Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle

By

EDWARD LEONARDO

TP058284

APU3F2211CS(DA)

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

BSc (Hons) in Computer Science Specialism in Data Analytics

Supervised by Dr. Murugananthan Velayutham

2nd Marker: Mr. Justin Gilbert A/L Alexius Silvester

July 2023

Acknowledgement

As my acknowledgement, I would like to give my deepest gratitude to couple of parties that have given their help and support for my Final Year Project (FYP). First and foremost, my FYP supervisor, Dr. Murugananthan Velayutham, who have help me during all the process in my FYP, including help checking my PPF, Ethics Form, Questionnaire, IR, and the FYP, pointing out any mistakes or rooms for improvement that can be done. I also would like to express my gratitude for all the time and attention he has dedicated to help me in this endeavour, And towards my second marker, Mr. Justin Gilbert A/L Alexius Silverster, that also provided some insights and opinions that helped my project. I would like to give an honourable mention to Mr. Dhason Padmakumar, as the FYP Project Manager for the Computing School, for all the detailed briefing about all aspects of the FYP that need to be done.

Next, I would like to show my gratitude towards my family who has supported my whole undergraduate journey, as without them, I would not have got the chance to experience studying and living overseas. All the experiences of studying, living alone, and becoming a more independent person is something that I am truly grateful with, and I would not be able to experience this without their sacrifices.

I would also like to express my gratitude to all the lecturers that have guided me during my undergraduate studies and my classmates who have work collaboratively with me during our learning process. I am also grateful towards all my friends, those at my home country, those I met during my studies, those I met within organisations, and the rest that I met along the way. All the above mentioned have teaches me to do my best and made me to become the best version of myself, and I am truly honoured to have met every single individual.

Finally, I am grateful towards the person that I have become. Throughout my three years studying overseas, away from home, away from friends, and away from my comfort zone. I have grown becoming more mature, more dedicated, more reliable, and more independent. I have learnt a lot throughout my stay in Malaysia, and hopefully all the lessons that I have learned, both in academics and in life, can be useful towards the next stage of my life.

DECLARATION OF THESIS CONFIDENTIALITY

Author's full name: **EDWARD LEONARDO**

IC No./Passport No.: **X994028**

Thesis/Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle

---

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: Edward Leonardo

Date: 18 July 2023

Supervisor's Name: Dr. Murugananthan Velayutham

Date: 18 July 2023

Signature: ……V. Murugananthan……

Figure 1: Confidentiality Document

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

First Name: Edward
Middle Name (only if applicable) :
Last Name: Leonardo
Title of the Final Year Project / Dissertation / Thesis : Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle
Abstract : Coronary Artery Disease (CAD) is a type of Heart Disease that is mainly caused by lifestyle choices. Even though the prevention actions that can be taken to reduce the risk of CAD, most of the population are still clueless regarding the danger of CAD and consequently did not take it seriously. As a result, CAD cases number continue to rise. The conducted project aims to reduce the cases number by creating an early warning detection and an education tool for CAD that can be used by the general population, utilizing Machine Learning and Data Visualization to realize it.
A few keywords associated with the work : Coronary Artery Disease, Heart Disease, Machine Learning, Predictive Modelling, Data Analysis
General Subject: Medical Analysis, Machine Learning, Predictive Modelling, Data Analysis
Date of Submission : 25th July 2023

Figure 2: Library Cataloguing Details

Table of Contents

Chapter 1: Introduction to the Study

1.1 Background to the Project

Heart disease is a type of disease that affects the heart or blood vessels (National Cancer Institute, n.d.). In 2019, before the COVID-19 pandemic started, Heart Disease is the disease with the highest mortality rate worldwide, with 19 million death, or 34 percent of all deaths in 2019 is caused by heart disease. Globally, 1 in 14 people are living with a type of heart disease, and the amount of people that are suffering heart disease has increased by 93 percent from 1990s to 2010s (British Heart Foundation, 2023). Moreover, the global number of deaths from heart diseases is still projected to rise.

From all the type of heart disease, Coronary Heart Disease (CHD), or more commonly known as Coronary Artery Disease (CAD) is the most common type of heart disease. CAD is the only heart disease type that is mainly caused due to poor lifestyle decision, such as smoking, unhealthy diet, excessive alcohol consumption, overweight and obesity, diabetes, and physical inactivity (CDC, 2022b). This unhealthy lifestyle choices can also be interpreted as Personal Key Indicators (PKIs). But the worldwide population did not realize that these PKIs can have the affect if they have the likelihood of heart disease, as most people in their early adulthood still not seriously taking care of their physical health, as majority of this people in the age group still adopt the "You Only Live Once" outlook towards life (Zhou, 2021). And when they start getting older, they will think that its already too late to start a healthy lifestyle, even though that it's as far from the fact, as a healthy lifestyle can be started by any age, with the benefits still valid for all (Johns Hopkins Medicine, 2021). But starting a healthy lifestyle is much easier as a child, as children have easier time to adopt any habits they started, compared to adults (Mock, 2019). This is why increasing awareness of the likelihood of Coronary Artery Disease is very important for any age groups, so more and more people can avoid it during later in life.

Machine Learning, or shorten into ML, is a branch of Artificial Intelligence (AI), where the field is focusing on the usage of data and algorithms so a machine can intelligently solve a problem that it faced. Machine Learning mimic how a human would learn, with the amount and quality of data can gradually have a positive impact to its accuracy (IBM, n.d.). This mean that Machine Learning does not require any explicit programming to operate (Kanade, 2022). The most common use case of Machine Learning is to do prediction, make a classification, or to uncover an insights and patterns in data mining. In a real-life application, Machine Learning is used for many things, such as Recommendation engines, Prediction model, Spam filtering,

malware threat detection, and many more (Burns, 2021). Machine learning has a lot of models with different use cases, that's why the project must correctly pick a suitable machine learning algorithm. The Machine Learning model performance also based on the data it is given, with more quantity of data can help with the training and testing to make sure it has a high accuracy, but the quality of the data also can help increasing the quality of the model itself (Miah, 2017). A good machine learning model will keep learning from any new data to increase its own accuracy. In this case, the data will be fed to a machine learning model to predict if a person have a likelihood to suffer from heart disease.

The aim of this project is to build a heart disease prediction model using machine learning to help the population to test if they have a risk to suffering from heart disease based on their current lifestyle. The prediction will be based on their Personal Key Indicators (PKIs). Other than that, the important PKIs will be studied to give the general masses a better understanding on which lifestyle choices can lead to a risk of heart disease.

1.2 Problem Statements

Heart disease is the disease with the highest mortality rate in the world (British Heart Foundation, 2023). But there are multiple types of heart diseases with different causes, such as Coronary Artery Disease (CAD), Irregular heartbeats (Arrhythmias), Birth Defects Heart Problem (Congenital Heart Defects), Heart muscle diseases, and Heart Valve Disease (Mayo Clinic, 2022c). From all these heart disease types, the only one that is based on a person's lifestyle is Coronary Artery Disease (CAD), whereas the other heart diseases are more commonly caused by other factors, such as birth defect, genetical inheritance, or an infection. CAD is a heart disease that is caused by a blockage of heart arteries, causing not enough oxygen-rich blood can reach the heart. The blockage can be caused by multiple factors, such as inflammation of the arteries, or more commonly is due to cholesterol deposits/fatty plaques clogging the arteries. CAD is a type of disease where it develops over decades, as it is very much based on a person's lifestyle (Mayo Clinic, 2022b).

The main problem that even though the cause of CAD is very much easily avoidable, widespread population is mainly ignorant about the issue, as 70 percent of them mistakenly assumed that there will be a warning sign beforehand (Reuters Staff, 2008). Even though CAD is more commonly found in older population, since their arteries are already narrowed and damaged due to their age (Mayo Clinic, 2022b). CAD is a disease that can happen at any age (CDC, 2021b), where people with high obesity and high blood pressure are in a greater risk of suffering from the disease earlier in life. Now considering that because of the COVID-19 Pandemic, research done by Hu et al. (2020) has shown that an increase in unhealthy lifestyle has been found due to the pandemic restriction, causing more people to stay at home and do fewer physical activities. This means that more and more people will have a risk to suffer from CAD soon.

Another problem is the lack of self-diagnosis alternatives to check if a person has a risk of suffering from CAD. Usually, most people only do a check-up after they feel the symptoms of CAD, and by that time, doing some of the preventive methods can be very challenging or impossible, since creating a new habit will need to take some time (Clear, 2020). There are also those who need to go to a clinic to do some diagnostic tests before they fully realized that they may have a risk of suffering from this disease, where they may have been ignorant towards any early symptoms. Moreover, the price of these tests is also not cheap, which make those who are struggling financially to missed out in these check-ups (Yetman, 2022).

1.3 Rationale

Identification of a Coronary Artery Disease (CAD) is not a hard thing to do, as a normal blood test, where checking certain elements, such as cholesterol level, in a person blood can already determine the risk of a CAD (Cleveland Clinic, 2022). Another way, although more extreme, is called Coronary Angiography, or also known as Coronary Catheterization, is known as a gold standard in CAD diagnosis, where a person under local anaesthesia, will be injected with a contrast medium, usually X-ray dye, via a catheter, then an X-ray camera will film the location and severity of an arteries blockage (UCSF, n.d.). With all the options mentioned, someone can easily get a diagnosis if they have a risk to suffer from CAD. But all these options require a medical professional help, whereas any easier method is still not available. This creates a gap where a person just wants to check if they have a likelihood of suffering from heart disease in the future, they do not have any easier or simpler methods, where based on their current lifestyle, they can already make the prediction. Then after the prediction, they could proceed with further medical help if they deemed is necessary.

Another thing is a lack of awareness from the population, which means there is no downward trend of the amount of people suffering from CAD. Even though most people understand the risk and danger of CAD, they are still clueless regarding on how to prevent suffering from the disease. Increasing awareness towards general population is one of the most effective ways of suppressing the trend of CAD, as encouraging a healthier lifestyle globally can reduce the risk of suffering from CAD. This means quitting smoking, reducing alcohol consumption, be more physically active, and maintaining a healthy weight. Other than reducing the risk of CAD, all lifestyle choices mentioned also have other great benefits, as being healthier in general, reducing the risk of other diseases, such as cancer (CDC, 2020), diabetes (CDC, 2022c) , and liver disease (CDC, 2022a).

1.4 Potential Benefits

1.4.1 Tangible Benefits

- Increasing the awareness of the population regarding on what lifestyle choices that can have a significant effect on the likelihood of heart disease.
- Reduce the amount of misconception regarding heart disease, especially the impact of personal lifestyle can have to the likelihood of heart disease.
- Efficient and reliable method of health self-diagnosis based on the Personal Key Indicators.

1.4.2 Intangible Benefits

- Promote healthy lifestyle towards the population to reduce the likelihood of suffering from heart disease.
- Reduce the amount of people suffering of heart disease.
- Become a Corrective Action guide for people that already suffered heart disease on which aspects of their lifestyle that contribute most towards heart disease.

1.5 Target Users

As the nature of the project focuses on which lifestyle choices contribute the most towards the likelihood of heart disease, the targeted users of this project will be mainly the common population, either those who have not suffered or already suffered from heart disease, as the result of the project can be used both as a preventive and a corrective action. Another target users are medical professionals since the information from this project can be used to as an educational tool.

1.6 Scope and Objectives

1.6.1 Aim

The aim of this project is to increase the awareness of the population on what Personal Key Indicators or lifestyle habits that can leads to heart disease, and how to do self-diagnosis based on those indicators using the machine learning and data analysis.

1.6.2 Objectives

1. To conduct an Exploratory Data Analysis (EDA) to identify and hidden correlations and trends of what Personal Key Indicators can cause heart disease.
2. To develop a predictive model using Machine Learning to predict if a person has a risk of suffering a heart disease based on their Personal Key Indicators.
3. To create an educational tool that can be used by the general population that covers both previous objectives.

1.6.3 Deliverables – Functionality of the proposed system

The key deliverables are stated as:

- A predictive machine learning model to predict if a person has a likelihood of suffering heart disease based on their personal lifestyle choices.
- A Functional Web Application that user can uses to predict their condition using the predictive machine learning model created.
- An Exploratory Data Analysis (EDA) result, which can be converted into an educational/informational tool that can be shared with the population.

1.6.4 Nature of Challenges

One of the challenges in executing this project is to select which programming language to be used for the data analysis and the prediction model, since there are multiple programming languages that support this task, a comparison between them should be done to determine which language will be used. Another challenge is the dataset used for the project, since the dataset must come from a reliable source, with good results and relevant information. The dataset should also be large in quantities to increase the number of variances of data, thus increasing the accuracy of the model itself. The developer needs to pick a dataset that matches all the criteria above to make sure the project outcome can be reached.

1.7 Overview of the Report

This Project will be divided into seven chapters. Chapter 1 is Introduction to the Study, where it gives focus on the background of the project, the problem that the project wants to solve, the rationale of why the project is important, the potential benefits the project could bring, the target users of the project, the scope, objectives, deliverables, and the challenges of the project. The first chapter will also include the Project Plan, where it shows a graph of the duration the project has taken to be completed. Chapter 2 is the literature review part of the project, where it gives a more in-depth explanation regarding the project, in this case about Coronary Artery Disease (CAD), what is the cause of CAD, and the complication that can happened because of CAD. It will also discuss about Machine Learning, and its various algorithm that can be used for the project. Another thing is a discussion regarding any similar systems that has been developed by another developer(s). Chapter 3 will be focused on the Technical Review, which give explanation regarding the technical aspects of the project. In this case, it will discuss in depth regarding what Programming Language, IDE, Operating Systems, and any library that is suitable, and can be used for the project. Chapter 4 is about the system development methodology used to develop the project, where it will give out the reasons and justifications on the chosen system development methodology. Chapter 5 is about Research Methodology, where it discusses about which research methods is best used to do data gathering for the project, as well as to understand the importance of the project. The data collected from the research will be analysed and translated in Chapter 6, where it will give validation regarding the problems that is the project is trying to tackle. Next, Chapter 7 will tackle the problems through data analysis, resulting in data visualizations and Machine Learning models created, utilizing the dataset that has been gathered. Chapter 8 will discuss the result of the Machine Learning model, how it performs compared to others' works, and the deployment of the results through Web Application. Finally, Chapter 9 will be a conclusion of the Project, where it gives out the summary and reflections regarding the whole project.

1.8 Project Plan

Final Year Project				
Task name	**Duration (Day)**	**Start Date**	**End Date**	**Status**
Chapter 1: Introduction to the Study	3	8th Feb	10th Feb	Done
1.1 Background to the project	1	8th Feb	8th Feb	Done
1.2 Problem Statements	1	9th Feb	9th Feb	Done
1.3 Rationale	1	9th Feb	9th Feb	Done
1.4 Potential Benefits	1	9th Feb	9th Feb	Done
1.5 Target Users	1	10th Feb	10th Feb	Done
1.6 Scope and Objectives	1	10th Feb	10th Feb	Done
1.7 Overview of the Report	1	10th Feb	10th Feb	Done
1.8 Project Plan	1	10th Feb	10th Feb	Done
Chapter 2: Literature Review	13	11th Feb	23rd Feb	Done
2.1 Introduction	1	11th Feb	11th Feb	Done
2.2 Domain Research	8	11th Feb	18th Feb	Done
2.3 Similar System(s)	5	18th Feb	22nd Feb	Done
2.4 Summary	1	23rd Feb	23rd Feb	Done
Chapter 3: Technical Research	4	23rd Feb	26th Feb	Done
3.1 Programming Language Chosen	2	23rd Feb	24th Feb	Done
3.2 IDE (Interactive Development Environment) Chosen	1	24th Feb	24th Feb	Done
3.3 Libraries Chosen	1	25th Feb	25th Feb	Done
3.4 Operating System Chosen	1	25th Feb	25th Feb	Done
3.5 Summary	1	26th Feb	26th Feb	Done
Chapter 4: Methodology	4	26th Feb	1st March	Done
4.1 Introduction	1	26th Feb	26th Feb	Done
4.2 Methods	3	27th Feb	1st March	Done
4.3 Summary	1	1st March	1st March	Done
Chapter 5: Research Methodology	8	2nd March	9th March	Done
5.1 Introduction	1	2nd March	2nd March	Done
5.2 Questionnaire Survey	1	2nd March	2nd March	Done

5.3 Design	1	9th March	9th March	Done
5.4 Summary	1	9th March	9th March	Done
Chapter 6: Requirement Validation	2	12th March	13th March	Done
6.1 Introduction	1	12th March	12th March	Done
6.2 Analysis of Questionnaire Data	1	13th March	13th March	Done
6.3 Summary	1	13th March	13th March	Done
Chapter 7: Data Analysis	32	16th June	18th July	Done
7.1 Introduction	1	16th June	16th June	Done
7.2 Initial Data Exploration	1	16th June	16th June	Done
7.3 Data Cleaning	5	17th June	21st June	Done
7.4 Data Visualization	3	22nd June	24th June	Done
7.5 Model Building	21	24th June	15th July	Done
7.6 Summary	1	18th July	18th July	Done
Chapter 8: Results and Discussion	3	16th July	18th July	Done
8.1 Introduction	1	16th July	16th July	Done
8.2 Model Building Results	2	16th July	17th July	Done
8.3 Evaluation on Model Building Results	1	18th July	18th July	Done
8.4 Deployment Results	1	18th July	18th July	Done
Chapter 9: Conclusion and Reflection	1	18th July	18th July	Done

Table 1: Project Plan

Chapter 2: Literature Review

2.1 Introduction

A literature review is research-based analysis to show a level of understanding of the current researchers based on the academic literature about a certain topic (Institute for Academic Development, 2022). Literature review act as a foundation of the investigation, where it provides the readers a solid background regarding the topic that the investigation wants to highlight. Literature review also can act as a support for the new insight that the researchers contributing towards the topic (UNC-Chapel Hill Writing Center, 2021). In this chapter, any past findings and research that has been done towards the topic will be studied and explored. It is integral to gain a better understanding of past findings and similar systems that has been developed by past researchers, which can help to improve the current project's quality.

