Diabetes Analysis Dashboard

By

MARCELL AGUNG WAHYUDI

TP058650

APU3F2211CS(DA)

A project submitted in partial fulfilment of the requirements of Asia Pacific University of Technology and Innovation for the degree of

BSc (Hons) in Computer Science with Specialism in Data Analytics

Supervised by Dr. Vazeerudeen Hameed

2nd Marker: Mr. Raheem Mafas

Acknowledgement

I would like to take a moment to express my gratitude to everyone who has supported me throughout my research.

First and foremost, I would like to thank my supervisor, Dr. Vazeerudeen Hameed for his guidance and support throughout the entire process. Their wisdom, knowledge, experience has been invaluable in helping me navigate through the challenges and complexities of this work.

Besides that, I'd like to salute our FYP project manager, Mr Dhason Padmakumar, who led FYP discussions to teach us how to use the FYP Bank, complete the PPF, PSF, Ethic Form, and IR, and submit these documents.

I would also like to thank my colleagues/classmates/friends who have provided me with their insights and feedback, and who have been a constant source of inspiration and motivation. Your collaboration has been instrumental in helping me achieve my goals.

Furthermore, I would like to acknowledge the institutions that have provided me with the resources necessary to carry out this research. Your support has enabled me to pursue my academic and professional aspirations.

Finally, I would like to express my appreciation to my family and loved ones for their unwavering support and encouragement. Your love and encouragement have been my anchor throughout this journey.

Thank you all for your support and encouragement.

Table of Contents

Table of Figures

CHAPTER 1: INTRODUCTION TO THE STUDY

1.1 Background to the project

In this current year and age, more and more people have been infected by many diseases and health conditions making life rougher for them and since some of these diseases does not have a complete full cure, and available medications are only available to prevent it to worsen. Although these diseases are not deadly, ignoring them completely is proven to be fatal for the diagnosed.

One disease that will be focused and discussed in this project is Diabetes. Diabetes is a long-lasting health condition that influences how a person's bodily system turns food into energy, normally, the human body gets energy by turning the eaten food and processes it into sugar and then delivering it to the bloodstream, but people suffering from Diabetes has to take medications in order for the said normal process to happen normally, they need to take medications for Diabetes type 1 or type 2 whichever one they suffer from (Centers for Disease Control and Prevention, 2022).

Diabetes is usually characterised as 3 types, diabetes type 1, type 2, and gestational diabetes or diabetes while pregnant (Centers for Disease Control and Prevention, n.d.). Type 1 diabetes is a chronic condition that was originally recognized as juvenile diabetes or insulin-dependent diabetes. The pancreas produces little or no insulin in this condition. Insulin is a hormone produced by the body that allows sugar (glucose) to move into the cell and produce energy. Type 1 diabetes can be caused by a variety of factors, including genetics and certain viruses. Although type 1 diabetes typically manifests in childhood and teenage years, it can still manifest in adults. Despite extensive research, there is no cure for type 1 diabetes. To avoid complications, treatment focuses on controlling blood sugar levels with insulin, diet, and lifestyle changes. Known symptoms of type 1 Diabetes are: feeling thirstier, more urination, hungrier, weight lost, more irritable, more tired and weaker, and blurry vision (Mayo Clinic, n.d.).

Type 2 diabetes is characterised by a malfunction inside the way the human body controls and utilises sugar (glucose) as an energy. This long-term (chronic) condition causes an excess of sugar to circulate in the bloodstream. High blood sugar levels could indeed ultimately cause circulatory, nervous, and immune system problems. There are mainly 2 interconnected issues at work in type 2 diabetes. The pancreas fails to produce sufficient insulin (Mayo Clinic, n.d.).

Diabetes diagnosed for the very first time during pregnancy is known as gestational diabetes (gestation). Gestational diabetes, like other types of diabetes, affects how the cells use sugar (glucose). Gestational diabetes causes high blood sugar levels, which has an effect on the pregnancy and the baby's health. Whereas any pregnancy complication is a cause for concern, there is some good news. The person can help manage gestational diabetes during pregnancy by consuming nutritious foods, exercising, and, if required, taking medication. Controlling the blood sugar levels can help the person and the baby remain healthy and avoid a tough delivery. If a person has gestational diabetes throughout your pregnancy, the blood sugar generally reverts to normal shortly after birth. However, if a person have had gestational diabetes, they are more likely to develop type 2 diabetes. they must be tested for changes (Mayo Clinic, n.d.).

Predictive analytics makes use of data analysis to forecast future predictable outcomes, enabling analysts to gain crucial knowledge and improve your business decisions. Predictive analytics can be utilized to acknowledge trends, forecast future occurrences, and determine specific risks and opportunities. Predictive analytics, regardless of the industry, can help anticipate business requirements and comprehend consumer trends (Bold BI, n.d.).

The method of employing statistical and/or logical methods to define and demonstrate, condense, and recapitulate, and evaluate data is known as data analysis. Various analytic procedures "provide a method for drawing inductive conclusions from information and defining the message (the occurrence of interest) from the noise (statistical fluctuations) existing in the data." (Shamoo & Resnik, 2003).

In this project, the researcher will be designing a prediction dashboard based on data analysis on a database of Diabetes consisting of various variable including age, sex, cholesterol level, body mass index, smoker or not, heart disease, physical activity, people’s diet, alcohol consumption, mental health, physical health, stroke, and blood pressure with the target variable being if the person has had diabetes or not alongside extensive literature review on diabetes. In the conclusion part of the project, the researcher will produce an algorithm predicting the chance of a person getting diabetes based on the variables shown in the database.

1.2 Problem statements

Diabetes affects approximately 537 million people in the age of 20-79 years old and more than 1.2 Million children and adolescents in the age of 0-19 years old are affected by type 1 diabetes in 2021, and the number of people affected by Diabetes is predicted to continue to rise amounting to be 643 million by 2030 and 83 million by 2045, Diabetes caused 6.7 Million deaths in 2021, meaning 1 in every 5 seconds (Federation, 2021).

Most people only know that Diabetes is genetic, meaning, once their parents have it, they will inherit it, and vice versa, if one’s parents does not have Diabetes, they think it is impossible for them to get Diabetes, although the chance of that is medium to low, it does still happen to many people especially with bad lifestyle choices.

Many people lack the knowledge of the correct lifestyle and eating habit to prevent said health conditions from happening, thus many people being affected by it. The correct lifestyle will be explained step by step at the outcome part of the project, results will be gathered from real patient data, and combined alongside of doctor’s/experts advise on Diabetes, the planned lifestyle should increase the health quality of any person.

Another problem is lack of information, Diabetes, which can affect people without having to show any symptoms, is predicted to have half of people affected by it to be clueless that they are affected by it.

1.3 Rationale

Diabetes is a challenging medical condition with no known medication to completely cure it, it is also one of the known medical conditions with rapid growth in the world, reporting approximately 537 million people worldwide, with nonstop rising numbers within the last decade, and still predicting growth to be around 783 million with diabetes by 2045, an increase of 46% (Zia Sherrel, 2022). From extensive research, many things are affecting diabetes, some of which are ethnicity, genetics, exercise, diet, and education, most of which can be controlled such as amount of exercise and diet, although ethnicity and genetics can’t be manipulated, these lifestyle choices can be improved to avoid getting diabetes. These lifestyle choices are still unknown whether how often one needs to exercise or how much sugar is the limit of one’s diet, or does the stress level of a person will affect them getting diabetes or not. These variables are the building blocks of the creation of predictive analysis dashboard which will calculate the right amount of exercise and diet, hopefully convincing more people to have a healthier lifestyle, reducing the number of diabetics.

1.4 Potential benefits

1.4.1 Tangible benefits

- Producing a lifestyle plan that includes how often a person should exercise and their diet plan.
- Giving insight of how lifestyle affects the chances of getting diabetes.
- Informing potential diabetics and providing information about diabetes.

1.4.2 Intangible benefits

- Improving the wellbeing of many people with bad diet.
- Potentially improve the well-fare of diabetics.

1.5 Target users

Due to the fact that diabetes affects the worldwide population, the target users of the prediction model are for all people with no exclusion, anyone may utilize the calculated lifestyle choices to improve their health as non-diabetics as well as diabetics.

1.6 Scope and objectives

1.6.1 Aim

The primary aim of this project is to analyse the current characteristics of around 100.000 records of patients, both having diabetes and diabetes-free using machine learning algorithms and displaying the result in a predictive analysis dashboard to find correlation between the variables and the target variables.

1.6.2 Objectives

- Determining useful variables that has correlation to the target variable.
- Applying exploratory data analysis to find hidden correlation and patterns might be useful for later research.
- Design and develop machine learning model to analyze the connection of the characteristics to diabetes.

1.6.3 Deliverables

- An analysis dashboard, visualizing correlation among various variables and its correlation with diabetes.

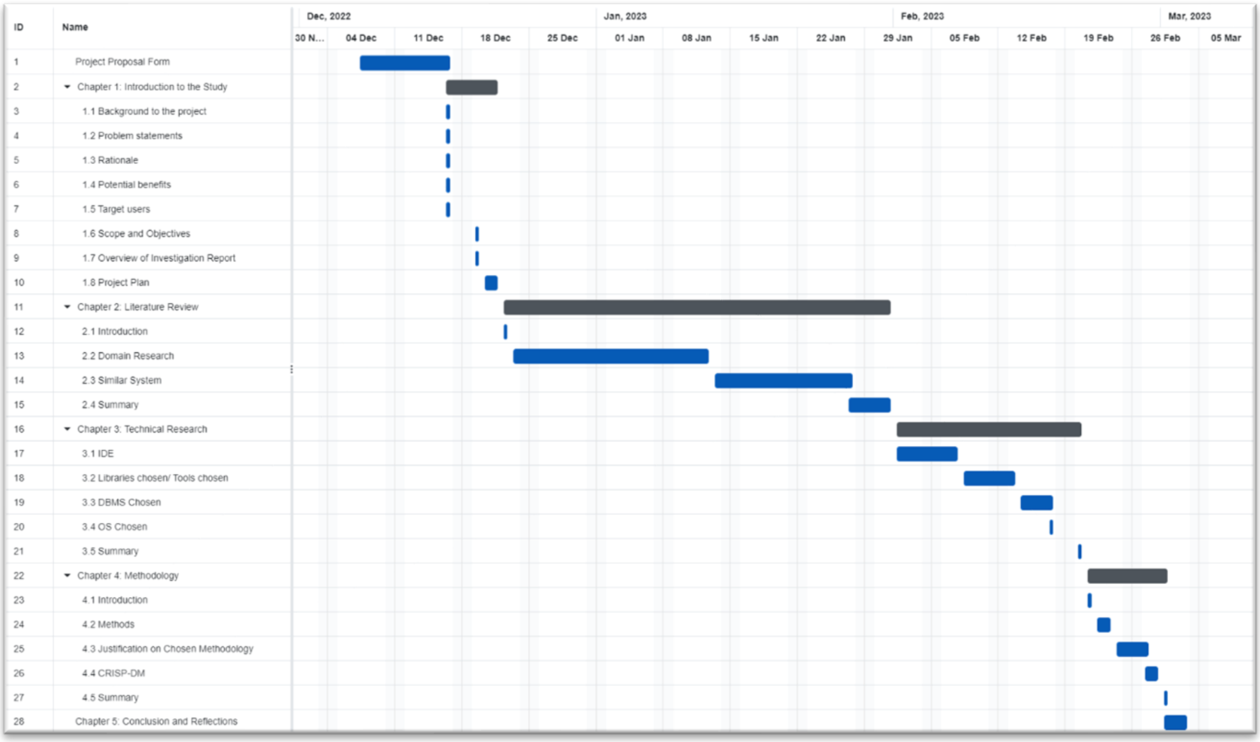
1.6.4 Nature of Challenges

The main challenge that determines the success of this project is the factor of genetics, ethnicity and other factors that cannot be predicted and manipulated. Genetic is known to have huge impact on diabetes itself, but in order to check the genome, the person needs to have extensive medical check-up and thus, it will not be included as a variable in the predictive model. As for ethnicity, although it can be guessed roughly by the person, it is still difficult to pinpoint the exact percentage of the ethnicity. Therefore, ethnicity will also not be included in the predictive analysis, That being said, both genetic and ethnicity will be discussed in the case if the person has knowledge and information of their own genetic and ethnicity, meaning they can have more insight information to whether they are susceptible to diabetes or not.

1.7 Overview of this report

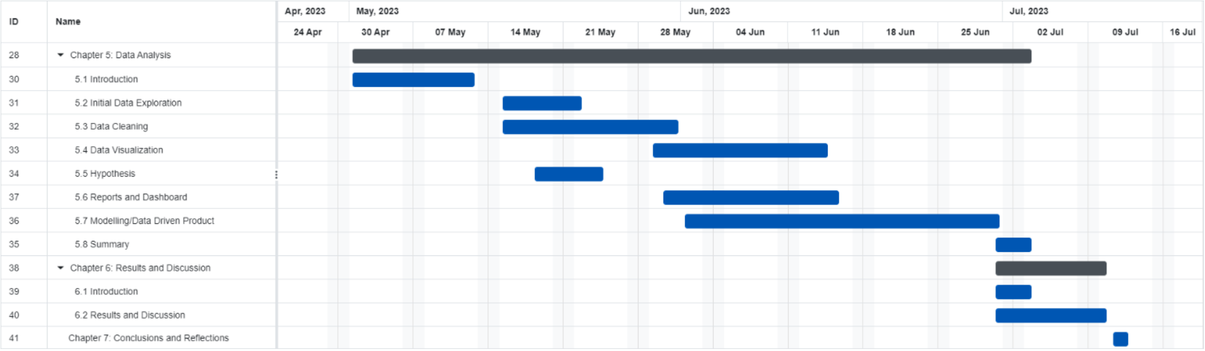
Included in this Investigation Report are five chapters. At the start of the report, Chapter 1, fully introduces and explains the background of the project, the problem statement, rationale, potential benefits, target users, and scope, aims and objectives. The next chapter, Chapter 2 includes literature review of correlating studies to back up found outcomes of the analysis. In this chapter, the researcher aims to find correlating studies of roles of genetics and ethnicity to diabetes and also supporting facts for the analysis. Chapter 3 consists of technical research, which are tools to be used in the production of the predictive analysis dashboard, including programming language used, IDE, with justification on why the chosen application is used. Chapter 4 is about methodologies used to develop the project, with justification and reasons for why to use the selected methodology. Chapter 5, the last chapter of the report, will conclude everything on the report, covering summaries and also reflection for the researcher, while also identifying gaps for future research about similar topic.

1.8 Project Plan



ID	Name
1	Project Proposal Form
2	▾ Chapter 1: Introduction to the Study
3	1.1 Background to the project
4	1.2 Problem statements
5	1.3 Rationale
6	1.4 Potential benefits
7	1.5 Target users
8	1.6 Scope and Objectives
9	1.7 Overview of Investigation Report
10	1.8 Project Plan
11	▾ Chapter 2: Literature Review
12	2.1 Introduction
13	2.2 Domain Research
14	2.3 Similar System
15	2.4 Summary
16	▾ Chapter 3: Technical Research
17	3.1 IDE
18	3.2 Libraries chosen/ Tools chosen
19	3.3 DBMS Chosen
20	3.4 OS Chosen
21	3.5 Summary
22	▾ Chapter 4: Methodology
23	4.1 Introduction
24	4.2 Methods
25	4.3 Justification on Chosen Methodology
26	4.4 CRISP-DM
27	4.5 Summary
28	Chapter 5: Conclusion and Reflections

Figure 1: Project Plan for FYP Semester 1



ID	Name
28	▾ Chapter 5: Data Analysis
30	5.1 Introduction
31	5.2 Initial Data Exploration
32	5.3 Data Cleaning
33	5.4 Data Visualization
34	5.5 Hypothesis
37	5.6 Reports and Dashboard
36	5.7 Modelling/Data Driven Product
35	5.8 Summary
38	▾ Chapter 6: Results and Discussion
39	6.1 Introduction
40	6.2 Results and Discussion
41	Chapter 7: Conclusions and Reflections

Figure 2: Project Plan for FYP Semester 2

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Literature review, can be considered as an examination of academic articles on a particular topic, which provides a snapshot of a current knowledge, allowing a researcher to expand their current research paper with correlating topic, giving them more insight on theories and methods other researcher have prepared and applying them to their own research paper (McCombes, How to Write a Literature Review, 2023).

2.2 Domain research

2.2.1 Diabetes Mellitus

Diabetes, or in technical term Diabetes mellitus, is a diverse group of disorders characterised by higher-than-normal blood glucose levels. Diabetes has been classified into four types: insulin-dependent diabetes mellitus (IDDM), non-insulin-dependent diabetes mellitus (NIDDM), gestational diabetes mellitus (GDM), and diabetes secondary to other conditions. IDDM is characterised by notable diabetes symptoms and severe hyperglycaemia. The presence of the classic symptoms and indicators of diabetes, as well as unambiguously elevated blood glucose levels; fasting plasma glucose (FPG) 2140 mg/dl; or venous plasma glucose 2200 mg/dl at 2 hours after a 75-g oral glucose challenge can all be used to diagnose NIDDM. GDM criteria vary, but all necessitate an oral glucose challenge and post-load plasma glucose measurement (Aubert, Ballard, & Barret-Connor, 1995).

In the United States, the incidence of insulin-dependent diabetes mellitus (IDMM) with onset at 30 years is estimated to be 120,000 individuals aged 20 years and 300,000-500,000 individuals of all ages. Adult-onset IDDM may affect up to 500,000 adults (onset at age 230 years). Every year, 30,000 new cases of IDDM are identified. The occurrence of IDDM in children differs widely by ethnicity and race. For example, in Colorado, the incidence rate per 100,000 was 8.8 for Hispanics, 12.1 for African Americans, and 17.3 for whites. Some European countries have experienced an increase in the incidence of IDDM over time. In the United States, incidence has remained steady for the past few decades, with the exception of brief spikes in certain years (Aubert, Ballard, & Barret-Connor, 1995).

In 1993, 90%-95% of the 7.8 million individuals living in the United States with diabetes seems to have non-insulin-independent diabetes mellitus. The incidence rates were 1.3% between the ages of 18 and 44, 6.2% between the ages of 45 and 64, and 10.4% between the ages of 265 and 265. Based on oral glucose tolerance testing in NHANES II, there is estimated one undiagnosed situation of NIDDM per each diagnosed case in addition to the confirmed instances of diagnosed NIDDM. As a result, the total incidence of NIDDM is approximated to be two times that of diabetes. When IGT is taken into account, total glucose intolerance rates can range from 9% at age 20-44 years to 42% at age 65-74 years. In the United States, the incidence of NIDDM is similar for men and women, but diabetes is much more frequent in blacks, Mexican Americans, Japanese Americans, and Native Americans than in non-Hispanic whites. According to the NHIS, 625,000 fresh diabetic cases are discovered in the United States each year (Aubert, Ballard, & Barret-Connor, 1995).

2.2.2 Diabetes and Ethnicity

Diabetes, particularly type 2 diabetes, has become more prevalent in all Western nations in recent years. Related trends have been noticed as well within these nations that are implementing a more 'western lifestyle'. These findings clearly indicate that external conditions associated with the urbanisation process have a massive effect on a large percentage of individuals globally. Studies have compared the incidence of diabetes in various ethnic groups who live in various environments support this. Diabetes was discovered to be pervasive in a small region of Japan in the 1970s, with an incidence of around 4%; a slightly greater incidence was discovered in Tokyo within the same age group. Incidence of diabetes has continued to increase in Japan, albeit with increasing values discovered in Japanese individuals residing in Hawaii, ultimately hitting a 21% incidence among those living in the continental US. In Korea, a similar trend has been noticed. Incidence rate was 2% and 4% in small towns and rural town areas, respectively, a low value when as compared to 13 and 15.9% in Seoul. Diabetes is at least 12 times more prevalent in African Americans than in Native African Black people (12 and 1%). It is widely known that Pima Indians have the highest incidence of type 2 diabetes, but Ravussin et al. found that Pima Indians in Arizona, who are genetically linked to those in Northern Mexico, have a much higher rate of diabetes than Mexican Pima Indians, with 54 and 37% vs. 6 as well as 11% for men and women. These studies demonstrate that environmental factors play an important role in the pathogenesis of diabetes and insulin resistance in all populations throughout the world. It has been discovered that the incidence and extent of change in the incidence of diabetes varies between many various ethnic groups. In fact, some ethnicities in multiracial communities not only experience a much greater rate of diabetes than a similar ethnic community who reside in their country of origin, but also an increased proportion when compared to other ethnic communities residing in the same surroundings. These findings suggest that when revealed to equivalent unfavourable environmental conditions, some ethnic groups are more likely than others to acquire type 2 diabetes. These data show that the 'western way of life' has a massive effect on the rising incidence of diabetes throughout all ethnic groups, but they also demonstrate that there is a cultural predisposition or susceptibility to developing diabetes. This proneness could be caused by genetic deficiencies in either insulin production or insulin action (L., et al., 2005).

2.2.3 Diabetes and Genetics

Even though some encouraging resistance genes, such as calpain-10, PPAR- and Kir 6.2.9-12, have been identified, the genetic base of type 2 diabetes mellitus remains unknown. The insulin signalling pathway begins with insulin binding to the -subunit of the insulin receptor and ends with insulin's biological effects in multiple tissues. Insulin stimulates glucose transport, which is one of its most important effects. Although other pathways stimulate glucose uptake, insulin receptor substrates and phosphatidylinositol 3-kinase are the primary regulators of glucose transport (PI3K). Single nucleotide mutations in genes that control insulin signalling and secretion have been discovered with varying frequency in the overall population, and many of these polymorphisms have been linked to a higher risk of type 2 diabetes. Polymorphisms in nearly any of the genes associated in the protein expression that govern insulin signalling could impair skeletal muscle glucose utilisation and play a role to insulin resistance. When multiple mutations or polymorphisms of genes that are individually affiliated with minor changes in insulin sensitivity are combined, the biological effects of insulin sensitivity may be drastically decreased. Several insulin receptor mutations have been identified; however, because the regularity of these mutations in the overall population is low, it is doubtful that they make a significant contribution to the pathogenesis of insulin resistance within the general population (L., et al., 2005).

2.2.4 Diabetes and Obesity

Obesity is a powerful indicator of type 2 diabetes development. Obesity rates have risen across many nations over the past few years. Diabetes is a consequence of the interaction of both genetic and environmental factors. Lack of physical activity, habitual energy consumption in relation to expenditure, macronutrient structure of the eating plan, and metabolic qualities are among these factors. Obesity is correlated with a rise in the prevalence of type 2 diabetes. According to one study. Although putting on weight throughout adolescence led directly to adult obesity. An increase in adult overweight and obesity is associated with a higher likelihood of type 2 diabetes. In a study performed by Shahraki and colleagues, poor education level was discovered to be a powerful indicator of overweight and obesity among adolescents. The Nurses Healthy Study showed that participants with Bmi values less than 21 had a reduced risk of diabetes. As the incidence of diabetes has already been linked to obesity, only a few research findings had strict conditions for age and gender. Those higher BMI individuals experience a greater incidence of type 2 diabetes at a relatively young age than lower BMI people, whose incidence rises as they get older. According to some research, waist measurement or waist-to-hip proportion is a strong predictor of the prevalence of diabetes as well as cardiovascular disease risk factors at various ages than BMI. According to some research, the dispersion of fat mass is a significant risk indicator for abdominal obesity or visceral fat. In Japanese American men, for example, intra-abdominal fat, as tested by CAT scan, was the best anthropometric forecaster of diabetes occurrence. Given the significance of central adiposity as a potential cause for diabetes, it is critical to understand the normal BMI (18.5-24.9 kg/m2) for all peoples. According to some studies, a normal BMI of 21 kg/m2 raises the likelihood of developing diabetes. Seeing as Pacific people have a larger percentage of muscle mass than Europeans, a higher BMI cut off might be appropriate for these inhabitants. Since data on waist size and waist-to-hip proportion are inconsistent, the WHO suggested BMI range (18.5- 24.9 kg/m2) and population mean of 21 kg/m2 should be used. (Rahati, Shahraki, Arjomand, & Shahraki, 2014).

2.2.5 Effects of Diet on Diabetes

Environments have an impact on the entire population that is exposed to them. By influencing the policies of corporations, governmental organizations, and other organisations whose choices have an impact. Many individuals might have the ability to modify the negative health environment, thus also shifting obesity on a population level. Furthermore, the fact that the present environment is not favourable to eating healthy and exercise may clarify why most regular exercise and healthy diets programmes have not been effectively maintained. Such education programmes and individual-level remedies will be ineffective if the surroundings make it difficult to follow the recommendations—for example, following a sensible diet is challenging if grocery stores do not make healthier foods plentifully and continuously accessible at good prices. The availability of nutritious foods and chances to exercise could be among the variables that contribute to the incidence of obesity in people with lower socioeconomic status. As a result, studying methods for altering the macroenvironment and thus alter dietary habits and regular exercise is an important new path for behavioural research. According to correlational data, environmental variables impact eating and exercising. Cheadle et al., for instance, discovered clear links among fat intake and the proportion of local grocery store shelf space dedicated to low-fat compared to regular meat and milk. Correspondingly, the concentration of physical exercise establishments in the neighbouring community and the amount of workout equipment in the residence have been linked to adult levels of physical activity. Children's physical activity levels have been demonstrated to be related to neighbourhood environment characteristics (Wing, et al., 2001).

2.2.6 Data Analysis

Data analysis refers to the act of cleaning, transforming, and refining data in order to productivity improvement, applicable information that assists businesses in making informed choices. The method helps minimize the hazards innate in judgement call by supplying important ideas and statistical data, often displayed in graphs, pictures, tabular, and charts. In order to produce desirable outcomes, few steps are required in order to do so, the first step to producing business decision is data requirement gathering, this involves understanding why the analysis is being done, the type of data needed, and what data is intended for analysis. The next is data collection, After identifying the requirements, it is necessary to collect data from various sources such as surveys, case studies, interviews, direct observation, questionnaires, and focus groups. Organizing the collected data for analysis is essential. Then to cleaning the data, since not all collected data will be useful. This step involves removing white spaces, duplicate records, and basic errors. Data cleaning is a must before sending the data for analysis. Then data analysis comes to analyse the data, data analysis tools such as R, Python, Excel, Looker, Rapid Miner, Chartio, Metabase, Redash, and Microsoft Power BI are utilized to interpret and comprehend the data to arrive at conclusions. After analysed, The results are interpreted to determine the best courses of action based on the findings, this is called Data Interpretation. And finally, to communicate those results, data visualization, Visualization is an essential aspect of data analysis, allowing the data to be graphically presented to enhance its understanding. Visualization methods include maps, charts, bullet points, and graphs, helping to identify relationships and compare datasets to draw valuable insights (Kelley, What is Data Analysis? Methods, Process and Types Explained, 2023).

