd_rl.pred)[1][1] #True Positive example: the actual value is 1, the predicted value is also 1
    except:
        TP = 0
    try:
        FP = pd.crosstab(pd_rl.true, pd_rl.pred)[1][0] #False positive example: actual value is 0, predicted value is 1
    except:
        FP = 0
    try:
        TN = pd.crosstab(pd_rl.true, pd_rl.pred)[0][0] #True negative example: the actual value is 0, and the predicted value is also 0
    except:
        TN = 0
    try:
        FN = pd.crosstab(pd_rl.true, pd_rl.pred)[0][1] #False Negative example: the actual value is 1, the predicted value is 0
    except:
        FN = 0
    #Model Accuracy、Precision、Sensitivity、Recall/Specificity、F1 value、cohen's kappa
    test_dict = {}
    print('Accuracy:' + str(model.score(test_x, test_y))) ; test_dict.update({'Accuracy' : model.score(test_x, test_y)}) #Model evaluation
    try:
        print('Precision:' + str(TP / (TP + FP))) ; test_dict.update({'Precision' : TP / (TP + FP)}) #Model evaluation, accuracy rate
    except:
        print('Precision:' + str(0)) ; test_dict.update({'Precision' : 0}) #Model evaluation, accuracy rate
    try:
    #Correctly judge the rate of patients;
        print('Sensitivity:' + str(TP / (TP + FN))) ; test_dict.update({'Sensitivity' : TP / (TP + FN)})
    except:
        print('Sensitivity:' + str(0)) ; test_dict.update({'Sensitivity' : 0}) #Model evaluation, recall/sensitivity
    #Correctly judge the rate of non-patients
    print('Specificity:' + str(TN / (TN + FP))) ; test_dict.update({'Specificity' : TN / (TN + FP)})
    print('F1: ' + str(2 * TP / (2 * TP + FN + FP))) ; test_dict.update({'F1' : TN / 2 * TP / (2 * TP + FN + FP)}) #Model evaluation, F1 
    p0 = (TP + TN) / (TP + FP + TN + FN)
    pe = ((TP + FN) * (TP + FP) + (TN + FP) * (TN + FN)) / (TP + FP + TN + FN) ** 2
    print("cohen's kappa:" + str((p0 - pe) / (1 - pe))) ; test_dict.update({"cohen's kappa" : (p0 - pe) / (1 - pe)})  #Model evaluation, c
    return test_dict
```

Split the dataset: select 20% data as the test data, 80% data as the train data. And choose the "ECQ020: Mother smoked when pregnant" as the risk factor.



```
#No risk factor:ECQ020
data1_x = data2[['ECD010', 'ECQ060', 'ECD070A','WHQ030E']] #Choose X variable
data1_y = data2[['MCQ080E']] #Choose Y variable
train1_x, test1_x, train1_y, test1_y = train_test_split(data1_x, data1_y, train_size = 0.8, random_state = 123) #20% Test; 80% Train Set

#With Risk Factor: ECQ020
data2_x = data2[['ECD010', 'ECQ060', 'ECD070A','WHQ030E', 'ECQ020']] #Choose X variable
data2_y = data2[['MCQ080E']] #Choose Y variable
train2_x, test2_x, train2_y, test2_y = train_test_split(data2_x, data2_y, train_size = 0.8, random_state = 123) #20% Test; 80% Train Set
```

Transfer the data into models and plot the ROC-AUC graph for each models in the same way below.



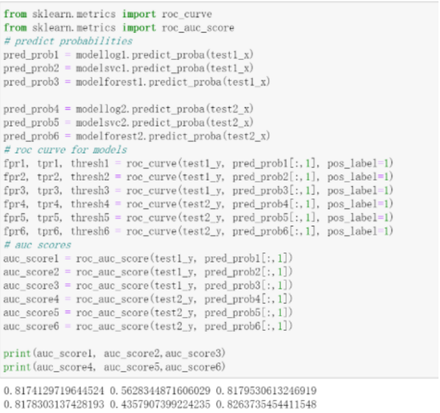
```
modellog1 = ModelLogic(train1_x, train1_y) #Build a logistic regression model
modelsvc1 = ModelSVC(train1_x, train1_y) #Build SVC Model1
modelforest1 = ModelForest(train1_x, train1_y) #Build a random forest model1

log1=ModelTest(modellog1, "LogicRegression Model (no risk)", test1_x, test1_y)
svc1 = ModelTest(modelsvc1, "SVM Model (no risk)", test1_x, test1_y)
forest1 = ModelTest(modelforest1, "Forest Model (no risk)", test1_x, test1_y)
```

```
Model slope:     [-0.02298014 -0.43034555  0.09436582 -1.50808294]
Model intercept: [1.86847385]
```

LogicRegression Model (no risk) ROC
1.0
0.8
0.6
0.4
0.2
0.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x
LogisticRegression (AUC = 0.82)

```
pred     0  1
true
0      552  6
1       71  2
Accuracy:0.8779714738510301
Precision:0.25
Sensitivity:0.0273972602739726
Specificity:0.989247311827957
F1: 0.04938271604938271
cohen's kappa:0.02715095208537758
```

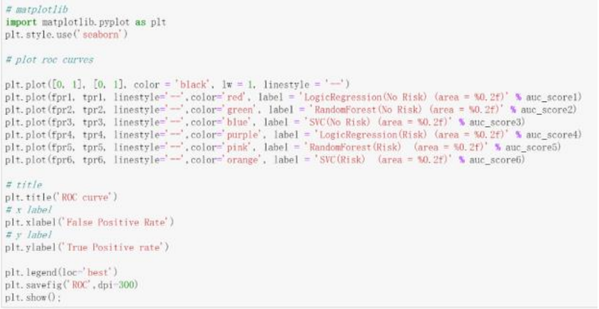
Print out the AUC values of all the models with and without risk factor.



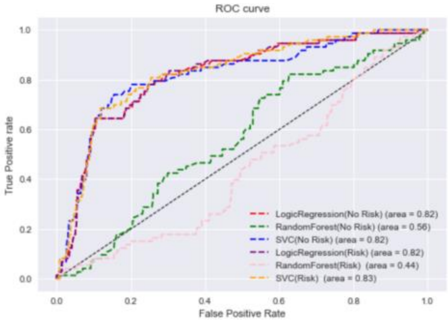
```
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
# predict probabilities
pred_prob1 = modellog1.predict_proba(test1_x)
pred_prob2 = modelsvc1.predict_proba(test1_x)
pred_prob3 = modelforest1.predict_proba(test1_x)

pred_prob4 = modellog2.predict_proba(test2_x)
pred_prob5 = modelsvc2.predict_proba(test2_x)
pred_prob6 = modelforest2.predict_proba(test2_x)
# roc curve for models
fpr1, tpr1, thresh1 = roc_curve(test1_y, pred_prob1[:,1], pos_label=1)
fpr2, tpr2, thresh2 = roc_curve(test1_y, pred_prob2[:,1], pos_label=1)
fpr3, tpr3, thresh3 = roc_curve(test1_y, pred_prob3[:,1], pos_label=1)
fpr4, tpr4, thresh4 = roc_curve(test2_y, pred_prob4[:,1], pos_label=1)
fpr5, tpr5, thresh5 = roc_curve(test2_y, pred_prob5[:,1], pos_label=1)
fpr6, tpr6, thresh6 = roc_curve(test2_y, pred_prob6[:,1], pos_label=1)
# auc scores
auc_score1 = roc_auc_score(test1_y, pred_prob1[:,1])
auc_score2 = roc_auc_score(test1_y, pred_prob2[:,1])
auc_score3 = roc_auc_score(test1_y, pred_prob3[:,1])
auc_score4 = roc_auc_score(test2_y, pred_prob4[:,1])
auc_score5 = roc_auc_score(test2_y, pred_prob5[:,1])
auc_score6 = roc_auc_score(test2_y, pred_prob6[:,1])

print(auc_score1, auc_score2,auc_score3)
print(auc_score4, auc_score5,auc_score6)
```

```
0.8174129719644524 0.5628344871606029 0.8179530613246919
0.8178303137428193 0.4357907399224235 0.8263735454411548
```

In order to enable users to compare the AUC of each model at a glance, the author tried to integrate the AUC curves of all models together and added a 45-degree straight line between the x-axis and y-axis to facilitate the comparison of the area under the curve.