2.2 Domain Research

2.2.1 Coronary Artery Disease

Coronary Artery Disease, or shorten as CAD, is the most common heart disease in the world, with 20.1 million adults (age 20 years old and older) globally has CAD (CDC, 2022d). CAD is also known as Coronary Heart Disease (CHD) and Ischemic Heart Disease (IHD) (CDC, 2021b). This heart disease is caused by a build-up of plaque in the arteries' walls that supply blood to the heart. The plaque that blocked the arteries consists of cholesterol and other substances, which make it harder and harder for the blood to reach the heart. Over time, this blockage can partially or totally block the blood flow, which is a process called Atherosclerosis. If there are too many plaques built-up, It will make the arteries to become narrower that it supposed to be. it will cause Angina, otherwise known as chest pain or discomfort, which is the most common symptoms of CAD. For most people, Angina will be their first clue of CAD. Other symptoms of CAD are shortness of breath, fatigue, and heart attack. People usually unrecognized these symptoms until the signs become more severe and/or frequents overtime (Mayo Clinic, 2022). CAD itself can be the cause to other diseases, which in medical fields called complications (National Institute of Health, 2023). The main complication from CAD is heart failure, which is the condition where the heart cannot pump the blood as usual (CDC, 2021b). Other complications include chest pain (angina), heart attack, and irregular heart rhythms (arrhythmias). From the above explanation, it can be summarized that CAD is a very dangerous disease, which if left uncared for can lead to many fatal diseases.

The risk factors that can caused CAD are unhealthy lifestyle choices. These key risk factors include high blood pressure, high blood cholesterol, and smoking. This key risk factors itself mainly caused by diabetes, obesity, unhealthy diet, physical inactivity, smoking, stress or bad mental health, and excessive alcohol consumption. Another risk factor is family history, such as close relatives having a CAD themselves can influence an individual's risk of developing CAD themselves (Mayo Clinic, 2022b). This means that CAD itself is a heart disease that is very individual based, since how much the risk factors increase is dependent on how a person's lifestyle is.

The main way to combat CAD is to have a healthier lifestyle habit. A healthy lifestyle can naturally keep the arteries strong and reduce the amount of plaque, or even clearing any plaque. This solution works both as prevention and treatment for CAD (Mayo Clinic, 2022b), but someone with an existing CAD should consult to a medical professional before doing any of these activities, as some activities can be dangerous if done excessively by someone with a pre-

existing condition. According to Watson (2020), there are multiple ways to prevent CAD naturally. The first one is to have a heart-healthy diet, this means reducing the amount of saturated fat, sugar, and salt in foods. Saturated fat is the type of fat that increase the levels of bad cholesterol, medically known as Low-Density Lipoprotein (LDL) Cholesterol, in turn increasing the number of plaques in the arteries (CDC, 2022e). Saturated fats usually come from animal sourced food, such as meat and dairy products. Another source of saturated fat is tropical food source, such as coconut and palm (American Heart Association, 2021). A substitute can be found in unsaturated fat, which increases the good cholesterol, medically known as High-Density Lipoprotein (HDL) Cholesterol, in turn help reduce any arteries blockage by absorbing cholesterol in the blood which then will be flushed out from the body (CDC, 2022e). The example of this is vegetable oils, nuts and seeds, and avocados. The next restriction is sugar, where too much sugar is the main culprit of a person developing diabetes, which itself one of the main causes of CAD. The last restriction is salt or sodium, as too much salt can raise blood pressure, thus developing into high blood pressure. High blood pressure will put an excessive force on the artery wall, which overtime will damage it, causing Atherosclerosis. Another healthy lifestyle habit is to increase physical fitness and losing weight. Exercising, either both aerobic or anaerobic type, can improve health and reduce the risk of cardiovascular diseases (Stinchcombe, 2022). CAD can be identified as a cardiovascular disease, since inactivity can be the leading cause of the disease (World Health Organization, 2019). Exercising can also reduce fat, lowers blood pressure, and increases HDL cholesterol. Next, weight loss is also an effective way to reduce the risk of CAD, since excess weight put extra tension on the heart and arteries. Decreasing weight can lower the blood pressure and LDL cholesterol. Weight loss itself can be achieved by following a healthy diet and increasing physical activities amount. An individual ideal weight may vary, but by using Body Mass Index (BMI) calculation, a person can target a healthy weight depending on their height and age. Normal weight will be between 18.5 - 25 kg/m^2 (Kilogram per Meter squared) (Calculator, n.d.). Another way to reduce the risk of CAD is limiting alcohol consumption and stop smoking. Cigarettes have a lot of chemical substances that can clogs the arteries and damages the heart, thus increasing the risk of Atherosclerosis, ends up resulting in CAD (Nunez, 2021). Alcohol consumption also influence the risk factors of CAD. Even though a moderate red wine consumption can have health benefits, due to its antioxidant, anti-inflammatory, and lipid-regulating effects can help the arteries (Smith, 2023). But excessive amount of alcohol can also contribute towards high blood pressure and obesity. Other miscellaneous ways of reducing the risk factors can be by reducing stress level and getting enough sleep.

There are also medication alternatives to combat these problems. If lifestyle changes are not enough, or if an individual have a medical restriction that blocks them from these changes, medication can also help in lowering cholesterol, lowering blood pressure, and clot-preventing. These types of drugs are usually doctor-prescribed, so a consultation with a doctor is required to get the suitable drug depending on each individual restriction and condition (Watson, 2020).

2.2.2 Coronary Artery Disease Diagnostic Tool

Diagnostic Tool can be defined as tools or equipment developed and used to identify patients' disease based on the results from said tools (The Free Dictionary, 2023). Based on Shahjehan and Bhutta (2022) works, there are multiple ways to diagnose Coronary Artery Disease (CAD).

The first way that will be discussed is Electrocardiogram (EKG). It is very basic yet helpful test, by measuring electrical activity in the cardiac conduction system (heart) using 10 leads attached to a predetermined locations at the skin, and when activated will gives out information about the condition of the heart. It will print the results on a piece of paper, where it will show the heart's rate, rhythm, and axis. Based on these results, a diagnosis can be pulled if the person has a CAD. EKG is a cost-effective and widely available.

The second method is called Echocardiography, which is using ultrasound to the heart. It is non-invasive method where an ultrasound scan of the heart will be captured, similar to pregnancy ultrasound. This method can show a great detail of the patient's heart condition but is considerably more costly than EKG.

The next test is called Stress Test, which is a relatively non-invasive test to evaluate CAD. Its function is to check if a person has a CAD based on any suspected Angina (chest pain). This means this test is dependent on the patient already facing one symptom of a CAD. During the test, the heart will be exposed to stress while being monitored using an EKG device. An EKG will be taken before, during, and after the stress test to check any anomalies. In this test, the patient will be running on a treadmill until they achieved 85% of age-predicted maximal heart rate. Age-predicted Maximal Heart Rate (HR Max) equation is 220 subtracted by the person's age (WebMD Editorial Contributors, 2022). If the patient develops any symptoms, the test will be stopped, and a CAD will be diagnosed.

Another method of CAD diagnosis is using Chest X-Ray, which is one of the first diagnostic tool of CAD available. This method uses X-Ray to produce film images of a patient chest, which will show their heart. The usual procedure is Standing Posteroanterior (PA, X-ray Imaging from the patient's back) (Merriam-Webster, 2023b), left lateral decubitus (the patient

lying on their side), or sometimes using anteroposterior (AP, the patient is lying down) (Merriam-Webster, 2023a) positions. The images then will be examined by a medical professional to do the diagnosis.

The most common method of diagnosis is Blood Work, which is usually the establishing diagnosis. The blood work will test a patient's blood to check on their blood's substances. Some of the most common blood works used to diagnose CAD is Cholesterol test, High-sensitivity C-reactive Protein test, and Troponin T test. Cholesterol test will check the Total Cholesterol and Low-density lipoprotein (LDL or the bad) and High-density lipoprotein (HDL or the good) cholesterols. High sensitivity C-reactive protein (CRP) test will check the mentioned protein amount in the blood, as CRP is the type of protein produced by the body as the response to injury or infection. Higher CRP levels are associated with risk of cardiovascular disease, including CAD. The final blood work test is Troponin T test, which will check the amount of Troponin T inside the blood. A high level of Troponin T has been associated with higher risk of heart disease, even though the patient has not experienced any symptoms (Mayo Clinic, 2022a).

The final method and the most accurate one is using Cardiac Catheterization. It is the gold standard method and the most accurate technique to assess CAD. It is an invasive procedure that can cause complications such as allergic reactions or a kidney injury. The method is done using local anaesthesia, where the patient will be injected with a contrast medium, usually X-ray dye via catheters tube. Then with X-ray camera films, the blood flow will show the location and the severity of Atherosclerosis in the patient.

2.2.3 Machine Learning in Medical Field

Machine Learning, or shorten as ML, is a branch of Artificial Intelligence (AI) field, where a computer can perform a complex task without being explicitly programmed (Burns, 2021). A Machine learning model/algorithm uses data to increase its own accuracy and effectiveness level. Machine learning most common usage is as a recommendation engine, business process automation, spam filtering, malware threat detection, and outcome prediction. In this modern era, Machine Learning has many applications in the real world, spread across different fields of study.

According to Sidey-Gibbons and Sidey-Gibbons (2019), the main application of ML in the medical field are for diagnosis and outcome prediction. One of the examples of this application is identifying risks of a disease, which can help to early diagnosis or prediction based on the

risk factors can give a person a chance of taking preventive methods of their diagnosed disease. In their case, using the data set from University of California Irvine (UCI), they accomplished to detect breast cancer based on the characteristics and features of cell nuclei taken from breast masses using the R language. They are using three ML algorithms, which are Logistic Regression using Generalised Linear Models (GLM), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). From their results, the three algorithms have an accuracy of .94 – .96, sensitivity .97 – .99, and specificity .85 – .94. From all three algorithms, SVM algorithms managed to produce the best accuracy compared with the others with .96 maximum accuracy.

Another study also has been done by Nandal et al. (2022). Using four different ML models, which are XGBoost, SVM, Naïve Bayes, and Logistic Regression to predict heart attack based on risk factors of heart disease. Based on their results, XGBoost classifier provided the best training and test scores, with .91 and .89 results, with .92 accuracy, with SVM providing the worst results, with .69 training and .64 test score.

Another case of Heart disease prediction using Machine Learning has been covered by Garg et al. (2021). Their prediction model utilizes attributes that need previous laboratory tests, such as cholesterol level. They chose two supervised ML algorithms to used, which are K-Nearest Neighbor (K-NN) and Random Forest. From their result, Both of algorithms got a good prediction accuracy percentage, with K-NN gets 86.885% and Random Forest got 81.967% accuracy.

Finally, Another research by Pal et al. (2022) is focusing on Cardiovascular Disease (CVD) risk prediction using ML algorithms. They used two techniques, Multi-layer perceptron (MLP) and K-nearest neighbour (K-NN). From their results, K-NN got an accuracy score of 73.77% with Area under Curve (AUC) 86.21% score. On the other hands, MLP got both higher accuracy and AUC score, with 82.47% and 86.41% respectively.

From all studies mentioned above, a note that all models' accuracy can be increased overtime. This means by using other data attributes, increasing the training and testing data amount and quality, improving data pre-processing, algorithm parameter tuning, and using other algorithms can have an effect towards improving the model accuracy to predict the outcome (Ray, 2020). Other studies and research will also be compared in the similar systems section.

2.3 Similar System(s)

Based on all past research that has been done on the topic of predicting heart diseases, especially Coronary Artery Disease (CAD) using Machine Learning Algorithms, a comparison table can be made to show the differences in results and algorithms used between each study.

Authors	Research Title	ML Algorithms	Results
Sidey-Gibbons, J. a. M., & Sidey-Gibbons, C. J. (2019)	Machine learning in medicine: a practical introduction	• Generalised Linear Models (GLM) • Support Vector Machine (SVM) • Artificial Neural Network (ANN)	Support Vector Machine has the best result, with 96% Accuracy with Area Under the Curve (AUC) 97%
Nandal, N., Goel, L., & Tanwar, R. (2022)	Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis	• XGBoost • SVM • Naïve Bayes • Logistic Regression	XGBoost provide the best results on both training and test, with 91% and 89% accuracy respectively
Garg, A., Sharma, B., & Khan, R. (2021)	Heart disease prediction using machine learning techniques	• K-Nearest Neighbor (K-NN) • Random Forest	Both Models produce good results, with K-NN getting 86.885% and Random Forest getting 81.967% accuracy.

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. (2022)	An artificial intelligence model for heart disease detection using machine learning algorithms	• Random Forest	Random Forest produce a result of approximately 83% accuracy over the training data
Nagavelli, U., Samanta, D., & Chakraborty, P. (2022)	Machine Learning Technology-Based Heart Disease Detection Models	• Naïve Bayes Weighted • 2 SVM's and XGBoost • SVM and Duality Optimization (DO) • XGBoost	• Naïve Bayes Weighted get 86% Accuracy • 2 SVM's and XGBoost get 94% Accuracy • SVM and DO get 89.4% Accuracy • XGBoost get 95.9% Accuracy
Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022)	Risk prediction of cardiovascular disease using machine learning classifiers	• K-Nearest Neighbor • Multi-Layer perceptron (MLP)	• K-NN get 73.77% accuracy with AUC of 86.21% • MLP get 82.47% accuracy with 86.41% AUC

Hossen, M. (2022)	Heart Disease Prediction Using Machine Learning Techniques	• Logistic Regression • Support Vector Machine • K-Nearest Neighbor • Random Forest • Gradient Boosting Classifier (GBC)	• Logistic Regression get 95% accuracy • SVM get 90% accuracy • K-NN get 87% accuracy • Random Forest get 79% accuracy • Gradient Boosting Classifier get 80% accuracy
Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021)	Heart disease prediction using machine learning algorithms	• K-Nearest Neighbor • Logistic Regression • Random Forest	• The algorithm gets average accuracy of 87.5%

Table 2: Similar Systems Comparsion

2.4 Summary

From the domain research above, It explain what Coronary Heart Disease (CAD) is, what its causes, symptoms, and how to prevent or treat it. Findings about all currently existing tools and tests to help diagnose CAD and regarding Machine Learning application in Medical Field, especially in Cardiovascular Disease that have been done. From all these findings, a Conclusion can be made that all CAD is a very dangerous disease, but because it develops in a very long time, it can very much be determined by a change in someone lifestyle habits. Currently, CAD diagnostic tools are still very much dependant on the medical professionals expertise. In hope to make self-diagnosis of CAD become easier for all population to use, A Machine Learning Algorithms can be utilized to predict if someone has a risk of developing CAD based on their Personal Key Indicators or their lifestyle habits. From all past similar studies that has been done, a pattern of recurring algorithms can be seen, such as Support Vector Machine, Random Forest, and K-Nearest Neighbor being used by different teams.

Chapter 3: Technical Research

3.1 Programming Language Chosen

3.1.1 Programming Language Comparison

Programming language have an essential role in developing this project. There are many programming languages, with different characteristics, purposes, and level of supports from the developers or communities. Keeping that in mind, choosing a suitable programming language for the project helps the developer to save on time, cost, and resources, without sacrificing on the project's results quality (Dhruv, 2019). Since this project is a data analytics-based project, a comparison between programming languages that are popular and most widely used in the data analytics space will be done, to figure out which one is the most suitable for this project. The table below will compare between three programming language, Python, R, and SAS, comparing it with different aspects in mind.