2.3 Similar System

2.3.1 Machine learning and deep learning predictive models for type 2 diabetes: a systematic review.

A machine learning and deep learning predictive models was developed by Luis Frergoso-Apricio, Julieta Noguez, Luis Monstesinos, and Jose A. Garcia-Garcia in 2021, the aim of the predictive model is to identify and report on potential areas for enhancing diabetes type 2 prediction utilising methods based on machine learning. They implemented two information sources, one is from PubMed, providing information on medical problems such as diabetes and potentially computer science solution, and Web of Science, providing the ability to find articles with high correlation with their search target. Firstly, after the articles were selected, they were First, the articles were classified into two groups depending on the kind of information provided (glucose forecasting or electronic health records). The first list comprises models that monitor blood glucose control levels, while the second group consists of models that predict diabetes relying on electronic medical records. The second classification was more thorough, utilising the Machine learning model, Validation parameter (sensitivity, accuracy, specificity, F1-score, AUC(ROC)), Data sampling (cross-validation, training-test set, complete data), Complementary techniques (complementary statistics and machine learning techniques for modelling), Description of the population (age, size, etc.). The initial search yielded 1327 results, including 925 from PubMed and 402 from Web of Science. When sorting by publication year (2017-2021) only 130 records were exempted. To narrow the outcomes, additional searches were carried out utilising fine-tuned search terms and alternatives for both databases. A fresh search was conducted using original key phrases but limiting the term 'diabetes' to the title, yielding 517 records from both databases. 51 duplicates were disposed. As a result, 336 records were chosen for further review. Thirty-seven records were excluded because the study was using non-omittable genetic features as input parameters, which was outside the topic of this review. Because they were review papers, 138 records were excluded. In total, 261 articles met the criteria and participated in the quality evaluation. The summary of their project is, given the heterogeneity, this evaluation was unable to recognise an accurate set of characteristics in the datasets. There are, however, some findings to report. First, the dataset has a significant impact on the model's performance: precision declined substantially as the dataset grew larger and more complex. Tidy, well-structured datasets with a small number of samples and features result in a better model. Even so, a small assortment of characteristics may not accurately reflect the actual complex nature of multifactorial diseases. The decision tree and random forest were the best-performing models, with accuracy comparable of 0.99 and AUC (ROC) of one. The strongest predictors for the accuracy metric on average were Swarm Optimization and Random Forest, both with a value of one. For AUC (ROC) decision tree with an AUC (ROC) of 0.98. Deep Neural Networks, tree-type (Gradient Boosting and Random Forest), and support vector machine learning were among the most used methods. Deep Neural Networks have the benefit of dealing well with large amounts of data, which is a compelling reason to use them frequently. These models have been employed in research with datasets containing over 70,000 findings. Furthermore, these models perform well with skewed data (Fregoso-Aparicio, Noguez, Montesinos, & Garcia-Garcia, 2021).

2.3.2 Data Visualization That “Fits”: Designing Effective Dashboards for Healthcare Providers, Patients, and Family Caregivers to Patients with Diabetes

The article discusses the design of data dashboards for individuals involved in the care of patients with diabetes. It emphasizes the importance of presenting health data in an easily understandable and accessible way for different users, such as healthcare providers, patients, and family caregivers. To identify the key data elements that should be included in a diabetes dashboard and the preferred visualizations for each group, the authors conducted a survey of healthcare providers, patients, and family caregivers. The survey results showed that each group has different information needs and preferences for data presentation. Healthcare providers, for example, may need detailed information on blood glucose levels and medication adherence to help them make clinical decisions. Patients, on the other hand, may prefer simpler visualizations that are easy to understand and act upon. Family caregivers may require a combination of both detailed and simplified information to help them understand and support the patient. Based on the survey results, the authors provide several recommendations for designing effective diabetes dashboards. These include using clear and simple visualizations that can be easily interpreted by different users, providing context for the data to help users understand its significance, and tailoring dashboards to specific user needs. The authors also suggest that effective dashboards should be user-friendly and accessible on multiple devices, such as smartphones and tablets, to increase their usefulness and usability. To conclude, the article stresses the importance of designing effective dashboards for different users in healthcare and provides valuable insights for dashboard design. By tailoring dashboards to specific user needs and preferences, healthcare providers, patients, and family caregivers can better understand and manage diabetes, leading to improved health outcomes for patients (Teves, 2015).

2.4 Summary

In conclusion, after extensive literature review about diabetes, it is seeming to be predictable to some degree, with diet being one of the major factor of diabetes, followed by genetics and ethnicity. This research will contribute to the thought of the predictive analysis to make sure the end model is not bias. The similar system, although discusses diabetes model, machine learning and deep learning was utilized to create the best model to predict diabetes and the results of the project gave valuable insight to be added on this investigation report. Unfortunately, the predictive model designed in this project will not be using machine learning, since exploratory data analysis is utilized.

CHAPTER 3: TECHNICAL RESEARCH

3.1 Web IDE (Interactive Development Environment) chosen

3.1.1 Google Colaboratory

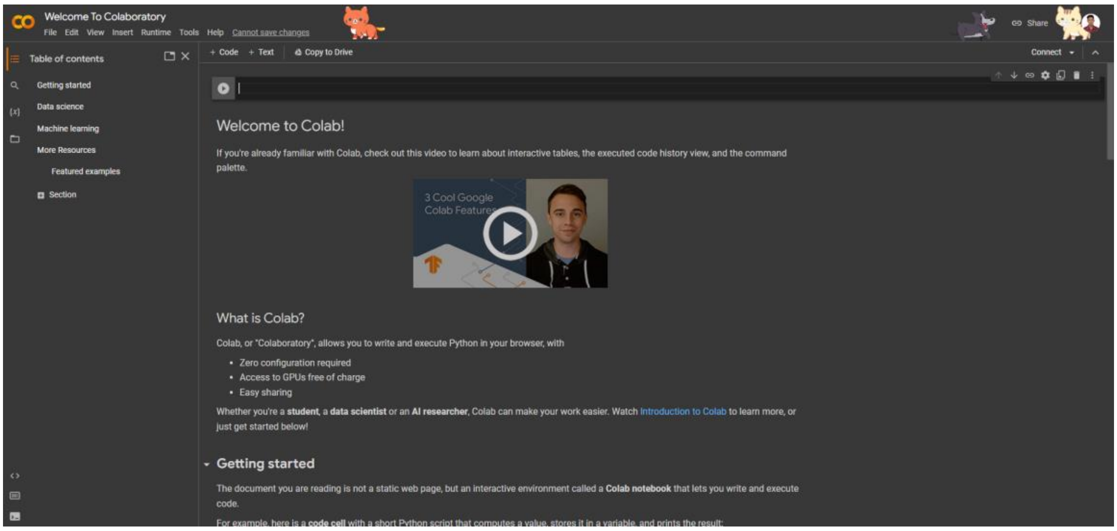
Google Colaboratory, sometimes known as Google Colab, is a Google-provided online platform for creating and distributing Jupyter notebook-based coding and data science projects. It provides a cloud-based environment in which users may write, execute, and collaborate with Python code without the need for any local installation or configuration (Google, n.d.).

Some benefits to using Google Colab are:

1. Cloud-Based Python Environment: Colab provides a virtual computer hosted in the cloud, complete with major data science libraries such as NumPy, Pandas, Matplotlib, and TensorFlow. Users can use a web browser to develop and execute Python code, eliminating the requirement for any local installation.
2. Jupyter Notebook Integration: Colab supports Jupyter notebooks, which are interactive documents that can include both code and rich text components such as headings, explanations, and visualisations. This enables users to write and distribute executable code alongside their explanations, analyses, and results.
3. Free access to computing resources: Colab provides free access to computational resources such as CPUs, GPUs, and even TPUs (Tensor Processing Units), which are very effective for speeding up deep learning applications. While resource

availability is limited, it provides an easy approach for customers to use powerful hardware for their applications.

4. Easy collaboration: Users can collaborate with others by sharing their Colab notebooks with them. Multiple users can operate on the same notebook at the same time, making it ideal for group projects or instructional purposes. Notebooks can also be distributed as static papers or as interactive reports.
5. Data access and storage: Colab integrates with Google Drive, allowing users to read and write data to and from their Drive storage. Access to Google Cloud Storage, BigQuery, and other data storage and processing services is also provided.
6. Rich ecosystem: Colab benefits from the huge Python ecosystem, which contains a plethora of modules and frameworks for data analysis, machine learning, deep learning, and other applications. Users can install extra packages and use the power of these tools within their Colab notebooks by installing these packages.



Welcome to Colab!

If you're already familiar with Colab, check out this video to learn about interactive tables, the executed code history view, and the command palette.

What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

Figure 3: Google Colab Interface

3.2 Libraries chosen / Tools chosen

3.2.1 Microsoft Power BI

Power BI is a Business Intelligence and Data Visualization instrument which converts data from various sources into immersive dashboards and analysis reports. Power BI offers cloud-based interactive visualisation facilities with an easy-to-use interface that enables users to create their own reports and dashboards. (Taylor, 2022).

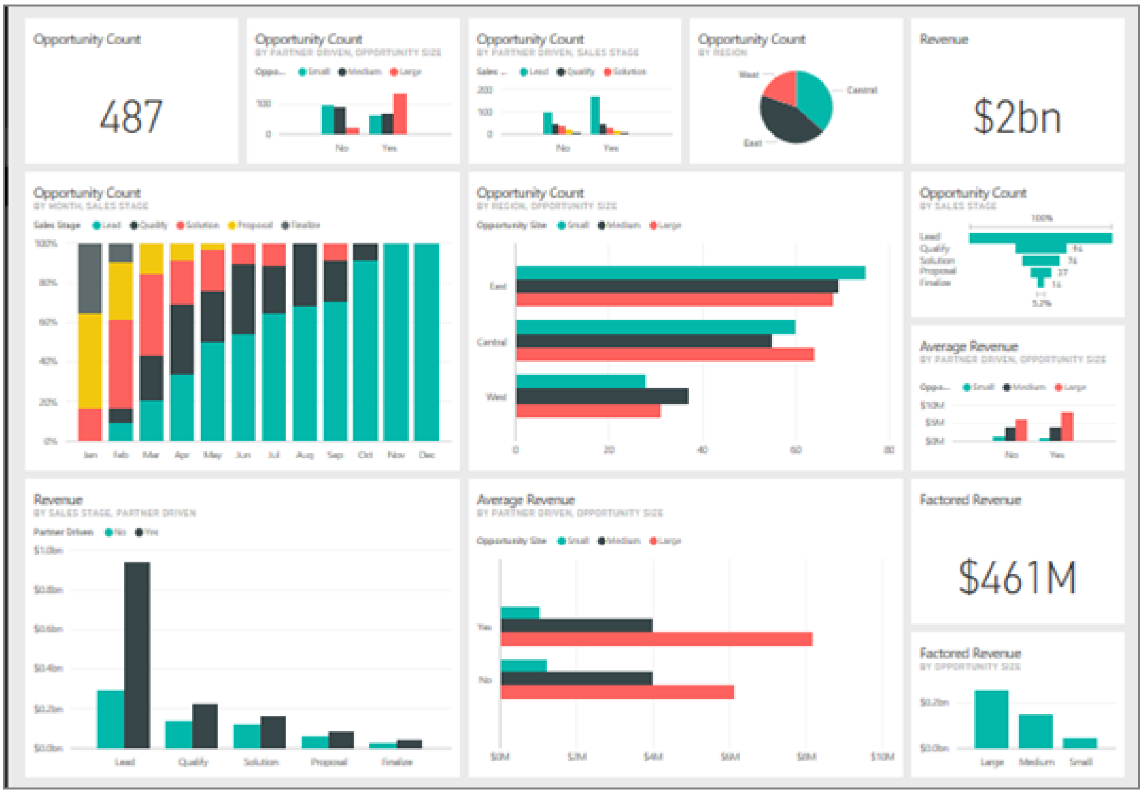


Figure 4: Example Power BI Dashboard (Microsoft, 2022).

3.2.2 Libraries

3.2.2.1 NumPy

NumPy, short for "Numerical Python," is a robust Python library widely used in scientific computing. It enables users to work with large, multi-dimensional arrays and matrices efficiently, offering an extensive collection of mathematical functions for various numerical computations. NumPy plays a crucial role in tasks like data analysis, machine learning, and image processing within the Python ecosystem. Its key features include support for multi-dimensional arrays, a comprehensive set of mathematical functions, broadcasting for convenient array operations, integration with other Python libraries, and performance optimization through C and Fortran implementations. NumPy is open-source and community-driven, ensuring continuous development and support (Travis E. Oliphant, 2006).

3.2.2.2 Matplotlib

Matplotlib is a widely used Python library for creating static, interactive, and animated data visualizations. It offers a versatile set of plotting options, allowing users to generate various types of plots such as line plots, scatter plots, bar plots, histograms, and more. The library is extensively used in data visualization, scientific research, engineering, and other fields that require data representation. Key features of Matplotlib include its flexibility and customization options, producing publication-quality output suitable for scientific journals, support for multiple graphical backends, and seamless integration with NumPy arrays. The pyplot module within Matplotlib provides a user-friendly interface that simplifies the process of generating simple plots quickly (matplotlib, 2007).

3.2.2.3 Seaborn

Seaborn is a Python data visualization library that builds on top of Matplotlib, designed to create visually appealing and informative statistical graphics. It offers a high-level interface, making it easy to visualize complex datasets. With its beautiful default styles and color palettes, Seaborn enhances the visual appeal of plots without requiring extensive customization. The library provides various high-level functions to generate different types of plots, including scatter plots, line plots, bar plots, heatmaps, and more. One of Seaborn's strengths is its ability to integrate with Pandas dataframes, enabling users to create plots with built-in statistical insights and easily handle categorical data. Moreover, Seaborn facilitates the creation of faceted plots, which help explore relationships in multidimensional data. While it offers appealing defaults, users can still customize plots using Matplotlib's syntax, making Seaborn flexible for creating personalized visualizations (VanderPlas, 2016).

3.2.2.4 Scikit-learn (Sklelarn)

Scikit-learn is an indispensable Python machine learning library that offers a wide array of tools and functionalities for various machine learning tasks. With its user-friendly interface, it simplifies the development of machine learning models, making it accessible to both beginners and experienced data scientists. By providing a rich set of algorithms for classification, regression, clustering, and more, Scikit-learn empowers users to tackle diverse data analysis challenges effectively. Furthermore, its seamless integration with other Python libraries like NumPy and Pandas facilitates efficient data manipulation and preparation, enabling users to preprocess data and extract relevant features effortlessly. Additionally, Scikit-learn provides critical utilities for model selection, cross-validation, and performance evaluation, ensuring users can select the best-fitted models and assess their accuracy reliably. Supported by an active and enthusiastic community, Scikit-learn continues to evolve, receiving frequent updates and improvements. Its extensibility allows users to incorporate custom algorithms and functionalities, making it adaptable to diverse research and industry applications (scikit-learn, n.d.).

3.2.2.5 Imbalanced-learn

imb-learn (imbalanced-learn) is a significant Python library that addresses the challenges posed by imbalanced datasets in machine learning applications. It provides a comprehensive set of techniques and algorithms tailored to handle imbalanced data, where one class dominates the others. This type of data distribution can lead to biased model performance and reduced accuracy in real-world applications. imb-learn offers an array of sampling techniques, including oversampling, under sampling, and synthetic sample generation, which aim to balance class distributions and improve model training. Additionally, the library supports ensemble methods and cost-sensitive learning algorithms, allowing users to effectively deal with imbalanced data in various scenarios. Moreover, imb-learn provides specialized performance evaluation metrics, ensuring that model evaluation considers the specifics of imbalanced datasets. By employing precision-recall curves, F1 scores, and other relevant metrics, users can better assess model performance and make informed decisions in real-world applications. With its rich functionality and support for imbalanced data handling, imb-learn has become an essential tool for data scientists, machine learning practitioners, and researchers. It plays a crucial role in improving the generalization and reliability of machine learning models, particularly when dealing with imbalanced datasets, and continues to contribute significantly to the advancement of the field (imbalanced-learn, 2023).

3.2.2.6 Pandas

Pandas is a widely used Python library designed for data manipulation, analysis, and cleaning tasks. It provides powerful data structures known as DataFrames and Series, which efficiently handle structured data and are fundamental to data science and analysis workflows. The library offers a wide range of functionalities, including data cleaning and preprocessing tools to handle missing data and reshape datasets. It allows users to perform various data manipulations such as filtering, grouping, aggregating, joining, and merging data, making it akin to working with SQL databases. Pandas supports data input and output in various formats, such as CSV, Excel, and SQL databases, enabling seamless integration with diverse data sources. One of its distinctive features is label-based indexing, allowing users to access and manipulate data using column names and row labels, providing an intuitive and flexible approach to data analysis. Built on top of NumPy, Pandas leverages its capabilities for numerical computations, and its high-performance nature makes it suitable for handling large datasets efficiently. Pandas has become an essential tool for data analysts, scientists, and researchers, empowering them to conduct data-driven research and decision-making effortlessly (VanderPlas, 2016).

3.3: Database Management System chosen

3.3.1 Microsoft SQL Server Management Studio

SQL Server Management Studio (SSMS) is a centrally managed environment that can be used to manage whatever SQL infrastructure. SSMS can connect, customise, handle, administrate, and improve all elements of Azure Synapse, SQL Server on Azure VM, Azure SQL Managed Instance, Azure SQL Database and SQL Server. SSMS is a solitary, thorough tool that integrates a broad range of graphical toolkits with a multitude of rich script editors and provide SQL Server access to developers/programmers and database administrators of all skill levels. (Microsoft, 2022).

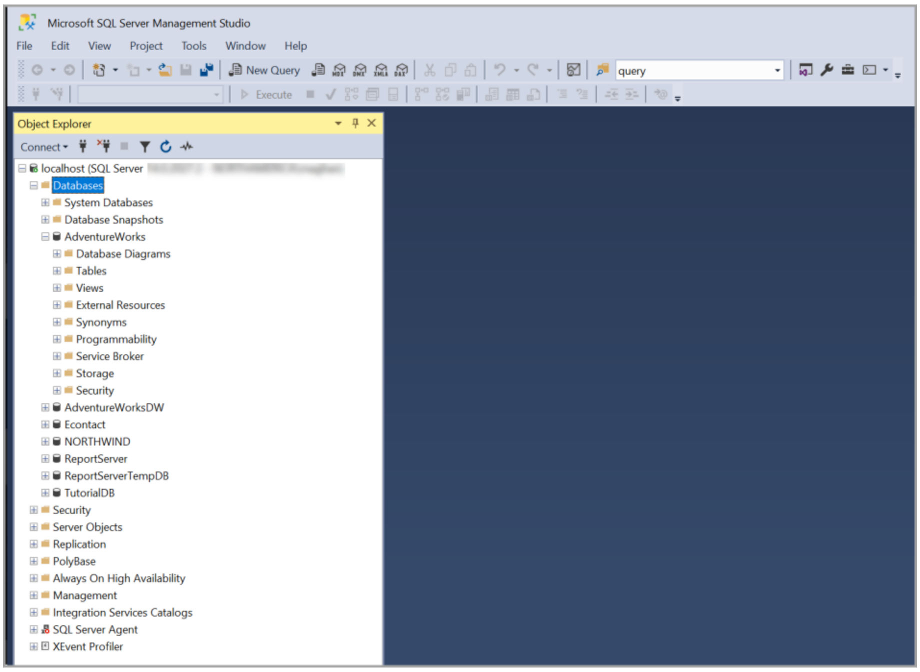
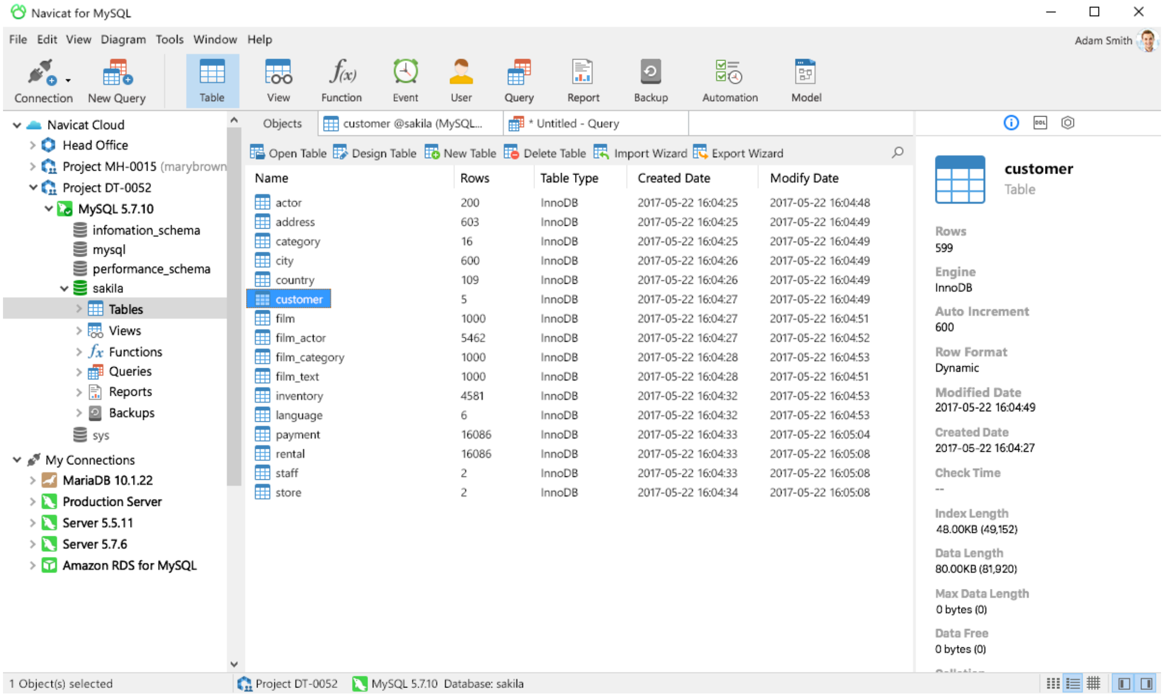


Figure 5: Microsoft SQL Server Management Studio Interface (Microsoft, 2022).

3.3.2. MySQL

MySQL is a client-server relational databases management system (RDBMS) which is free software. RDBMS is a bit of software, or a provider is used to build and manage relational databases (B., 2023).



Name	Rows	Table Type	Created Date	Modify Date
actor	200	InnoDB	2017-05-22 16:04:25	2017-05-22 16:04:48
address	603	InnoDB	2017-05-22 16:04:25	2017-05-22 16:04:49
category	16	InnoDB	2017-05-22 16:04:25	2017-05-22 16:04:49
city	600	InnoDB	2017-05-22 16:04:26	2017-05-22 16:04:49
country	109	InnoDB	2017-05-22 16:04:26	2017-05-22 16:04:49
customer	5	InnoDB	2017-05-22 16:04:27	2017-05-22 16:04:49
film	1000	InnoDB	2017-05-22 16:04:27	2017-05-22 16:04:51
film_actor	5462	InnoDB	2017-05-22 16:04:27	2017-05-22 16:04:52
film_category	1000	InnoDB	2017-05-22 16:04:28	2017-05-22 16:04:53
film_text	1000	InnoDB	2017-05-22 16:04:28	2017-05-22 16:04:51
inventory	4581	InnoDB	2017-05-22 16:04:32	2017-05-22 16:04:53
language	6	InnoDB	2017-05-22 16:04:32	2017-05-22 16:04:53
payment	16086	InnoDB	2017-05-22 16:04:33	2017-05-22 16:05:04
rental	16086	InnoDB	2017-05-22 16:04:33	2017-05-22 16:05:08
staff	2	InnoDB	2017-05-22 16:04:33	2017-05-22 16:05:08
store	2	InnoDB	2017-05-22 16:04:34	2017-05-22 16:05:08

customer
Table

Rows
599

Engine
InnoDB

Auto Increment
600

Row Format
Dynamic

Modified Date
2017-05-22 16:04:49

Created Date
2017-05-22 16:04:27

Check Time
--

Index Length
48.00KB (49,152)

Data Length
80.00KB (81,920)

Max Data Length
0 bytes (0)

Data Free
0 bytes (0)

Figure 6: MySQL Interface ({Coding} Sight, 2019)

3.3.3 Comparison

	MySQL	MSSQL Server
OS Compatibility	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux (But certain specific functions are not available when on Mac OS X and Linux (Plesky, 2022).)
Quality Support	• All MSSQL Server Supported Languages + • Perl • Eiffel • Haskel • Tcl.	• Go • Delphi • Virtual Basic • PHP • Python • Java • Ruby • C++
Affordability	Open Source (Free)	Licenses Required
IDE Tools	Enterprise Manager	Management Studio
Binary Collections	Allows developers to use binaries to manipulate database files while they are operating, and database files may additionally be manipulated at runtime by other processes (Plesky, 2022).	Prevents developers from accessing and changing binaries or database files (Plesky, 2022).
Data Backup	Takes more time than MSSQL	Easier and Faster backup and restore huge amounts of data
Execution Flexibility	Query cannot be cancelled after it begins running, need to kill whole process to	Query able to be cancelled while running without

	cancel a query (Plesky, 2022).	needing to kill the whole process (Plesky, 2022).

From the comparison above, the more suitable DBMS for the project is MSSQL Server Management Studio in the case that some queries may be needed to be cancelled in the midst of running a process.

3.4: Operating System chosen

An operating system is a computer program which runs additional software applications in a device right after being loaded into the device itself. The application programmes interact with the operating system by asking services via a predefined application programme interface (API). Furthermore, users could also interact directly with the operating system via an user interface, such as a command-line interface (CLI) or a graphical user interface (GUI) (Bigelow, 2001).