```
# matplotlib
import matplotlib.pyplot as plt
plt.style.use('seaborn')

# plot roc curves

plt.plot([0, 1], [0, 1], color = 'black', lw = 1, linestyle = '--')
plt.plot(fpr1, tpr1, linestyle='--',color='red', label = 'LogicRegression(No Risk) (area = %0.2f)' % auc_score1)
plt.plot(fpr2, tpr2, linestyle='--',color='green', label = 'RandomForest(No Risk) (area = %0.2f)' % auc_score2)
plt.plot(fpr3, tpr3, linestyle='--',color='blue', label = 'SVC(No Risk) (area = %0.2f)' % auc_score3)
plt.plot(fpr4, tpr4, linestyle='--',color='purple', label = 'LogicRegression(Risk) (area = %0.2f)' % auc_score4)
plt.plot(fpr5, tpr5, linestyle='--',color='pink', label = 'RandomForest(Risk)  (area = %0.2f)' % auc_score5)
plt.plot(fpr6, tpr6, linestyle='--',color='orange', label = 'SVC(Risk)  (area = %0.2f)' % auc_score6)

# title
plt.title('ROC curve')
# x label
plt.xlabel('False Positive Rate')
# y label
plt.ylabel('True Positive rate')

plt.legend(loc='best')
plt.savefig('ROC',dpi=300)
plt.show();
```



As Chapter 2.4.3 discussed, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease[63]. Therefore, the author chooses the ROC-AUC as the criteria to figure out the most suitable model in our research: the higher the AUC, the better performance of the model .

From the figure we can clearly compare the auc values of all the machine learning models , obviously, the Logic Regression Model(AUC-scores are 0.82) and SVC model(auc-score: 0.82, 0.83) has better performance. Therefore, the author would suggest the user to choose either Logic Regression model or SVC model in this case.

### 4.7 Conclusions

When designing applications for users with different backgrounds, developers must carefully choose the type of dashboard and follow the best design plan for the dashboard.

The author fully understands the pressure and challenges of daily busy work for medical staff, as well as the difficulties in choosing ML models and statistical methods across disciplines. To help doctors reduce their cognitive burden and make correct decisions, the author makes full use of visualization technology to mine and analyze data, develops and designs three scenarios, and displays the analysis results of different fields on the dashboard in a more intuitive way. For example, the dashboard can deepen the patient's understanding of their disease status through the average value of a certain index with other patients of the same type, or help doctors choose the best M1 model or statistical method, and so on.

The author developed several prototypes by using Axure, and finally determined which prototype is more intuitive to use through the interview in Chapter 6, helping users independently experience the fun of using the program for data analysis to the greatest extent.

In this chapter, the author also describes the functional requirements of the automatic health analysis dashboard, describes how the system should be operated; also lists a series of non-functional requirements, such as whether the appearance is beautiful, whether it is easy to maintain and use, etc.

# Chapter 5 The Application of the dashboard

## 5.1 Introduction

Clinical dashboards are increasingly used to provide information to clinicians in visualized format, under the assumption that visual display reduces cognitive workload[114]. Normally, a dashboard provides the user a global overview, with access to the most important data, functions and controls[115].

In this chapter, the author describes the design decisions of the application based on the knowledge of various disciplines listed in Chapter 2 to solve our research problems. And will explain the design of the application and explain its function in detail.

## 5.2 Design principles for dashboard

In the design of the autonomous health analytic dashboard, the author insist on using best practice in the visualisation of medical data.

### 5.2.1 Determine user’s requirement

Most dashboard design principles revolve around the needs of the end users[21]. Before starting the design process, analyse the user’s psychology, anticipate their needs and expectations, and understand their requirements before moving ahead with the design part [21], [116].

Surveys and user interviews can be conducted with the users, clients, or stakeholders to select relevant key performance indicators (KPIs) [21] [92]. The data that you need to display in the dashboard helps shape the rest of the design process[21]. Once you set the KPIs, you can begin selecting the right type of dashboard and visualization tools[21]. In our research the author has summarized several key points for each window and construct the page accordingly to satisfy the user's requirements.

### 5.2.2 Select the right dashboard type and element

One of the most overlooked dashboard design principles is the need to select the right visualization tools[21]. As we already determined the user’s requirement and now it is time to define the purpose of our dashboard. In fact, there are three types of dashboards, the Operational Dashboard, the Analytical Dashboard and the Strategic Dashboard. They designed for different purposes:

**Operational dashboards**

Operational dashboards aims to help the user see what’s happening right now[117]. If we want to help the user to take actions or make decisions based on data, an operational dashboard is more suitable.

**Analytical dashboards**

Analytical dashboards aims to give the user a clear view of performance trends and potential problems[117]. In contrast to Operational dashboards, Analytical dashboards provide the user with at-a-glance information used for analysis and decision making[118]. A primary goal of this kind of dashboard is to help users make the best sense of the data, analyse trends and drive decision making[118]. If the objective is to simply present data, we can use analytical dashboards.

**Strategic Dashboards**

Strategic dashboards are used to indicate performance against a set of key performance indicators (KPIs), aims to let the user track their main strategic goals via KPIs[117].A strategic dashboard should reflect how the user is performing against their strategic goals [117].

For the data visualization elements, in fact, many junior designers feel confused which charts to use as there are so many graphs. When selecting visualization methods, avoid cluttering the screen with a wide variety of charts and tables[21]. Stick to a minimal selection of visualization types to maintain a consistent look throughout the dashboard[21]. For the choice of graphics, please refer to the Chapter 2.4 which has clearly described the benefits and limitations of each graph.

### 5.2.3 Keep the layout simple

Creating a simple interface is one of the most important dashboard design principles[21]. Make it as easy as possible for users to analyse the information on the screen[21]. the right dashboard design process, irrespective of its type — should avoid cognitive overload, i.e., displaying too much information in one-go that it becomes confusing to interpret and derive data[116]. To design an easier to use web app, we can follow below principles[115], [119]:

- Use consistent design language and colour scheme. The dashboard do not have to be beautiful, but they should provide clear visibility, straightforward navigation, and eye-catching good looks.
- Group related or relevant data. Find ways to show datasets that are easier to grasp when shown combined in a single card but be careful not to confuse the user.
- Chose the right representation for the data. Data representation is a difficult issue, especially as we will want to display several sorts of information in a dashboard, whether static or dynamic changes over time—this can be difficult. Using the incorrect

chart type or defaulting to the most prevalent kind of data visualization may mislead users or lead to data misunderstanding.

### 5.3 A Walk through the health analytic dashboard

#### 5.3.1

Analytical dashboards should be data-centric, and show as many relevant data views as is feasible[115].

#### 5.3.2

#### 5.3.3 Personalized dashboard Applications

**Personalized dashboard:**

**Scene 1:**

**Objects to Be Remembered:**

### 5.4 Conclusions

## Chapter 6. Case study, Results and Discussion

### 6.1 Introduction

Case studies is a detailed study of a specific topic to explain, describe or explore events or phenomena in the everyday contexts in which they occur, such as a person, a group, a place, an event, an organization, or a phenomenon [96], [120]. Case study design usually involves qualitative methods, which are very suitable for describing, comparing, evaluating and understanding all aspects of the research problem [120].

As previously outlined, the primary research question is ‘How to create an autonomous analysis health dashboard to the users in different stage?’ The secondary research question is ‘Does the autonomous analysis health dashboard helps the user to choose appropriate approach to do data analytics?’

The author explains the methodology used to finalize the design of the health app, the experiments done for the case studies, and the findings and discussion from these case studies in this chapter.

### 6.2 Finalization of the Design

In this chapter, the author will take a random sampling method to conduct 3 times semi-structured interviews with the public, with the aim of understanding the needs and doubts of users in different stage.