Aspects	Python	R	SAS
Background	Python was created by Guido van Rossum on February 20, 1991. Currently It is being maintained by Python Software Foundation (Python Institute, n.d.).	R was developed by Ross Ihaka and Robert Gentleman at University of Auckland in 1991. It is based on the S language from 1976 (Kumar, 2021).	SAS was created in 1971, and in 1976, the SAS Institute Inc. released the SAS software as it first products (SAS, 2021a).
Purpose	Python can be used for : • Data Analysis • Data Visualisation • Machine Learning • Back-end Web App Development • Automation/Scripting • Software Testing and Prototyping (Coursera, 2022b)	R is mainly used for: • Data Analysis • Data Visualisation • Statistical Analysis • Machine Learning (Simplilearn, 2022)	SAS is used for: • Data Analysis • Business Intelligence • Data Visualisation • Statistical Analysis • Machine Learning (SAS, 2021c)

Market Presence	Python is regarded as the most popular programming language for Data Analysis (Narang, 2023). Since Python is also designed for other tasks, Python become one of the most versatile programming language available (Luna, 2022).	R is a Programming language more focused on Data Analytics and visualisation. In the Data Analytics and Science field, R is considered one of the best programming languages (Luna, 2022).	Since SAS require a license to use, SAS is more popular in large business (Narang, 2023).
IDE	• IDLE • PyCharm • Visual Studio Code • Sublime Text 3 • Jupyter Notebook (Programiz, n.d.)	• RStudio • JupyterLab • R Tools for Visual Studio • PyCharm • Eclipse StatET (Chia, 2023)	• SAS Studio (SAS, 2021b)
Cost	Python is an open-source language and free for everyone (Jain, 2017).	R is an open-source language and free for everyone (Jain, 2017).	SAS is a commercial software, meaning its closed-source and can be very expensive for individual, thus mainly used by private organizations (Jain, 2017).
Difficulty	Python is a high-level programming language, and famously known for its syntax simplicity and ease of use (Jain, 2017).	R has the steepest learning curve. It is a low-level programming language, making a simple task requiring a more complex code (Jain, 2017).	SAS is considered the easiest to use since it has GUI interface and drag-and-drop feature. It also uses PROC SQL language for manual coding (Jain, 2017).

Accessibility	Python is open-source software, meaning new features will be available for free, with community support creating custom libraries (Shahid, 2021).	R is open-source software, meaning new features will be available for free, with community support creating custom libraries (Shahid, 2021).	SAS is a closed-source software, meaning the users need to purchase new license or software for any new features (Shahid, 2021).
Data Visualization	Python has libraries such as Plotly, matplotlib, and seaborn for visualization, but Python's visualization is not as strong as R's visualization capabilities (Khete, 2022).	R has a famous library called ggplot2, which have the capabilities to create all types of charts for data visualization, making R the best programming language with visualization capability (Khete, 2022).	SAS has functions to create graphical visualization, but merely a functional one. To do any customization on the plots, it requires a deep understanding of SAS Graph Package (Jain, 2017).
Machine Learning	For Machine Learning, Python has libraries such as: • NumPy • Pandas • Scikit-learns • TensorFlow • PyTorch • Matplotlib • Keras • Seaborn (Coursera, 2022a)	For Machine Learning, R has libraries such as: • Caret • DataExplorer • Dplyr • Ggplot2 • kernLab • Plotly • randomForest • rpart (Choudhury, 2020)	For Machine Learning, SAS utilize different software, such as: • SAS Viya • SAS Enterprise Miner (Lawson, 2021)

Deep Learning	Python has some deep learning libraries: • TensorFlow • Keras (Jain, 2017)	R has some deep learning libraries: • KerasR • Keras These libraries are imported version of the Python package. (Jain, 2017)	Currently, SAS still have not any support for Deep Learning task, as it still under development phase. (Jain, 2017)

Table 3: Comparison of Python, R, and SAS

3.1.2 Justification on Programming Language Chosen

Based on the comparison on the previous segment, Python is the most suitable programming language to be used in this project. The first reason of this pick is Python is widely used for data analysis and machine learning tasks. Python also has a lot of powerful libraries that can be used for this project. This means that the developer can use the libraries to improve the work efficiency and performance of the project. The next reason is Python is also a freeware and open-source software, which means there no cost required for the project development. Python is also a high-level programming language, meaning it is easier to understand the code written. Moreover, Python is also famous for its syntax being very simple, increasing the readability of the codes, especially for beginners learning coding. Another reason why Python is picked is because of its strong communities presences. Python has a large, active, and supportive communities spread across the internet that help each other if an individual is facing a problem. This means the developer can also gain help/knowledge from these communities, especially from previous questions threads that has been solved by communities, such as Stack Overflow.

3.2 IDE (Interactive Development Environment) Chosen

An Integrated Development Environment, or commonly shorten as IDE, is a software that helps developer to build application, with built-in tools in one application (Red Hat, 2019). An IDE can increase a developer productivity with its built-in tools and enhancements, such as editing source code, building executables, syntax highlighting, autocomplete, suggestions, and debugging (Codecademy Team, n.d.). It is vital for the developer to pick the correct IDE for the project, as IDEs also built for different tasks in mind.

For Python-Based Data Analytics and Machine Learning applications, a different type of file, named "Interactive Python Notebook" or shortened as ".ipynb" is more commonly used. Interactive Python Notebook are developed by Jupyter Notebook which utilizes Python characteristic of being interpretive, as it allows code execution by cells, which allows code separation into smaller cells (Jeba, 2023). This means that during experimentation phase, like pre-processing or model building, using notebook can be more convenience, since not all the code needs to be executed. Jupyter Notebooks also can store a markdown, which covers normal text, images, or links that can be used for explanation purposes (Faccioni, 2022). Other than that, Jupyter Notebooks also save the results of each cell (Saturn Cloud, 2023). For all these advantages, Notebooks provides a more suitable format for Data Analytics and Machine Learning applications.

For IDE chosen for this project, Visual Studio Code (VS Code) has been chosen for multiple reasons. The first reason is VS Code supports all python file types, including normal .py type, while allow easier process during deployment development phase. VS Code also shows the execution time for each cell execution, which can be helpful for documentation purposes. Another quality-of-life feature in VS Code is the Outline Mode, where all Headings created using Markdown, it will show each section of the code in a separate window, which allow traversal of the notebook become more convenience. Another convenience is through Jupyter Variable Explorer, which saves all variables created from the notebook, allowing easier tracking and debugging in a case of an error. Finally, VS Code also provides plugins in their marketplace, with many useful plugins that can help the developer during the development of the project (Moffitt, 2021).

The developer has decided to change from Jupyter-native IDE, such as Jupyter Notebook or JupyterLab into VS Code due to the convenience and advantages provided by the latter IDE.

3.3 Libraries Chosen

3.3.1 Data Pre-Processing and Data Analysis

In Python, Data Pre-Processing and Data Analysis requires multiple libraries to do the tasks more efficiently. The most famous libraries for these tasks are NumPy (Numerical Python) and Pandas (Python Data Analysis Library). NumPy is mainly used for scientific computing. NumPy is known for its more optimized version of arrays compared to Python regular arrays. NumPy is also provides multidimensional array object and its variations, such as masks and matrices. NumPy also compatible and can be used in conjunction with other libraries, such as pandas and matplotlib (ActiveState, 2022a). NumPy also provides the basis for other libraries, such as SciPy (Scientific Python) and scikit-learn (Duggal, 2023).

Pandas is a must have Python libraries for data analyst and data scientist (Duggal, 2023). It is built based on NumPy, Pandas make repetitive tasks in data analytics process, such as data cleansing, data filling, data normalization, and other data pre-processing tasks simpler, making it more time efficient (ActiveState, 2022c). Pandas also provides essential data analytics tasks, such as Extract, Transform, and Load (ETL) jobs, for example, loading a CSV file into data frame format (Duggal, 2023).

3.3.2 Data Visualization

For Data Visualization, two most famous Python libraries will be utilized for this project. Matplotlib is famous for its capability to do data visualization in Python. It was originally created to be an open-source alternative for MATLAB, thus it allows to create visualization that similar in style to MATLAB. Matplotlib is also compatible with NumPy library (ActiveState, 2022b). The only problem that matplotlib is based on MATLAB, which was released circa 1999. This means matplotlib's function is relatively low level, meaning that its not the most code efficient. It also has some compatibilities issues with newer libraries, such as pandas' DataFrame format. All these problems are solved by another library, called Seaborn. Seaborn is a data visualization library built on top of matplotlib, with pandas data frame integration built in. Seaborn is basically an enhanced version of matplotlib (Katari, 2021). This project will utilize both libraries for data visualization.

3.3.3 Machine Learning

For Machine Learning tasks, Python has some libraries that can be utilized for these tasks. First library is Scikit-learn, which is a Python library specialized in machine learning. It has the capability to create a bunch of different machine learning models, such as classification, regression, clustering, and dimensionality reduction (Jain, 2020). Scikit-learn, or usually also shorten as sklearn, is designed to be used with NumPy and SciPy (Duggal, 2023). But since sklearn can only create Machine Learning models, other libraries need to be utilized to do a deep learning task.

3.3.4 Deployment

Last, for Deployment of the created Machine Learning model, Streamlit library will be utilized to create the web application for the deployment. Streamlit is a free, open-source framework that can be used for deploying Python code into a web application without any previous front-end development experience. Streamlit is mainly targeted for Data Scientist and Artificial Intelligence Engineer, where they can deploy the results in Python, using Python code (Mhadhbi, 2021).

The advantage of Streamlit is its convenience, where there is no need for any frontend knowledge, such as HTML, CSS, or JavaScript. It is also compatible with many data-based libraries, such as Pandas, Seaborn, Matplotlib, Scikitlearn, and Keras, which is also utilized in this project (Mhadhbi, 2021). Streamlit model also can be deployed into the web using its provided community cloud feature (Streamlit, n.d.). Due to this, the developer has decided to pick Streamlit as its deployment method.

3.4 Operating System Chosen

An Operating System (OS) is the first software that a computer will load by its boot program. It controls everything regarding the computer, such as network connection, storage management, RAM and CPU cores utilization, and loading other software that the developer may use (Bigelow, 2020). This means the OS is the base of all software that it loaded during it runtime. Some of the most popular OS available currently is Microsoft Windows, macOS, and Linux. For this project, the developer chooses Microsoft Windows 10 as the OS of their choice. Microsoft Windows 10 first released by Microsoft on 29th July 2015 as the successor of Microsoft Windows 8.1 (Lifewire, 2022). Even though Microsoft Windows 10 is not the latest version of Windows that has been released, the OS is still supported by Microsoft until 14th October 2025 (Microsoft, n.d.). Moreover, the newest Microsoft Windows 11 is known for having a lot of problems and bugs (Wawro, 2023), which may cause issue during the execution of the project. Microsoft Windows 10 also still supports all the software needed for the project; thus Microsoft Windows 10 is still a good choice for the project.

3.5 Summary

For the rundown, after much comparison and deliberation, Python programming language is chosen to develop the project, due to its capability that suit the task of Data Analytics (DA) and Machine Learning (ML) with the help of its powerful libraries, moreover, Python is free and have a lot of communities support. With the language of the project has been chosen, then the IDE chosen is Visual Studio Code using Python Notebook, with its line-per-line execution and output, making it more favourable for this project. The libraries that have been chosen for this task are NumPy and Pandas for Data Pre-processing and Analysis, Matplotlib and Seaborn for Data Visualization, and Scikit-learn and for Machine Learning and models. Lastly, Windows 10 has been chosen to be the Operating System of the project.

3.5.1 Hardware

The hardware that the developer use to create this project is as follows:

- Central Processing Unit (CPU): AMD Ryzen 9 5900HX (8 Cores, 16 Threads, Base Clock 3.30 GHz, Max Turbo Clock 4.60 GHz)
- Graphical Processing Unit (GPU): NVIDIA RTX 3070 Laptop (8GB GDDR6, TDP 145 W), AMD Radeon Graphics
- Random Access Memory (RAM): 2x8 GB, 3200 MHz DDR4

3.5.2 Software

The software that the developer use to create this project is as follows:

- IDE: Visual Studio Code
- Programming Language: Python
- Documentation: Microsoft Word
- Dataset: Microsoft Excel
- Web Browser: Google Chromes

Chapter 4: Methodology

4.1 Introduction

Methodology refers to a framework that dictates how someone can do something, based on the field or study (Cambridge Dictionary, 2023). In Data Analytics field, Methodology can be a set of principle that developers can follow to make sure the result of the project can be reached with success. Thus, in this field, there are 3 Data Mining methodologies that are well known, they are CRISP-DM, SEMMA, and KDD (Quantum, 2021). A comparison will be done first so the developer can decide on which methodology suits the project.

4.2 Methods

4.2.1 Methodologies Comparison

Aspects	CRISP-DM	KDD	SEMMA
Overview	• CRISP-DM stands for Cross Industry Standard Process for Data Mining. (Hotz, 2023b)	• KDD stands for Knowledge Discovery in Databases (Hotz, 2023a)	• SEMMA Stands for Sample, Explore, Modify, Model, and Assess (Hotz, 2023c)
Developed By	CRISP-DM Consortium, an association between companies under the EU's Framework IV R&D initiative (Khabaza, 2010)	Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (Valadkhani, 2018)	The SAS Institute (Hotz, 2023c)
Number of Phases	6 Phases	5 Phases	5 Phases
Phases	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modelling 5. Evaluation 6. Deployment (Hotz, 2023b)	1. Selection 2. Pre-processing 3. Transformation 4. Data Mining 5. Interpretation or Evaluation (Hotz, 2023a)	1. Sample 2. Explore 3. Modify 4. Model 5. Assess (Hotz, 2023c)
Accessibility	• CRISP-DM is very flexible, as it provides many benefits from agile methodology • CRISP-DM allow phase reversal	• Cover the end-to-end process in Data Mining • KDD can work better in a larger team project	• Similar phases with KDD • Designed to work in conjunction with SAS

	• CRISP-DM is still applicable to be used for beginner until advanced level data science tasks. • Can be implemented with minimum Data Mining knowledge • Suitable for small team or individual project. (Hotz, 2023b) (Quantum, 2021)	• KDD has combination of phases and process, which can cause confusion • Takes more time as it is not as flexible as CRISP-DM • Require Data Mining Knowledge (Hotz, 2023a) (Quantum, 2021)	Enterprise Miner • Has a Cyclic nature, meaning higher flexibility compared to KDD, but not as flexible as CRISP-DM (Hotz, 2023c) (Quantum, 2021)

Table 4: Comparison of CRISP-DM, KDD, and SEMMA

4.2.2 Justification on Methodology Chosen

Based on the comparison above, CRISP-DM is chosen to be the methodology that the developer will be using to do this project. CRISP-DM provide both a rigid yet flexible methodology, meaning that it has a very intensive and detailed documentation process, like waterfall methodology, but very flexible in how to move between phases, as going back and forth between phase is an encouraged behaviour, which is a characteristic of agile methodology. This means that the developer can work following the phases stated and do any backtracking if new issues arise from previous phases, making it less prone to project failure. CRISP-DM is also the only methodology with Business Understanding and Deployment phase, meaning the developer is required to gain an understanding about the requirements of the project, and the results of the project need to be deployed, so it can be used by the public, which covers both requirement of this project. CRISP-DM is also the most popular data mining methodology used, based on a poll spanning 12 years, starting from 2002 (Hotz, 2023b).

4.2.3 Chosen methodology explanation – CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) is a methodology model that is most popular in the data science field, especially for data mining and analytics projects. It has 6 phases, which are:

Figure 1: CRISP-DM Methodology Phases

Phase 1: Business Understanding

Business Understanding focuses on understanding the objectives and requirements of the project. In this case, the developer needs to understand the problems that the project wants to tackle, which is why Coronary Artery Disease (CAD) case number is still very high, and what is the Personal Key Indicators (PKIs) that may have effect on the risk of CAD. The developer needs to understand the benefits, the aims and objectives, and the risk and challenges of the project itself. With this, the data mining process can work based on the business needs, thus increasing the chance of accomplishing the goals and objectives (Hotz, 2023b). In this project, Business Understanding is being done during the literature review process.

Phase 2: Data Understanding

The second phase is Data Understanding, that focuses on understanding the data that will be worked on for the project. First, the developer will need to acquire the data needed for the project, in this case, the dataset will be about heart disease and the PKIs or lifestyle that has affect to the risk of heart disease. This data can be obtained through a health institute survey,

where the data will have a high variance. Next, the developer needs to examine and explore the data, making sure the data is viable to use, understanding what each variables means, and making sure all variables are in the right format. The developer will also check the quality of the data, and what pre-processing steps are required to clean the data (Hotz, 2023b). In this project, Data Understanding is being done during the Data Exploration and Initial Data Viewing, where all pre-processing steps are being considered for the data.

Phase 3: Data Preparation

Before the data is visualized and used for analysis and machine learning, the data need to be pre-processed first, to make sure no faulty and inaccurate data can have an effect towards the analysis results and the machine learning model. This can be done by data cleaning, reformatting the data, integrating data from another dataset to increase variance and data amount, constructing new variable from existing variables, filling missing data, and partitioning the data for training and testing purposes. All these steps purpose is to make sure the data is well prepared and ready to be used to create an effective analysis and models in the next phase (Hotz, 2023b). Based on the project, the developer will need to do the appropriate pre-processing steps to make sure the data that will be used for the next step is optimized. In this project, Data Preparation covers the Data Pre-Processing process, where the data is being modified to be the most optimal data for the project's objective.