3.4.1 Hardware

The minimum hardware requirements:

- Hard Disk: 6 GB
- Monitor: Super-VGA (800x600) or Higher Resolution Monitor
- Internet: Internet Functionality
- RAM: 4GB
- Processor Speed: x64 Processor: 1.4 GHz, Recommended 2.0GHz or Faster
- Processor Type: x64 Processor

3.4.2 Software

The minimum software requirements:

- Operating System: Windows 10 or greater, Windows Server 2016 or greater.
- Dataset: Microsoft Excel 2016
- Documentation: Microsoft Word 2016
- IDE: Microsoft Visual Studio
- Tools: Microsoft Power BI

3.5: Summary

In conclusion, no programming language are needed for this project since the data is already clean, thus not needing the use of data cleaning language such as Python or R. However, Microsoft Visual Studio is needed as the IDE for connecting the variables and also Power BI for visualizing the connected variables from Microsoft Visual Studio.

Hardware requirement listed in the previous chapter is a must if the project were to be repeated, since it is the minimum hardware requirement for several application such as Microsoft Visual Studio, Microsoft Power BI, and MS SQL Server.

CHAPTER 4: METHODOLOGY

4.1 Introduction

The gathering and analyzing of data methodologies used in research are discussed and explained in methodology. The method section, which is a crucial component of thesis, dissertation, or research paper, describes what users did as well as how they did it, enabling viewers to assess the validity and reliability of the research and thesis topic (McCombes & George, What Is a Research Methodology, 2023).

4.2 Methods (Comparison of Data Mining Methodologies)

The table below will be comparing three different methodologies, which are KDD, SEMMA, and CRISP-DM.

Aspects	**KDD**	**SEMMA**	**CRISP-DM**
Overview	Acronym for Knowledge Discovery in Databases	Acronym for Sample, Explore, Modify, Model, Asses.	Acronym for Cross Industry Standard Process for Data Mining
Developer	Fayyad, Piateksky-Saphiro and Smyth	SAS Institute	CRISP-D consortium
Year of Introduced	1996	1997	1996
Number of Phases	5	5	6
Phases	1. Selection 2. Data Pre-Processing 3. Data Transformation 4. Data Mining 5. Interpretation/ Evaluation	1. Sample 2. Explore 3. Modify 4. Model 5. Assess	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modelling 5. Evaluation 6. Deployment

Flexibility	• High Flexibility • Iterative and Interactive • Repeatable, able to move forwards or backwards. • Allows Feedbacks	• Low Flexibility • Iterative • Repeatable, able to move forwards or backwards.	• Highest Flexibility • Iterative • Repeatable, able to move forwards or backwards. • Allows mistakes at the beginning of the model. • Allows changes.
Focus Area	• Beginning to end (Entire Process)	• Focus on technicality. • No Deployment Phase	• Focus on data science process
Documentation	• No Documentation	• Less documentation than CRISP-DM	• Heavy Documentation
Required Knowledge	• Data Mining Knowledge	• Easily understandable • Need knowledge of utilizing SAS Enterprise Miner	• Easily understandable • Needs data mining knowledge.

4.3 Justification on Chosen Methodology: CRISP-DM

Based on the research on multiple methodologies, the suitable methodology that matches the project is CRISP-DM, since it offers an organized structure and guidelines for data mining project planning, management, and execution (Hotz, 2023). CRISP-DM is chosen because the primary requirement for the product is data analysis, which is the strong point of CRISP-DM, making clearer path for the process of data analysis compared to other methodologies. CRISP-DM is also considered very flexible meaning development has more freedom, hence also reducing the chance of the project failing.

4.4 CRISP-DM

CRISP-DM, short for Cross Industry Standard Process for Data Mining, is a methodology for influencing Data Mining projects that was formed in 1996. A Data Mining task is conceptualized in 6 phases, with repetitions based on the developers' needs. The steps involved are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Rodrigues, 2020).

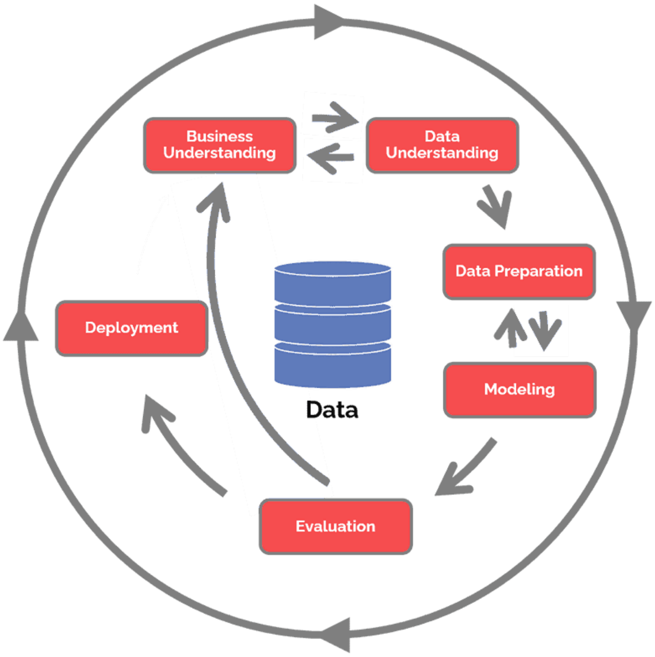


Figure 7: CRISP-DM Diagram (Data Science Process Alliance, n.d.)

Step 1. Business Understanding

The first step's goal is to offer background for the aims and details to ensure the user understands the importance of information in that specific business model. The goal of this step is to ensure that everyone on the team understands the context of the project (Rodrigues, 2020).

Step 2. Data Understanding

The second step is largely focused on figuring out what may be anticipated and accomplished by using information. It explores the data's level of quality in terms of completeness, distribution, and data governance adherence. This step is critical to the project because it defines how precise the end scores can be. In this step, group members discover the best methods for extracting the most useful information from the information. If the team is unsure of the application or worth of a particular piece of knowledge, they could go back to discover about the business so they can profit from the data given. Rodrigues (Rodrigues, 2020).

Step 3. Data Preparation

The third step is Data Preparation, which contains the ETLs or ELTs procedure, where it uses algorithms and processes to transform raw data it in to something utilizable (Rodrigues, 2020). Data preparation is vital to the effective modelling process. If a mistake is committed in this step, the next step will generate no meaningful results. As a result, this stage should be as focused and characterized as possible (Wijaya, 2021).

Step 4. Modelling

The fourth step is modelling, which is the most fun thing of any machine learning task. This step generates results which should accomplish or assist in the accomplishment of the project's primary objective. It is not only the most thrilling component of the project, but it also tends to take the shortest time to complete because, if the prior steps are followed correctly, there won't be much to modify. If the outcomes can be improved, the methodology will revert back to data preparation and enhance the existing data (Rodrigues, 2020). The decision tree model will be used in the project. A decision tree is really a non-parametric supervised algorithm for learning that can be used for both regression and classification tasks. It has a hierarchical tree structure with a root of the tree, branches, internal nodes, and leaf nodes. (IBM, n.d.).

Step 5. Evaluation

Step five, Evaluation, entails the procedure of guaranteeing that the finished version is feasible and beneficial. If the outcomes are not useable, the method enables for a review the initial step to determine why the findings are not useable. Depending on the task and the context, different methods are available for use (Rodrigues, 2020).

Step 6. Deployment

The sixth step, Deployment, is merely providing the final result in an advantageous and insightful manner in order to achieve the project's goals. This is the sole move in the loop which isn't repeatable. (Rodrigues, 2020).

4.5 Summary

To choose the most appropriate methodology for the current project, a comparison was made between three data mining methodologies: CRISP-DM, SEMMA, and KDD. As a result, CRISP-DM was selected because it is a comprehensive and structured approach that includes a Deployment phase and is flexible and iterative.

The CRISP-DM methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Data Science Process Alliance, n.d.). In the Business Understanding phase, the developer determines the project goals and objectives to find some ideas to the cause of diabetes. In the Data Understanding phase, exploratory data analysis is conducted to evaluate the lifestyles and other variables of diabetics and non-diabetics. The Data Preparation phase involves data pre-processing techniques to clean and transform the data. In the Modelling phase, suitable machine learning algorithms are selected to develop the diabetes prediction model. In the Evaluation phase, the model's results are evaluated and compared to the literature review. Finally, in the Deployment phase, the model is deployed with the selected prediction model, and the dashboard visualizes the important variables causes diabetes.

In sum, the methodology chosen for this project is CRISP-DM. CRISP-DM process allows the creation of a long-term tactic through quick repetitions at the start of a project. During the implementation of this change, the developer can develop a basic and simple prototype loop that can be quickly enhanced in successive iterations. This principle enables the improved performance of a previously defined plan after receiving new data and insights, while also having high flexibility (Zavgorodniy, 2018)

CHAPTER 5: Data Analysis

5.1 Introduction

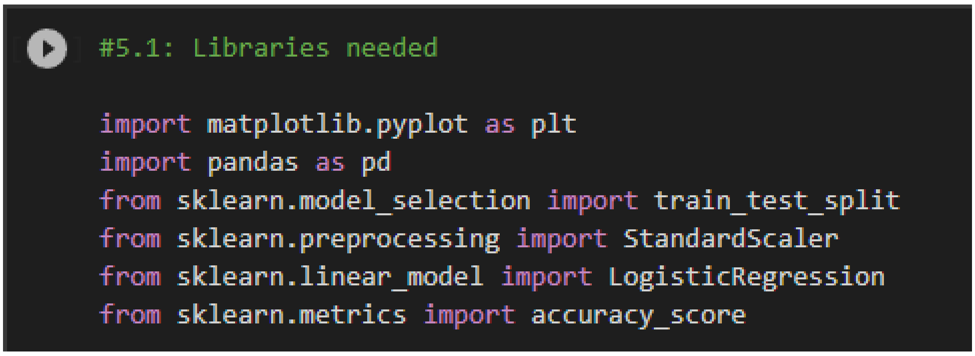
The process of collecting, modelling, and analyzing data using various statistical and logical methodologies and procedures is known as data analysis. Analytics techniques and tools are used by businesses to gather insights that help strategic and operational decision-making. (Calzon, 2023).

The data analysis process typically encompasses several steps. Firstly, relevant data is collected from diverse sources, ensuring accuracy and representativeness. Next, the data is cleansed and preprocessed to address errors, inconsistencies, and missing values. Exploratory data analysis techniques are then used to gain an initial understanding of the data, detect outliers, and develop hypotheses. Statistical analysis techniques are employed to explore relationships, dependencies, and trends within the data, providing quantitative insights and validating hypotheses (Kelley, What is Data Analysis? Methods, Process and Types Explained, 2023).

Data mining and machine learning techniques play a role in extracting hidden patterns, identifying trends, and constructing predictive models. This involves the application of algorithms to uncover patterns, cluster similar data points, classify data, or make predictions based on historical data. Interpretation of the analysis findings is essential, drawing conclusions, making inferences, and generating actionable insights within the context of the original problem or research question (Kelley, What is Data Analysis? Methods, Process and Types Explained, 2023).

Visualizations such as charts, graphs, dashboards, and reports are often used to present the results of data analysis, aiding in effective communication of complex information. Data analysis is widely applicable across various domains, including business, finance, healthcare, marketing, and social sciences, supporting customer behavior understanding, process optimization, fraud detection, product improvement, and data-driven decision-making (Kelley, What is Data Analysis? Methods, Process and Types Explained, 2023).

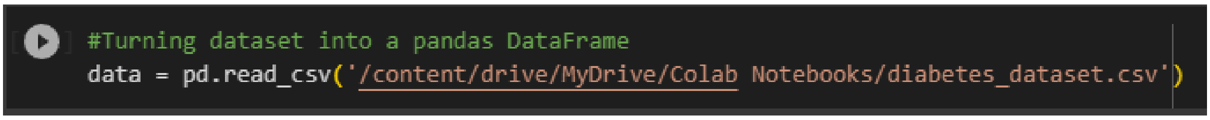
To begin data analysis in this project, the following libraries are needed:



```
#5.1: Libraries needed

import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Figure 8: Libraries Needed

Then, the following code is used to load the dataset into pandas DataFrame:

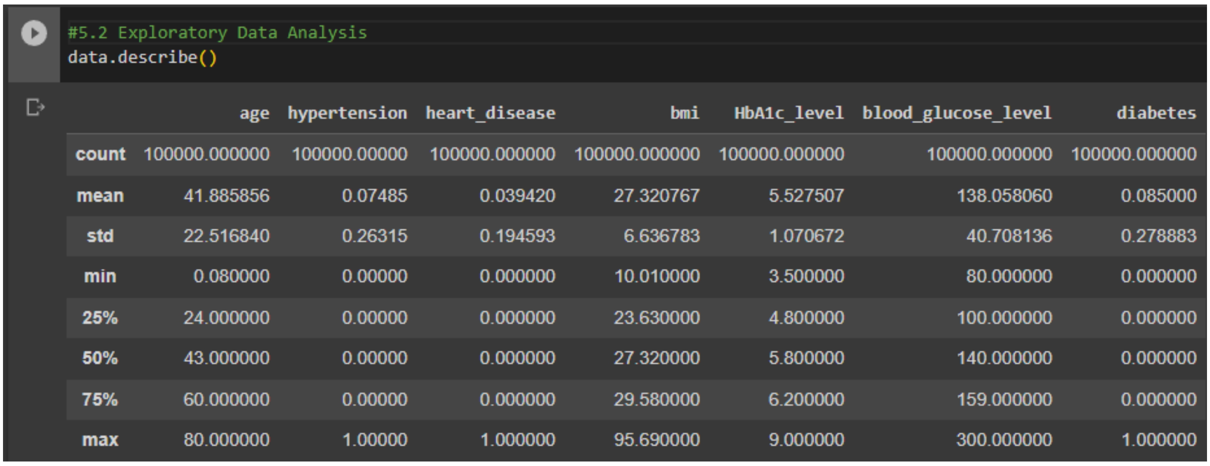


```
#Turning dataset into a pandas DataFrame
data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/diabetes_dataset.csv')
```

Figure 9: Reading dataset

5.2 Initial Data Exploration

To check the data' numbers, the describe() function is utilized:

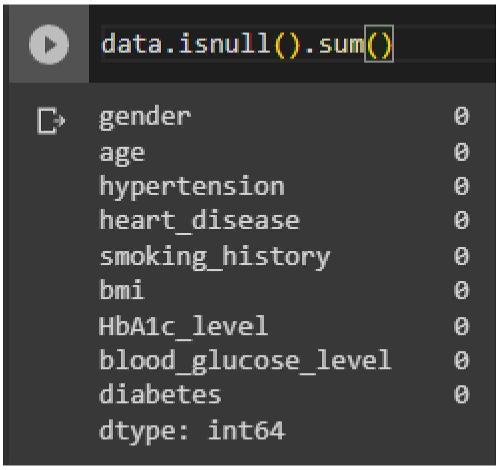


```
#5.2 Exploratory Data Analysis
data.describe()
```

```
                 age  hypertension  heart_disease            bmi     HbA1c_level  blood_glucose_level       diabetes
count  100000.000000  100000.00000  100000.000000  100000.000000  100000.000000        100000.000000  100000.000000
mean       41.885856       0.07485       0.039420      27.320767       5.527507           138.058060       0.085000
std        22.516840       0.26315       0.194593       6.636783       1.070672            40.708136       0.278883
min         0.080000       0.00000       0.000000      10.010000       3.500000            80.000000       0.000000
25%        24.000000       0.00000       0.000000      23.630000       4.800000           100.000000       0.000000
50%        43.000000       0.00000       0.000000      27.320000       5.800000           140.000000       0.000000
75%        60.000000       0.00000       0.000000      29.580000       6.200000           159.000000       0.000000
max        80.000000       1.00000       1.000000      95.690000       9.000000           300.000000       1.000000
```

Figure 10: Data Describe()

To check missing value, isnull() function and sum() function is used:



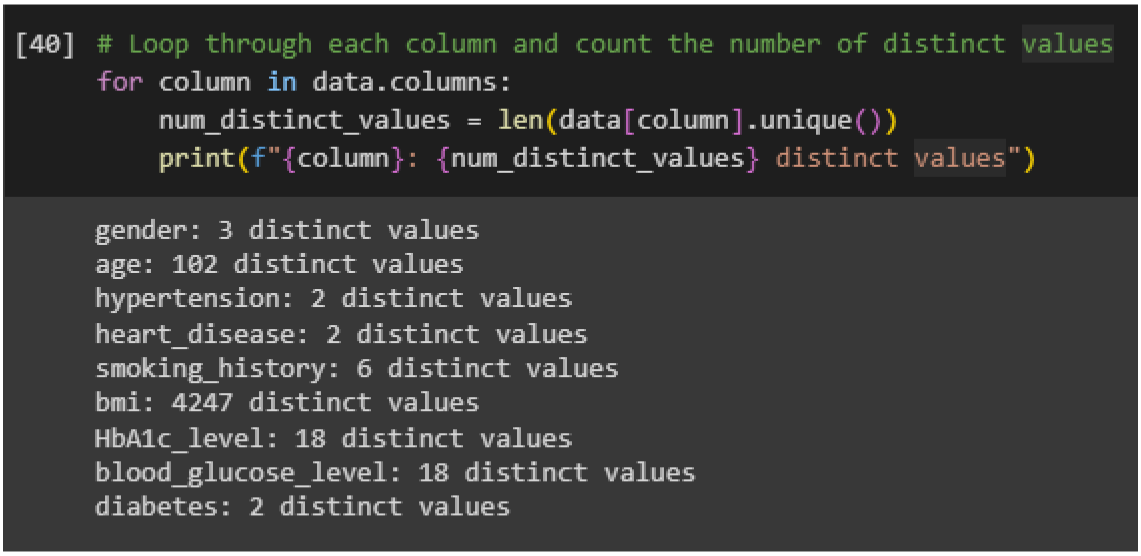
```
data.isnull().sum()
```

```
gender                 0
age                    0
hypertension           0
heart_disease          0
smoking_history        0
bmi                    0
HbA1c_level            0
blood_glucose_level    0
diabetes               0
dtype: int64
```

Figure 11: Check Null Value

Seen from the output, no missing values are present in the dataset.

To check for unique values, the following code can be used:

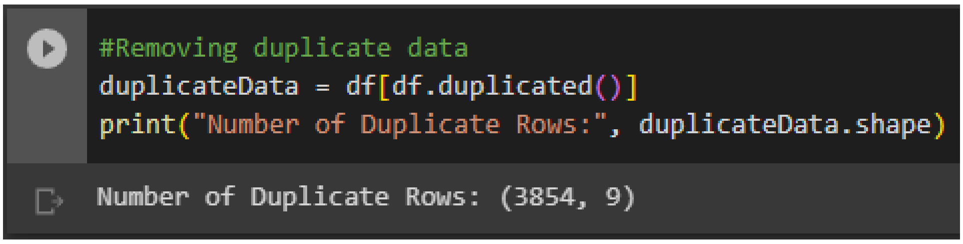


```
[40] # Loop through each column and count the number of distinct values
     for column in data.columns:
         num_distinct_values = len(data[column].unique())
         print(f"{column}: {num_distinct_values} distinct values")
```

```
gender: 3 distinct values
age: 102 distinct values
hypertension: 2 distinct values
heart_disease: 2 distinct values
smoking_history: 6 distinct values
bmi: 4247 distinct values
HbA1c_level: 18 distinct values
blood_glucose_level: 18 distinct values
diabetes: 2 distinct values
```

Figure 12: Check Unique Values

To check duplicate data, duplicated() function can be used, the following code is to print out how many duplicates is present in the dataset:



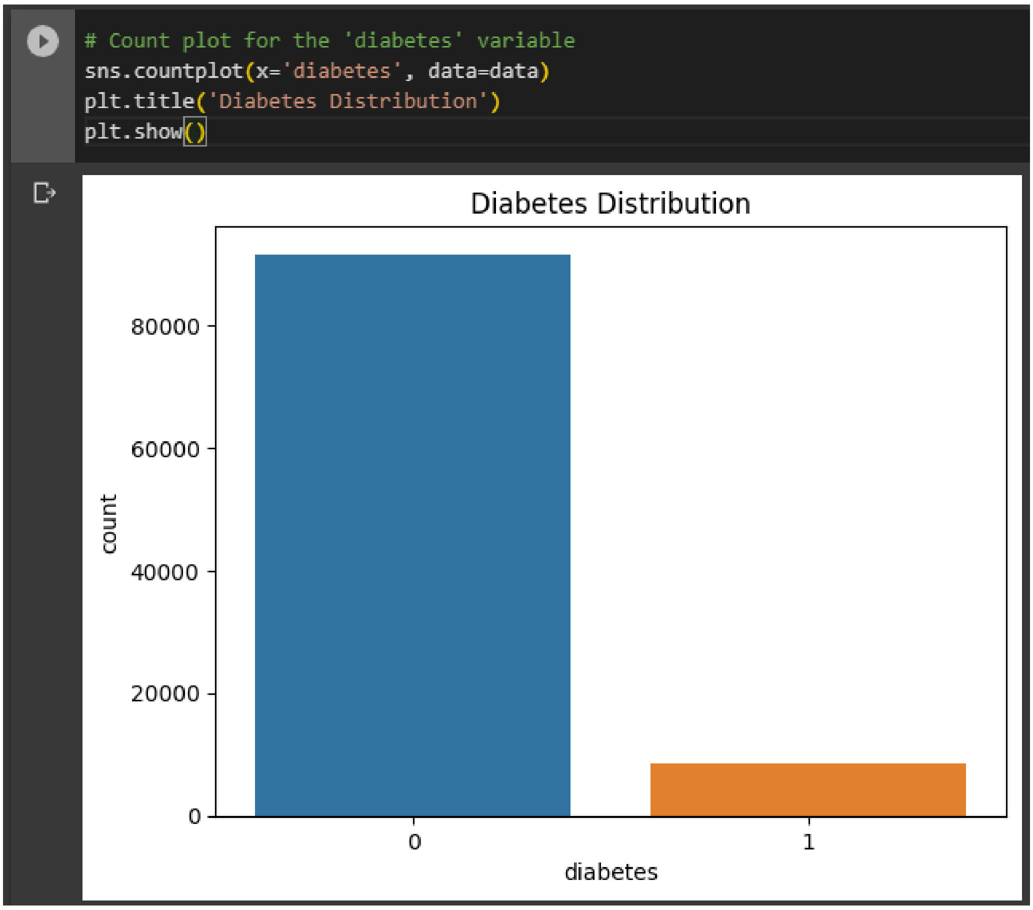
```
#Removing duplicate data
duplicateData = df[df.duplicated()]
print("Number of Duplicate Rows:", duplicateData.shape)
```

```
Number of Duplicate Rows: (3854, 9)
```

Figure 13: Remove Duplicate Data

The next chapter will show how to remove the duplicate rows.

The following graph shows distribution of diabetes in the dataset.

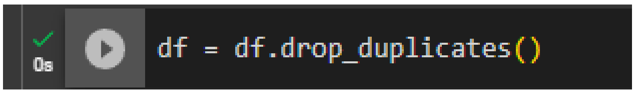


```
# Count plot for the 'diabetes' variable
sns.countplot(x='diabetes', data=data)
plt.title('Diabetes Distribution')
plt.show()
```

Figure 14: Diabetes Distribution CountPlot

5.3 Data Cleaning

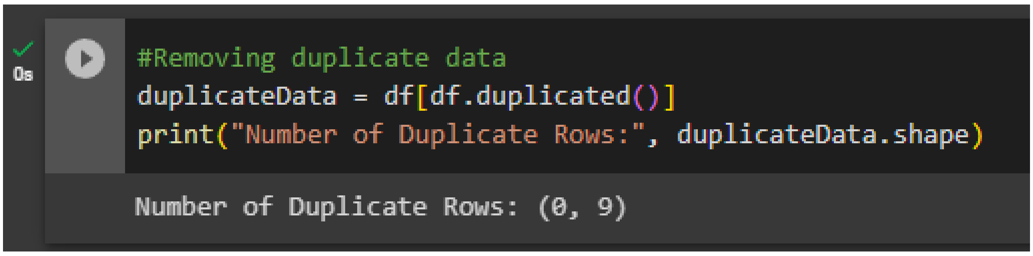
The following code is to remove duplicate data from the dataset as mentioned in the previous chapter:



```
df = df.drop_duplicates()
```

Figure 15: Drop Duplicates()

Result:



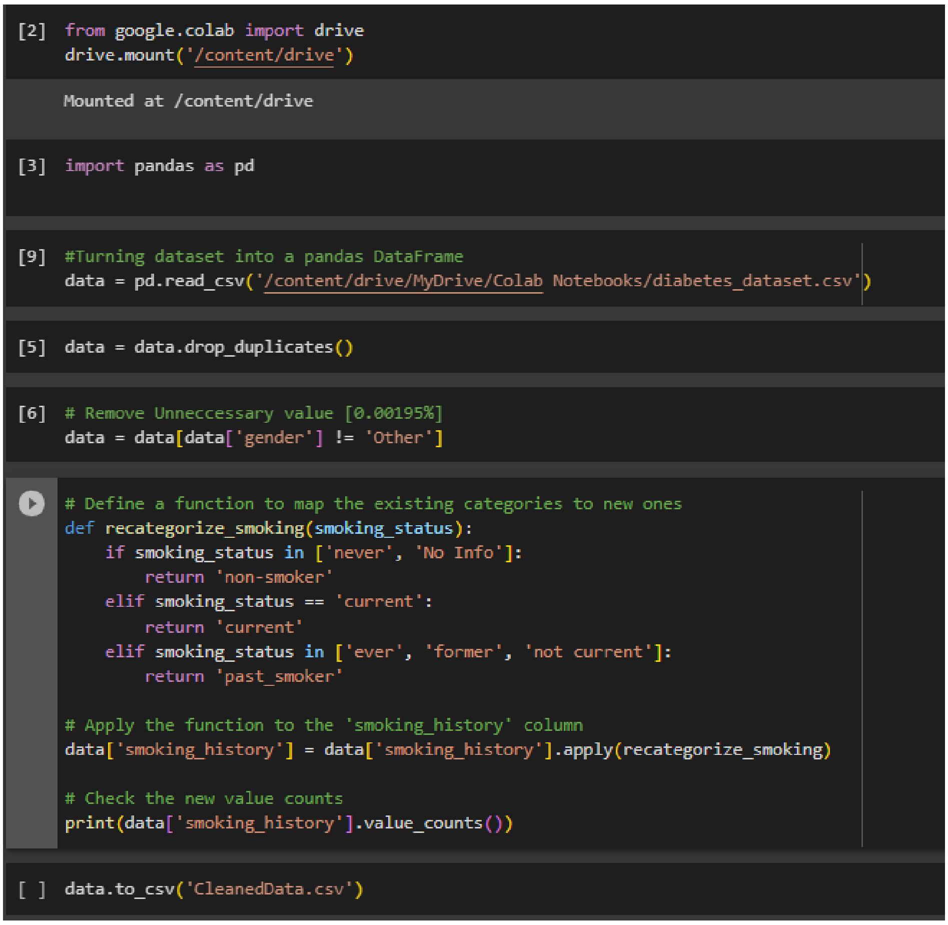
```
#Removing duplicate data
duplicateData = df[df.duplicated()]
print("Number of Duplicate Rows:", duplicateData.shape)
```