In order to enhance the product design, the author adopted the RITE (Rapid Iterative Testing and Evaluation)) method in the development phase. It's advantage is adopting regular and early tests with RITE reduces the cost of fixing issues when compared to identifying them later in the product development process[121]. Designers need to carefully record interviews to collect feasible opinions from target customers and modify the app. The author organized four group meetings, each with three participants. After each group of meetings, the author also conducted semi-structured interviews, as follows:

1. Focus group session 1: First Scene-- AI Model selection.
2. Focus group session 2: Second Scene—Statistical Approach selection.
3. Focus group session 3: Third Scene—Compare with the average.
4. Focus group session 3: Fourth Scene—Interpret the analysis results.

In order to test the specific part of the solution, the author designed a prototype. Use prototypes to sell new ideas, motivate buy-in from internal or external stakeholders, or inspire markets toward radical new ways of thinking and doing[122]. Due to a rapid redesign, the last focus group conducted a final review of the application to ensure that all functional requirements were correctly implemented and resolved. The questions raised by each interview and focus group are summarized as follow, please be aware the author always insists on adopting a people-oriented design (See Chapter 4.2).

According to the best practice of the dashboard design, the focus group went through the prototype from below aspects (see Appendix A):

***1. Was the app easy to navigate? If not, how so?***

**Navigation**. Participants commented they cannot clearly distinguish the difference of each function [Focus Group 1]. To address, the author tried to add three buttons to highlight each function and also attached the partition table of our dashboard in the background appendix (Appendix A). Each button appears different graphics and the intepretion of the field that could help the user easy to navigate and distinguish the difference of the functions.

***2. Was the design of the dashboard clear and visible? If not, how so?***

**First iteration to fix the navigation issue.** In order to deepen the user's understanding, the author has added a lot of content, information overload can confuse patients [130] [Focus Group 2]. That is to say no matter how interesting this information is, it's not relevant[123]. Creating a simple interface is one of the most important dashboard design principles[21].Simplify content and reduce visual elements to only the most critical pieces, [129] .Therefore, unnecessary information has been removed after the first focus group meeting, only reserves the information immensely relative with the research topic

***3. Was the layout of the dashboard designed reasonably? If not, how so?***

**Second iteration to reduce contents on the dashboard**. Once you know what information the users need, group related data next to each other[124]. The author's original design was to integrate all functions (ML, statistics) into one interface, but considering that the content that needs to be presented is too complicated and difficult for users to quickly consume and understand [Focus Group 2], the second focus group insisted to divide the main interface into three based on its function. One interface with one button function to control and display different analysis results makes the user know better of the functional use of the dashboard.

***4. How do you think the colour of the dashboard?***

**Third iteration to perfect the layout of the dashboard**. At the beginning, the author was too eager for beauty and added a lot of colors (Some of them are quite similar) when presenting various information, that is not user friendly to distinguish the differences [Focus Group 3]. Color is widely used as a way of representing and distinguishing information, it is also a key factor in user decision-making[10]. Ignored similar colors may be difficult to distinguish from one another[124]. Therefore, the author has adopted contrasting colors to highlight the information because it makes the message clearer. And also bear in mind that always use six or fewer colors in visualizations or it becomes difficult for users to see the differences[124].

***5. Was the information on each scene clear and easy to read? If not, how so?***

**Fourth iteration to improve the theme font**. Several participants said the prototype was not ideally laid out and some words were vague to read [Focus Group 4]. After trying several theme fonts, most of the participants agreed to use the 'Arial' words. In dashboard design, the general rule is to use a single font type and no more than three sizes in that type[124].

***6. Did you feel uncomfortable***?

Some participants found the waiting time to refresh the app page was a bit longer, the elder and youngest were very easy to lose patience [Focus Group 3]. There are various experiments showing that a response time of less than 0.1 second is perfect because the user does not notice any delay and can interact with the system smoothly[10].

***7. Were there any functions, you would expect, missing?***

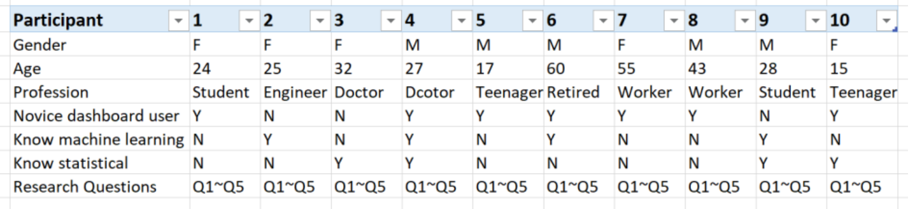
After the last iteration, all participants were happy with the functionality presented and had no further comments [Focus group 1, 2 ,3 and 4]. All these difficulties listed above were addressed in future design iterations to ensure that the main menu material is brief and clear, and the colour and layout are just perfect to showcase the theme. Beginning with the fourth focus group meeting, all participants are pleased with the design and placement of the interface and may easily modify the dashboard for an engaging experience.
After the utility tests were completed and the application was ready, the case studies and their associated experiments were carried out.

### 6.3 Participants

As introduced in Chapter 3.4, the Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample[84]. This advantage is especially obvious in our experiments-random sampling can ensure that the diversity of users can deeply understand the actual needs of people of different backgrounds for automatic health analysis dashboard.

In order to ensure the diversity of the participants, the author has tried best to select people randomly from the public, ranging in age from 14 to 65 years old. Many factors were taken into consideration such as the professional or educational background, gender, and age of the participants etc. In order to ensure the accuracy of the experiment, the author randomly separate all the participants into two groups, the participants number are same. The first group as the experiment group was taught how to use the dashboard. The second group was set as a control group, they will be shown the analysis in traditional methods such as in text or in table without any visualization technology, the most important the result were presented statically. These two groups will be given the same background information (See Appendix) and they were given 15 mins to get familiar with all the concepts and adapt themselves into the experiment situation. The author as the host observes and records the performance of each group without any unnecessary interference.

For the protection of participants' privacy, sensitive information of participants will not be disclosed here, only the author knows.



| Participant | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Gender | F | F | F | M | M | M | F | M | M | F |
| Age | 24 | 25 | 32 | 27 | 17 | 60 | 55 | 43 | 28 | 15 |
| Profession | Student | Engineer | Doctor | Dcotor | Teenager | Retired | Worker | Worker | Student | Teenager |
| Novice dashboard user | Y | N | N | Y | Y | Y | Y | Y | N | Y |
| Know machine learning | N | Y | N | Y | N | Y | N | N | Y | N |
| Know statistical | N | N | Y | Y | N | N | N | N | Y | Y |
| Research Questions | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 | Q1~Q5 |

Table 6.1: An overview of the participants.

To address the 1-4 research questions, the author conducted comparative experiments (experimental group and control group) for four research questions and conducted one-to-one personality assessment with all participants in the experiment. At the same time, unnecessary intervention is avoided, and the main task is to observe and record the answer of the participants.

As per best practice, same background information of this research was provided to each participant no matter what the participant's profession background are (see Appendix ). To ensure the accuracy of the experiment, the author has minimized potential intervention. And also in response to the government's isolation policy due to COVID, the author borrowed an empty room in university, the room is quiet and clean as in healthy adults, environmental noise adversely affects many cognitive domains[126]. As all the participants were separated in different rooms, they can take any postures and methods to relax themselves to stay comfortable, so as to reduce the influence of cognition and prejudice on the experiment.

### 6.4 Research Question 1

How to show a patient how they sit compared to the average? As per best practice, two experiments were carried out, details are listed as below:

#### 6.4.1 Experiment

**Pre-experiment:** Background information of this research was provided to each participant (Appendix A). All the participants were trained to use the personalized health analytic dashboards without importing any dataset. When the experiment starts, please be aware they were only allowed to view the "Compare with the average" page.

**Post experiment:** The author set a 15 mins time clock and allowed them to import the dataset and start the real experiment. During this period, the participants were requested to try best to understand the data presented by visualization technique. After using the app, participants were asked to answer the following questions:

Q1: "What is the overall mean value of the baby's weight?"

Q2: "How many pounds is the heaviest baby? What is the age of the mother accordingly?"

Q3: "How many pounds is the lightest baby? What is the age of the mother accordingly?"

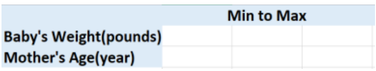
Q4: "Please recall the baby's weight correspondence with the mother's age as much as you can"

Q5: "Please describe the weight's trend of the baby with mother's age of giving birth."

**Observation:** The author observed the participants and record the time stamp as soon as the participants open the dashboard app on the website. The author observes how the participants interact with the dashboard and record whether they met any issues. The author also responsible to evaluate their understanding of Q5 and record whether the participants gave the right answer of Q1~Q4.