Phase 4: Modelling

After the data has been pre-processed, the next phase will be modelling, where the data is used to analysed, interpreted, and create a machine learning model based on it. Usually for interpretation, the data will be transformed into a graph or chart, for easier viewing and showcasing any correlation between two or more variables. Another modelling case is creating a machine learning model that suit the data and the project's objectives. In this case, the model also will be trained using the prepared training dataset, with the intention of increasing the accuracy result of the machine learning model created. With CRISP-DM methodology, modelling become more iterative as the methodology give the flexibility to create model and analysis until the developer found the best results possible (Hotz, 2023b). For this case, the developer needs to create a Machine Learning Model and Exploratory Data Analysis based on the data that has been prepared. In this project, Modelling covers both Data Visualization step and Model Building Step.

Phase 5: Evaluation

In this phase, the results from the previous phases will be reviewed and evaluated based on the predetermined objectives of the project, with the result of this phase will determine the next step of the project. If the result of the whole project has not reached satisfactory result or the required requirement, then the developer can backtrack to a previous phase and iterate more to make sure the results already met the requirements. On the other hand, if the results already met the project's requirements with great results, then the project can proceed to Deployment phase (Hotz, 2023b). The developer will need to evaluate based on the results of modelling phase, considering all possible outcomes, and making sure that the results published will be the most optimized one. In this project, Evaluation is covered during Model Evaluation step.

Phase 6: Deployment

The final phase in CRISP-DM methodology, it covers how the created results can reach the intended target users. This means creating a plan to deploy the model and analysis results, monitoring and providing maintenance for the model if required, producing the final report and documentation of the whole project, and reviewing the project as a retrospective look towards the whole project (Hotz, 2023b).

4.3 Summary

In conclusion, after comparing between all three data mining-focused methodologies, CRISP-DM is picked to be the methodology of this project, due to its flexible and iterative-friendly nature, making it a good match with the project nature, where multiple iteration of models creation can be beneficial to create the most optimized result. Following CRISP-DM methodology, a business understanding will be made first, based on the project's aim and objectives. The next step will be gaining data understanding of the selected dataset, understanding what each variable means, and choosing all pre-processing that may be required. The third step will be Data Preparation, where the data will be cleaned, processed, and transformed, if necessary, to make sure all data used is well optimized for analysis and machine learning model creation. After that, the modelling part of the project starts, where the tasks are to creating the most optimized machine learning model(s) and creating the analysis of the data. This part will likely take multiple iteration before the most optimized and accurate model can be created. Before ending the project, An evaluation of the whole process leading to the end will be done, making sure all the process has been done correctly, without leaving any possible improvements or possibilities on the table. If evaluation has reached satisfaction results, then The project can be finished by the deployment of the project's result.

Chapter 5: Research Methodology

5.1 Introduction

Research Methodology is a detailed explanation of a researcher's approach to gaining reliable data for their research. It shows how a researcher designs a research technique to acquire reliable and valid results, depending on their project goals (Indeed Editorial Team, 2021). The design should show what data the researcher wants to collection and how to get it.

Depending on the project, the researcher will need to pick between two types of research methodologies, which will affect on how they can do their data gathering, and what data that they need to collect. The first methodology is quantitative research. It focuses on confirming something, thus collecting numerical data is the focus of this methodology, with large participants size. The other methodology is qualitative research, where it focuses on non-numerical data, where the data gathering will consume more time, and using a much smaller, carefully chosen participants. It is more focused on gaining an understanding and exploring of the research's subjects (Mcleod, 2023). There are various data collection methods that can be done to fulfil the requirement of a research methodology. Examples are Survey, Observation, Interview, Questionnaire, and case study (Iterators, 2021).

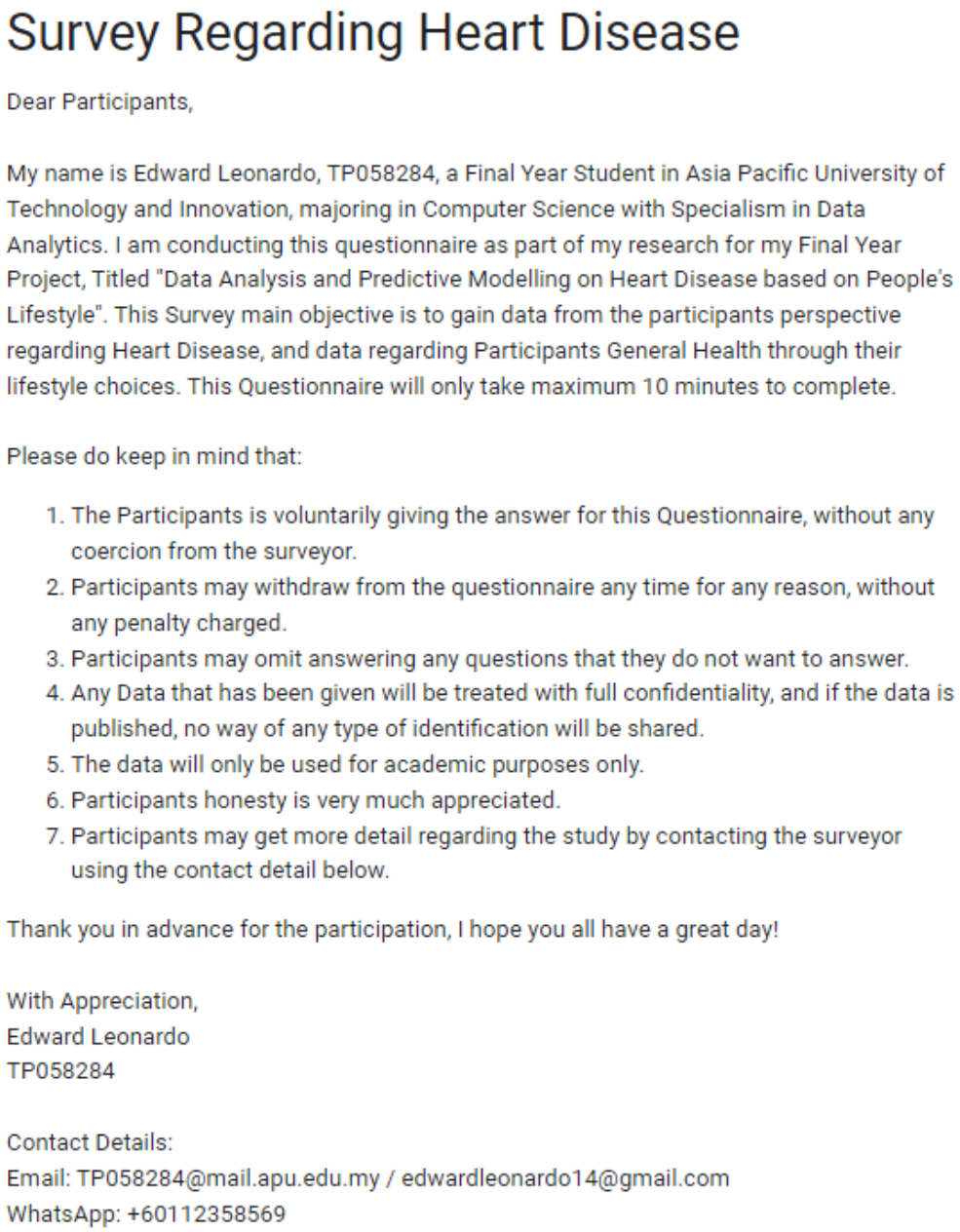
5.2 Questionnaire Survey

For this project, quantitative research will be carried out since the data gathered will be in numerical format, with the main objective is to gain validation that the project is relevant with the target users and to gain data from the participants. Therefore, Questionnaire Survey will be conducted as the data collection method. Questionnaire Survey is a type of research tool to do data gathering that consists of questions that is used to gain information from the participants (Bhat, 2023). The main reason of choosing Questionnaire as the main data gathering tool is because the method is free, easy to be conducted, and the reachability capability. Since the Questionnaire will via online using Google Form, it is easy to distribute the questionnaire, as a shareable public link address is all that is required. Another advantage of using questionnaire is because its effectiveness on gathering high amount of data for less time compared to other techniques. Finally, the survey's results can be easily visualised using Google Form result page (Bhat, 2023).

The questionnaire that will be used consists of 4 sections with 30 questions total, with 13 of them is compulsory, with the rest of them are optional. The type of questions are closed-ended questions, where a predetermined responses has been set for the participants to choose (Bhat, 2023). Except for one question, asking for participants feedback regarding the project. The objective of this project is to validate the importance of developing a Heart Disease Prediction model based on Personal Key Indicators. Based on the participants' perspectives, it can show the participants demand regarding this model. Another objective of the questionnaire is to do data gathering that can be used as data to be fed into the model. In the next section, the questionnaire design will be explained based on each section. Since the developer is originally from Indonesia, the questionnaire will be translated into Bahasa Indonesia version, where all the questions and options are identical, just in a different language.

5.3 Design

5.3.1 First Page



Survey Regarding Heart Disease

Dear Participants,

My name is Edward Leonardo, TP058284, a Final Year Student in Asia Pacific University of Technology and Innovation, majoring in Computer Science with Specialism in Data Analytics. I am conducting this questionnaire as part of my research for my Final Year Project, Titled "Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle". This Survey main objective is to gain data from the participants perspective regarding Heart Disease, and data regarding Participants General Health through their lifestyle choices. This Questionnaire will only take maximum 10 minutes to complete.

Please do keep in mind that:

1. The Participants is voluntarily giving the answer for this Questionnaire, without any coercion from the surveyor.
2. Participants may withdraw from the questionnaire any time for any reason, without any penalty charged.
3. Participants may omit answering any questions that they do not want to answer.
4. Any Data that has been given will be treated with full confidentiality, and if the data is published, no way of any type of identification will be shared.
5. The data will only be used for academic purposes only.
6. Participants honesty is very much appreciated.
7. Participants may get more detail regarding the study by contacting the surveyor using the contact detail below.

Thank you in advance for the participation, I hope you all have a great day!

With Appreciation,
Edward Leonardo
TP058284

Contact Details:
Email: TP058284@mail.apu.edu.my / edwardleonardo14@gmail.com
WhatsApp: +60112358569

Figure 2: Questionnaire - First Page

The first page of questionnaire shows the information regarding the surveyor, the questionnaire objectives, the participant's rights regarding their responses, and the terms and conditions of the questionnaire.

5.3.2 Section A – Basic Information

Section A covers the basic information of the participants. It consists of gender, age, and employment status. The purpose of this section is to get demographic information of the participants.

Question 1:

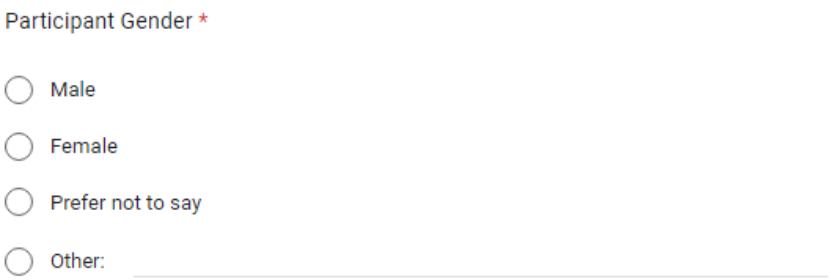


Figure 3: Questionnaire Section A – Question 1

The first question covers the participant's gender. Its purpose is to gain an understanding of how different gender's perspective regarding the questionnaire.

Question 2:

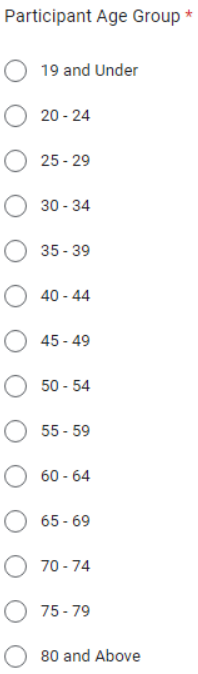


Figure 4: Questionnaire Section A – Question 2

The second question covers the participant age group. Dividing the age group into 5 years gaps can give detailed understanding of how different age groups view heart disease.

Question 3:

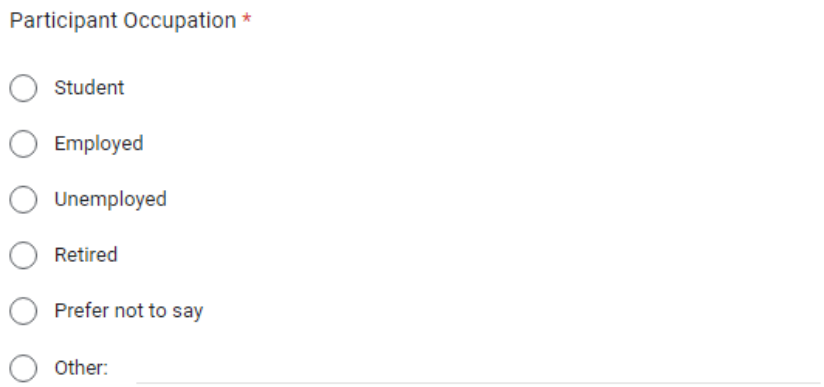


Figure 5: Questionnaire Section A – Question 3

Third question covers the participant's occupation. This can give insight into how different occupations think about heart disease.

5.3.3 Section B – Research Identification

Section B covers the research identification. This section collects information from participants perspective regarding heart disease. It covers basic their understanding of Heart Disease knowledge, what form of tool the project needs to realize, and feedback or suggestions regarding the whole project.

Question 1:

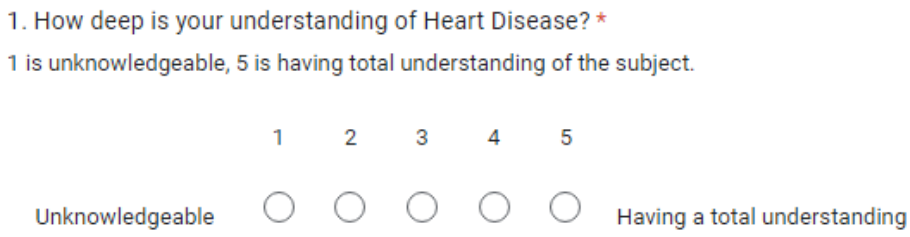


Figure 6: Questionnaire Section B – Question 1

The First question in this section covers the participant's general understanding of heart disease. The answer will be in form of linear scale, where 1 is having no knowledge regarding heart disease, while 5 means having a total understanding of the subject.

Question 2:

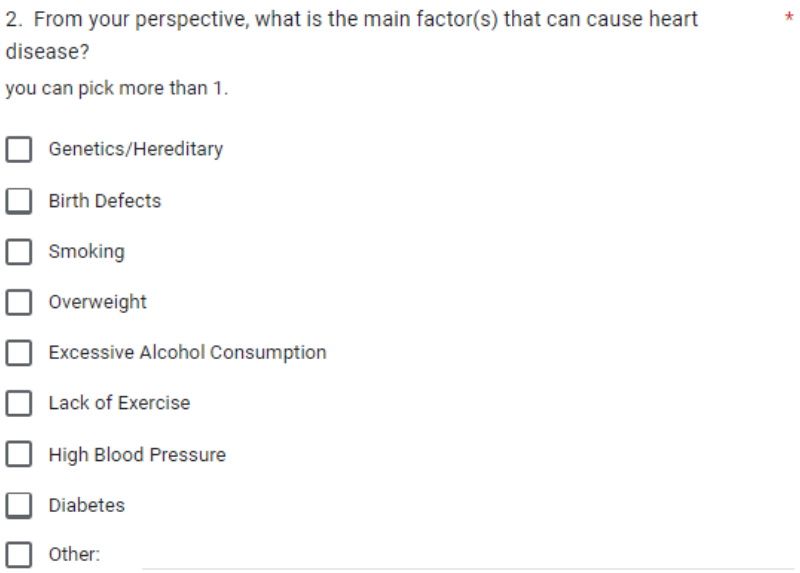


Figure 7: Questionnaire Section B – Question 2

The second question asks the participants which factors can cause heart disease. The question covers multiple factors from different type of heart diseases, to gain insight into how much the factors the participants know. In this question, the participants can pick more than one answer.

Question 3:

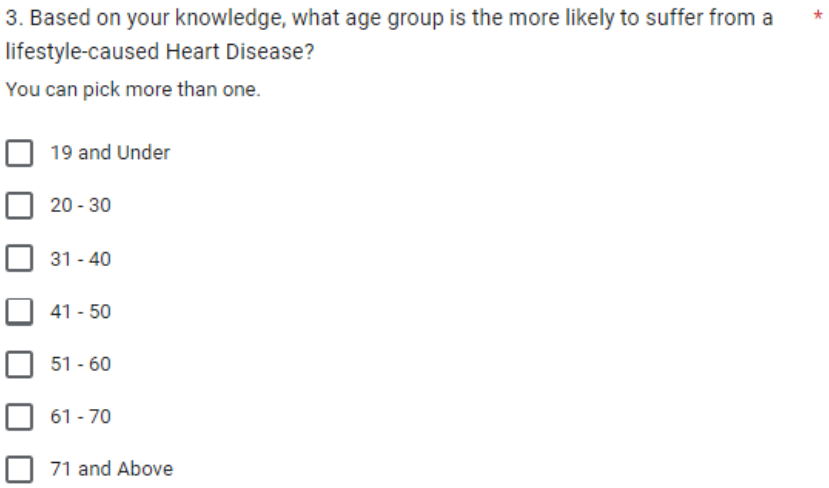


Figure 8: Questionnaire Section B – Question 3

The third question covers which age groups do the participants think have a higher risk of suffering from lifestyle-caused heart disease. In this question, the participants can pick more than one answer.