```
Number of Duplicate Rows: (0, 9)
```

Figure 16: Print Duplicate Data

5.3.1 Data Cleaning for Dashboard

Data cleaning for dashboard simply removing duplicates, removing gender ‘Other’ and categorizing smoking status, then export to csv:



```
[2] from google.colab import drive
    drive.mount('/content/drive')

    Mounted at /content/drive
```

```
[3] import pandas as pd
```

```
[9] #Turning dataset into a pandas DataFrame
    data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/diabetes_dataset.csv')
```

```
[5] data = data.drop_duplicates()
```

```
[6] # Remove Unneccessary value [0.00195%]
    data = data[data['gender'] != 'Other']
```

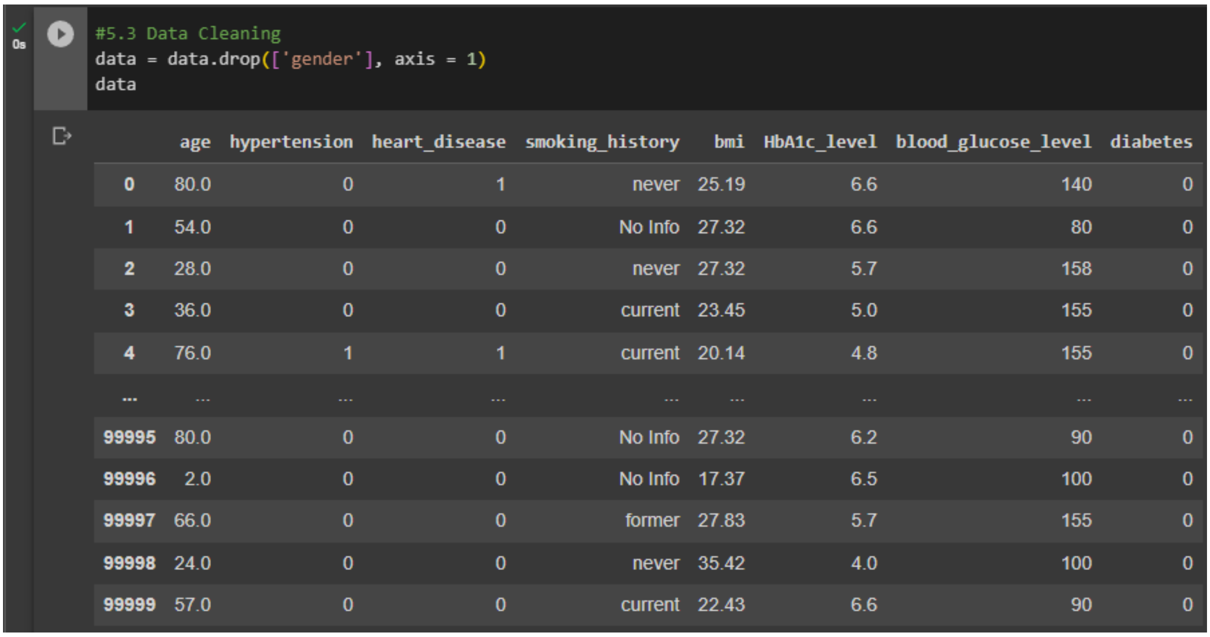
```
# Define a function to map the existing categories to new ones
def recategorize_smoking(smoking_status):
    if smoking_status in ['never', 'No Info']:
        return 'non-smoker'
    elif smoking_status == 'current':
        return 'current'
    elif smoking_status in ['ever', 'former', 'not current']:
        return 'past_smoker'

# Apply the function to the 'smoking_history' column
data['smoking_history'] = data['smoking_history'].apply(recategorize_smoking)

# Check the new value counts
print(data['smoking_history'].value_counts())
```

```
[ ] data.to_csv('CleanedData.csv')
```

Figure 17: Data Cleaning for Dashboard

5.3.2 Data Cleaning for Logistic Regression Machine Learning Model

For the sake of accuracy and non-redundant output, the gender variable is decided to be removed from the later discussed logistic regression model:

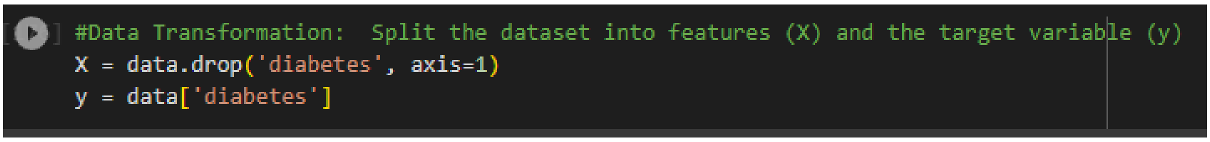


```
#5.3 Data Cleaning
data = data.drop(['gender'], axis = 1)
data
```

	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	80.0	0	1	never	25.19	6.6	140	0
1	54.0	0	0	No Info	27.32	6.6	80	0
2	28.0	0	0	never	27.32	5.7	158	0
3	36.0	0	0	current	23.45	5.0	155	0
4	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...
99995	80.0	0	0	No Info	27.32	6.2	90	0
99996	2.0	0	0	No Info	17.37	6.5	100	0
99997	66.0	0	0	former	27.83	5.7	155	0
99998	24.0	0	0	never	35.42	4.0	100	0
99999	57.0	0	0	current	22.43	6.6	90	0

Figure 18: Drop Gender Feature

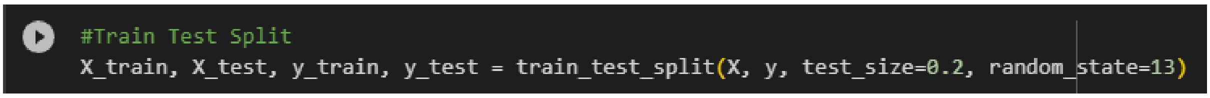
Before making a machine learning model, the dataset has to be split into features and target variable:



```
#Data Transformation:  Split the dataset into features (X) and the target variable (y)
X = data.drop('diabetes', axis=1)
y = data['diabetes']
```

Figure 19: Splitting Dataset

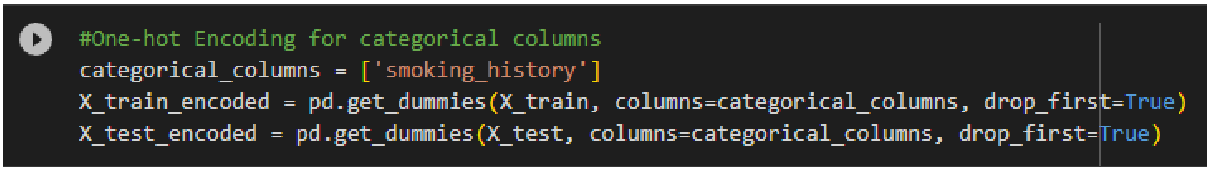
Next up, is to use train_test_split() function to split the data into training data and test data, as well as training target variables and test training target variables, test_size=0.2 means that it is using 20% of the entire dataset:



```
#Train Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=13)
```

Figure 20: Train Test Split (Logistic Regression Model)

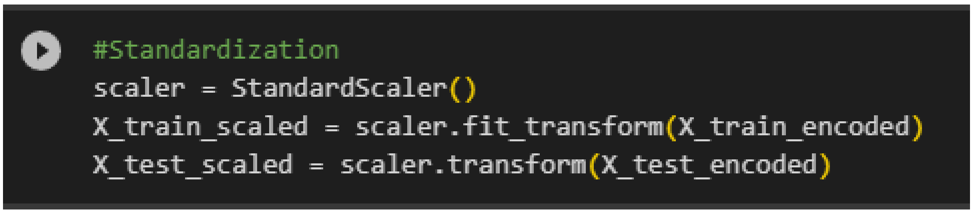
Since there are categorical data, which is not understandable by the program, One-hot encoding must be done, the following is the code for One-hot encoding:



```
#One-hot Encoding for categorical columns
categorical_columns = ['smoking_history']
X_train_encoded = pd.get_dummies(X_train, columns=categorical_columns, drop_first=True)
X_test_encoded = pd.get_dummies(X_test, columns=categorical_columns, drop_first=True)
```

Figure 21: One-hot Encoding [Smoking History]

The next step is standardization:



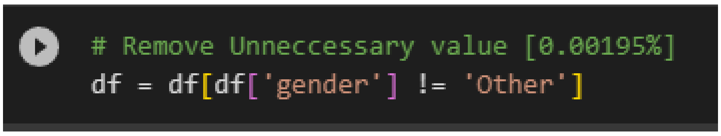
```
#Standardization
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_encoded)
X_test_scaled = scaler.transform(X_test_encoded)
```

Figure 22: Standardization for Logistic Regression Model

After standardization is complete, the model can then be built, which is discussed in chapter 5.7 (Model).

5.3.3 Random Forest

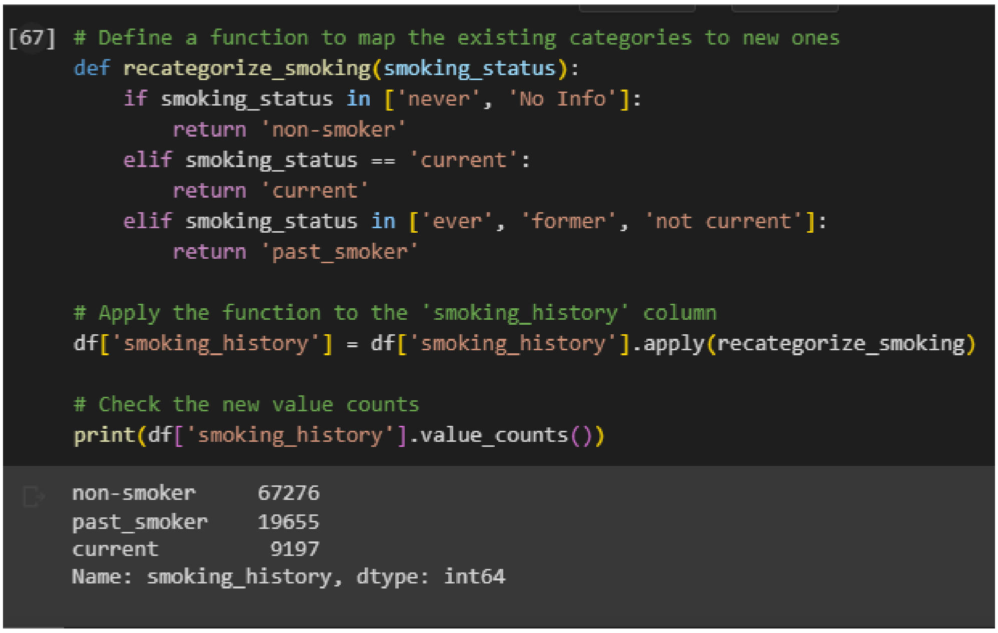
In random forest, the gender variable will be used unlike the logistic regression model, which is why the gender of ‘Other’ needs to be filtered out first because it is considered unnecessary.



```
# Remove Unneccessary value [0.00195%]
df = df[df['gender'] != 'Other']
```

Figure 23: Data Cleaning [Gender != Other]

Then the smoking status will be grouped to only 3 categories, grouping ‘Never’, and ‘No Info’ to Non Smoker, ‘Current’ to Current, and ‘Ever’, or ‘Former’, or ‘Not Current’, as Past Smoker.

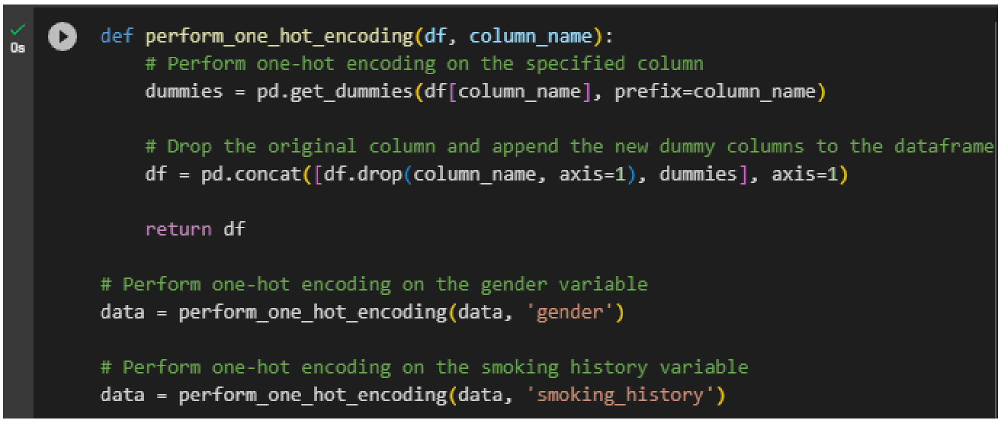


```
[67] # Define a function to map the existing categories to new ones
     def recategorize_smoking(smoking_status):
         if smoking_status in ['never', 'No Info']:
             return 'non-smoker'
         elif smoking_status == 'current':
             return 'current'
         elif smoking_status in ['ever', 'former', 'not current']:
             return 'past_smoker'

     # Apply the function to the 'smoking_history' column
     df['smoking_history'] = df['smoking_history'].apply(recategorize_smoking)

     # Check the new value counts
     print(df['smoking_history'].value_counts())
```

```
non-smoker     67276
past_smoker    19655
current         9197
Name: smoking_history, dtype: int64
```

Figure 24: Categorizing Smoking History Feature

One-hot encoding is then done to encode categorical data (gender, and smoking history).



```
def perform_one_hot_encoding(df, column_name):
    # Perform one-hot encoding on the specified column
    dummies = pd.get_dummies(df[column_name], prefix=column_name)

    # Drop the original column and append the new dummy columns to the dataframe
    df = pd.concat([df.drop(column_name, axis=1), dummies], axis=1)