### 6.4.2 Control Experiment

**Pre-experiment:** Same as the experiment group, the background information of this research was provided to each participant (see Appendix F). This group of participants have been given time to read and understand the empty table (Appendix) before experiment. Once the real experiments start, they were told to use whatever method they want to try best understanding the information on the table.



| | Min to Max | | |
|---|---|---|---|
| Baby's Weight(pounds) | | | |
| Mother's Age(year) | | | |

**Observation:** The author recorded the first-time stamp when giving the participants the empty table (Appendix). All the participants could use their own styles to remember and understand the values in the table. When they were ready, the author recorded the second time stamp and calculated the time of duration to get familiar with the table form.

**Post-experiment:** The author adopted the traditional information presenting method that is summarised all the research data of the babies in the table (Appendix ). Same as the experiment group, the participants of the control group were given 15 mins to remember and understand as much as possible of the average weight at a specific age of the mother. The participants also requested to answer below questions:

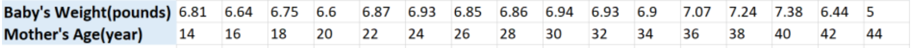
Q1: "What is the overall mean value of the baby's weight?"

Q2: "How many pounds is the heaviest baby? What is the age of the mother accordingly?"

Q3: "How many pounds is the lightest baby? What is the age of the mother accordingly?"

Q4: "Please recall the baby's weight correspondence with the mother's age as much as you can"

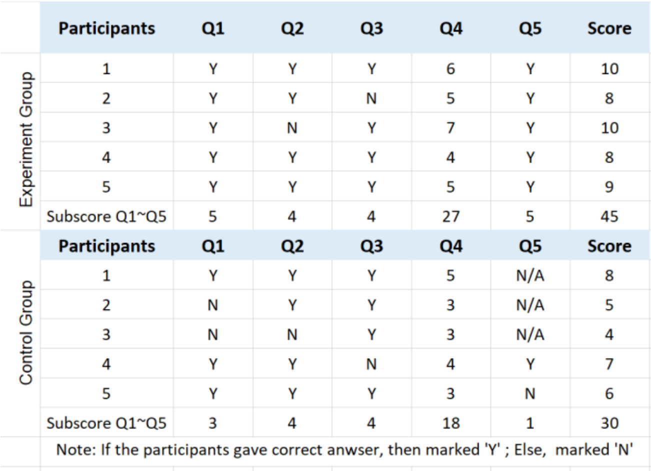
Q5: "Please describe the weight's trend of the baby with mother's age of giving birth."



| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baby's Weight(pounds) | 6.81 | 6.64 | 6.75 | 6.6 | 6.87 | 6.93 | 6.85 | 6.86 | 6.94 | 6.93 | 6.9 | 7.07 | 7.24 | 7.38 | 6.44 | 5 |
| Mother's Age(year) | 14 | 16 | 18 | 20 | 22 | 24 | 26 | 28 | 30 | 32 | 34 | 36 | 38 | 40 | 42 | 44 |

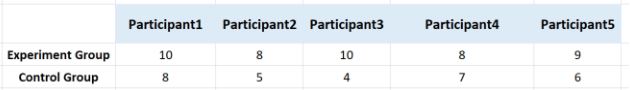
### 6.4.3 Research Question 1: Results

During the observation, the author has collected two groups answer into below table:



| | Participants | Q1 | Q2 | Q3 | Q4 | Q5 | Score |
|---|---|---|---|---|---|---|---|
| Experiment Group | 1 | Y | Y | Y | 6 | Y | 10 |
| | 2 | Y | Y | N | 5 | Y | 8 |
| | 3 | Y | N | Y | 7 | Y | 10 |
| | 4 | Y | Y | Y | 4 | Y | 8 |
| | 5 | Y | Y | Y | 5 | Y | 9 |
| | Subscore Q1~Q5 | 5 | 4 | 4 | 27 | 5 | 45 |
| | Participants | Q1 | Q2 | Q3 | Q4 | Q5 | Score |
| Control Group | 1 | Y | Y | Y | 5 | N/A | 8 |
| | 2 | N | Y | Y | 3 | N/A | 5 |
| | 3 | N | N | Y | 3 | N/A | 4 |
| | 4 | Y | Y | N | 4 | Y | 7 |
| | 5 | Y | Y | Y | 3 | N | 6 |
| | Subscore Q1~Q5 | 3 | 4 | 4 | 18 | 1 | 30 |
| | Note: If the participants gave correct anwser, then marked 'Y' ; Else, marked 'N' | | | | | | |

Table 6.2: Experiment results collection towards Research Question 1



| | Participant1 | Participant2 | Participant3 | Participant4 | Participant5 |
|---|---|---|---|---|---|
| Experiment Group | 10 | 8 | 10 | 8 | 9 |
| Control Group | 8 | 5 | 4 | 7 | 6 |

Table 6.3:

Based on their experience in the automated health analysis dashboard, participants were asked to fill out a questionnaire. Each question can be scored 1~5 points, 1: Disagree, 5: Strongly agree. The author collects and organizes these data into Table 6.2 after completing the questionnaire (see Appendix B).

**Was the app easy to use?** 98% participants said the app was very easy to use and they can easily read the information from the visualized graphics.

**Did you like the layout?** 99% participants said they enjoyed the layout which are clearly distinguished the function of each part. Only participant 4 said the layout looks not perfect enough, but he cannot figure out better solutions.

**Did you enjoy using the app?** 94% of the participants highly rated their enjoyment of using the app. Only Participant 10 said its boring as it is not as funny as the computer games and Participant 6 said he is not too interested in the computer due to bad eyesight as he is an elder. But these are due to their personal perceptual factors, not against this dashboard. Therefore, they could be treated as abnormal values.

***Which part did you enjoy most?*** Participant 3 and 5 interactive function did add a lot of interest, when they got confused of how to use it, the interactive design did give them a lot of guidance.

***If not, what is your dislike?*** 95% participants were highly rated the dashboard, but they do not have any interested in machine learning or statistical technologies.

**Would you like to use it again?** 94% participants showed their willingness of using the app again.

**Would you introduce to your friends?** 98% participant said they would like to introduce to others as a daily tool for the clinicians.

### 6.4.4 Research Question 1: Discussion

After calculating how many questions the experiment and control group can answer correctly based on the table X, the author made a data visualization to compare the two groups. Obviously, the experiment group can answer more and more accurate questions in comparison with the control group. And the author also noticed the experiment group does have a better understanding of the data itself as they can easily and accurately predict the trend of data(Q5). It is worth mentioning that most of them recognized the usefulness and ease of use of the dashboard.

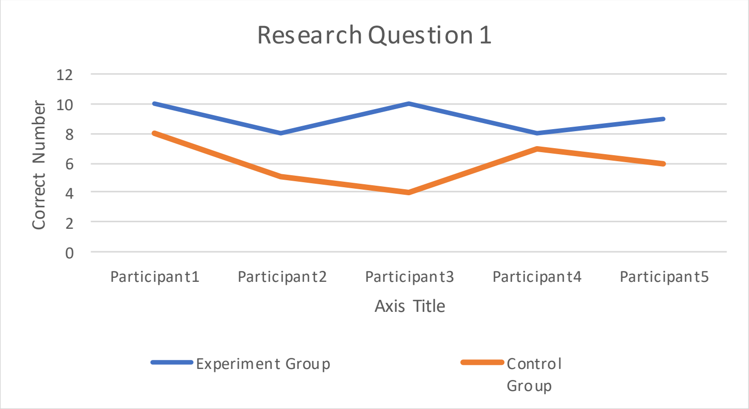


Figure 6.1: Illustrates the experiment group have a better performance compared with the control group.

### 6.5 Research Question 2

How to support doctors in selecting the 'best' statistical approaches?

#### 6.5.1 Experiments

**Pre-experiment:** Statistical information of this research was provided to each participant (Appendix C). All the participants were given 15 mins to understand the appendix and 10 mins to get trained of how to use the scene 2 of the personalized health analytic dashboards with empty dataset. When the experiment starts, please be aware they were only allowed to view the "Statistical Approach" interface.