Question 4:

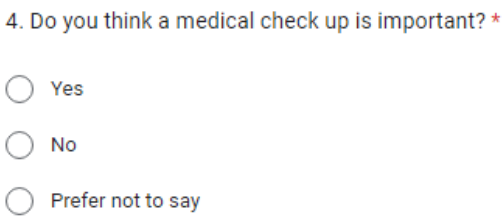


Figure 9: Questionnaire Section B – Question 4

The fourth question asks the participants their opinion regarding the importance of medical check-up. This question is connected to the next question.

Question 5:

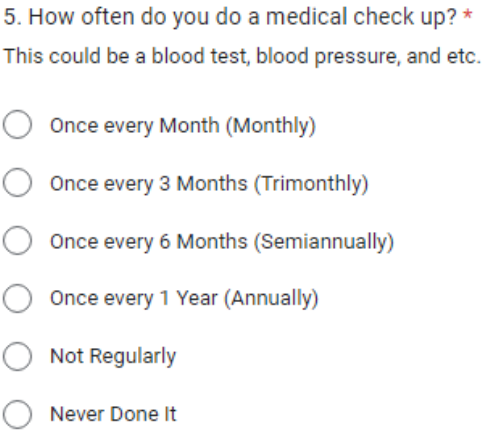


Figure 10: Questionnaire Section B – Question 5

Following up from the previous question, the fifth question asks the participants how often they do a medical check-up. This question is also connected to the next question.

Question 6:

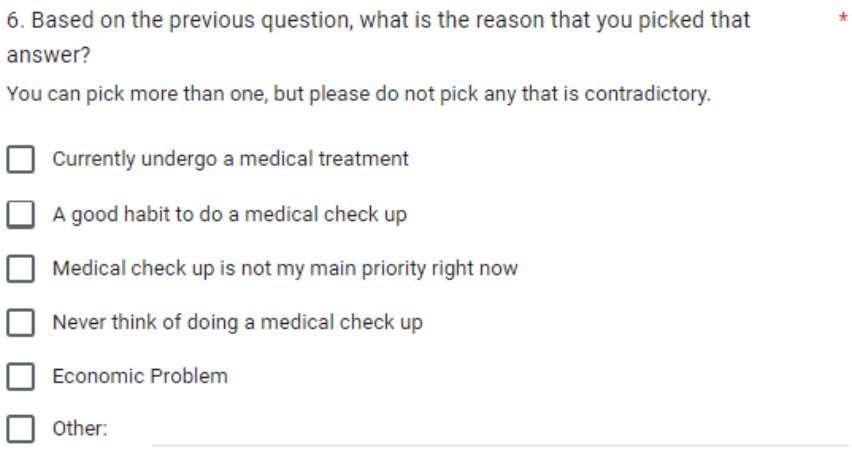


Figure 11: Questionnaire Section B – Question 6

The sixth question ask the reason of why participants picked the fifth question's answer. For this question, the participants can pick more than one reason, but the participants need to pick reason that is not contradictory.

Question 7:

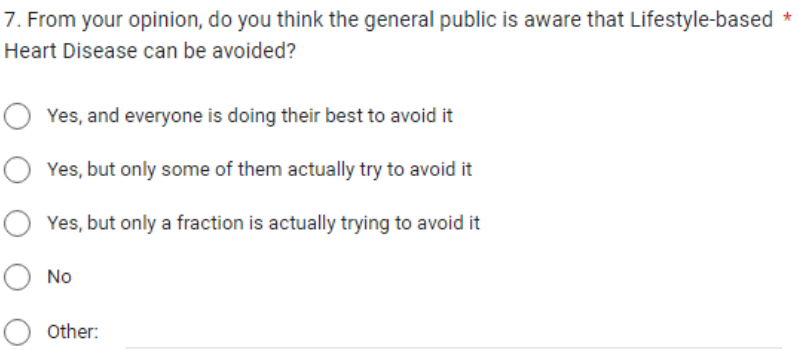


Figure 12: Questionnaire Section B – Question 7

The Seventh question ask participants opinion about the general awareness of lifestyle-based Heart disease. It asks if the general population knew that lifestyle-based heart disease can be avoided.

Question 8:

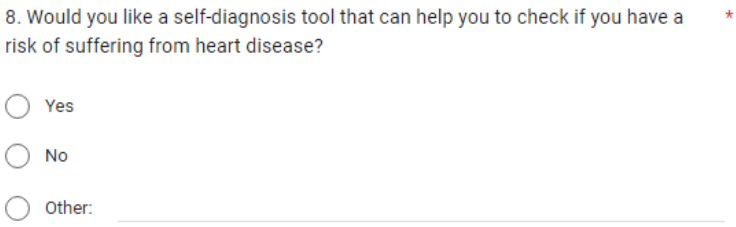


Figure 13: Questionnaire Section B – Question 8

This eighth question asks the participants if the yearn for a self-diagnosis tool can be used to check the risk of suffering from heart disease.

Question 9:

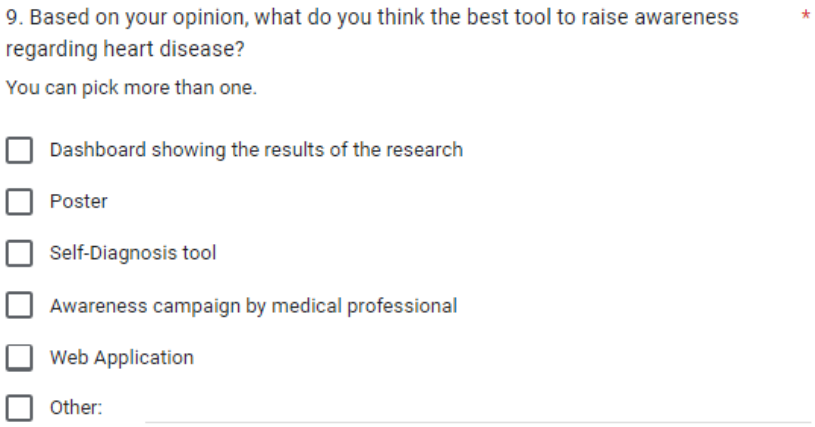


Figure 14: Questionnaire Section B – Question 9

This ninth question ask the participants opinions in what form an awareness-raising tool would they like. The participants can choose one or more forms or suggest other forms.

Question 10:

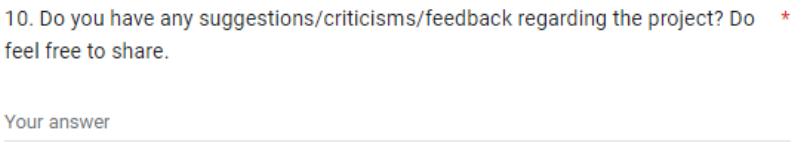


Figure 15: Questionnaire Section B – Question 10

The final question of this section asks for participants' suggestions, criticisms, or feedback regarding the whole project.

5.3.4 Section C – Data Gathering

Section C is data gathering section. This whole section is optional for participants, so if participants could skip this section if they deemed it unnecessary for them to answer. The data gathered in this section will be used as additional data for Exploratory Data Analysis and Predictive Model creation.

Question 1:

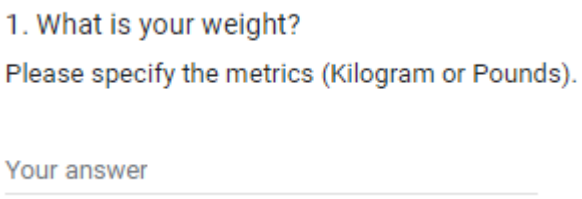


Figure 16: Questionnaire Section C – Question 1

The first question asks the participants for their body weight. The participants need to describe the metrics they used to calculate their body weight. This question will give Body Mass Index (BMI) data if combined with the second question.

Question 2:

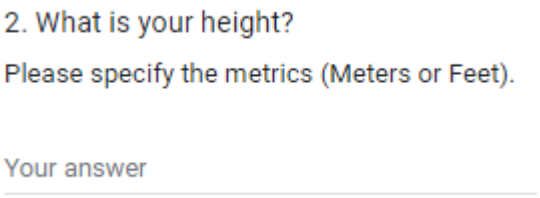


Figure 17: Questionnaire Section C – Question 2

The first question asks the participants for their body height. The participants need to describe the metrics they used to calculate their body height. This question will give Body Mass Index (BMI) data if combined with the first question.

Question 3:

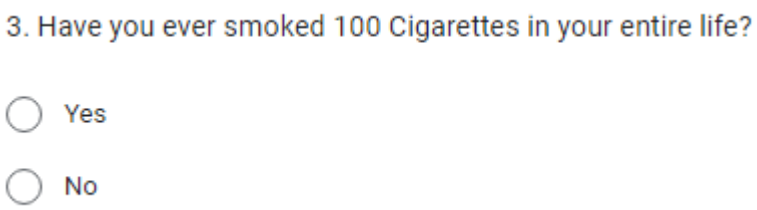


Figure 18: Questionnaire Section C – Question 3

The third question ask if the participant is or was a smoker, as someone can be considered as a smoker if they have smoked 100 cigarettes in their lifetime (CDC, 2017).

Question 4:

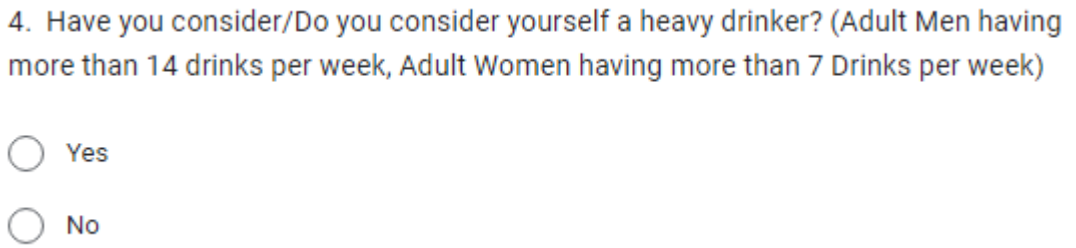


Figure 19: Questionnaire Section C – Question 4

The fourth question is to gain data if the participant is a heavy drinker. A person can be considered as heavy drinker if they consume more than 14 drinks per week for adult male, and more than 7 drinks per week for adult female (NIAAA, 2023).

Question 5:

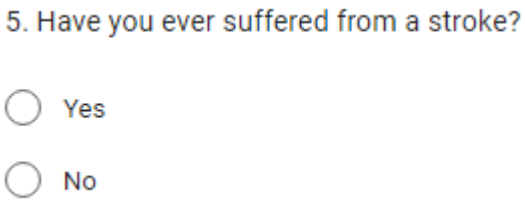


Figure 20: Questionnaire Section C – Question 5

The fifth question asks the participant if they have ever suffered from a stroke before.

Question 6:

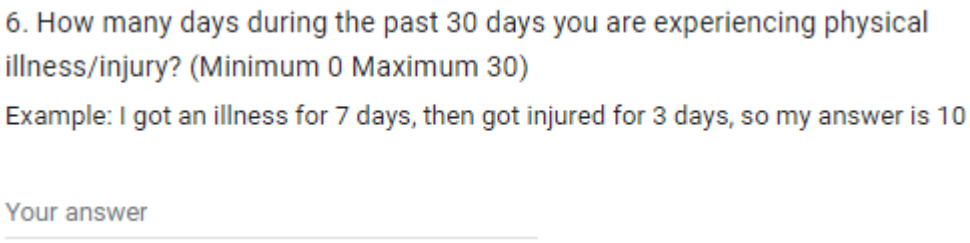


Figure 21: Questionnaire Section C – Question 6

The sixth question asks the participant for the amount of days in a month they have suffered from a physical illness or injury. The participant can answer based on the amount of days they suffered from physical illness.

Question 7:

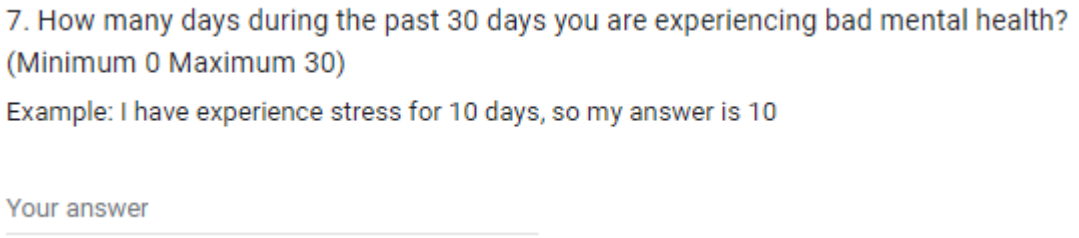


Figure 22: Questionnaire Section C – Question 7

The seventh question asks the participant for the amount of days in a month they have suffered from a bad mental health. The participant can answer based on the amount of days they suffered from bad mental health.

Question 8:

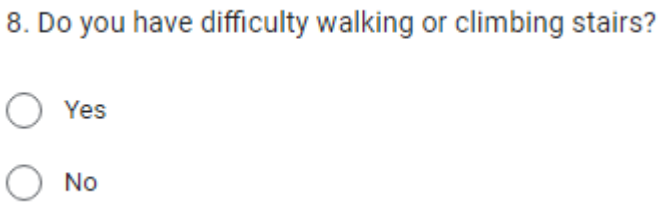


Figure 23: Questionnaire Section C – Question 8

The eighth question ask the participant if they have difficulty walking or climbing stairs. It can be a measurement of the participant's physical condition.

Question 9:

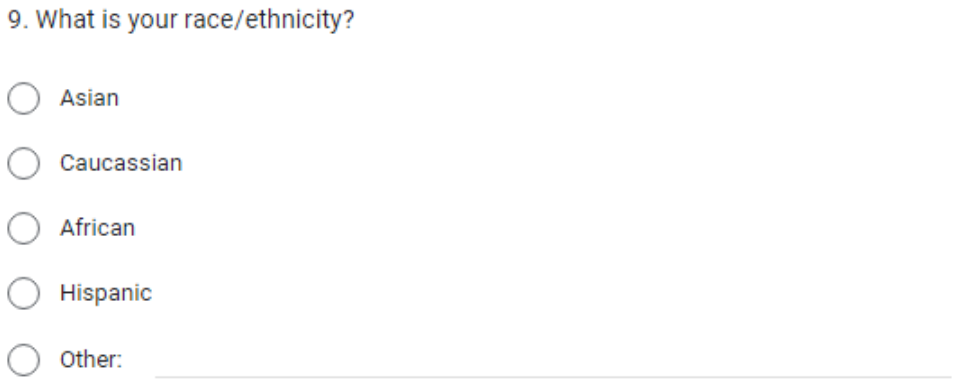


Figure 24: Questionnaire Section C – Question 9

The ninth question asks the participant for their ethnicity or race.

Question 10:

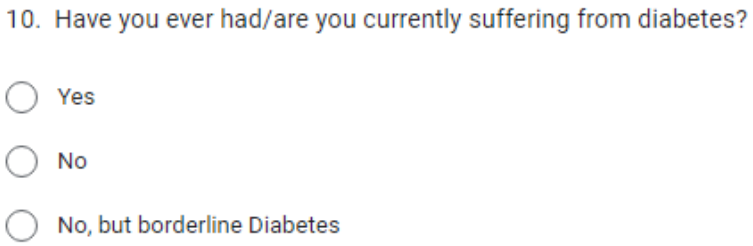


Figure 25: Questionnaire Section C – Question 10

The tenth question ask if the participant is currently, have suffered, or borderline (on the edge of) suffering from diabetes.

Question 11:

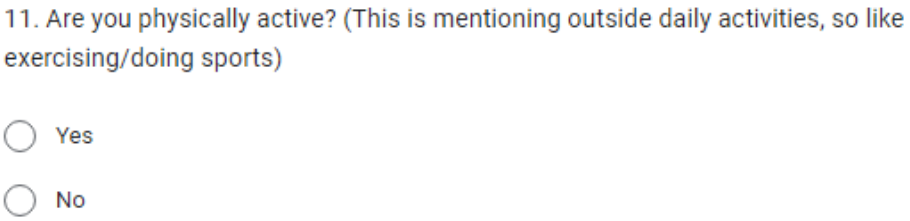


Figure 26: Questionnaire Section C – Question 11

The eleventh question as if the participant considered themselves physically active. The measurement of active here is if the participant dedicates some of their time to do some exercise or sports, outside of their daily activities.