    return df

# Perform one-hot encoding on the gender variable
data = perform_one_hot_encoding(data, 'gender')

# Perform one-hot encoding on the smoking history variable
data = perform_one_hot_encoding(data, 'smoking_history')
```

Figure 25: One-Hot Encoding Gender and Smoking History

After all the preparations are done, the modelling phase will be discussed in chapter 5.7 (Model)

5.4 Data Visualization (Univariate, Bivariate, Multivariate)

5.4.1 Sum of Diabetes and Sum of Hypertension

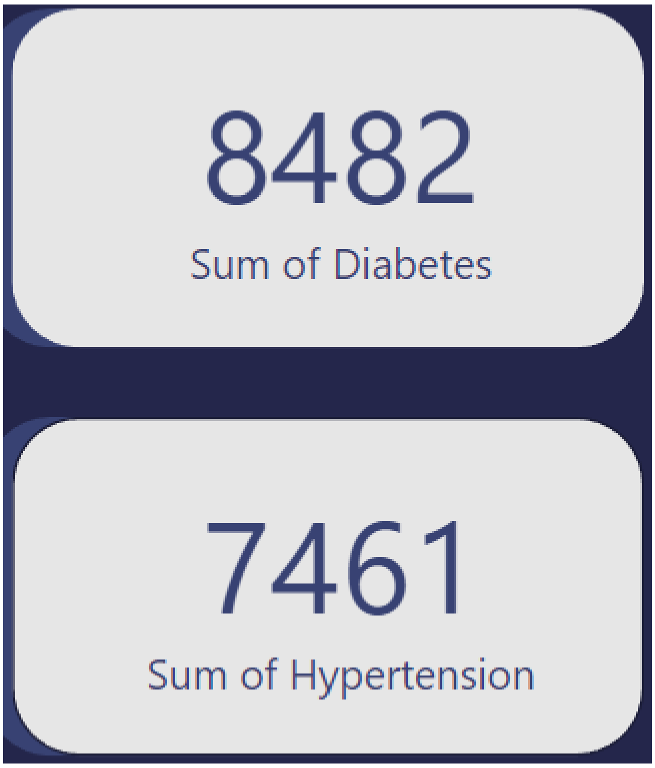


Figure 26: Univariate Visual of Diabetes and Hypyertension

The above card both shows sum of diabetes and hypertension, both of these cards can change the amount shown based on the filters applied.

5.4.2 Sum of Diabetes by Hypertension

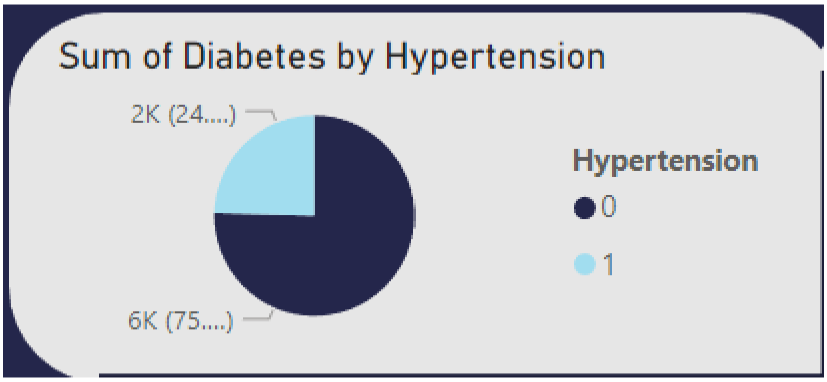


Figure 27: Sum of Diabetes by Hypertension

The sum of diabetes by hypertension chart shows 6412 (75.4%) diabetics do not suffer from hypertension while 2088 (24.5%) diabetics suffer from hypertension, meaning that having hypertension might increase one’s chances of getting diabetes .

5.4.3 Sum of Diabetes by Heart Disease

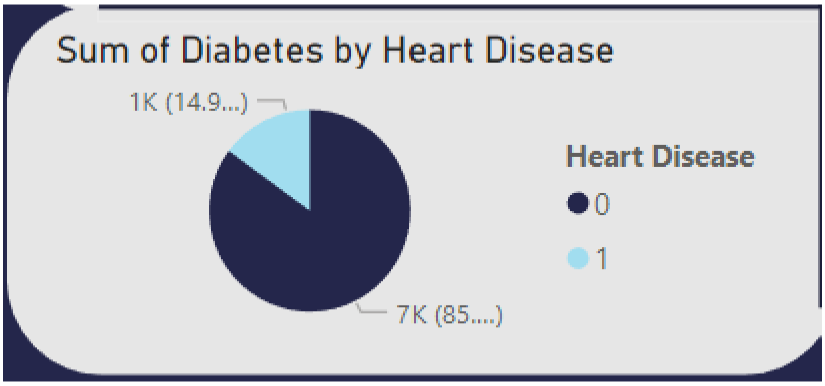


Figure 28: Sum of Diabetes by Heart Disease

In the chart above, the sum of diabetes by heart disease chart shows 7233 (85%) diabetics do not have heart disease while 1267 (14.9%) diabetics has heart disease.

5.4.4 Sum of Diabetes by BMI

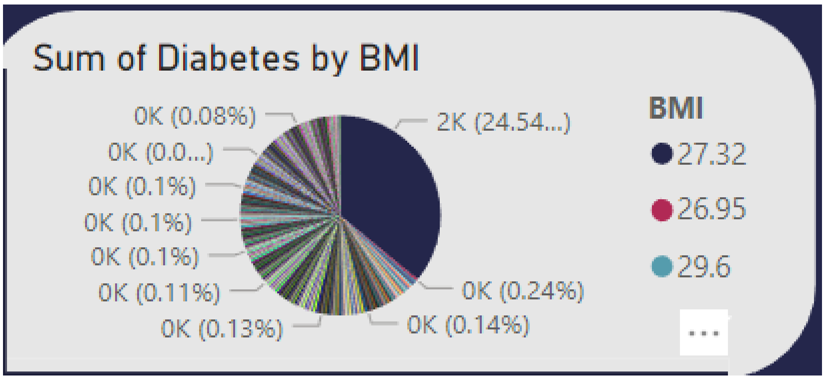


Figure 29: Sum of Diabetes by BMI

The pie chart shown shows the many small percentages of BMI of many people who have diabetes, and from the pie chart, the number that has huge effects on diabetes is having BMI of 27.32, having 1531 (25.54%) diabetics.

5.4.5 Sum of Diabetes by Blood Glucose Level

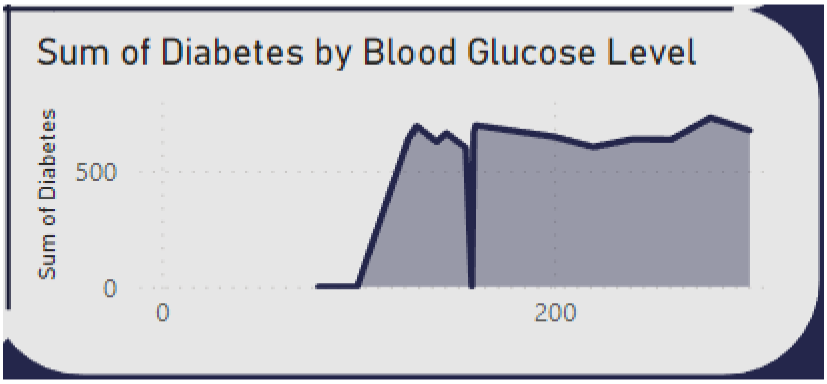


Figure 30: Sum of Diabetes by Blood Glucose Level

The area chart shows levels of Blood Glucose Level of diabetics.

5.4.6 Sum of Diabetes by Age

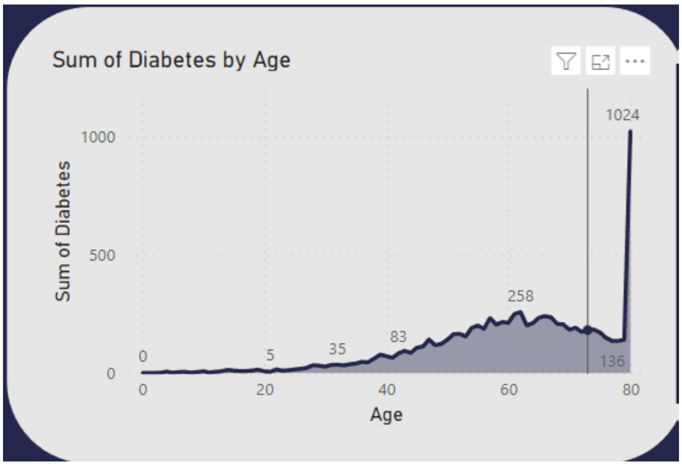


Figure 31: Sum of Diabetes by Age

The shown area chart provides information about the ages of diabetics in the dataset.

5.4.7 Sum of HbA1c_Level

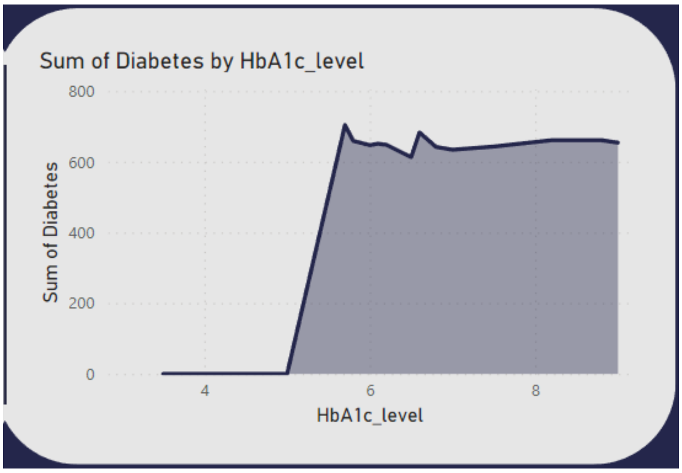


Figure 32: Sum of Diabetes by HbA1C Level

The above area chart shows the HbA1c Levels of diabetics in the dataset.

5.4.8 Filters (Multivariate Visuals)

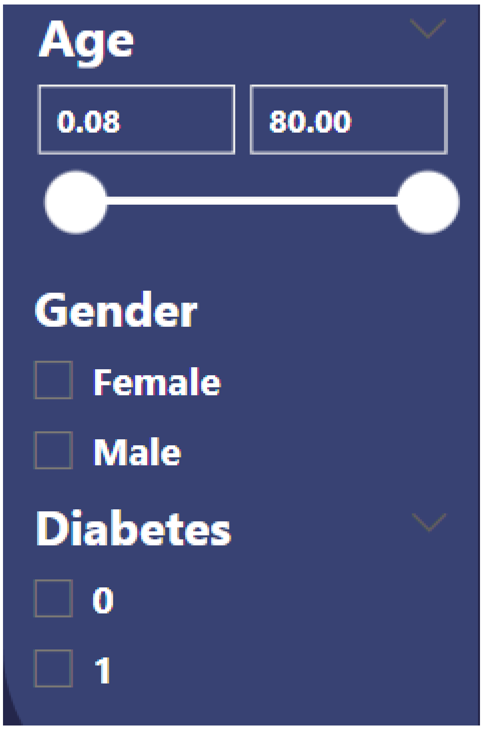


Figure 33: Filters for Multivariate Data Visual

The above snippet is filters for the dashboard. By selecting one or more filters, the chart shown previously, it will alter the chart based on the selected filters, user can choose between age groups, gender, and diabetics or not. For example, selecting Male, will not only change the sum of diabetes to 4035 (from 8500), but also change the shape of all the charts:

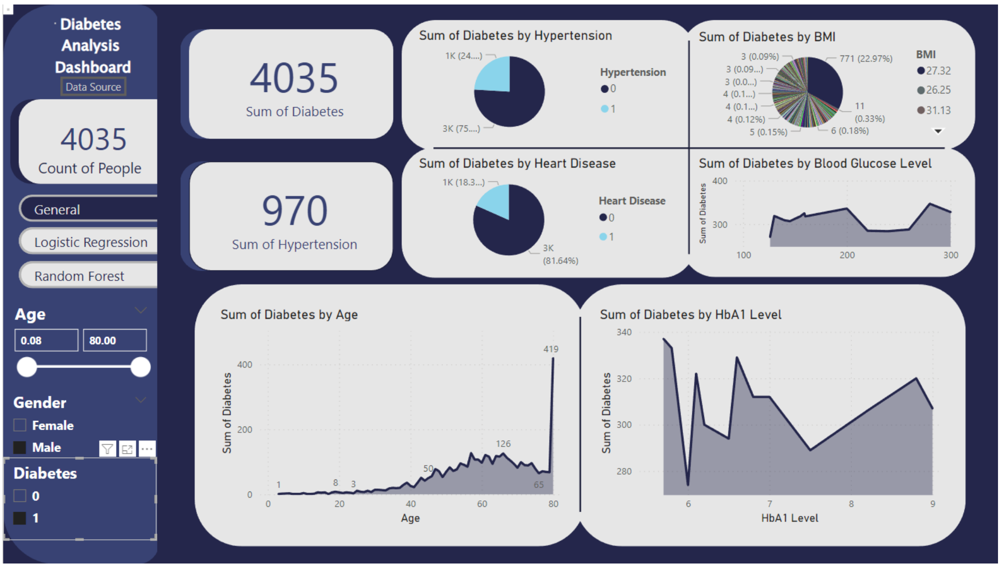


Figure 34: Data Dashboard after Filtering

5.5 Hypothesis

1. It is hypothesized that heart disease and hypertension have a high correlation with diabetes.
2. It is hypothesized that age, although does not have high correlation, is still somewhat correlated to diabetes.
3. It is hypothesized that HbA1C level is highly correlated to having diabetes.

5.6 Reports and Dashboard

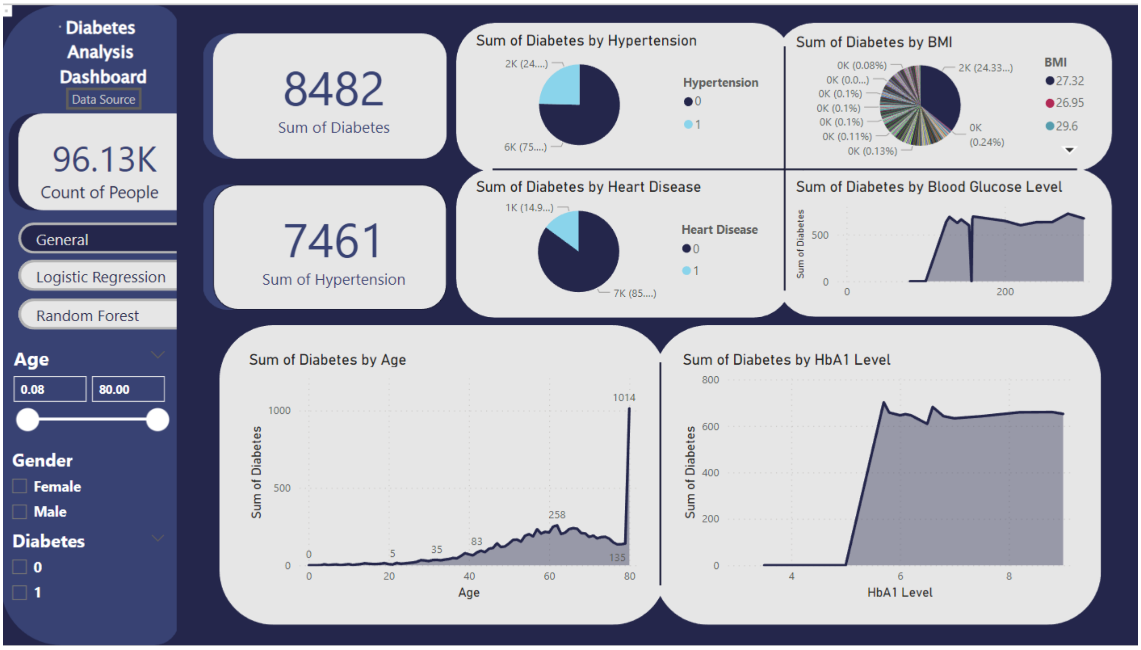
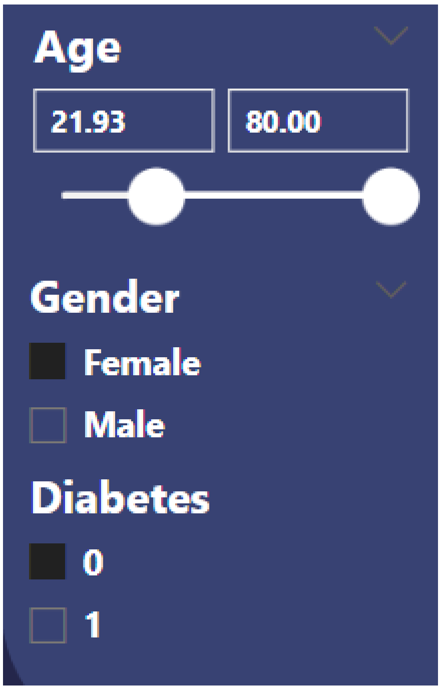


Figure 35: Data Dashboard Overview



Age

21.93 | 80.00

Gender

■ Female

□ Male

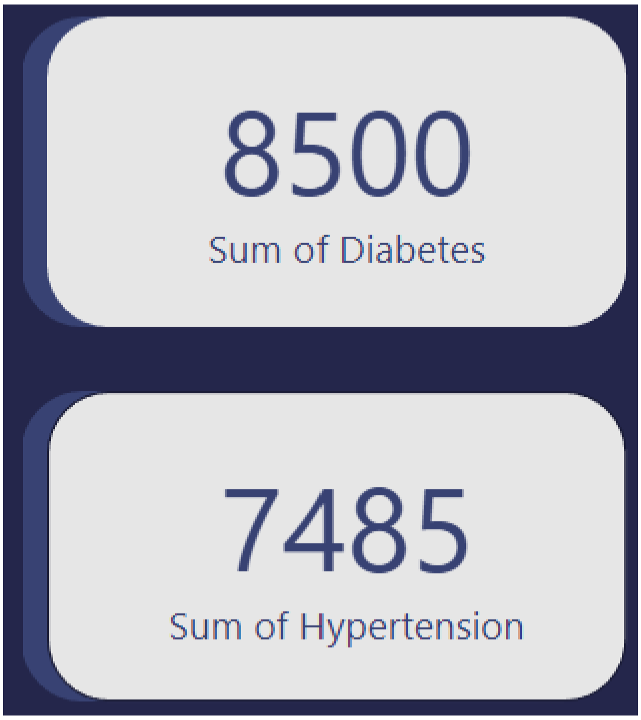
Diabetes

■ 0

□ 1

Figure 36: Data Dashboard Filters

Aside from the charts, the dashboard has a slider at the bottom left of the dashboard, which can be used to filter the dashboard based on used needs.

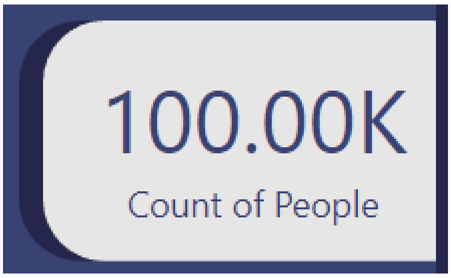


8500

Sum of Diabetes

7485

Sum of Hypertension



100.00K

Count of People

Figure 37: Count of People Card

Figure 38: Sum of Diabetes and Hypertension Card

The first thing in the dashboard is the sum of diabetes and hypertension cards. This in it of itself does not provide much information, but with some combination and filtration, it will show how from the diabetes count out of how many people.

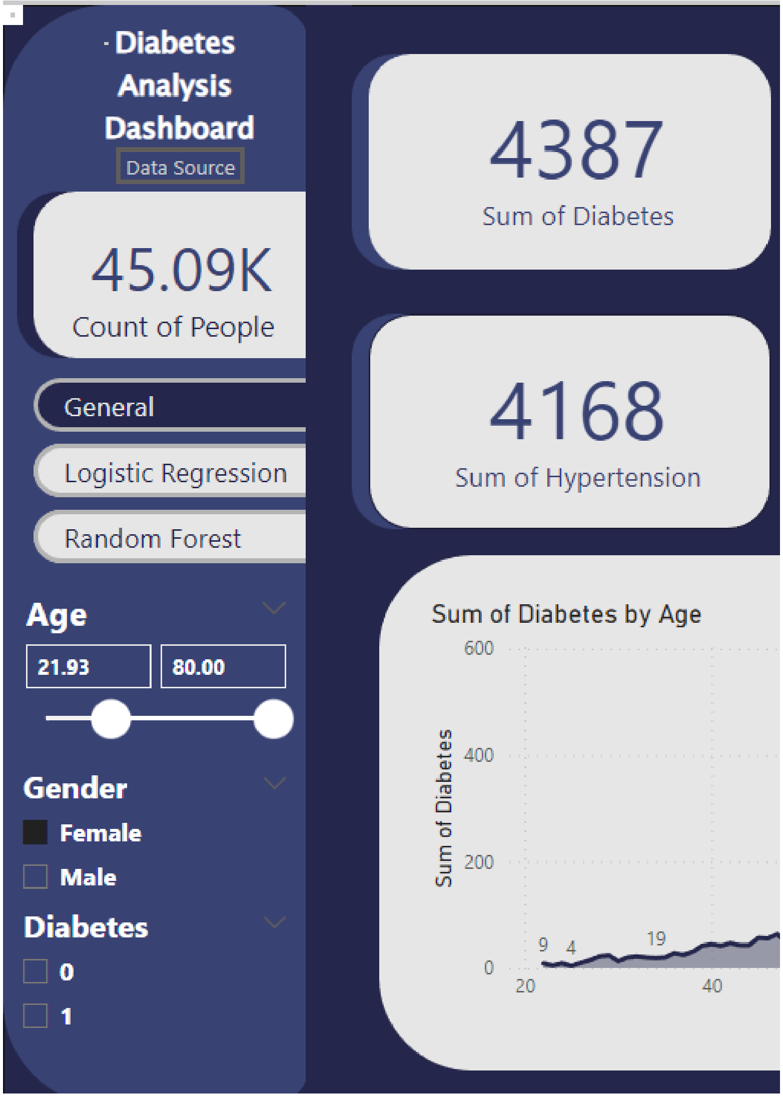


Figure 39: Dashboard Filter Function Outcome

For instance, just by filtering the dashboard to show Female gender only, the dashboard shows there are 45.09K of males with 4387 of them havinng diabetes and 4168 of them having Hypertension

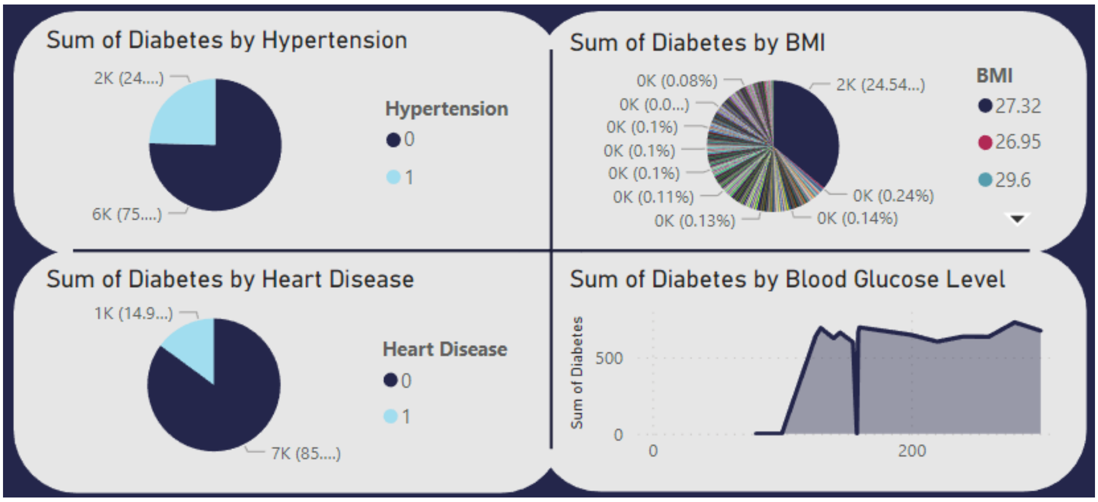


Figure 40: Dashboard Features 1

The next four charts indicats general correlation between diabetes and other variables. In this case, sum of diabetes by hypertension, sum of diabetes by heart disease, sum of diabetes by BMI, and sum of diabetes by blood glucose level.

From the chart, many information can be gathered, such as 75% of diabetics have hypertension, and 85% of diabetics having heart disease, while also describing the BMI of people with highest diabetes percentage is 27.32, and from the fourth chart, it can concluded that people that start to develop diabetes has blood glucose level of 126 ~ 300

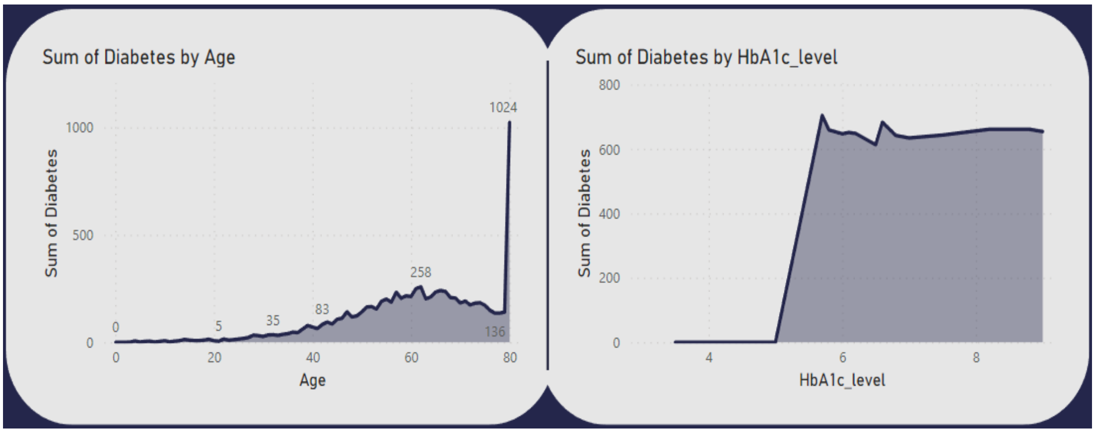


Figure 41: Dashboard Features 2

The next two charts represent sum of diabetes by age, and also sum of diabetes by H bA1c level, which stands for Hemoglobin A1C, which is a measurement of a person’s average blood sugar level over the past 3 months

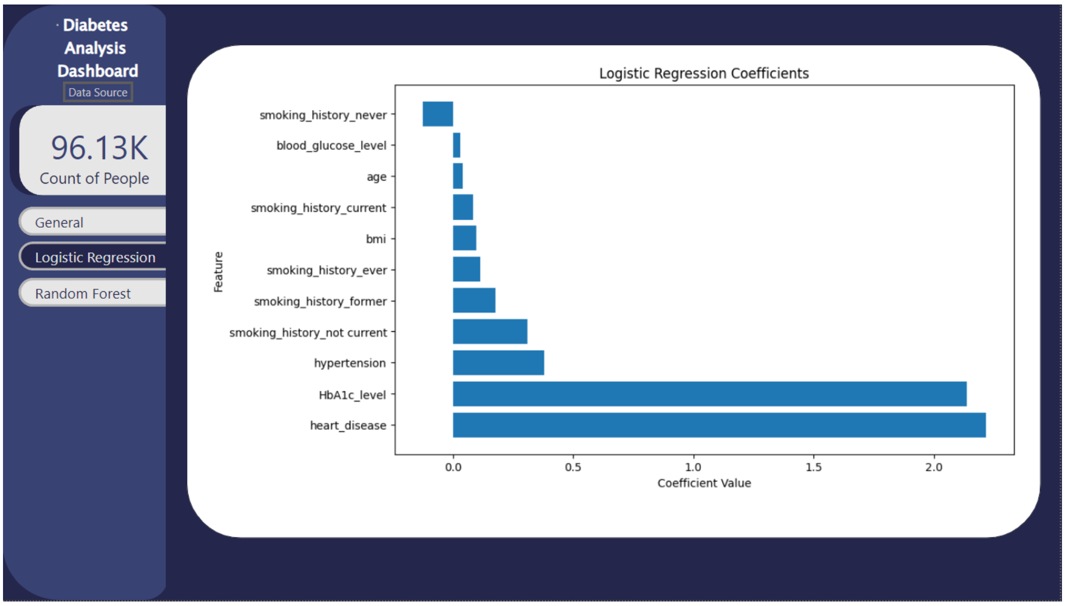


Figure 42: Page 2 of Dashboard: Logistic Regression Coefficients

The second page of the analysis dashboard contains a logistic regression model. From this chart, correlation can be drawn by looking at the graph, for instance, if a person never smoked their entire life, it will drastically lower their chances of having diabetes, and on the contrary, if a person has a high HbA1C level, or has a heart disease, it drastically increases their chances of having or getting diabetes.

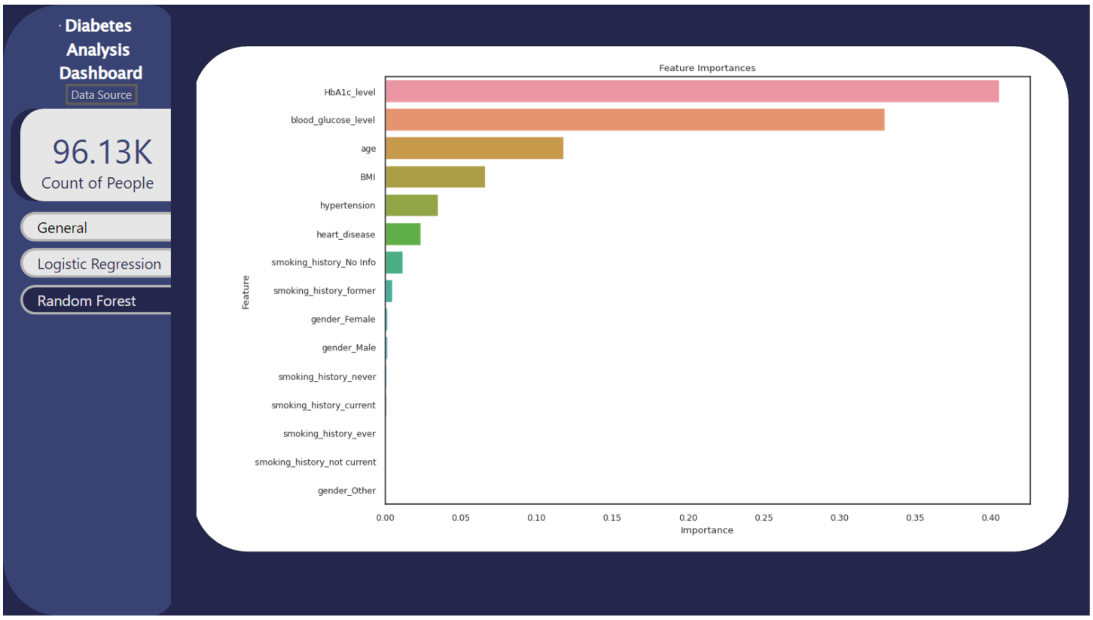


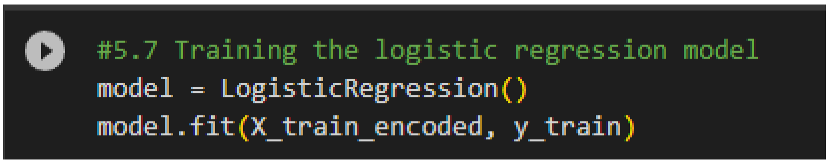
Figure 43: Page 3 of Dashboard: Random Forest Feature Importance

The third page contains Random Forest feature importance bar plot

5.7 Modelling

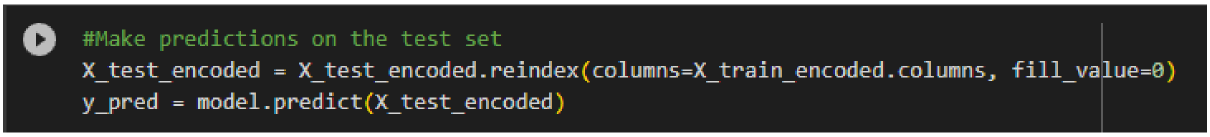
5.7.1 Logistic Regression Machine Learning

The machine learning model chosen for this project is logistic regression model because of its simplicity, efficiency, and interpretability:



```
#5.7 Training the logistic regression model
model = LogisticRegression()
model.fit(X_train_encoded, y_train)
```

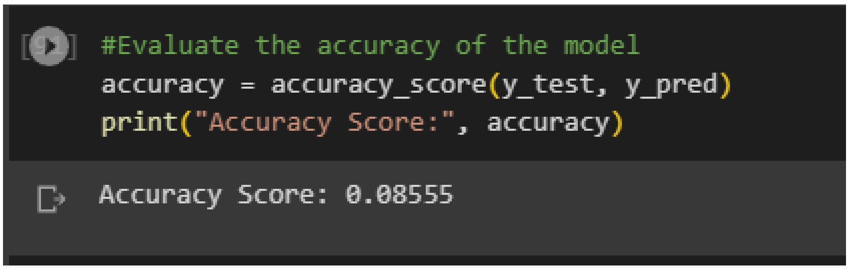
Figure 44: Logistic Regression Model Training



```
#Make predictions on the test set
X_test_encoded = X_test_encoded.reindex(columns=X_train_encoded.columns, fill_value=0)
y_pred = model.predict(X_test_encoded)
```

Figure 45: Logistic Regression Prediction

The next step to choose whether to use this model or not is based on the accuracy score, in this case, it got 95.7%.

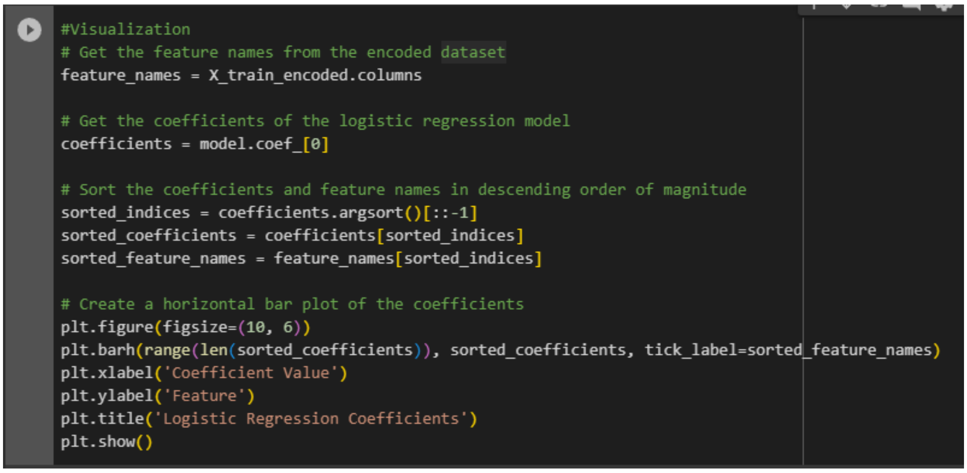


```
#Evaluate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy Score:", accuracy)
```

```
Accuracy Score: 0.08555
```

Figure 46: Prediction Accuracy Score

The model is then visualized:



```
#Visualization
# Get the feature names from the encoded dataset
feature_names = X_train_encoded.columns

# Get the coefficients of the logistic regression model
coefficients = model.coef_[0]

# Sort the coefficients and feature names in descending order of magnitude
sorted_indices = coefficients.argsort()[::-1]
sorted_coefficients = coefficients[sorted_indices]
sorted_feature_names = feature_names[sorted_indices]

# Create a horizontal bar plot of the coefficients
plt.figure(figsize=(10, 6))
plt.barh(range(len(sorted_coefficients)), sorted_coefficients, tick_label=sorted_feature_names)
plt.xlabel('Coefficient Value')
plt.ylabel('Feature')
plt.title('Logistic Regression Coefficients')
plt.show()
```

Figure 47: Visualization Code

Output:

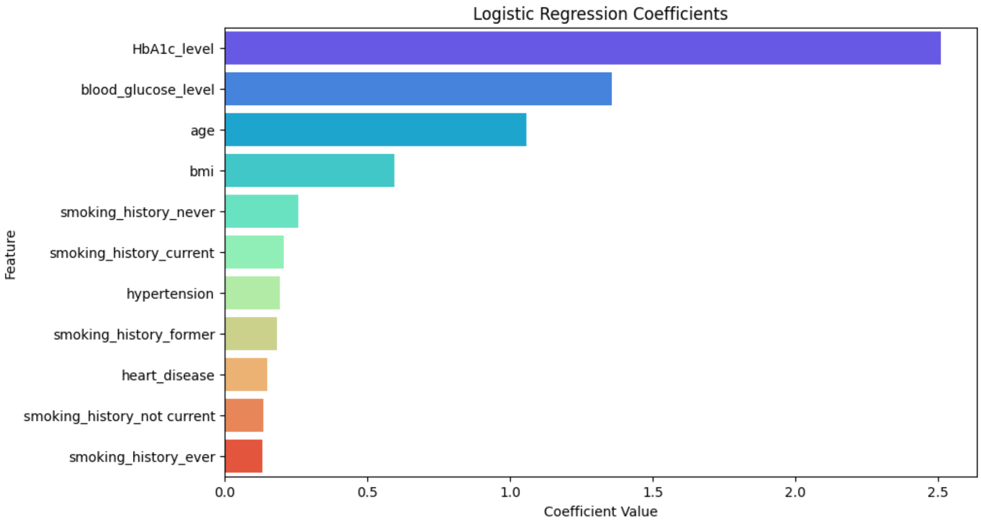
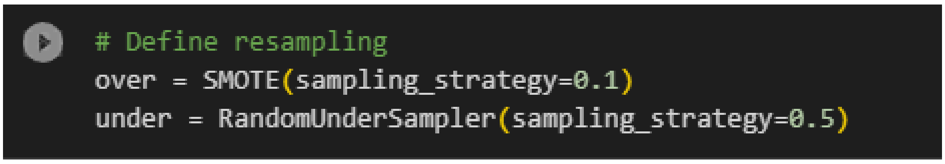


Figure 48: Barplot Output

5.7.2 Random Forest Machine Learning

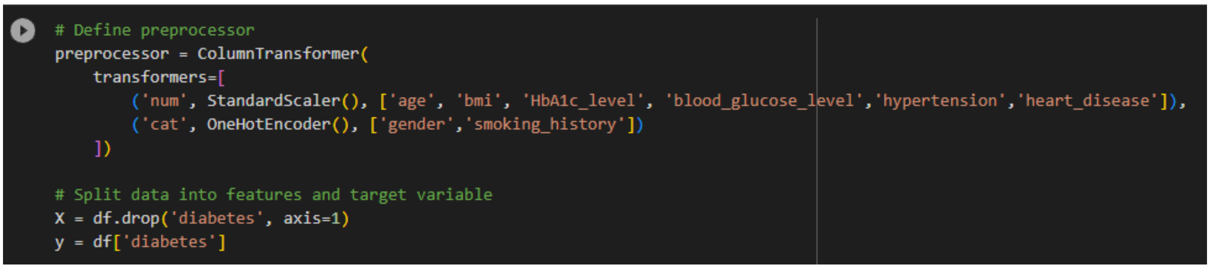
After the data cleaning steps are done in chapter 5.3 (Data Cleaning), SMOTE which stands for Synthetic Minority Over-sampling Technique will be used because from the exploratory data analysis, the dataset is not balanced being 9% positive for diabetes while 91% being negative for diabetes, it is done to make sure that the model will not get biased to most of the class.