**Post experiment:** When the experiment start, all the participants allowed to import the dataset into the dashboard. The dashboard after analysis would prompt the picture with key KPIs. The participants were requested to try best to understand the information presented on the screen and follow the rules (Appendix C) to find out the 'best' statistical approach. When they were ready, the participants need to raise their hands and answer below questions:

Q1: "Please suggest the best statistical approach?"

Q2:" Please clarify why you would suggest it as the best statistical approach?"

**Observation:** The author observed how they used the app and collected their answers as well as recording how long time they need to find out the 'best' statistical method.

#### 6.5.2 Control Experiment

**Pre-experiment:** This group of participants were also told to follow the rules to find out the most suitable statistical approach for the research (see Appendix C). They were given 15 mins to review the appendix and should be proficient with all the statistical concepts and what is the purpose of using the concept or parameter.

**Post experiment:** The participants were given a complete dataset and they were told to use whatever method they like to understand and figure out the 'best' statistical approach. The author also allowed them to search the internet as long as they can find the most suitable statistical approach. They can also ask for all the related information of the research. When they were ready, the participants need to raise their hands and answer below questions:

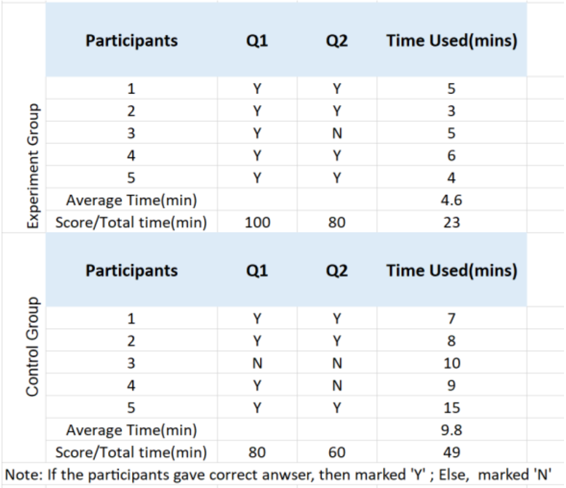
Q1: "Please suggest the best statistical approach?"

Q2:" Please clarify why you would suggest it as the best statistical approach?"

**Observation:** The author collected their answers and record how long time they need to find out the 'best' statistical method.

### 6.5.3 Research Question 2: Results

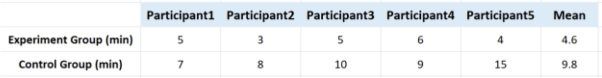
Table 6.5.3 illustrates the results from the two experiments. We can see the overall result of the experiment group is much better compared with the control group as the first group used little time to answer Q1 and Q2. The experiment group only used 4.6 mins average time to answer Q1 and Q2 and their correct rate can reach to 100% and 80%. While the control group need to spend 9.8 mins in average to understand the data and figure out the best statistical approach which almost need double time compared with the experiment group.



| | Participants | Q1 | Q2 | Time Used(mins) |
|---|---|---|---|---|
| Experiment Group | 1 | Y | Y | 5 |
| | 2 | Y | Y | 3 |
| | 3 | Y | N | 5 |
| | 4 | Y | Y | 6 |
| | 5 | Y | Y | 4 |
| | Average Time(min) | | | 4.6 |
| | Score/Total time(min) | 100 | 80 | 23 |
| Control Group | Participants | Q1 | Q2 | Time Used(mins) |
| | 1 | Y | Y | 7 |
| | 2 | Y | Y | 8 |
| | 3 | N | N | 10 |
| | 4 | Y | N | 9 |
| | 5 | Y | Y | 15 |
| | Average Time(min) | | | 9.8 |
| | Score/Total time(min) | 80 | 60 | 49 |

Note: If the participants gave correct anwser, then marked 'Y' ; Else, marked 'N'

Table 6.3: Overview of the results generated from research question 2.



| | Participant1 | Participant2 | Participant3 | Participant4 | Participant5 | Mean |
|---|---|---|---|---|---|---|
| Experiment Group (min) | 5 | 3 | 5 | 6 | 4 | 4.6 |
| Control Group (min) | 7 | 8 | 10 | 9 | 15 | 9.8 |

Table 6.4: Time cost when answering Q1 and Q2.

### 6.5.4 Research Question 2: Discussion

Results has been visualized into below bar char, from the chart we can see that the experiment did have a better performance and understanding of the dataset as they need less time to understand the questions and won higher score. This shows our dashboard can help the user to select the ‘best’ statistical method in a shorter time compared with the traditional way that summarizes results from the tables.

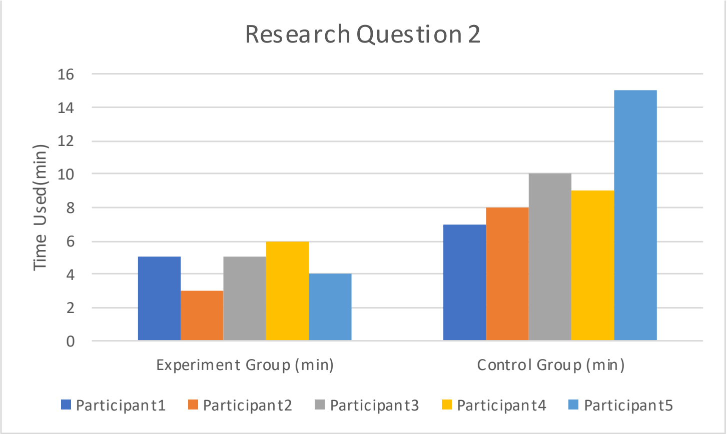


Figure 6.2: Illustrates the experiment group cost less time to figure out best statistical approach compared with the control group.

## 6.6 Research Question 3

How to help doctors in selecting the ‘best’ Machine learning approaches?

### 6.6.1 Experiment

**Pre-experiment:** Machine Learning technology provided to each participant (Appendix C). All the participants were given 15 mins to understand the appendix and 10 mins to get trained of how to use the scene 3 of the personalized health analytic dashboards with empty dataset. When the experiment starts, please be aware they were only allowed to view the “Suggest Machine Learning Model” interface.

**Post experiment:** When the experiment start, all the participants allowed to import the dataset into the dashboard. The dashboard after analysis would prompt the picture with key KPIs. The participants were requested to try best to understand the information presented on the screen and follow the rules (Appendix D) to find out the ‘best’ machine learning model. When they were ready, the participants need to raise their hands and answer below questions:

Q1: “Please suggest the best machine learning model?”

Q2:” Please clarify why you would suggest it as the best machine learning model?”

**Observation:** The author observed how they used the app and collected their answers as well as recording how long time they need to find out the 'best' machine learning model.

### 6.6.2 Control Experiment

**Pre-experiment:** This group of participants were also told to follow the rules to find out the most suitable statistical approach for the research (see Appendix D). They were given 15 mins to review the appendix and should be proficient with all the machine learning concepts and what is the purpose of using the concept or parameter.

**Post experiment:** The participants were given a complete dataset and they were told to use whatever method they like to understand and figure out the 'best' machine learning model. The author also allowed them to search the internet as long as they can find the most suitable model. They can also ask for all the related information of the research. When they were ready, the participants need to raise their hands and answer below questions:

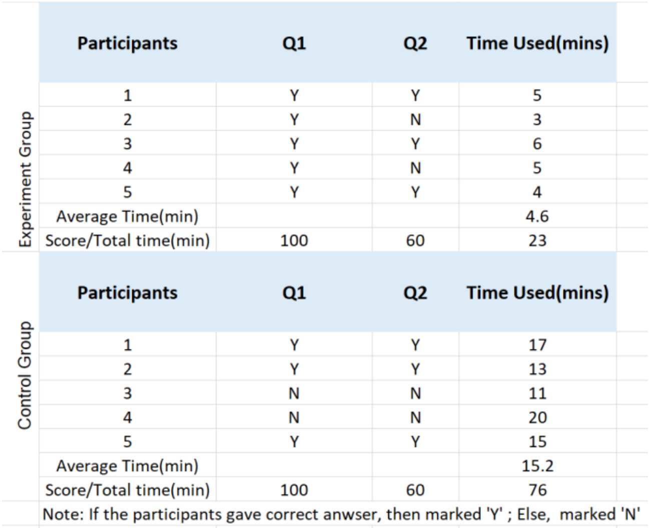
Q1: "Please suggest the best machine learning model?"