Question 12:

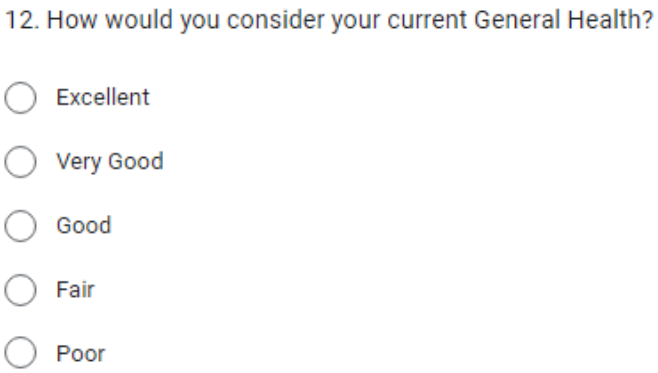


Figure 27: Questionnaire Section C – Question 12

The twelfth question asks for the participant's general health condition according to their self-evaluation. The participant can pick from Excellent, Very good, Good, Fair, or Poor.

Question 13:

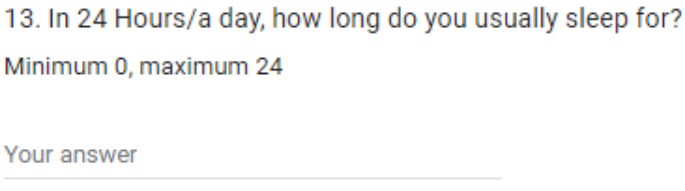


Figure 28: Questionnaire Section C – Question 13

The thirteenth question asks for the participant's amount of sleep measured in hours. The participant can enter a number between 0 (no sleep) to 24 (one day full sleep).

Question 14:

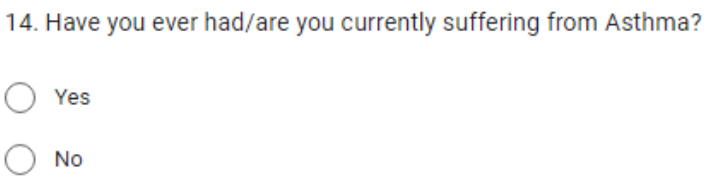


Figure 29: Questionnaire Section C – Question 14

The fourteenth question ask if the participant is currently or have suffered from Asthma.

Question 15:

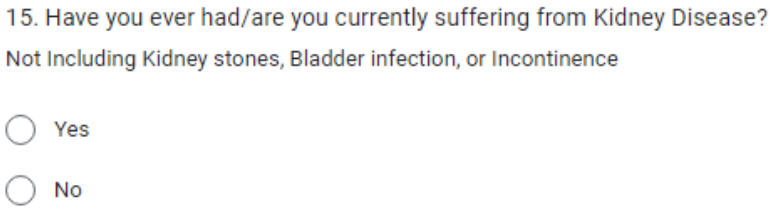


Figure 30: Questionnaire Section C – Question 15

The fifteenth question asks if the participant has ever suffered or currently suffering from a type of kidney disease, with exception of Kidney Stones, Bladder Infection, or Incontinence.

Question 16:

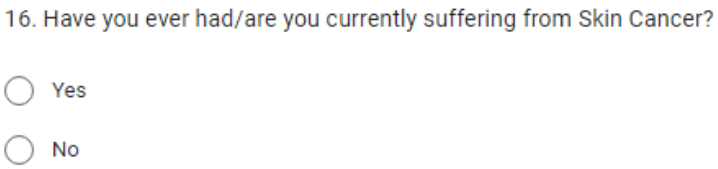


Figure 31: Questionnaire Section C – Question 16

The sixteenth question asks if the participant has ever suffered or currently suffering from any type of skin cancer. A study shown that cancer patient had a higher risk of developing Cardiovascular Disease (Hatch, 2022)

Question 17:

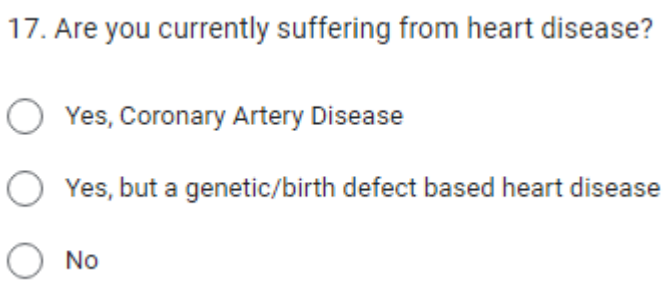


Figure 32: Questionnaire Section C – Question 17

The final question asks if the participant is currently suffering from a heart disease. The participant could pick if they suffered from Coronary Artery Disease (CAD), a genetic/birth defect-based heart disease, or they are not suffering from a heart disease.

5.4 Summary

For conclusion, the research methodology that is used in this project will be questionnaire survey, since it is easy to be conducted, and can gain many variances of data. The questionnaire will be created using Google Forms, where two versions will be created, with one version is using English Language, and the other one is using Bahasa Indonesia (Indonesia Language). This is due to the developer originally came from Indonesia, and data from Indonesia can have more variance. The questionnaire will be distributed via links, where participants can access and answer the survey. The survey itself is divided into 4 sections. The first section covers the rules of the survey and acknowledgement from participants. The second section covers the basic information regarding the participants. The third section covers the participants general knowledge regarding the topic and the participants' opinions regarding the project. The final section is a data gathering, that asks for participants' data, where it can be used combined with data set to do the Exploratory Data Analysis and Machine Learning Model Creation. The fourth section is entirely optional for participants.

Chapter 6: Requirement Validation

6.1 Introduction

Based on the results gathered from the questionnaire, an analysis will be done to understand the importance of developing the Machine Learning model to detect Coronary Artery Disease based on the questionnaire's responses. In this chapter, the developer will go through the responses, analyse, and give an interpretation based on the results. In this Questionnaire Survey, there will be 2 results per question since there are 2 version of the Questionnaire with different language. The results from each version will be interpreted both separately and combined. Since the final section of the questionnaire survey is only for data gathering for model creation and exploratory data analysis, it will not be analysed in this chapter. Question 10 from Section B, which asks for respondents' suggestion or feedback will also not be analysed in this chapter. The English Language version questionnaire got 48 responses, while in the Bahasa Indonesia version, the questionnaire got 51 responses.

6.2 Analysis of Questionnaire Data

6.2.1 Section A – Basic Information

Section A covers the Basic Information of the participants, such as gender, age group, and employment status.

Question 1:

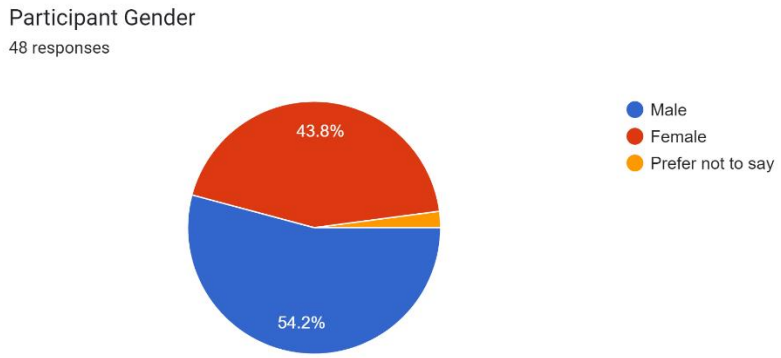


Figure 33: Questionnaire Section A – Question 1 – English Version

In the English version, it shows that the participants are male dominant, where 54.2% participants identifies as male, 43.8% identifies as female, while 1 participant prefer not to share their gender

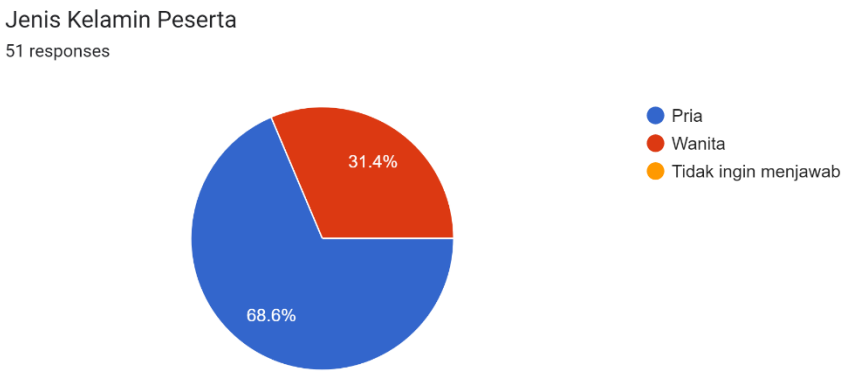


Figure 34: Questionnaire Section A – Question 1 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows the same results that majority of the participants are male with 68.6%, while 31.4% of the participants are female.

Thus, from both results, it shows that majority of the participants are male.

Question 2:

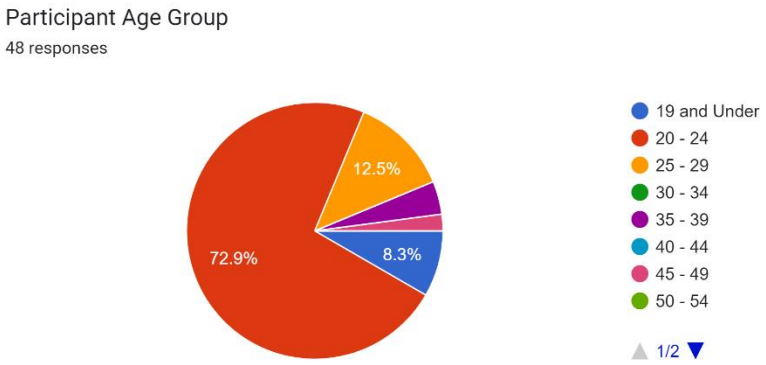


Figure 35: Questionnaire Section A – Question 2 – English Version

In the English version, it shows that most of the participants are in the 20-24 age group with 72.9% participants are from that age group. The second biggest is 25-29 with 12.5%, and the third largest is 19 and under, with 8.3% percentage share.

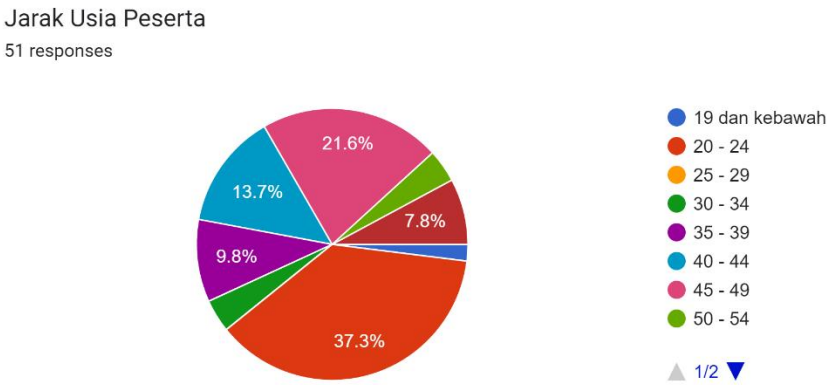


Figure 36: Questionnaire Section A – Question 2 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows that 20-24 age group is still the majority with 37.3%, but the second largest is 45-49 with 21.6%, third largest is 40-44 with 13.7% percentage share. It shows that the Bahasa Indonesia version's participants are more diversed age group wise.

From both results, the 20-24 age group is the majority age group, but Bahasa Indonesia's version have more variance in age group compared to the English version.

Question 3:

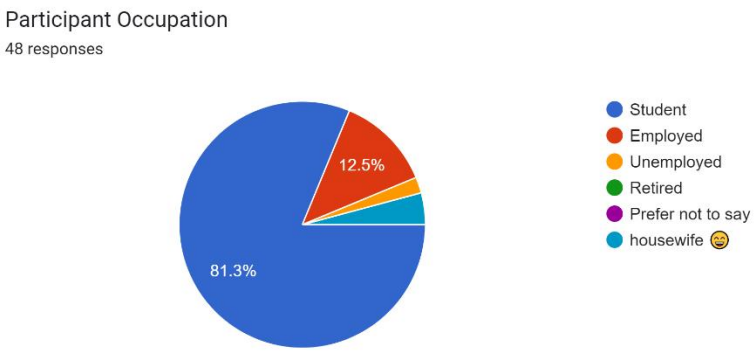


Figure 37: Questionnaire Section A – Question 3 – English Version

In the English version, it shows that the questionnaire is dominated by students, since 81.3% of the participants are students. The next biggest occupation is employed with 12.5% percentage share.

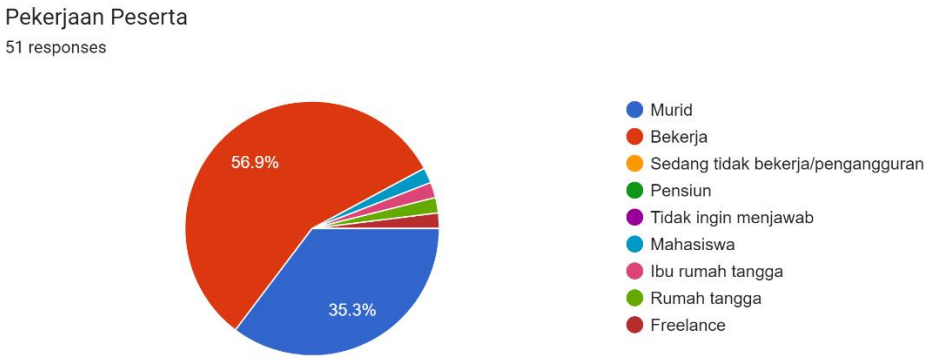


Figure 38: Questionnaire Section A – Question 3 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows different result, with more than half of the participants are already employed with 56.9% share. The next biggest is students with 35.3% share, with other types of occupation takes up small shares, around 10% of the total responses.

From the results above, it shows that the majority of the respondents' occupation are either student or employed.

6.2.2 Section B – Research Identification

Question 1:

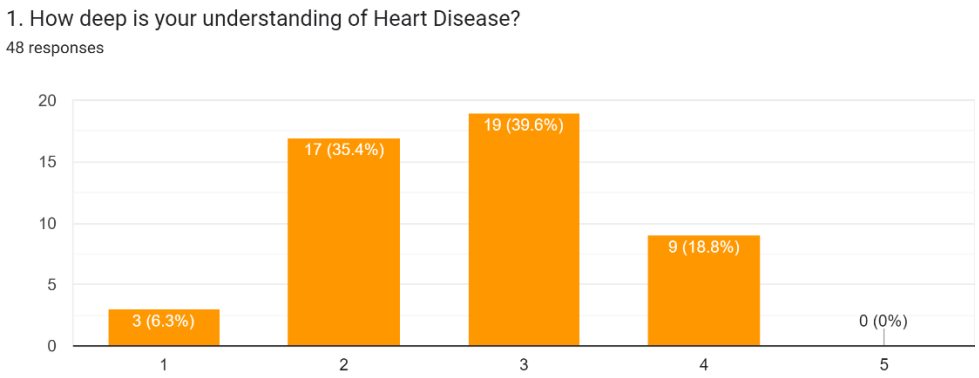


Figure 39: Questionnaire Section B – Question 1 – English Version

In the English version, it shows that the majority of the participants have none to a minimum basic understanding heart disease, since the result bar graph is more negative skewed. This shows that majority of the English's respondents, who is students, does not have a good understanding about heart disease.

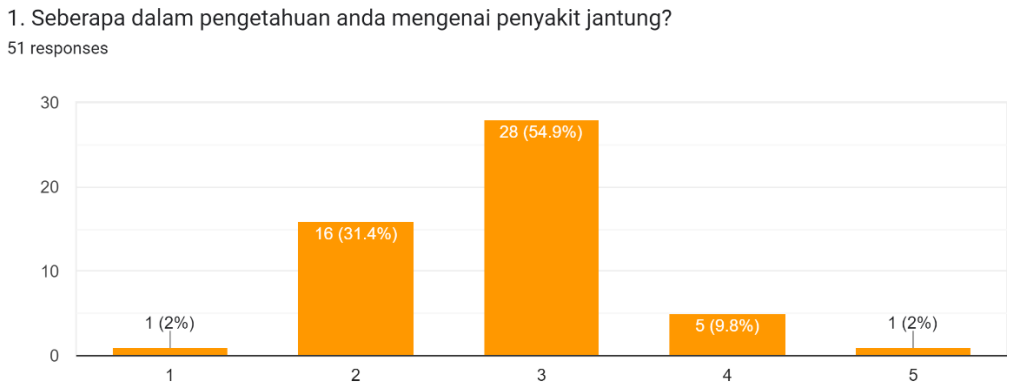


Figure 40: Questionnaire Section B – Question 1 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows the same results that majority of the participants only have a basic understanding of heart disease.

Thus, from both results, it shows that majority of the participants only have a basic knowledge of heart disease, with both bar graph has negative skew distribution, with more people have lower understanding of heart disease.