```
# Define resampling
over = SMOTE(sampling_strategy=0.1)
under = RandomUnderSampler(sampling_strategy=0.5)
```

Figure 49: SMOTE resampling

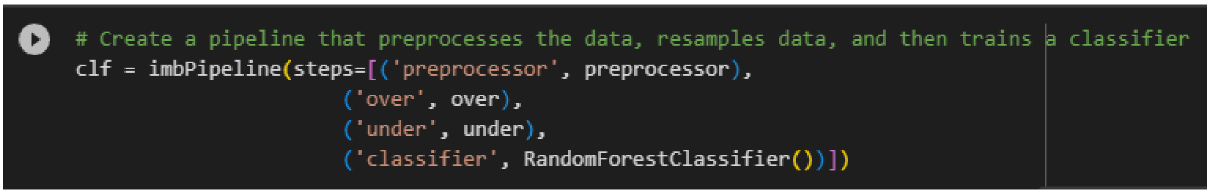
Standardization is the next step, this step also includes one-hot encoding the categorical data (gender and smoking history), this step also includes splitting the data into features and target variable similar to the last chapter.



```
# Define preprocessor
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), ['age', 'bmi', 'HbA1c_level', 'blood_glucose_level','hypertension','heart_disease']),
        ('cat', OneHotEncoder(), ['gender','smoking_history'])
    ])

# Split data into features and target variable
X = df.drop('diabetes', axis=1)
y = df['diabetes']
```

Figure 50: One-Hot Encoding and Standardization

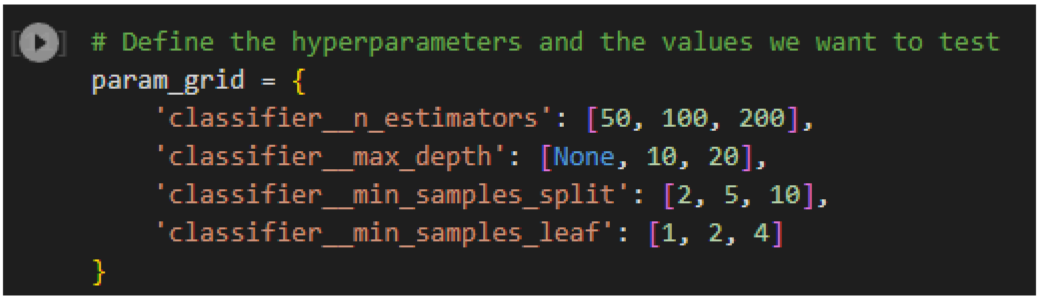
Then, using the following code, a classifier will then be trained.



```
# Create a pipeline that preprocesses the data, resamples data, and then trains a classifier
clf = imbPipeline(steps=[('preprocessor', preprocessor),
                         ('over', over),
                         ('under', under),
                         ('classifier', RandomForestClassifier())])
```

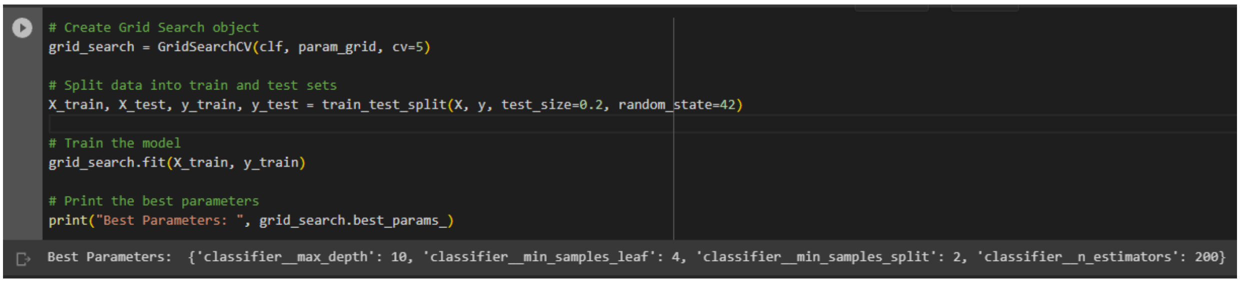
Figure 51: Pipeline Creation

A pipeline is built that initially performs preprocessing procedures before training a model on the data. RandomForestClassifier is employed, which is a popular and a powerful classification technique. GridSearchCV is used to tune the model's hyperparameters, which executes an exhaustive search over the provided parameter values for the estimator. Cross-validation is used to choose the best-performing model.



```
# Define the hyperparameters and the values we want to test
param_grid = {
    'classifier__n_estimators': [50, 100, 200],
    'classifier__max_depth': [None, 10, 20],
    'classifier__min_samples_split': [2, 5, 10],
    'classifier__min_samples_leaf': [1, 2, 4]
}
```

Figure 52: Hyperparanmeter Definition



```
# Create Grid Search object
grid_search = GridSearchCV(clf, param_grid, cv=5)

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train the model
grid_search.fit(X_train, y_train)

# Print the best parameters
print("Best Parameters: ", grid_search.best_params_)
```

```
Best Parameters:  {'classifier__max_depth': 10, 'classifier__min_samples_leaf': 4, 'classifier__min_samples_split': 2, 'classifier__n_estimators': 200}
```

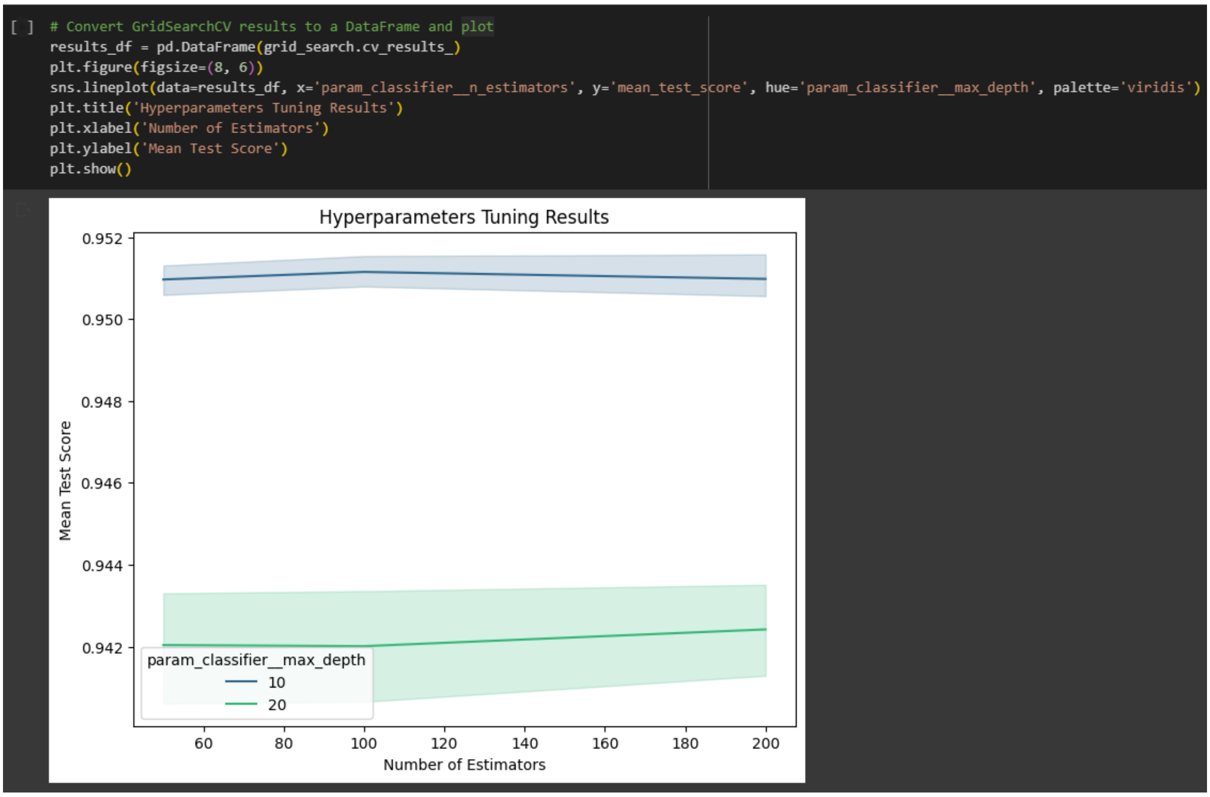
Figure 53: Grid Search Best Parameters

Output Explanation:

1. max_depth of 10: This suggests that the trees in the forest have a maximum depth of 10 levels. Limiting the depth of the tree aids in the reduction of overfitting. Based on this finding, it appears that a medium-complexity tree works well for our data. A tree with too much complexity (a deeper tree) may catch noise, whereas a tree with too little complexity (a shallower tree) might fail to identify the underlying structure of the data.
2. min_samples_leaf of 2: Specifies that each leaf (the terminal node of a decision tree where predictions are created) must have at least two samples. Like max_depth, this option is employed to control overfitting. The model avoids fitting to outliers or noise in the training data by requiring at least two samples to make a prediction.

3. min_samples_split of 2: The min_samples_split of 2 indicates that a node must have at least two samples in order to be split (to produce two child nodes). This, like the min_samples_leaf parameter, can aid in the management of overfitting.
4. n_estimators of 50: The number of decision trees in the forest is indicated by n_estimators of 50. The Random Forest algorithm reduces overfitting and variation by averaging the forecasts of multiple decision trees to generate a final prediction. In this situation, it appears that having 50 trees in the forest yields the best results.

The following code is used to visualize the result of hyperparameter using grid search with cross-validation, it is inportant since it is used to evaluate and tune hyperparameters for random forest classifier.

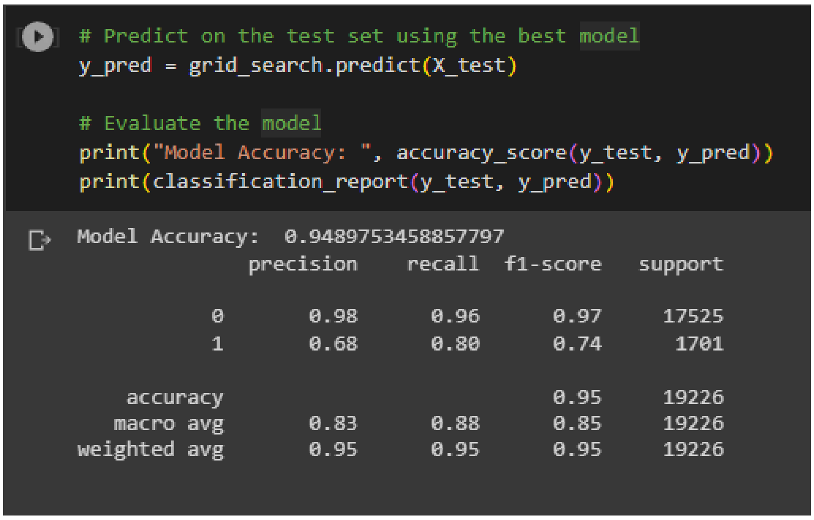


```
[ ] # Convert GridSearchCV results to a DataFrame and plot
    results_df = pd.DataFrame(grid_search.cv_results_)
    plt.figure(figsize=(8, 6))
    sns.lineplot(data=results_df, x='param_classifier__n_estimators', y='mean_test_score', hue='param_classifier__max_depth', palette='viridis')
    plt.title('Hyperparameters Tuning Results')
    plt.xlabel('Number of Estimators')
    plt.ylabel('Mean Test Score')
    plt.show()
```

Figure 54: Hyperparameter Tuning

The line plot created by the code visualizes the relationship between two important hyperparameters for a random forest classifier

The next step is to find out the model accuracy score.



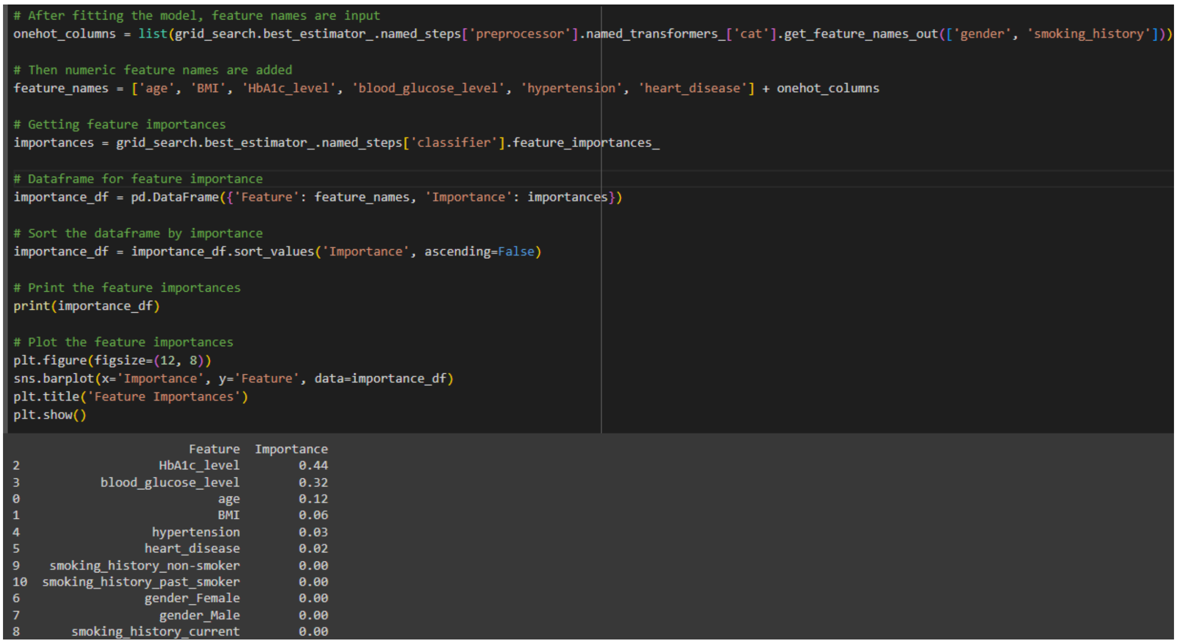
```
# Predict on the test set using the best model
y_pred = grid_search.predict(X_test)

# Evaluate the model
print("Model Accuracy: ", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Model Accuracy:  0.9489753458857797
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     17525
           1       0.68      0.80      0.74      1701

    accuracy                           0.95     19226
   macro avg       0.83      0.88      0.85     19226
weighted avg       0.95      0.95      0.95     19226
```

Figure 55: Random Forest Model Accuracy

As seen from the output, the accuracy score is around 94.9% which is considered to be good.

The following code is the last step of the modelling process, which is coming up with the feature importance.



```
# After fitting the model, feature names are input
onehot_columns = list(grid_search.best_estimator_.named_steps['preprocessor'].named_transformers_['cat'].get_feature_names_out(['gender', 'smoking_history']))

# Then numeric feature names are added
feature_names = ['age', 'BMI', 'HbA1c_level', 'blood_glucose_level', 'hypertension', 'heart_disease'] + onehot_columns

# Getting feature importances
importances = grid_search.best_estimator_.named_steps['classifier'].feature_importances_

# Dataframe for feature importance
importance_df = pd.DataFrame({'Feature': feature_names, 'Importance': importances})

# Sort the dataframe by importance
importance_df = importance_df.sort_values('Importance', ascending=False)

# Print the feature importances
print(importance_df)

# Plot the feature importances
plt.figure(figsize=(12, 8))
sns.barplot(x='Importance', y='Feature', data=importance_df)
plt.title('Feature Importances')
plt.show()
```