Q2:" Please clarify why you would suggest it as the best machine learning model?"

**Observation:** The author observed how they used the app and collected their answers as well as recording how long time they need to find out the 'best' machine learning model.

### 6.6.3 Research Question 3: Results

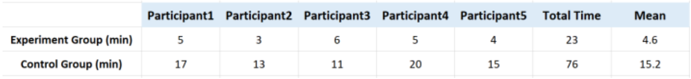
Table 6.6.3 illustrates the time cost from the two experiments. The control group took triple time compared with the experiment group. And the experiment of scores for Q1 was 100 which was much better than the control group's scores 60. However, for Q2 which designed to check the participant's understanding of the machine learning, both groups gave same result at score 60. This result shows it seems our dashboard does not make significant help on the understanding area of our research question which needs further investigation.



| | Participants | Q1 | Q2 | Time Used(mins) |
|---|---|---|---|---|
| Experiment Group | 1 | Y | Y | 5 |
| | 2 | Y | N | 3 |
| | 3 | Y | Y | 6 |
| | 4 | Y | N | 5 |
| | 5 | Y | Y | 4 |
| | Average Time(min) | | | 4.6 |
| | Score/Total time(min) | 100 | 60 | 23 |
| Control Group | **Participants** | **Q1** | **Q2** | **Time Used(mins)** |
| | 1 | Y | Y | 17 |
| | 2 | Y | Y | 13 |
| | 3 | N | N | 11 |
| | 4 | N | N | 20 |
| | 5 | Y | Y | 15 |
| | Average Time(min) | | | 15.2 |
| | Score/Total time(min) | 100 | 60 | 76 |

Note: If the participants gave correct anwser, then marked 'Y' ; Else, marked 'N'

Table 6.4: Overview of the results generated from research question 3.



| | Participant1 | Participant2 | Participant3 | Participant4 | Participant5 | Total Time | Mean |
|---|---|---|---|---|---|---|---|
| Experiment Group (min) | 5 | 3 | 6 | 5 | 4 | 23 | 4.6 |
| Control Group (min) | 17 | 13 | 11 | 20 | 15 | 76 | 15.2 |

Table 6.5: Time cost when answering Q1 and Q2.

### 6.6.4 Research Question 3: Discussion

Here, the author made a visualization of these two groups which can clearly illustrate these two groups participant’s performance of the research question 3.

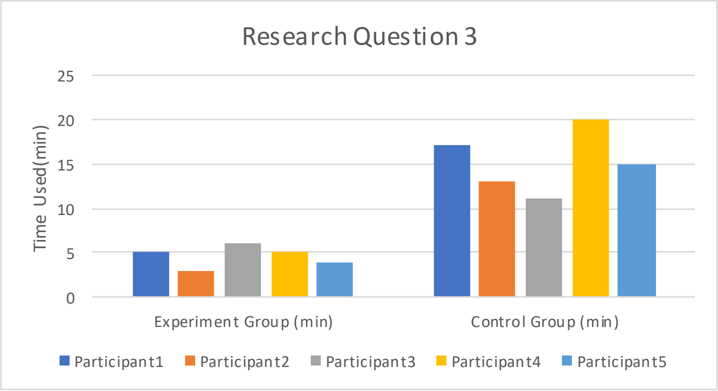


Figure:

### 6.7 Research Question 4

How to use best practice in data visualization to support doctors in the interpretation of the results from the Machine Learning and statistical approaches.

#### 6.7.1 Experiment

**Pre-experiment:** The author randomly chose five graphics from our analysis and printed them. The participants need to observe those graphics and try best to explain what information these pictures convey.

**Post-experiment:** The participants need to write down their understandings on the graph paper as much as they can.

**Observation**: The author needs to observe the participant's reaction and record the time stamp before and after the experiment. The author also needs to check whether the participants have explained the information correctly and clearly.

#### 6.7.2 Control Experiment

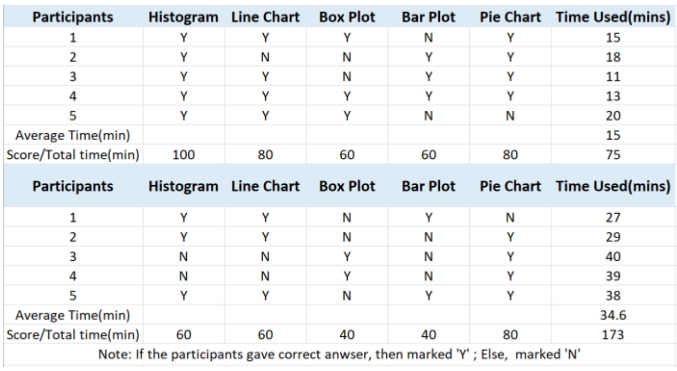
**Pre-experiment:** Same as the experiment group, the control group gave same information, but the difference is the data was shown in a table not in the form of visualization. The author provided them five tables of same information (see Appendix E) and asked the participants to make a summary themselves.

**Post-experiment:** The participants need to observe the data table carefully and explain the meaning of the data, for example, what is the trend of the data, how is the occupation of different data etc.

**Observation**: The author needs to observe the participant's reaction and record the time stamp before and after the experiment. The author also needs to check whether the participants have explained the information correctly and clearly.

#### 6.7.3 Research Question 4: Results

Table shows the experiment result between the two groups, as the two groups are reviewing same information in different methods. The whole experiment group only used 75 mins to digest the information while the control group need double time which was 173 mins to understand the same information. And the first group has won higher score in answering the questions compared with the control group.



| Participants | Histogram | Line Chart | Box Plot | Bar Plot | Pie Chart | Time Used(mins) |
|---|---|---|---|---|---|---|
| 1 | Y | Y | Y | N | Y | 15 |
| 2 | Y | N | N | Y | Y | 18 |
| 3 | Y | Y | N | Y | Y | 11 |
| 4 | Y | Y | Y | Y | Y | 13 |
| 5 | Y | Y | Y | N | N | 20 |
| Average Time(min) | | | | | | 15 |
| Score/Total time(min) | 100 | 80 | 60 | 60 | 80 | 75 |

| Participants | Histogram | Line Chart | Box Plot | Bar Plot | Pie Chart | Time Used(mins) |
|---|---|---|---|---|---|---|
| 1 | Y | Y | N | Y | N | 27 |
| 2 | Y | Y | N | N | Y | 29 |
| 3 | N | N | Y | N | Y | 40 |
| 4 | N | N | Y | N | Y | 39 |
| 5 | Y | Y | N | Y | Y | 38 |
| Average Time(min) | | | | | | 34.6 |
| Score/Total time(min) | 60 | 60 | 40 | 40 | 80 | 173 |

Note: If the participants gave correct anwser, then marked 'Y' ; Else, marked 'N'

Table : Overview of the results generated from research question 4.

### 6.7.4 Research Question 4: Discussion

From below figure we can see that the same information presented in different ways leads to different result. The experiment group was reviewing the information in the form of data visualization and this group used little time to understand the data. In comparison, the experiment has much better performance as they won higher score. Therefore, we may make a conclusion regards this experiment, by the data visualization technique our dashboard did help the user to make correct decisions and save time to understand the difficult terminologies or concepts.

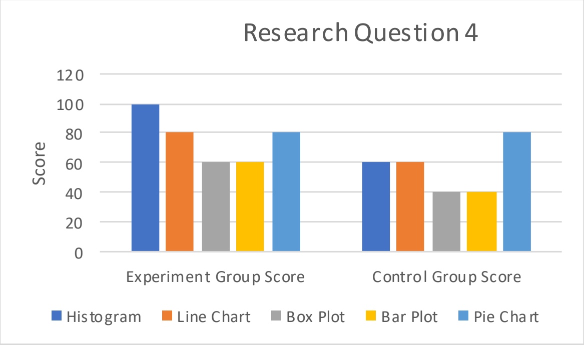


Figure:

## Chapter 7. Conclusions and Future recommendations

### 7.1 Introduction

Data visualization is the presentation of data in a pictorial or graphical format [127]. There are many forms of visualization, such as pie charts, histograms, heat maps, etc. The advantage of visualization is that by using concrete graphics, it can help users better understand the data and even make predictions, so that they can make the best decision in a shorter time. However, the current market lacks a visualization method that reduces communication gaps and supports cross-domain sharing. Therefore, the author considers developing a dynamic automatic analysis visualization dashboard to eliminate these gaps.