Question 2:

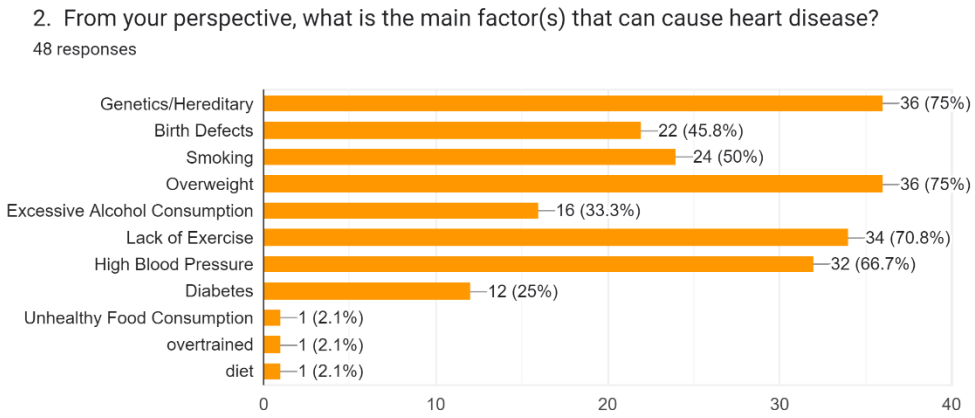


Figure 41: Questionnaire Section B – Question 2 – English Version

In the English version, it shows that participants think that the major risk factors that can cause heart disease are Genetics/Hereditary, being Overweight, Lack of Exercise, High Blood Pressure, and Smoking.

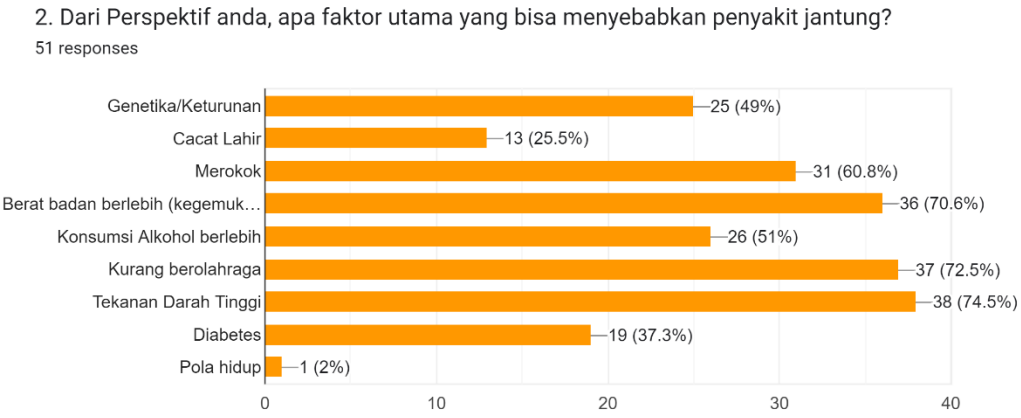


Figure 42: Questionnaire Section B – Question 2 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows that participants think that the major risk factors that can cause heart disease are Smoking, Being overweight, Excessive Alcohol Consumption, Lack of exercise, and High Blood Pressure.

From both results, the the major risk factors are Smoking, High Blood Pressure, Being overweight, Lack of exercise. It shows that majority of participants have a good knowledge of the risk factors that can cause Coronary Artery Disease.

Question 3:

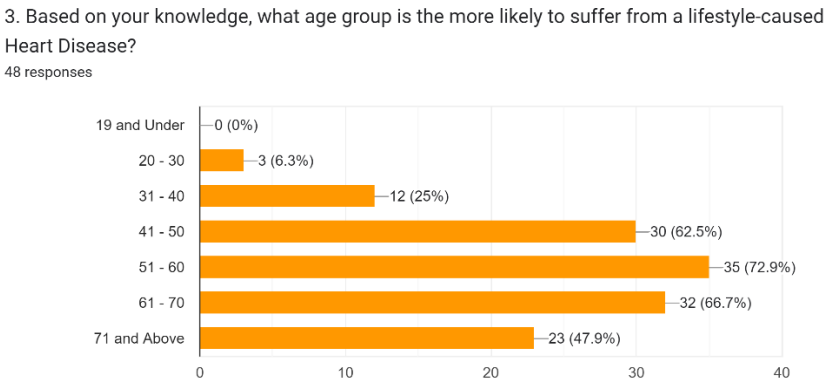


Figure 43: Questionnaire Section B – Question 3 – English Version

In the English version, it shows that the age groups where Coronary Artery Disease starts is 41-50, with 62.5% of participants answer this age group. Then the responses amount keep rising until 51-60 with 72.9% of participants answer this age group. The 61-70 and 71 and above still have a pretty high percentage, with 66.7% and 47.9% respectively.

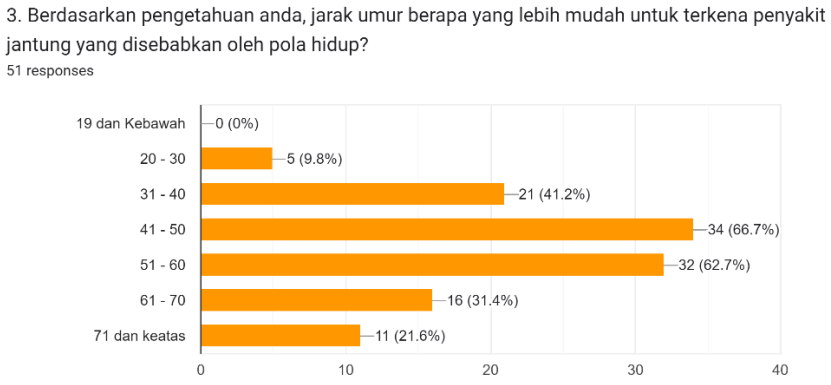


Figure 44: Questionnaire Section B – Question 3 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows different result, 31-40 age group already having a bigger number of responses, with 41.2% of participants answering this age group. The trends then continue similar to the English version, with the 61-70 and 71 and above getting less responses.

Based on results above, it shows that Bahasa Indonesia's participants has better knowledge regarding which age group is likely to suffer from lifestyle-caused heart disease or Coronary Artery Disease, with the answers is more evenly spread out on between all age groups.

Question 4:

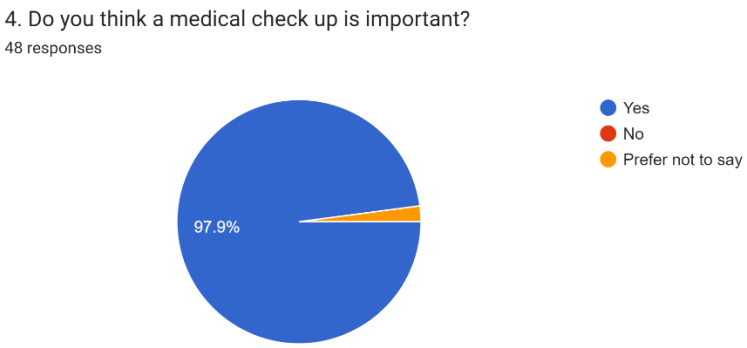


Figure 45: Questionnaire Section B – Question 4 – English Version

In the English version, it shows that almost all participants, with 97.9% of them agree that medical check-up is important, with a very small, 2.1% opting not to answer.

Figure 46: Questionnaire Section B – Question 4 – Bahasa Indonesia Version

Similar to the English counterpart, Bahasa Indonesia version of the questionnaire also shows an overwhelmingly 96.1% agreeing to the importance of medical check-up, with 3.9% preferring not to answer.

From both results, a safe conclusion can be made that majority considers medical check-up is very important.

Question 5:

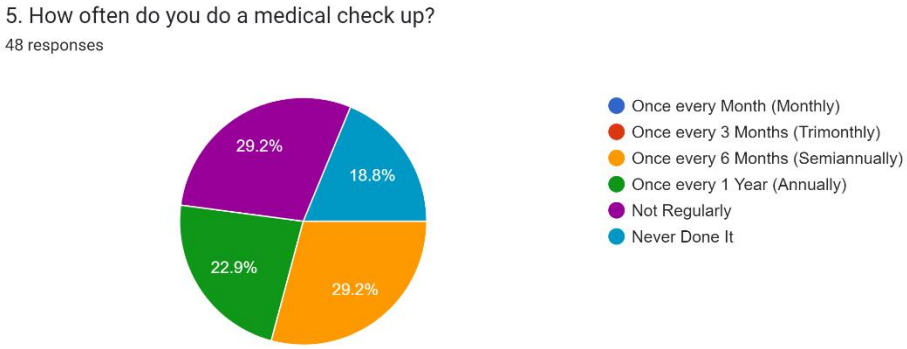


Figure 47: Questionnaire Section B – Question 5 – English Version

In the English version, it shows that the participants has been split 4 groups. The biggest group is those who do a medical check-up semianually (every 6 months) and those who do not do a medical check-up regularly. The third group is those who do a check-up anually, while the rest are those who never done any medical check-up.

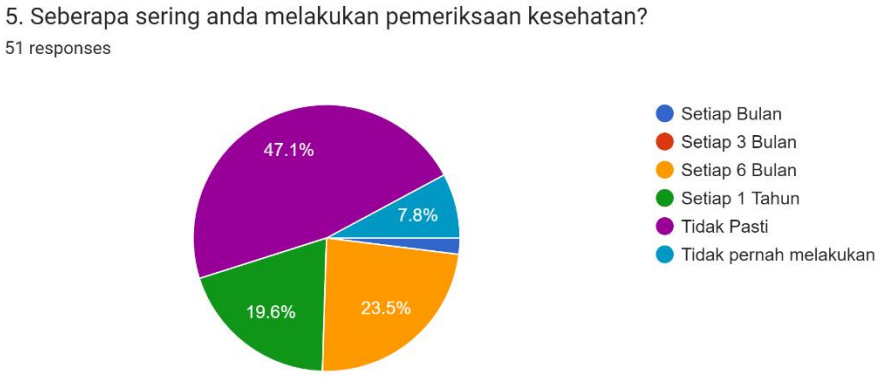


Figure 48: Questionnaire Section B – Question 5 – Bahasa Indonesia Version

In the Bahasa Indonesia version, It shows a very different results. The majority, with 47.1% are those that do not do a medical check-up regularly. The next biggest are those who are semiannually with 23.5%, with the next biggest are those who do it anually. 7.8% of respondents never done any medical check-up, and a very tiny percentage do a check-up every month.

From both results, It can be concluded that majority of respondents either do a medical check-up semianually, or do not do it regularly.

Question 6:

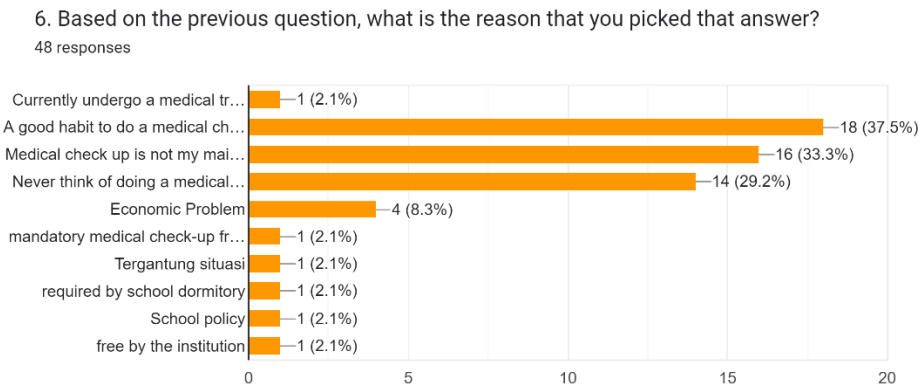


Figure 49: Questionnaire Section B – Question 6 – English Version

In the English version, it shows that the most respondents know that medical check-up is a good habit to do, with 37.5% percentage share. But it also shows that a lot of respondents never considered to do any medical check-up, or a medical check-up is not their priority. Some participants also pick economic problem as their reason of not doing medical check-up. 4 participants also put an institution policy to do a medical check-up.

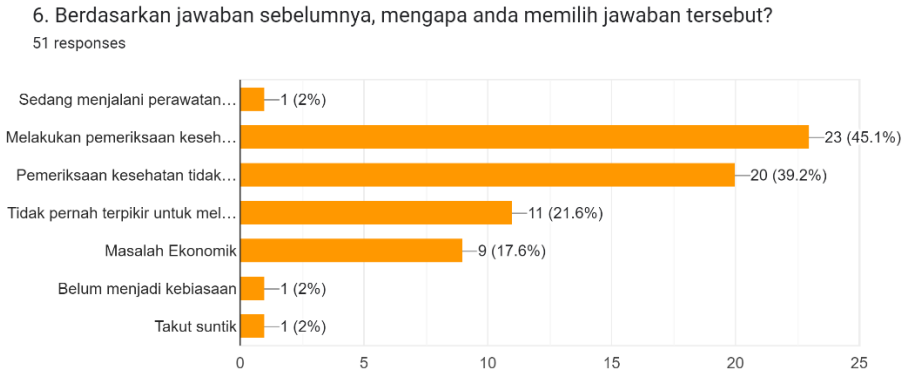


Figure 50: Questionnaire Section B – Question 6 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it a very similar results, with 45.1% majority also agrees that medical check-up is a good habit to do. A lot of respondents also never considered to do any medical check-up, or a medical check-up is not their priority. An economic problem has a bigger share, with 17.6% respondents pick this reason. There is also one participant said that they have a fear of needle.

From the results above, it shows that majority of respondents shows an understanding that medical check-up is a good habit to do. But it also shows that a lot of people still not considering a medical check-up to be their main priority, or even never thinking about it. A small group of respondents also pick economic problem, which shows that a medical check-up price can be still too much.

Question 7:

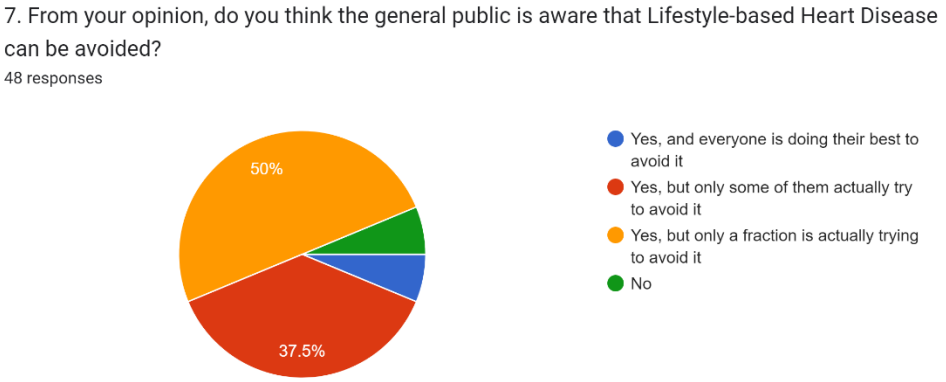


Figure 51: Questionnaire Section B – Question 7 – English Version

In the English version, it shows that 87.5% of respondents have an opinion that only some people is trying to avoid lifestyle-based heart disease or Coronary Artery Disease (CAD).

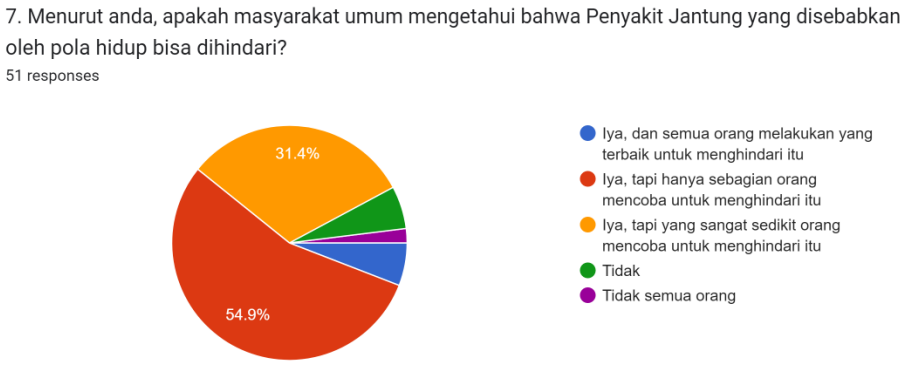


Figure 52: Questionnaire Section B – Question 7 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows the same results with similar percentage share with the English counterpart.

From this question, it can be concluded that eventhough general population understands that CAD is avoidable, but only a fraction of or only some of them are actually trying to avoid it.

Question 8:

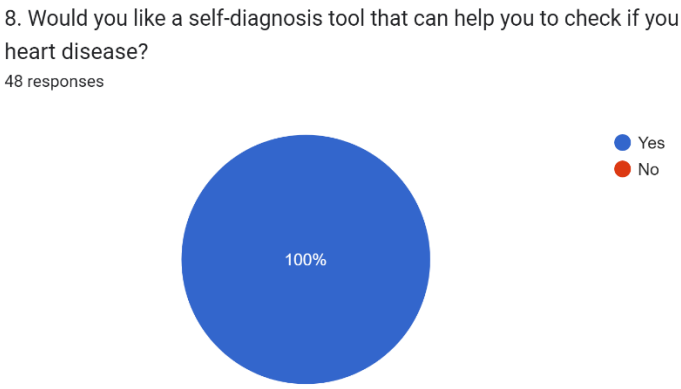


Figure 53: Questionnaire Section B – Question 8 – English Version

In the English version, it shows that all participants wants a self-diagnosis tool that can help check if they have a risk of suffering from CAD.