```
                      Feature  Importance
2                 HbA1c_level        0.44
3         blood_glucose_level        0.32
0                         age        0.12
1                         BMI        0.06
4                hypertension        0.03
5               heart_disease        0.02
9   smoking_history_non-smoker       0.00
10  smoking_history_past_smoker      0.00
6               gender_Female        0.00
7                 gender_Male        0.00
8     smoking_history_current        0.00
```

Figure 56: Random Forest Feature Importance

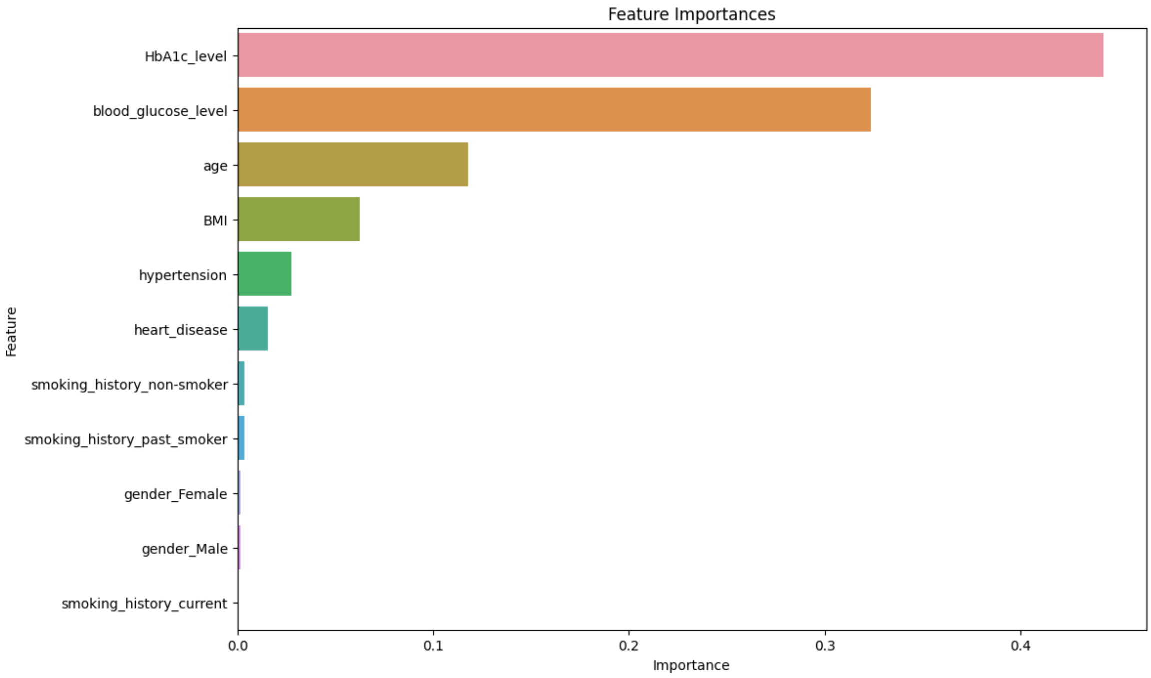


Figure 57: Random Forest Feature Importance Barplot

Based on the output, it is clear that HbA1C level has the highest feature importance out of all features, and the rest of the output will be discussed in chapter 6 (Results and Discussion)

5.8 Summary

In summary, in this chapter, the diabetes dataset has been fully explored and cleaned before exporting into a dashboard, then visualized using PowerBI in a form of dashboard producing univariate, bivariate and multivariate data visuals. Then the same dataset is being re-cleaned for the purpose of two machine learning models, the first one being logistic regression model, which the model accuracy score reached 85.5%, and the second one, which is random forest model, which yields model accuracy score of 95.7%. Both model gives deeper insight from the dataset on what variable effects diabetes, which will be discussed in chapter 6 (results and discussion).

CHAPTER 6: Results and Discussion

6.1 Introduction

Chapter 6 is the chapter in which mainly the outcome of the model will be discussed, coefficient values from logistic regression model and also feature importance from random forest model.

6.2 Results and discussion

The following bar plot shows the logistic regression coefficients with X being the coefficient value representing the estimated effect or contribution of Y features.

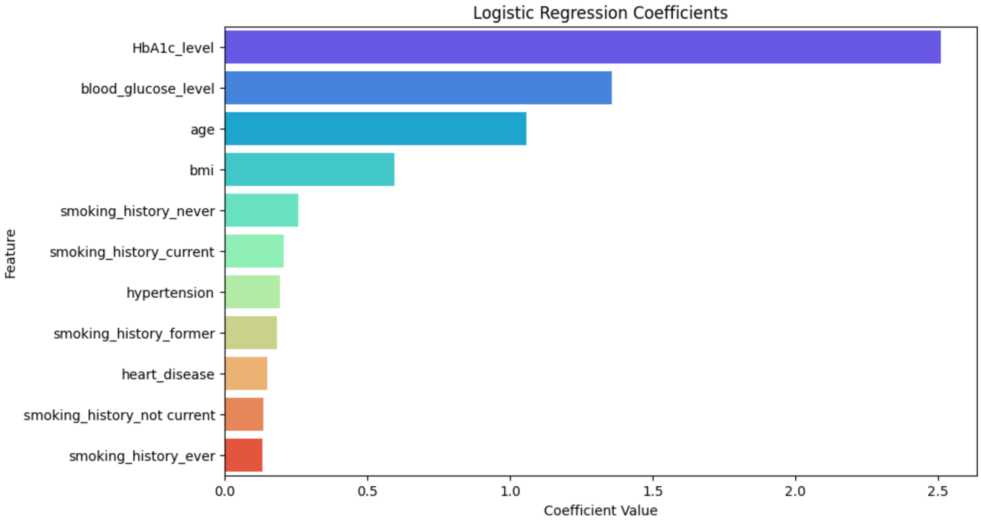


Figure 58: Logistic Regression Discussion

The following is the interpretation of each feature which are considered important in the plot:

- The four most important and highest feature coefficient value out of all features are BMI, Age, Blood glucose level and HbA1C level.
- Followed by smoking histories, and hypertension.

The second machine learning model, Random tree, produces the following bar plot showing feature importances, meaning how important a feature is for getting diabetes.

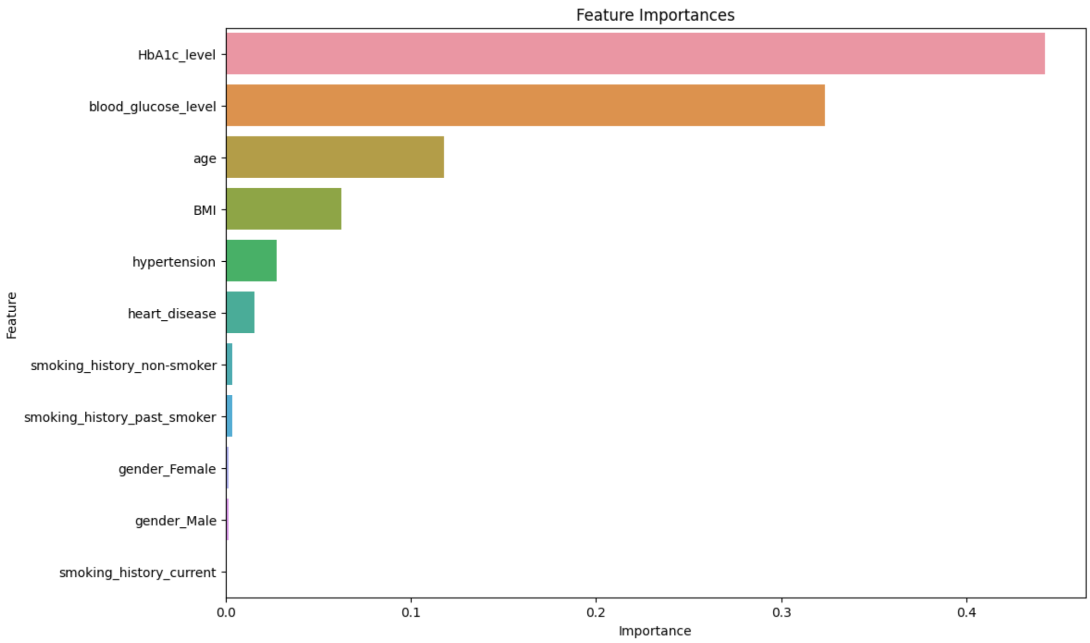


Figure 59: Random Forest Discussion

The following is the interpretation according to the barplot:

- The four most important and highest feature coefficient value out of all features are BMI, Age, Blood glucose level and HbA1C level.
- BMI has 0.06 feature importance meaning it has a relatively medium correlation to diabetes.
- Age twice the feature importance than BMI being 0.12, meaning the older someone gets, the higher the risk of the person getting diabetes.
- Blood glucose level, which is the main cause of diabetes itself, has the 2nd highest feature importance, being 0.32.
- The highest out of all features is HbA1C level, being 0.44, meaning if a person has high HbA1C level, it is very dangerous for them since they have a high chance of manifesting diabetes.

Both models has clashes and similarities, it is taken that HbA1C level is the most important feature to be maintained since it reached a high feature importance and also high coefficient value from both models, followed by hypertension and heart disease.

CHAPTER 7: CONCLUSIONS AND REFLECTIONS

At the end of the project, the dataset has been successfully analysed with expected outcome, being able to draw connections among other variables to the target variable being diabetes utilizing machine learning models while also visualizing said model and implementing it to the dashboard.

Investigations and research are done thoroughly in order to finish the project, as for gaps in the project, mainly, the project does not discuss about genetics in diabetics which were discussed in the investigation report, the reason is because the dataset provided is limited to patients' testing records and also characteristics. If there were information about patients' deeper test results, such as genetics and small lifestyle choices, there might be some areas of the project that can be improved to even predict diabetes as high as 99.9%, but as of now, since complete cure of any types of diabetes has not been discovered yet, it is very difficult to make any kinds of predictions.

Data exploration, data cleaning, and data modelling are all done using Colaborator Google which is a hosted Jupyter Notebook service which is provided free and without any setups. Data visualizations are done through PowerBI, a dashboard building tool.

For problems and limitations, the analysis of diabetes could not be completely accurate for all people since variables analysed is only a handful whereas in the real world, there could be thousands of variables that may have higher relation to diabetes, the analysis also utilize 2 machine learning models and in the case of future enhancement, should use more.

As for reflections, the student has deepened their knowledge of research and mainly, data analysis, in all aspect, from choosing dataset, to data exploration, data cleaning, data visualization, machine learning modelling and building dashboard. All was done in hopes that the outcome may prove beneficial to all people, no matter gender and age.

In this project, the student is able to overcome their issues, problems, errors, confusions, and desperation thanks to the lecturers and mainly to the FYP supervisor Dr. Vazeerudeen Hameed.

References

{Coding} Sight. (10 Sept, 2019). *Coding Sight*. Retrieved from 10 Best MySQL GUI Tools: https://codingsight.com/10-best-mysql-gui-tools/

Aubert, R. E., Ballard, D. J., & Barret-Connor, E. (1995). *Diabetes in America 2nd Edition.* National Institutes of Health.

B., R. (31 Jan, 2023). *HostingerTutorials*. Retrieved from What is MySQL: MySQL Explained For Beginners: https://www.hostinger.com/tutorials/what-is-mysql

Bigelow, S. J. (June, 2001). *Operating System*. Retrieved from operating system (OS): https://www.techtarget.com/whatis/definition/operating-system-OSv

Bold BI. (n.d.). *Analytics*. Retrieved from Bold BI for Predictive Analytics: https://www.boldbi.com/dashboard-examples/analytics/predictive-analytics

Calzon, B. (3 Mar, 2023). *datapine*. Retrieved from Your Modern Business Guide To Data Analysis Methods And Techniques: https://www.datapine.com/blog/data-analysis-methods-and-techniques/#data-analysis-definition

Centers for Disease Control and Prevention. (7 July, 2022). *What is Diabetes?* Retrieved from Diabetes: https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=With%20diabetes%2C%20your%20body%20either,releases%20it%20into%20your%20bloodstream.

Centers for Disease Control and Prevention. (n.d.). *Diabetes*. Retrieved from What is Diabetes?: https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=There%20are%20three%20main%20types,diabetes%20(diabetes%20while%20pregnant).

Data Science Process Alliance. (n.d.). *What is CRISP DM?* Retrieved from Data Sceince Process Alliance: https://www.datascience-pm.com/crisp-dm-2/

Diabetes UK. (n.d.). *what is hba1c?* Retrieved from Diabetes UK: https://www.diabetes.org.uk/guide-to-diabetes/managing-your-diabetes/hba1c

Federation, I. D. (9 December, 2021). *Diabetes facts & figures*. Retrieved from International Diabetes Federation: https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html#:~:text=The%20IDF%20Diabetes%20Atlas%20Tenth,and%20783%20million%20by%202045.

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & Garcia-Garcia, J. A. (2021). Machine learning and deep learning. *Diabetology & Metabolic Syndrome*, 1-22.

GeeksforGeeks. (04 Sep, 2019). *Introduction to Visual Studio*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/introduction-to-visual-studio/

Google. (n.d.). *Google Colab*. Retrieved from Welcome to Colab!: https://colab.research.google.com/notebooks/intro.ipynb

Hotz, N. (19 Jan, 2023). *Data Science Process Alliance*. Retrieved from What is CRISP DM?: https://www.datascience-pm.com/crisp-dm-2/

IBM. (n.d.). *What is a Decision Tree?* Retrieved from IBM: https://www.ibm.com/my-en/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.

imbalanced-learn. (08 Jul, 2023). *imbalanced-learn documentation*. Retrieved from imbalanced-learn: https://imbalanced-learn.org/stable/

Kelley, K. (07 Feb , 2023). *What is Data Analysis? Methods, Process and Types Explained*. Retrieved from simplilearn: https://www.simplilearn.com/data-analysis-methods-process-types-article

Kelley, K. (9 Jun, 2023). *What is Data Analysis? Methods, Process and Types Explained*. Retrieved from simplilearn: https://www.simplilearn.com/data-analysis-methods-process-types-article

L., C., S., R., S., L., I., C., P., L., & N. (2005). diabetes, genetics and ethnicity. *Alimentary Pharamacology & Therapeutics*, 16-19.

matplotlib. (2007). *Matplotlib Official Documentation*. Retrieved from matplotlib: https://matplotlib.org/

Mayo Clinic. (n.d.). *Diabetes & Conditions*. Retrieved from Type 1 diabetes: https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20353011

Mayo Clinic. (n.d.). *Diabetes & Conditions*. Retrieved from Type 2 diabetes: https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Mayo Clinic. (n.d.). *Diabetes & Conditions*. Retrieved from Gesttional Diabetes: https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/symptoms-causes/syc-20355339#:~:text=Gestational%20diabetes%20is%20diabetes%20diagnosed,pregnancy%20and%20your%20baby's%20health.

McCombes, S. (2 January, 2023). *How to Write a Literature Review*. Retrieved from How to Write a Literature Review | Guide, Examples, & Templates: https://www.scribbr.com/methodology/literature-review/

McCombes, S., & George, T. (30 January, 2023). *What Is a Research Methodology*. Retrieved from What Is a Research Methodology | Steps & Tips: https://www.scribbr.com/dissertation/methodology/

Microsoft. (26 Mar, 2022). *Introduction to dashboards for Power BI designers*. Retrieved from Microsoft: https://docs.microsoft.com/en-us/power-bi/create-reports/service-dashboards

Microsoft. (29 Jan, 2022). *What is SQL Server Management Studio (SSMS)?* Retrieved from Microsoft: https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver15

Nita Gandhi Forohi, N. J. (2010). Epidemiology of diabetes. *Medicine*, 602-606.

Plesky, E. (27 Mar, 2022). *plesk*. Retrieved from MySQL vs MSSQL: Comparing Similarities and Differences: https://www.plesk.com/blog/various/mysql-vs-mssql/#:~:text=MySQL%20lets%20developers%20utilize%20binaries,accessing%20binaries%20or%20database%20files.

Rahati, S., Shahraki, M., Arjomand, G., & Shahraki, T. (2014). Food Pattern, Lifestyle and Diabetes Mellitus. *Food Pattern, Lifestyle and Diabetes Mellitus*, 1-5.

Rodrigues, I. (17 Feb, 2020). *CRISP-DM methodology leader in data mining and big data*. Retrieved from TowardsDataScience: https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781

scikit-learn. (n.d.). *scikit-learn Machine Learning in Python*. Retrieved from scikit-learn: https://scikit-learn.org/stable/

Shamoo, & Resnik. (2003). Responsible Conduct of Research. Oxford University Press.

Taylor, D. (18 May, 2022). *Power BI Tutorial: What is Power BI? Why Use? DAX Examples*. Retrieved from Guru99: https://www.guru99.com/power-bi-tutorial.html

Teves, J. P. (2015). Data Visualization that "Fits": Designing Effective Dashboards for Healthcare Providers, Patients, and Family Caregivers to Patients with Diabetes. *Journal of Health & Medical Informatics*, 1-131.

Travis E. Oliphant, P. (2006). *Guide to NumPy.* Treelgol Publishing.

VanderPlas, J. (2016). *Python Data Science Handbook.* O'Reilly Media.

Wijaya, C. Y. (27 Apr, 2021). *CRISP-DM Methodology For Your First Data Science Project*. Retrieved from TowardsDataScience: https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c

Wing, R. R., Goldstein, M. G., Acton, K. J., Birch, L. L., Sallis, F. J., Sallis, J. F., . . . Surwit, R. S. (2001). Behavioral Science Research in Diabetes. *Lifestyle changes related to obesity, eating behavior, and physical*, 117-123.

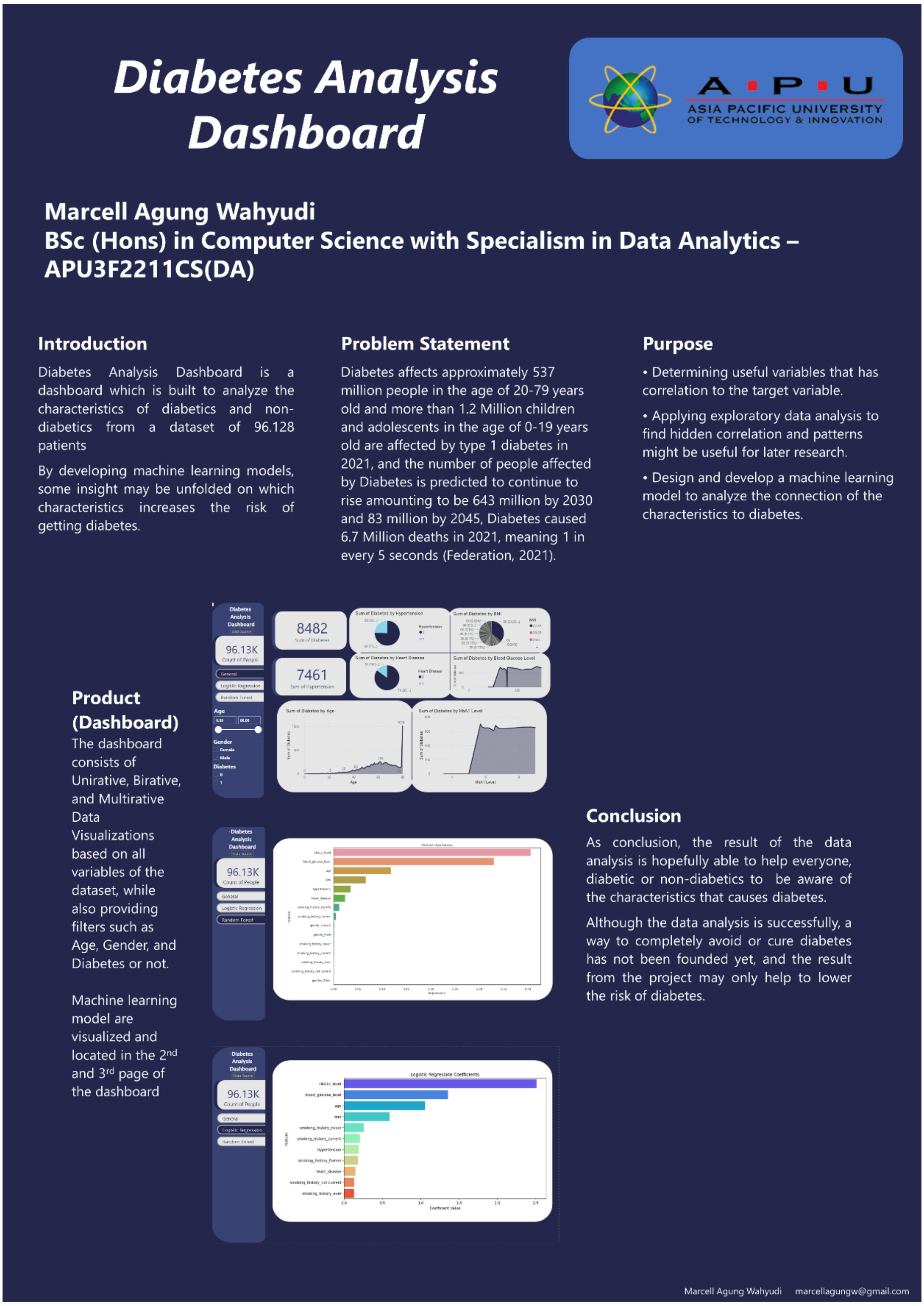
World Heart Federation. (n.d.). *Hypertension*. Retrieved from World Heart Federation: https://world-heart-federation.org/what-we-do/hypertension/#:~:text=Hypertension%20is%20the%20leading%20preventable,10%20million%20people%20every%20year.

Zavgorodniy, A. (11 Jan, 2018). *MANAGING DATA SCIENCE PROJECT*. Retrieved from LinkedIn: https://www.linkedin.com/pulse/managing-data-science-project-aleksey-zavgorodniy

Zia Sherrel, M. (30 August, 2022). *Medical News Today*. Retrieved from How do diabetes rates vary by country?: https://www.medicalnewstoday.com/articles/325592#takeaway

APPENDICES

FYP Poster


A • P • U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION
Diabetes Analysis Dashboard
96.13K
Count of People
General
Logistic Regression
Random Forest
Age
Gender
Female
Male
Diabetes
8482
Sum of Diabetes
7461
Sum of Hypertension
Sum of Diabetes by Hypertension
Sum of Diabetes by BMI
Sum of Diabetes by Heart Disease
Sum of Diabetes by Blood Glucose Level
Sum of Diabetes by Age
Sum of Diabetes by HbA1c Level
Logistic Regression Coefficients
HbA1c_level
blood_glucose_level
age
bmi
smoking_history_never
smoking_history_current
hypertension
smoking_history_former
heart_disease
smoking_history_not current
smoking_history_ever
Coefficient Value

Diabetes Analysis Dashboard

Marcell Agung Wahyudi
BSc (Hons) in Computer Science with Specialism in Data Analytics – APU3F2211CS(DA)

Introduction

Diabetes Analysis Dashboard is a dashboard which is built to analyze the characteristics of diabetics and non-diabetics from a dataset of 96.128 patients

By developing machine learning models, some insight may be unfolded on which characteristics increases the risk of getting diabetes.

Problem Statement

Diabetes affects approximately 537 million people in the age of 20-79 years old and more than 1.2 Million children and adolescents in the age of 0-19 years old are affected by type 1 diabetes in 2021, and the number of people affected by Diabetes is predicted to continue to rise amounting to be 643 million by 2030 and 83 million by 2045, Diabetes caused 6.7 Million deaths in 2021, meaning 1 in every 5 seconds (Federation, 2021).