The research purpose of this article is to build a dynamic and visual dashboard that allows interdisciplinary doctors to participate, experience the fun of data exploration, and at the same time be able to gain in-depth understanding of patient information and choose the best solution. At the same time, interpreting the analysis results to enhance the understanding of the disease data.

The application has gone through five times of focus group discussions on the prototype and has improved and improved the deficiencies of the analytical dashboard. In this chapter, the author summarizes the findings and innovations of the research in detail and gives some suggestions for improvement in the future.

### 7.2 Conclusions from the Primary Research

As the primary research question is 'How to create an autonomous analysis health dashboard to the users in different stage?'. The primary purpose of data visualization is to make it simpler to spot patterns, trends, and outliers in massive data sets by utilizing images to make data easier for the human brain to interpret and extract insights from [128]. To address the primary research, the author has designed an autonomous health dashboard and add four scenes in it. Scene one is designed to help the doctor to select the 'best' statistical approach. Scene two aims to help the doctor especially who has no IT backgrounds to select a 'best' machine learning model. The reason for the design of Scenario 3 is that most clinicians report that if the application can show the average value better and more clearly, it will be of great help to the diagnosis of the disease. Scene four was designed to add more graphics that may help the doctors to get more information from the graphics and may add some interaction interest towards the users.

The author designed a prototype using Axure tools at the beginning of development and introduced three focus groups. After each group meeting, the author conducted a semi-structured interview with the focus group and recorded all feedback so that it can be updated in the next iteration. The content of the interview is recorded as follows:

- The interface needs to be strictly functionally divided according to user needs.
- The design of the dashboard must be as simple and easy to navigate as possible.
- Highlight key content and delete unnecessary content as much as possible so as not to increase the cognitive burden of users and cause confusion.
- The layout should be consistent, and high-contrast colors should be used appropriately to highlight differences in content.
- The font should be as large as possible, the black Arial font has a better visual effect.

The main results of the research question are as follows:

**Research Question 1:** *How to show a patient how they sit compared to the average?*

In order to solve the problem, the author conducted experimental experiments, one group of experimental groups and one experimental group. Series of questions to check how they understand the data. The only difference is that the experimental group is a dashboard designed by the author, instead of using a traditional data table format. One mention is that the author separately gave this aspect before the start of the experiment. Study 15 minutes to understand the dashboard and record tabular form to reduce the distracting factor data.

The analysis of the experimental results concluded that the data that the experimental group can remember is about 1.5 times that of the control group. And the experiment group hold a positive attitude of the dashboard.

**Research Question 2:** *How to support doctors in selecting the ‘best’ statistical approaches?*

To address this question, two experiments were carried out. Same backgrounds information was given to the experiment group and the control group from which the two groups can all fetch the same information such as the concepts, the technologies would be used during the experiment. To exclude the external interference factors, both the control group and experiment group has 15 mins to get familiar with the tools, for the experiment, the tool is the dashboard while for the control group, the tool is the table which is about to fulfill the same information as the dashboard.

The conclusion is the control group cost nearly double time to choose the 'best' statistical approach and their scores were 20% higher compared with the control group.

**Research Question 3:** *How to help doctors in selecting the 'best' ML approaches?*

Same as the research question 2, two experiments were carried out with two groups. Each group were given same background information and technologies related to the machine learning and some KPIs that were selected as the criterial to choose the 'best' machine learning model.

Then 15 mins were given to each group to be familiar with the tool they were going to use. When experiment starts, the author observes and records how long time they need to figure out the most suitable machine learning model and also asks several questions to see their understandings.

The conclusion of this research question 3 is, regards the machine learning part, the control group took triple time compared with the experiment group and the experiment group has a better understanding of the machine learning concepts.

**Research Question 4:** *How to use best practice in data visualization to support doctors in the interpretation of the results from the ML/statistical approaches?*

The author divided the participants into two groups: the experiment group and the control group. Due to the COVID block down policy by the government and the author has very limited chance to get in touch with the doctor and real patients. Therefore, the number of participants is relatively small.

Use of data visualization is thought to improve comprehension and reduce cognitive load, leading to more effective decision making[114].

------------------

Smoking during pregnancy increases the risk of maternal complications, pre-term delivery and low birth weight[129]. Low birth weight is associated with a variety of important adult health problems including coronary heart disease, type II diabetes and obesity [130]. Tobacco smoking during pregnancy is relatively common, especially in low to middle income populations but is strongly associated with poverty, low educational attainment, poor social support and mental health issues [131]. In Scotland, a strong relationship between smoking during pregnancy and deprivation exists, with records at first booking of pregnant women by maternity services showing smoking prevalence rates ranging from 5.8% in the most affluent communities to 29.4% in the poorest, in the year ending March 2009 [132]. In addition, parental smoking in the home has direct, substantial and immediate impacts on children's health from inhaled environmental tobacco smoke and also potentiates the likelihood of children becoming adult smokers as a consequence of behaviour-modelled parental smoking [133]. Evidence shows when mothers smoke during pregnancy may contribute to overweight beginning later in childhood[134].

With increasing evidence that the intrauterine milieu influences disease development later in life ,cigarette smoking during pregnancy is associated with reduced fetal growth and an increased risk for intrauterine growth retardation[135]–[137], with increased obesity as early as age 5 years [138], and with increased diabetes risk when babies reach adulthood[139]

### 7.2 Future recommendations

## Reference

## Appendix A

Regards the experience of using the app. Semi-structured interview completed after each focus group.

1. Was the app easy to navigate? If not, how so?

2. Was the design of the dashboard clear and visible? If not, how so?

3. Was the layout of the dashboard designed reasonably? If not, how so?

4. How do you think the colour of the dashboard?

5. Was the information on each scene clear and easy to read? If not, how so?

6. Did you feel uncomfortable?

7. Were there any functions, you would expect, missing?

## Appendix B

Questionnaire 1: To be completed after the participants first session.

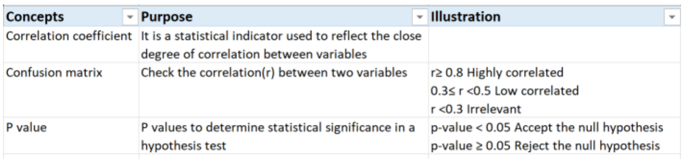
| | Strongly Disagree | | Neutral | | Strongly Disagree |
|---|---|---|---|---|---|
| Was the app easy to use? | 1 | 2 | 3 | 4 | 5 |
| Did you like the layout? | 1 | 2 | 3 | 4 | 5 |
| Did you enjoy the interaction function of the app? | 1 | 2 | 3 | 4 | 5 |
| Was the app easy to use? | 1 | 2 | 3 | 4 | 5 |
| Did you enjoy using the app? | 1 | 2 | 3 | 4 | 5 |
| If not, what did you not like? | | | | | |
| Any other likes or dislikes? | | | | | |
| Would you use this application again? | 1 | 2 | 3 | 4 | 5 |
| If not, why not? | | | | | |
| Would you introduce to your friends? | 1 | 2 | 3 | 4 | 5 |
| If not, why not? | | | | | |

## Appendix C

All the related concepts of the research are summarized as below:

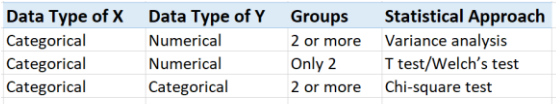
For the control group, they must follow below rules to select the 'best' machine learning model and the statistical approach.

## Statistical:



| Concepts | Purpose | Illustration |
|---|---|---|
| Correlation coefficient | It is a statistical indicator used to reflect the close degree of correlation between variables | |
| Confusion matrix | Check the correlation(r) between two variables | r≥ 0.8 Highly correlated<br>0.3≤ r <0.5 Low correlated<br>r <0.3 Irrelevant |
| P value | P values to determine statistical significance in a hypothesis test | p-value < 0.05 Accept the null hypothesis<br>p-value ≥ 0.05 Reject the null hypothesis |

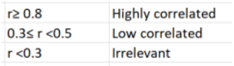
Before asking the participants to find out the best statistical methods, they must use 15 mins time to remember the logic of choosing a most suitable approach (see Table 1 below).