Purpose

- Determining useful variables that has correlation to the target variable.
- Applying exploratory data analysis to find hidden correlation and patterns might be useful for later research.
- Design and develop a machine learning model to analyze the connection of the characteristics to diabetes.

Product (Dashboard)

The dashboard consists of Unirative, Birative, and Multirative Data Visualizations based on all variables of the dataset, while also providing filters such as Age, Gender, and Diabetes or not.

Machine learning model are visualized and located in the 2nd and 3rd page of the dashboard

Conclusion

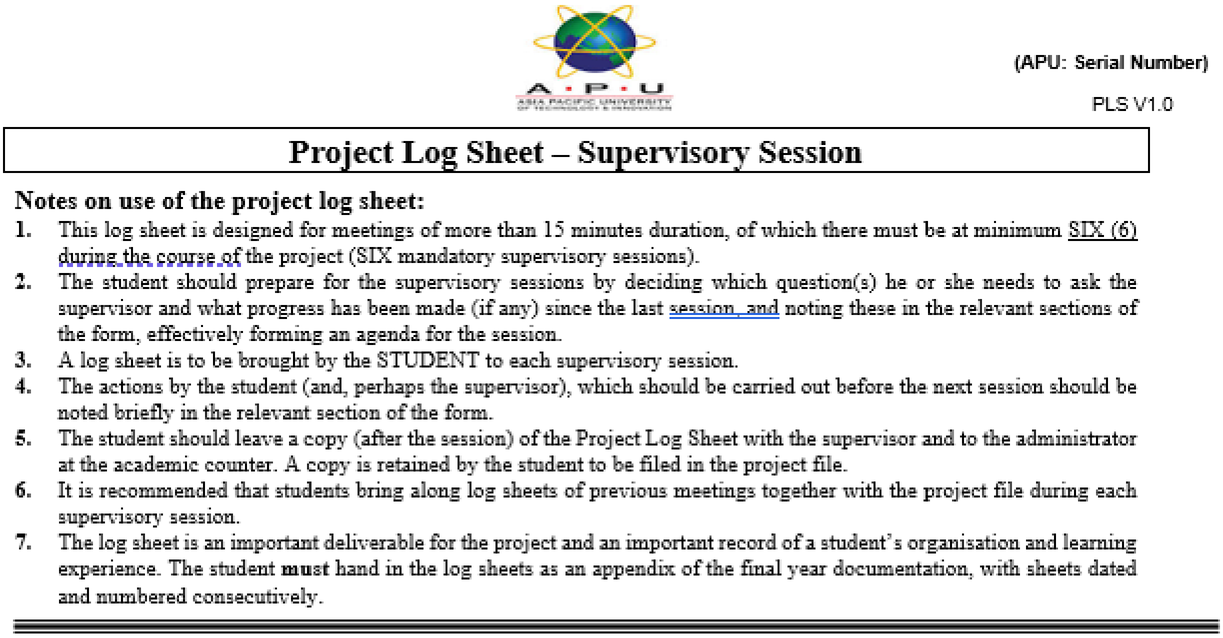
As conclusion, the result of the data analysis is hopefully able to help everyone, diabetic or non-diabetics to be aware of the characteristics that causes diabetes.

Although the data analysis is successfully, a way to completely avoid or cure diabetes has not been founded yet, and the result from the project may only help to lower the risk of diabetes.

Marcell Agung Wahyudi marcellagungw@gmail.com

Figure 60: FYP Poster

Log sheets



A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: Marcell Agung Wahyudi   Date: 16th February 2023 Meeting No: 1

Project title: Diabetes Prediction Dashboard Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Vazeerudeen Hameed   Supervisor's signature: VAZ

Items for discussion (noted by student before mandatory supervisory meeting):

1. **Topic Discussion**
2. **Topic Acceptance**
3. **Format of the IR**
4.

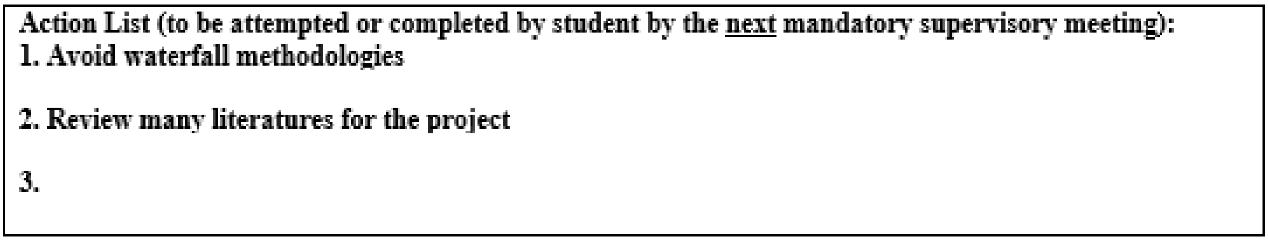
Record of discussion (noted by student during mandatory supervisory meeting):

1. What to be careful of (Don't use waterfall methodologies)

2. What should be focused more (Methodology Chapter)

3. How to do Literature Review correctly

4.

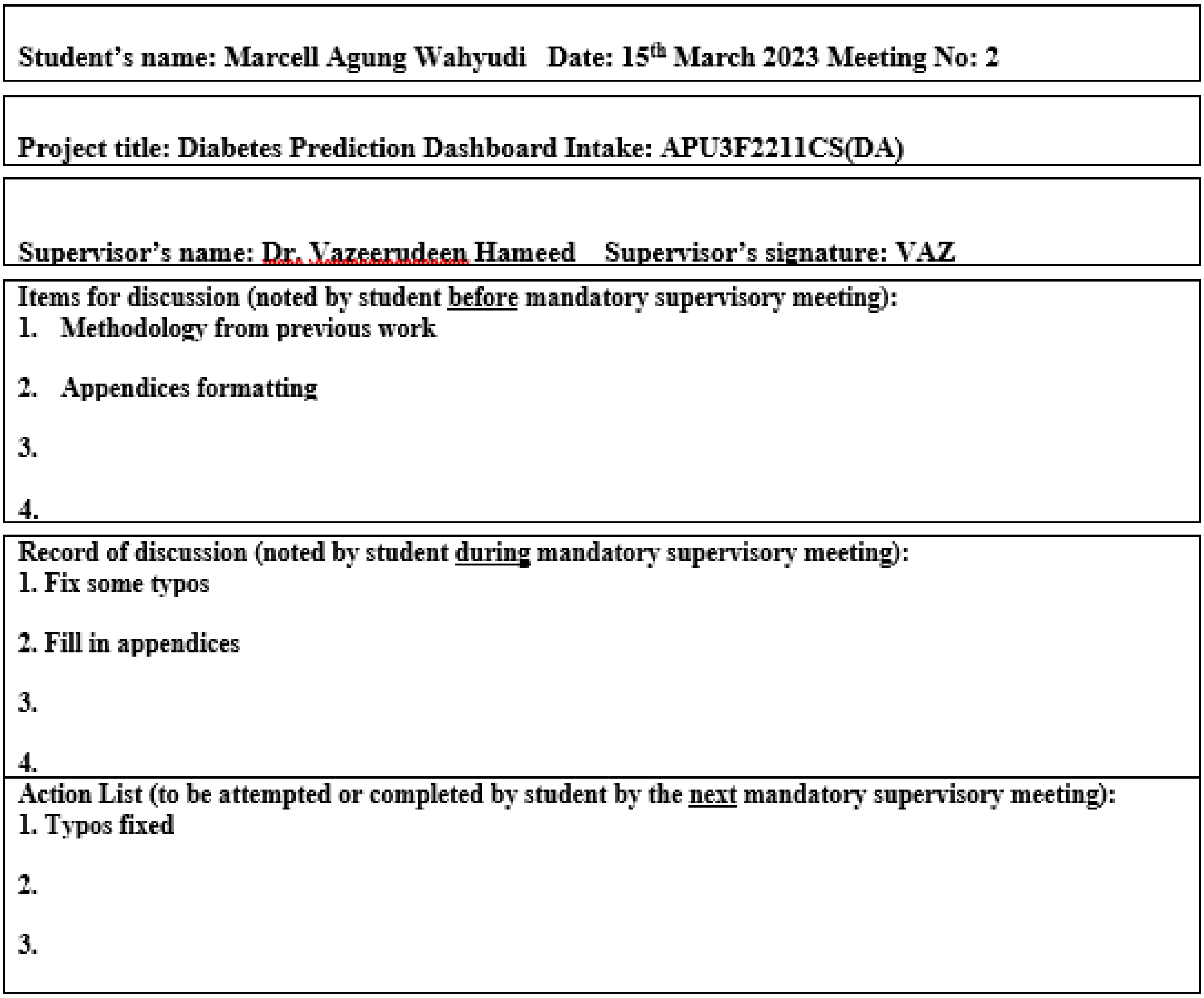


Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Avoid waterfall methodologies

2. Review many literatures for the project

3.



Student's name: Marcell Agung Wahyudi Date: 15th March 2023 Meeting No: 2

Project title: Diabetes Prediction Dashboard Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Vazeerudeen Hameed Supervisor's signature: VAZ

Items for discussion (noted by student before mandatory supervisory meeting):

1. Methodology from previous work

2. Appendices formatting

3.

4.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Fix some typos

2. Fill in appendices

3.

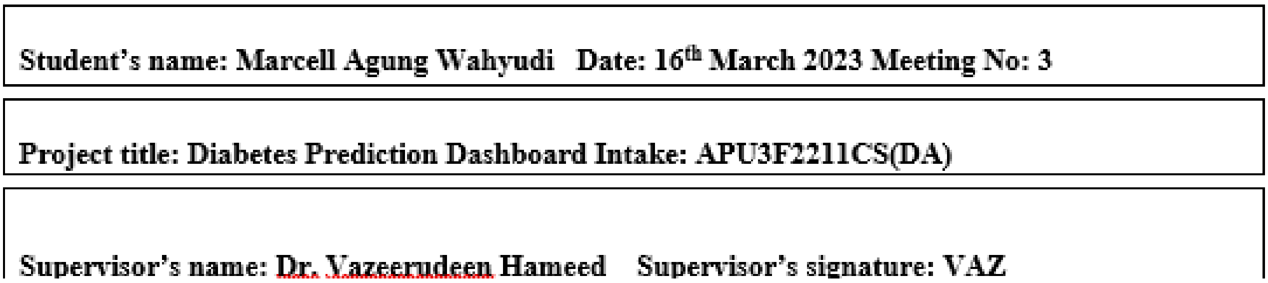
4.

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Typos fixed

2.

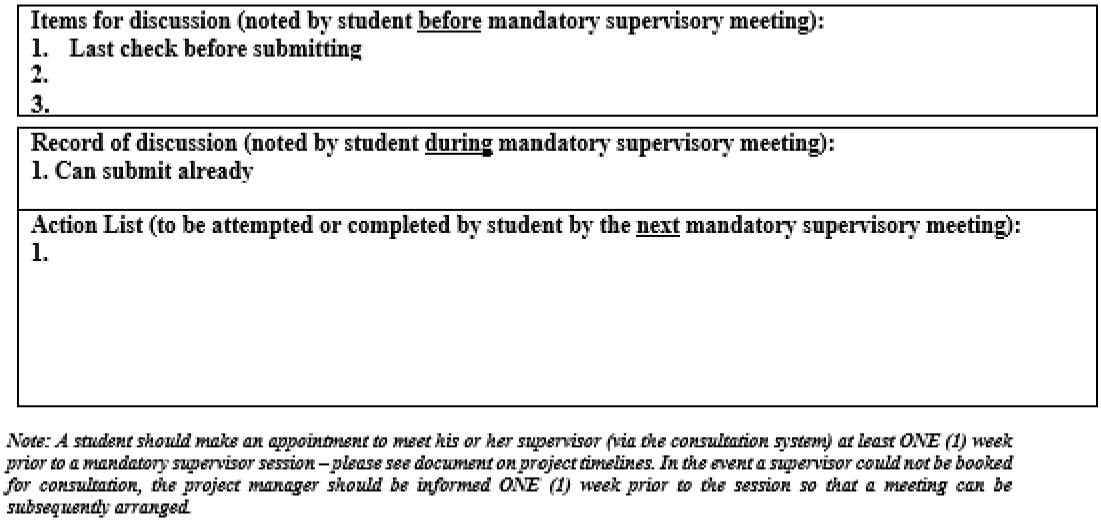
3.



Student's name: Marcell Agung Wahyudi Date: 16th March 2023 Meeting No: 3

Project title: Diabetes Prediction Dashboard Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Vazeerudeen Hameed Supervisor's signature: VAZ



Items for discussion (noted by student before mandatory supervisory meeting):

1. Last check before submitting
2.
3.

Record of discussion (noted by student during mandatory supervisory meeting):

1. Can submit already

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: Marcell Agung Wahyudi Date: 26th June 2023 Meeting No: 4

Project title: Diabetes Prediction Dashboard Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Vazeerudeen Hameed Supervisor's signature: VAZ

Items for discussion (noted by student before mandatory supervisory meeting):

1. **Dashboard Progress**
2. **Dataset Problem**

Record of discussion (noted by student during mandatory supervisory meeting):

1. Biased Dataset

2. Work on data modelling

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Find another dataset

2. Try to start on data modelling

3.

Student's name: Marcell Agung Wahyudi Date: 5th July 2023 Meeting No: 5

Project title: Diabetes Prediction Dashboard Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Vazeerudeen Hameed Supervisor's signature: VAZ

Items for discussion (noted by student before mandatory supervisory meeting):

1. **Project Completion**

2. Items missing from completion

3.

4.

Record of discussion (noted by student <u>during</u> mandatory supervisory meeting):

1. Complete Documentation

2. Decorate Dashboard

3.

4.

Action List (to be attempted or completed by student by the <u>next</u> mandatory supervisory meeting):

1. Complete the Documentation

2. Complete the Dashboard

3.

Student's name: Marcell Agung Wahyudi Date: 13th July 2023 Meeting No: 6

Project title: Diabetes Prediction Dashboard Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Vazeerudeen Hameed Supervisor's signature: VAZ

Items for discussion (noted by student <u>before</u> mandatory supervisory meeting):

1. **Last check before submitting**
2. **Machine Learning Model Comparison**
3.

Record of discussion (noted by student <u>during</u> mandatory supervisory meeting):

1. Need to change some things

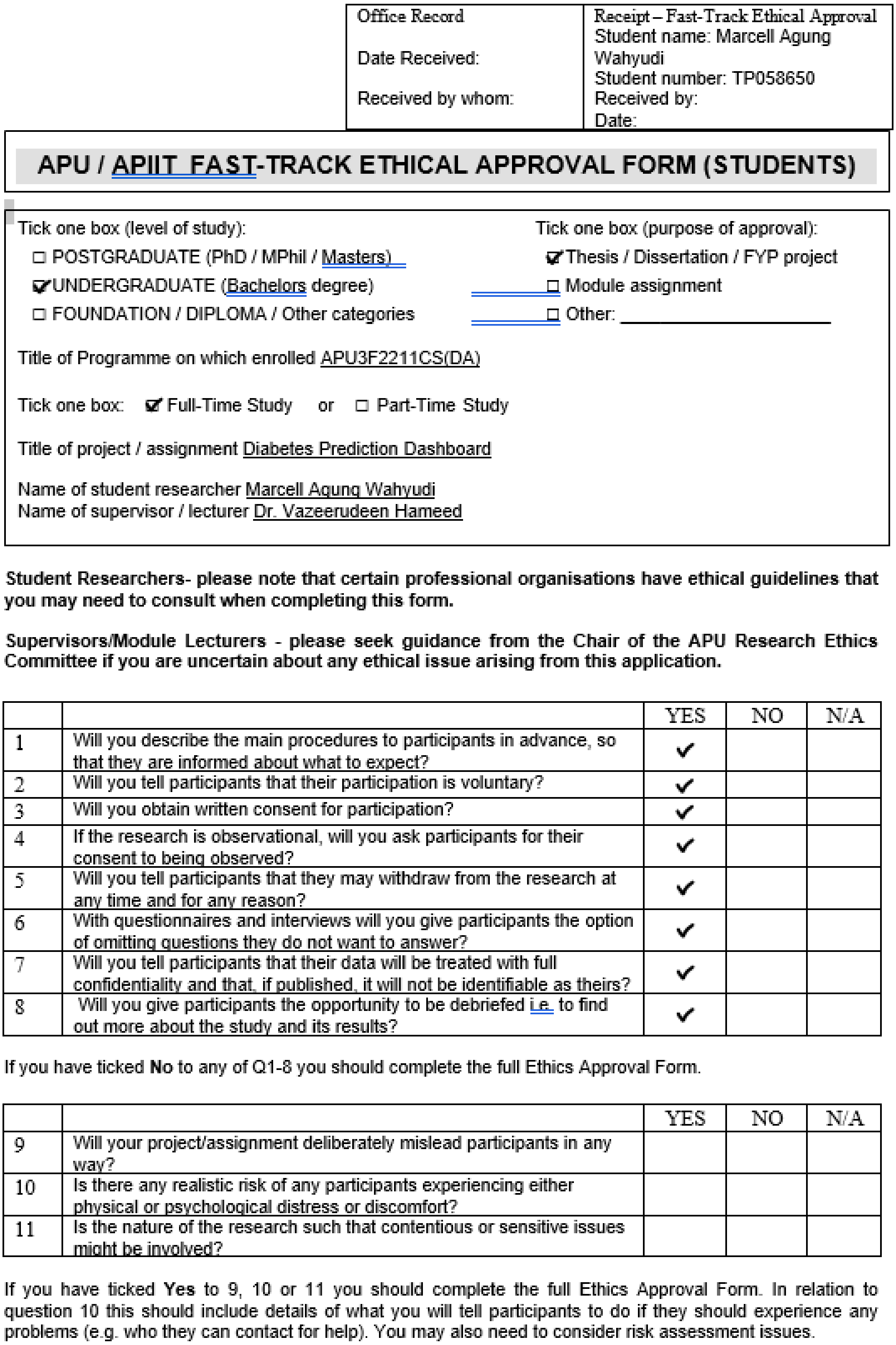
2. The Project is finished and can be proceeded into presentation already

Action List (to be attempted or completed by student by the <u>next</u> mandatory supervisory meeting):

1. IDE have been changed in the document

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Fast-Track Form



Office Record Date Received: Received by whom:	Receipt – Fast-Track Ethical Approval Student name: Marcell Agung Wahyudi Student number: TP058650 Received by: Date:

APU / APIIT FAST-TRACK ETHICAL APPROVAL FORM (STUDENTS)

Tick one box (level of study):

- □ POSTGRADUATE (PhD / MPhil / Masters)
- ☑ UNDERGRADUATE (Bachelors degree)
- □ FOUNDATION / DIPLOMA / Other categories

Tick one box (purpose of approval):

- ☑ Thesis / Dissertation / FYP project
- □ Module assignment
- □ Other: ____________________

Title of Programme on which enrolled APU3F2211CS(DA)

Tick one box: ☑ Full-Time Study or □ Part-Time Study

Title of project / assignment Diabetes Prediction Dashboard

Name of student researcher Marcell Agung Wahyudi
Name of supervisor / lecturer Dr. Vazeerudeen Hameed

Student Researchers- please note that certain professional organisations have ethical guidelines that you may need to consult when completing this form.

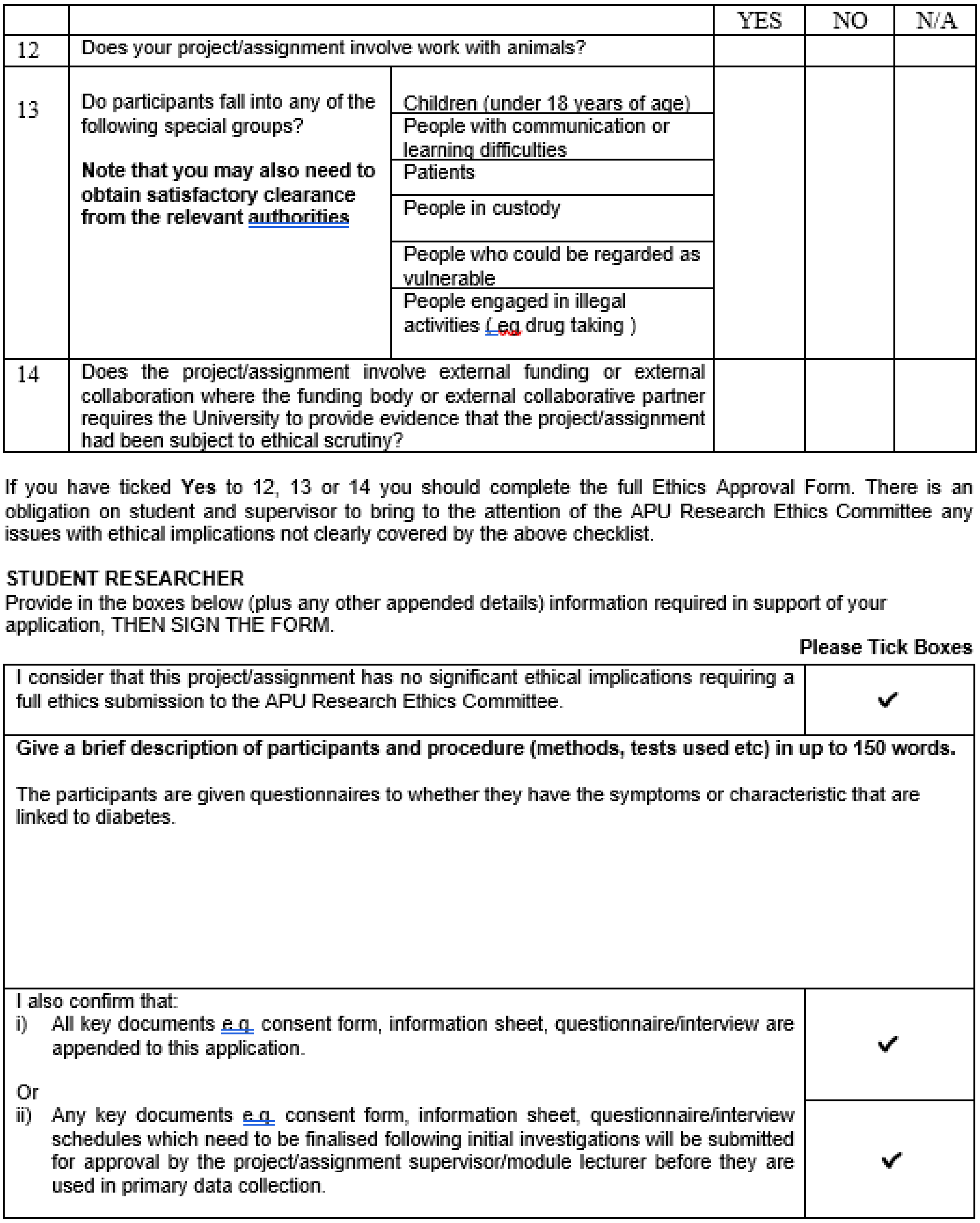
Supervisors/Module Lecturers - please seek guidance from the Chair of the APU Research Ethics Committee if you are uncertain about any ethical issue arising from this application.

		YES	NO	N/A
1	Will you describe the main procedures to participants in advance, so that they are informed about what to expect?	✓		
2	Will you tell participants that their participation is voluntary?	✓		
3	Will you obtain written consent for participation?	✓		
4	If the research is observational, will you ask participants for their consent to being observed?	✓		
5	Will you tell participants that they may withdraw from the research at any time and for any reason?	✓		
6	With questionnaires and interviews will you give participants the option of omitting questions they do not want to answer?	✓		
7	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?	✓		
8	Will you give participants the opportunity to be debriefed i.e. to find out more about the study and its results?	✓		

If you have ticked **No** to any of Q1-8 you should complete the full Ethics Approval Form.

		YES	NO	N/A
9	Will your project/assignment deliberately mislead participants in any way?			
10	Is there any realistic risk of any participants experiencing either physical or psychological distress or discomfort?			
11	Is the nature of the research such that contentious or sensitive issues might be involved?			

If you have ticked **Yes** to 9, 10 or 11 you should complete the full Ethics Approval Form. In relation to question 10 this should include details of what you will tell participants to do if they should experience any problems (e.g. who they can contact for help). You may also need to consider risk assessment issues.



			YES	NO	N/A
12	Does your project/assignment involve work with animals?				
13	Do participants fall into any of the following special groups? **Note that you may also need to obtain satisfactory clearance from the relevant authorities**	Children (under 18 years of age)			
		People with communication or learning difficulties			
		Patients			
		People in custody			
		People who could be regarded as vulnerable			
		People engaged in illegal activities (eg drug taking)			
14	Does the project/assignment involve external funding or external collaboration where the funding body or external collaborative partner requires the University to provide evidence that the project/assignment had been subject to ethical scrutiny?				

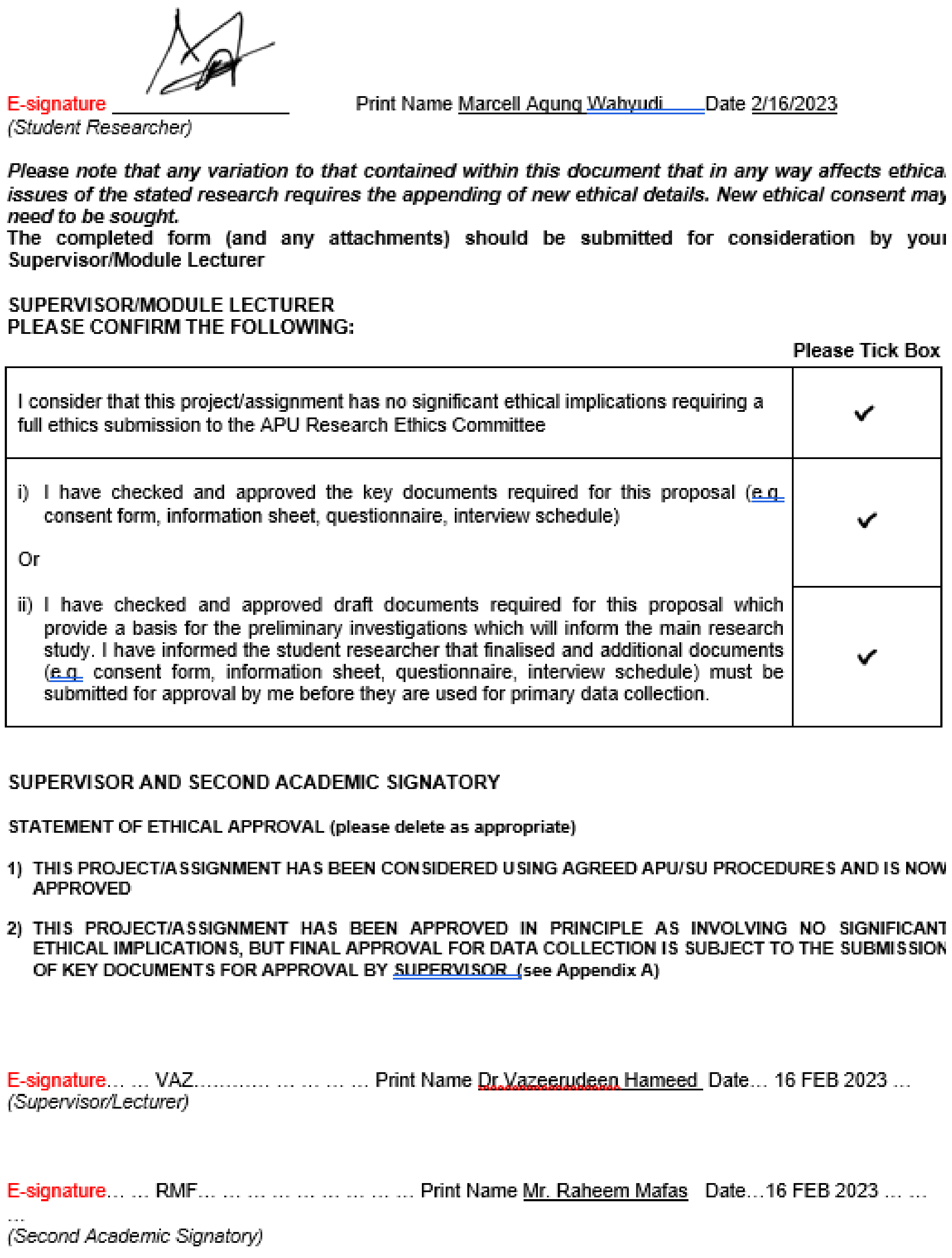
If you have ticked **Yes** to 12, 13 or 14 you should complete the full Ethics Approval Form. There is an obligation on student and supervisor to bring to the attention of the APU Research Ethics Committee any issues with ethical implications not clearly covered by the above checklist.

STUDENT RESEARCHER

Provide in the boxes below (plus any other appended details) information required in support of your application, THEN SIGN THE FORM.

Please Tick Boxes

I consider that this project/assignment has no significant ethical implications requiring a full ethics submission to the APU Research Ethics Committee.	✓
Give a brief description of participants and procedure (methods, tests used etc) in up to 150 words. The participants are given questionnaires to whether they have the symptoms or characteristic that are linked to diabetes.	
I also confirm that: i) All key documents e.g. consent form, information sheet, questionnaire/interview are appended to this application.	✓
Or ii) Any key documents e.g. consent form, information sheet, questionnaire/interview schedules which need to be finalised following initial investigations will be submitted for approval by the project/assignment supervisor/module lecturer before they are used in primary data collection.	✓



E-signature ________________ Print Name Marcell Agung Wahyudi Date 2/16/2023
(Student Researcher)

Please note that any variation to that contained within this document that in any way affects ethica issues of the stated research requires the appending of new ethical details. New ethical consent may need to be sought.
The completed form (and any attachments) should be submitted for consideration by you Supervisor/Module Lecturer

SUPERVISOR/MODULE LECTURER
PLEASE CONFIRM THE FOLLOWING:

Please Tick Box

I consider that this project/assignment has no significant ethical implications requiring a full ethics submission to the APU Research Ethics Committee	✓
i) I have checked and approved the key documents required for this proposal (e.g. consent form, information sheet, questionnaire, interview schedule) Or	✓
ii) I have checked and approved draft documents required for this proposal which provide a basis for the preliminary investigations which will inform the main research study. I have informed the student researcher that finalised and additional documents (e.g. consent form, information sheet, questionnaire, interview schedule) must be submitted for approval by me before they are used for primary data collection.	✓

SUPERVISOR AND SECOND ACADEMIC SIGNATORY

STATEMENT OF ETHICAL APPROVAL (please delete as appropriate)

1) **THIS PROJECT/ASSIGNMENT HAS BEEN CONSIDERED USING AGREED APU/SU PROCEDURES AND IS NOW APPROVED**

2) **THIS PROJECT/ASSIGNMENT HAS BEEN APPROVED IN PRINCIPLE AS INVOLVING NO SIGNIFICANT ETHICAL IMPLICATIONS, BUT FINAL APPROVAL FOR DATA COLLECTION IS SUBJECT TO THE SUBMISSION OF KEY DOCUMENTS FOR APPROVAL BY SUPERVISOR (see Appendix A)**

E-signature… … VAZ.……… … … … … Print Name Dr Vazeerudeen Hameed Date… 16 FEB 2023 …
(Supervisor/Lecturer)

E-signature… … RMF… … … … … … … … … … Print Name Mr. Raheem Mafas Date…16 FEB 2023 … … …
(Second Academic Signatory)

Library Form

Please fill in **all** the following details for library cataloguing purposes.

First Name: Marcell

Middle Name (only if applicable): Agung

Last Name: Wahyudi

Title of the Final Year Project / Dissertation / Thesis :

Diabetes Analysis Dashboard

Abstract :

The project consists of analyzing dataset of Diabetic patients, with variables including patients' test results and characteristics, a dashboard is built from said variables, providing charts and visualizations according to user needs, machine learning models are also built and visualized at the dashboard.

A few keywords associated with the work :

Diabetes, Analysis, Data Visualization.

General Subject: Data Analytics
Date of Submission: 24th of July 2023

DECLARATION OF THESIS CONFIDENTIALITY

Author's full name: **MARCELL AGUNG WAHYUDI**

IC No./Passport No.: **C8092526**

Thesis/Project title: **DIABETES ANALYSIS DASHBOARD**

I declare that this thesis is classified as:

□ CONFIDENTIAL

□ RESTRICTED

☑ OPEN ACCESS

I acknowledged that Asia Pacific University of Technology & Innovation (APU) reserves the right as follows:

1. The thesis is the property of Asia Pacific University of Technology & Innovation (APU).
2. The Library of Asia Pacific University of Technology & Innovation (APU) has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Author's Signature:

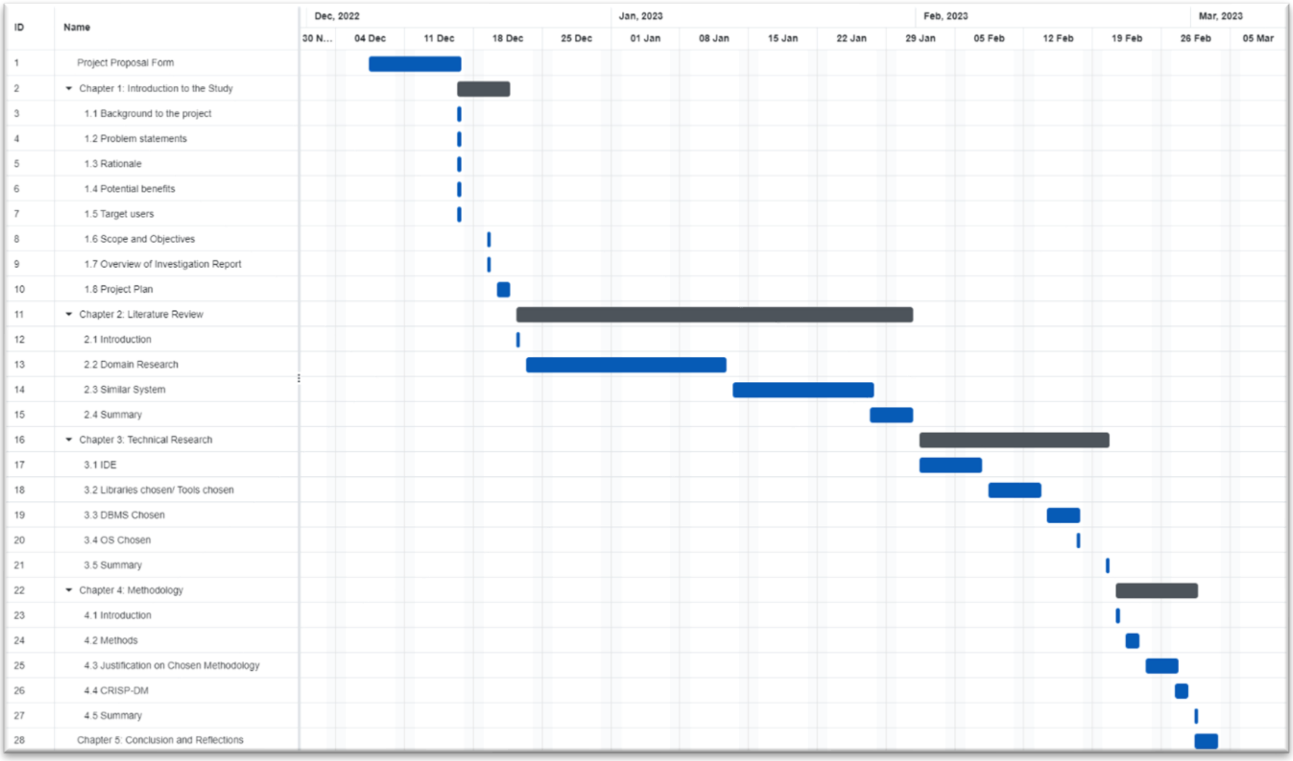
Date: 27 September 2015

Supervisor's Name: **DR. VAZEERUDEEN HAMEED**

Date: 27 September 2015

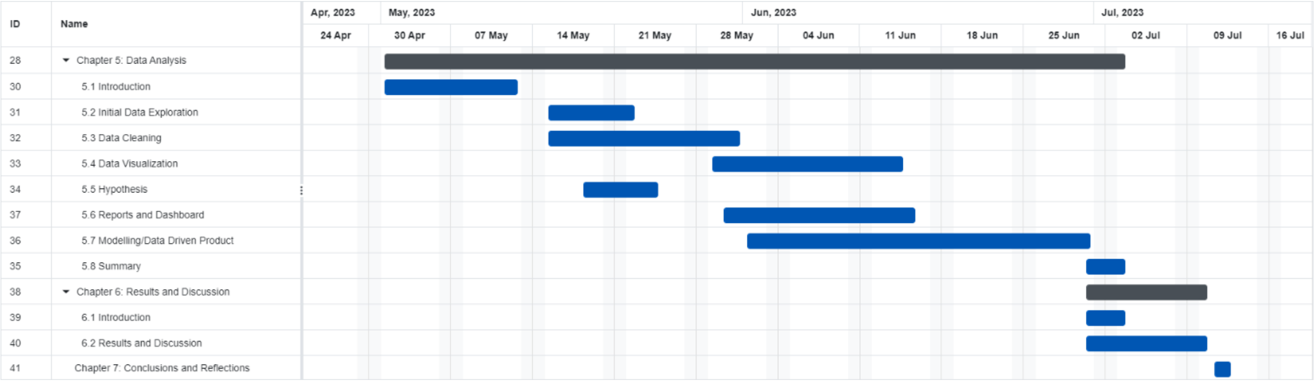
Signature: VAZ

Gantt chart for Investigation Report



ID	Name
1	Project Proposal Form
2	▾ Chapter 1: Introduction to the Study
3	1.1 Background to the project
4	1.2 Problem statements
5	1.3 Rationale
6	1.4 Potential benefits
7	1.5 Target users
8	1.6 Scope and Objectives
9	1.7 Overview of Investigation Report
10	1.8 Project Plan
11	▾ Chapter 2: Literature Review
12	2.1 Introduction
13	2.2 Domain Research
14	2.3 Similar System
15	2.4 Summary
16	▾ Chapter 3: Technical Research
17	3.1 IDE
18	3.2 Libraries chosen/ Tools chosen
19	3.3 DBMS Chosen
20	3.4 OS Chosen
21	3.5 Summary
22	▾ Chapter 4: Methodology
23	4.1 Introduction
24	4.2 Methods
25	4.3 Justification on Chosen Methodology
26	4.4 CRISP-DM
27	4.5 Summary
28	Chapter 5: Conclusion and Reflections

Gantt chart for FYP Documentation



ID	Name
28	Chapter 5: Data Analysis
30	5.1 Introduction
31	5.2 Initial Data Exploration
32	5.3 Data Cleaning
33	5.4 Data Visualization
34	5.5 Hypothesis
37	5.6 Reports and Dashboard
36	5.7 Modelling/Data Driven Product
35	5.8 Summary
38	Chapter 6: Results and Discussion
39	6.1 Introduction
40	6.2 Results and Discussion
41	Chapter 7: Conclusions and Reflections