| Data Type of X | Data Type of Y | Groups | Statistical Approach |
|---|---|---|---|
| Categorical | Numerical | 2 or more | Variance analysis |
| Categorical | Numerical | Only 2 | T test/Welch's test |
| Categorical | Categorical | 2 or more | Chi-square test |

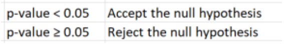
Table 1: Logic to choose a statistical approach.

To check the correlation(r) between two variables, Confusion matrix is involved.



| | |
|---|---|
| $r\geq 0.8$ | Highly correlated |
| $0.3\leq r <0.5$ | Low correlated |
| $r <0.3$ | Irrelevant |

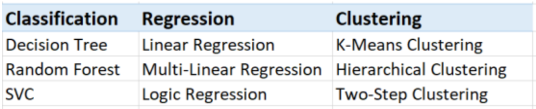
To check the statistical significance p value is involved in a hypothesis test.



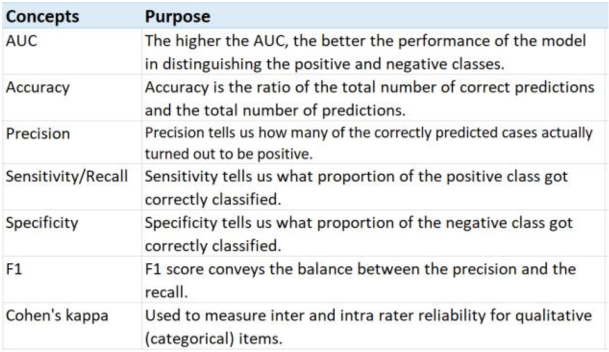
| | |
|---|---|
| p-value $< 0.05$ | Accept the null hypothesis |
| p-value $\geq 0.05$ | Reject the null hypothesis |

Then the participants will be shown below table of the concepts we just discussed:

**Machine Learning**



| Classification | Regression | Clustering |
|---|---|---|
| Decision Tree | Linear Regression | K-Means Clustering |
| Random Forest | Multi-Linear Regression | Hierarchical Clustering |
| SVC | Logic Regression | Two-Step Clustering |



| Concepts | Purpose |
|---|---|
| AUC | The higher the AUC, the better the performance of the model in distinguishing the positive and negative classes. |
| Accuracy | Accuracy is the ratio of the total number of correct predictions and the total number of predictions. |
| Precision | Precision tells us how many of the correctly predicted cases actually turned out to be positive. |
| Sensitivity/Recall | Sensitivity tells us what proportion of the positive class got correctly classified. |
| Specificity | Specificity tells us what proportion of the negative class got correctly classified. |
| F1 | F1 score conveys the balance between the precision and the recall. |
| Cohen's kappa | Used to measure inter and intra rater reliability for qualitative (categorical) items. |

## Appendix D

Background information and app description will be  supplied to all the participants for the experiment.

## Background:

Nowadays, there are various techniques are applied to do data visualization, for example, the machine learning techniques. As doctors are very busy in the daily work, and most of them lack professional AI and statistical knowledge, which makes it difficult to choose the most accurate and effective method to assist treatment.

At the same time, both patients and doctors found it useful by comparing with the average value (such as the conversion of thyroid-stimulating hormone in certain populations), and it gives intuitively understanding for their current health status. Therefore, the author designed and developed this interactive, automatic analysis health dashboard and recommended the most suitable models and statistical methods for inspection or prediction through scientific calculations.

## What is an autonomous health analytic dashboard?

It is an app that can help the user to choose the best model and statistical method as well as interpret the analysis result by data visualization.

- Suggested machine learning model:  three models are involved, SVC model, random forest model and logic regression model. By comparing the AUC (Area Under ROC Curve) the app can tell which model can best distinguish the patients with disease and no disease.
- Suggested statistical approaches: four test methods are involved, T-test and Chi-square test, Welch's test. Please be aware that the t-test has three types: one sample t-test, two independent samples t-test and a paired t-test.

## What is the purpose of this app?

1. To help the doctor select the 'best' machine learning model.

2. To help the doctor select the 'best' statistical approach.

3. Calculate the average value of a specific indicator (including the average value of the control group) and present to the users.

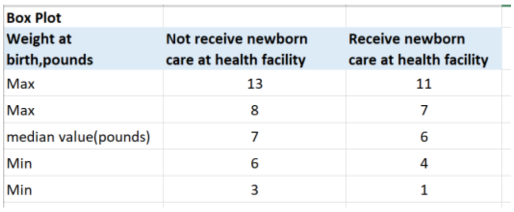
4. To interpret the analysis results by data visualization with various graphs.

## What is looks like?

1. The Main window:

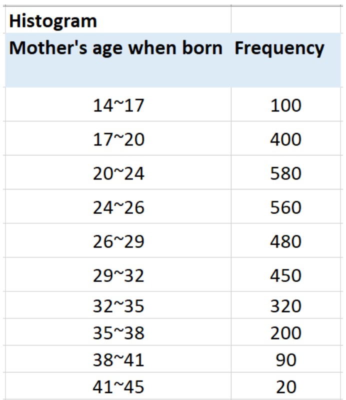
## Appendix E

Below table were designed for the control group of research question 4:



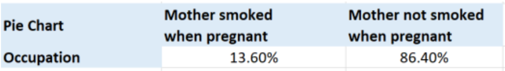
| Box Plot | | |
|---|---|---|
| Weight at birth,pounds | Not receive newborn care at health facility | Receive newborn care at health facility |
| Max | 13 | 11 |
| Max | 8 | 7 |
| median value(pounds) | 7 | 6 |
| Min | 6 | 4 |
| Min | 3 | 1 |

Table 1: Box Plot in table



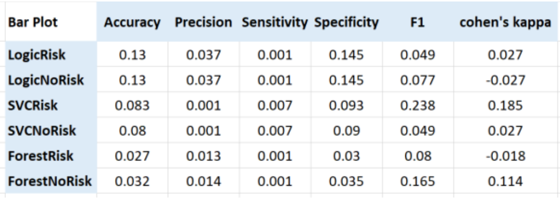
| Histogram | |
|---|---|
| Mother's age when born | Frequency |
| 14~17 | 100 |
| 17~20 | 400 |
| 20~24 | 580 |
| 24~26 | 560 |
| 26~29 | 480 |
| 29~32 | 450 |
| 32~35 | 320 |
| 35~38 | 200 |
| 38~41 | 90 |
| 41~45 | 20 |

Table 2: Histogram in table



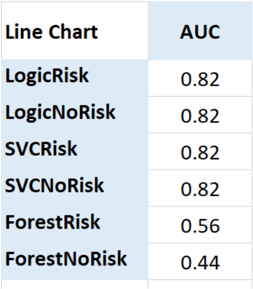
| Pie Chart | Mother smoked when pregnant | Mother not smoked when pregnant |
|---|---|---|
| Occupation | 13.60% | 86.40% |

Table 3: Pie Chart in table



| Bar Plot | Accuracy | Precision | Sensitivity | Specificity | F1 | cohen's kappa |
|---|---|---|---|---|---|---|
| LogicRisk | 0.13 | 0.037 | 0.001 | 0.145 | 0.049 | 0.027 |
| LogicNoRisk | 0.13 | 0.037 | 0.001 | 0.145 | 0.077 | -0.027 |
| SVCRisk | 0.083 | 0.001 | 0.007 | 0.093 | 0.238 | 0.185 |
| SVCNoRisk | 0.08 | 0.001 | 0.007 | 0.09 | 0.049 | 0.027 |
| ForestRisk | 0.027 | 0.013 | 0.001 | 0.03 | 0.08 | -0.018 |
| ForestNoRisk | 0.032 | 0.014 | 0.001 | 0.035 | 0.165 | 0.114 |

Table 4: Bar Plot in table



| Line Chart | AUC |
|---|---|
| LogicRisk | 0.82 |
| LogicNoRisk | 0.82 |
| SVCRisk | 0.82 |
| SVCNoRisk | 0.82 |
| ForestRisk | 0.56 |
| ForestNoRisk | 0.44 |

Table 5: Line Chart in table