Figure 54: Questionnaire Section B – Question 8 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it shows similar results with the English version, with 96.1% agreeing. There is only 1 participant picking no, and 1 participant answer “only if the price is affordable for the middle-to-lower economic class.

From both results, It shows that almost all participants want to have a self-diagnosis tool to check if they have a risk of suffering from CAD.

Question 9:

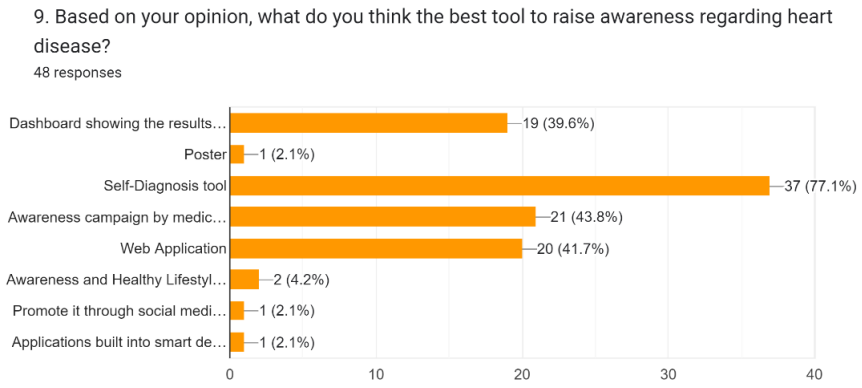


Figure 55: Questionnaire Section B – Question 9 – English Version

In the English version, it shows that the preferred tools to raise awareness regarding CAD are Self-diagnosis tool, an awareness campaign by medical professionals, a web application, and a dashboard showing the research's results.

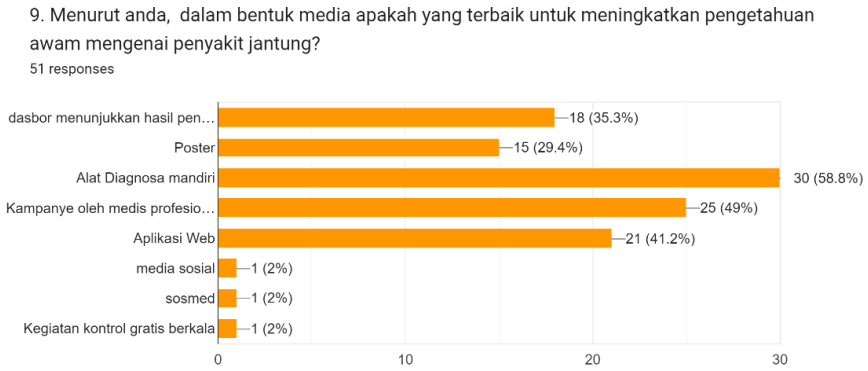


Figure 56: Questionnaire Section B – Question 9 – Bahasa Indonesia Version

In the Bahasa Indonesia version, it also shows similar results, with Self-diagnosis tool, an awareness campaign by medical professionals, a web application, and a dashboard showing the research's results as the preferred tools. It also shows Poster as a preferred tool, and 2 participants also mentioned social media as an awareness raising tool.

From the results above, it can be concluded that a lot of people prefer a self-diagnosis tool, that can be in a form of web application, an awareness campaign by medical professionals, and a dashboard showing research's results.

6.3 Summary

Based on all results above from a total of 99 participants, a conclusion can be made that majority of participants understands about heart disease, especially about Coronary Artery Disease (CAD). But it also shows that a lot of them are ignorant regarding heart disease, as a lot of people still think that medical check-up is not important enough to be their priority, and a big majority also still ignoring the precautions that they can do right now to actively avoid the risk of suffering from CAD. It also shows that a self-diagnosis tool is in-demand, with a web application version seems to be the most popular. It also shows that an awareness campaign by medical professional and a dashboard showing the project's results also can be alternative tools to spread and raise awareness regarding CAD.

Chapter 7: Data Analysis

7.1 Introduction

Data Analysis is a pre-requisite step in Data Analytics, Business Intelligence, and Machine Learning Modelling. Data Analysis itself can be described as the process of inspecting, cleaning, processing, and transforming the data gathered, as part of the preparation for further process (Kelley, 2023). Data Analysis covers both Exploratory Data Analysis (EDA) and Data Pre-Processing. In this case of this project, which focuses on Machine Learning Modelling, Data Analysis covers the data preparation step, means that the data is prepared so it become a clean, well-curated data, which can be used effectively during the Machine Learning Modelling step. Data Analysis and Preparation is a very crucial step in Machine Learning Modelling, since if the data inputted into the model during its training is dirty (missing, inconsistent, poorly formatted), it can produce misleading outcome or a less accurate predictive model (DataRobot, 2023).

7.2 Initial Data Exploration

Initial Data Exploration covers the pre-requisite Data Exploration and Analysis before any Pre-processing steps are conducted. Initial Data Exploration, or Exploratory Data Analysis (EDA), covers the initial investigations of the dataset to discover any anomalies, patterns, and any pre-processing that may be needed before the data is used (Patil, 2022).

In this project, there are three datasets that are going to be used. The first data, refers as the main data, is a dataset originally collected from CDC as part of Behavioral Risk Factor Surveillance System (BRFSS) that is collected in 2020. Originally consists of 279 columns or attributes, it has been reduced to 18 columns, that is more focused on the Personal Key Indicators of Heart Disease (Pytlak, 2022).

The other two datasets used are the surveys responses, both in English and in Bahasa Indonesia Version. The surveys consist of two parts, where the first part covers the requirement validation of the project, which is covered in Chapter 6, while the second part covers the data gathering part. In this case, the second part of both datasets are the one used for this process, where it will be combined into the main dataset, after some pre-processing steps.

7.2.1 Library Import

```
# Data Manipulation
import pandas as pd
import numpy as np
import re

# Visualization
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(font_scale = 2)
import missingno as msno

import warnings
warnings.filterwarnings('ignore')
from collections import Counter

# Encoding
from sklearn.preprocessing import LabelEncoder

# Sampling
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler

# Data Splitting
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, StandardScaler

# Model Metrics
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report,precision_score,recall_score,f1_score,roc_auc_score,roc_curve

# Models
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier,GradientBoostingClassifier
from sklearn.model_selection import cross_val_score,GridSearchCV, RandomizedSearchCV, StratifiedKFold
from sklearn.svm import LinearSVC
from sklearn.neural_network import MLPClassifier
import tensorflow
import keras_tuner as kt
```

Figure 57: Library Import

Library Import covers all libraries and packages that are being used for the project. The only exception is deployment package, Streamlit, since it will be separated into different Python file.

7.2.2 Data Loading

```
# Installing the initial Data, and both surveys results (Indonesia and English ver)
data = pd.read_csv("heart_2020_final.csv")
surveyIna = pd.read_csv("Survey Bahasa Indonesia Version.csv")
surveyEng = pd.read_csv("Survey English Language Version.csv")
```

Figure 58: Data Loading

Data loading covers all data import, where it will be saved into a dataframe variable, so it can be accessed and manipulated for the data pre-processing step. The main dataset is saved under the variable “data”, while each survey is saved in “surveyIna” for the Bahasa Indonesia version, and the “surveyEng” for the English Language version.

7.2.3 Initial Data Viewing

```
#initial viewing of the Dataset
data.head()
```

✓ 0.0s Python

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDis
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes	Very good	5.0	Yes	
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes	Very good	7.0	No	
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes	Fair	8.0	Yes	
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No	Good	6.0	No	
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes	Very good	8.0	No	

Figure 59: Initial Data Viewing Part 1

```
data.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      319795 non-null  object
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object
 3   AlcoholDrinking   319795 non-null  object
 4   Stroke            319795 non-null  object
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object
 8   Sex               319795 non-null  object
 9   AgeCategory       319795 non-null  object
 10  Race              319795 non-null  object
 11  Diabetic          319795 non-null  object
 12  PhysicalActivity  319795 non-null  object
 13  GenHealth         319795 non-null  object
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object
 16  KidneyDisease     319795 non-null  object
 17  SkinCancer        319795 non-null  object
dtypes: float64(4), object(14)
memory usage: 43.9+ MB
```

Figure 60: Initial Data Viewing Part 2

For Initial Data Viewing, .head() and .info() is mainly used, where .head() shows the first 5 data inside the dataframe, while .info() shows all attributes inside the dataframe, the amount of existing data for each attribute, and the Data type of each attribute.

Attribute Name	Variable Types	Description
Heart Disease	Nominal	Indicates if the person is suffering from Heart Disease
BMI	Continuous	The person's BMI
Smoking	Nominal	Indicates if the person has at least smoked 100 cigarettes in their entire life
AlcoholDrinking	Nominal	Indicates if the person is a heavy drinker (adult men drink more than 14 drinks per week, adult women drink more than 7 drinks per week)
Stroke	Nominal	Indicates if the person is suffering/have suffered from Stroke
PhysicalHealth	Continuous	How many days in the past 30 days have the person suffered from a physical illness
MentalHealth	Continuous	How many days in the past 30 days have the person suffered from a mental illness
DiffWalking	Nominal	Indicates if the person is having difficulty walking
Sex	Nominal	The person's gender
AgeCategory	Ordinal	The person's age category
Race	Nominal	The person's race
Diabetic	Nominal	Indicates if the person is suffering/have suffered from Diabetes
PhysicalActivity	Nominal	Indicates if the person is physically active outside of their daily life (Exercise)
GenHealth	Ordinal	The person's general health condition
SleepTime	Continuous	The number of hours of sleep in a day
Asthma	Nominal	Indicates if the person is suffering/have suffered from Asthma
KidneyDisease	Nominal	Indicates if the person is suffering/have suffered from Kidney Disease
SkinCancer	Nominal	Indicates if the person is suffering/have suffered from Skin Cancer

Table 5: Attribute Detail

7.3 Data Cleaning

Data Cleaning is the step where any data that is inconsistent, incorrectly formatted, duplicates, or missing data is fixed. The main objective of Data cleaning is to prepare the data for Data visualization or Model Building, where any incorrect data may have an affect towards the visualization results or the Model performance. Data Cleaning have no exact steps, as the process can be completely different based on the data initial condition (Tableau, 2023).

In the case of this project, Data Cleaning can be separated into two main steps, where pre-processing is being done on Survey data, so it can be assimilated into the main data first. After that, the combined data can be pre-processed at once.

7.3.1 Survey Data – Unused Data Removal

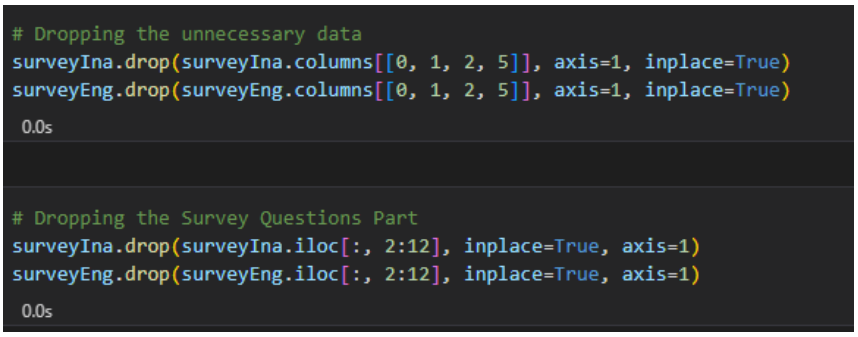


Figure 61: Unused Data Removal

The first step of the Survey Data Pre-processing is to remove all unneeded attributes. This covers the timestamp, participant's username, participant's consent, and participant's occupation. This part also removes the Requirement Validation part of the survey, so the remaining data is only the data that can be used for the project's deliverable.

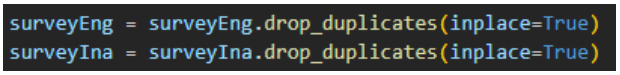


Figure 62: Survey Data Duplicates Dropping

The next step of survey data cleaning is to drop all duplicates from both versions.

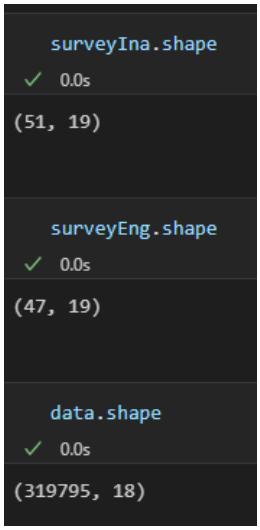


Figure 63: Datasets Shape Checking

From the .shape, it shows that survey data and main data has different column count. This is because the survey data does not have the column BMI but have the column BodyWeight and BodyHeight. Using both attributes, BMI, with measuring unit of kg/m^2 can be calculated (NHS, 2023).

Attribute Name	Variable Types	Description
BodyWeight	Continuous	The person's body height
BodyHeight	Continuous	The person's body weight

Table 6: Additional Survey's Attribute Detail

7.3.2 Survey Data – Data Standardization

```
# Creating the Attributes name mapping following the main data attributes' name
newAttributeName = ['Sex', 'AgeCategory', 'BodyWeight', 'BodyHeight',
                    'Smoking', 'AlcoholDrinking', 'Stroke', 'PhysicalHealth',
                    'MentalHealth', 'DiffWalking', 'Race', 'Diabetic',
                    'PhysicalActivity', 'GenHealth', 'SleepTime', 'Asthma',
                    'KidneyDisease', 'SkinCancer', 'HeartDisease']

# Creating a map dictionary to connect the old names to the new names
attributeMappingIna = {oldName: newName for oldName, newName in zip(surveyIna.columns, newAttributeName)}
attributeMappingEng = {oldName: newName for oldName, newName in zip(surveyEng.columns, newAttributeName)}

# Rename the columns for both surveyIna and surveyEng dataframe
surveyIna.rename(columns=attributeMappingIna, inplace=True)
surveyEng.rename(columns=attributeMappingEng, inplace=True)
```

Figure 64: Survey Data Attribute Renaming

The first step of data standardization is to merge between both version of the survey. Thus, attribute name can be replaced since the attribute name is still following the google form's result. With attribute name replacement, the data can be merged easier.

```
# Check all the unique values for each columns in surveyIna
for column in surveyIna.columns:
    unique_values = surveyIna[column].unique()
    print(f"Unique values in column '{column}':")
    count_nan = surveyIna[column].isna().sum()  # Count the number of NaN values
    print(f"NaN: {count_nan}")
    for value in unique_values:
        if pd.notna(value):  # Exclude NaN values from individual value counts
            count = surveyIna[column].value_counts().get(value, 0)  # Get value count, default to 0 if value not found
            print(f"{value}: {count}")
    print()
```

Figure 65: Function Print all Unique Values

```
surveyIna = surveyIna.replace(['Tidak'], 'No')
surveyIna = surveyIna.replace(['Pria'], 'Male')
surveyIna = surveyIna.replace(['Wanita'], 'Female')
surveyIna = surveyIna.replace(['Iya'], 'Yes')
surveyIna = surveyIna.replace(['Baik'], 'Good')
surveyIna = surveyIna.replace(['Sangat Baik'], 'Very Good')
surveyIna = surveyIna.replace(['Jelek'], 'Poor')
surveyIna = surveyIna.replace(['Ya'], 'Yes')
surveyIna = surveyIna.replace(['Tidak, tapi saya diambang diabetes'], 'No, borderline diabetes')
surveyIna = surveyIna.replace(['19 dan kebawah'], '19 and Under')
surveyIna = surveyIna.replace(['Asia'], 'Asian')

# Since the main dataframe only includes the Coronary Artery Disease problem, then any Birth Defect Heart Problem is considered as a No
surveyIna = surveyIna.replace(['Iya, Penyakit Jantung dari genetika/cacat lahir'], 'No')
```

Figure 66: Survey Data Language Translation

The next step is to translate the Bahasa Indonesia version into English, so it become the same language. The first figure shows the code snippet to print all unique values for each attribute in surveyIna. The second figure shows the code snippet to translate the values. The rest of the data that is not translated will be replaced into missing data in a later step.

```
survey = pd.concat([surveyIna, surveyEng])
```

Figure 67: Survey Data Merging

Now that the Bahasa Indonesia version has been translated, it can be merged with the English version, where it is saved into “survey” variable. After that, all unique values inside “survey” dataframe will be printed, so each inconsistency be spotted, evaluated, and acted accordingly.

```
survey = survey[survey.Sex != 'Prefer not to say']
```

Figure 68: One Outlier Removal

Based on the results, there is only one “Prefer not to say” in the “Sex” attribute. Thus, the data can be removed entirely.

```
# To easily remove the measuring unit, make all the units into lowercase.
survey['BodyWeight'] = survey['BodyWeight'].str.lower()
survey['BodyHeight'] = survey['BodyHeight'].str.lower()
```

Figure 69: Data Lowercase Conversion

```
# BodyWeight
# 1. BodyWeight are all in kilograms, thus we only need to remove the units
# 2. Some values contain a hypen (-), thus we need to get both values, then get the average between the two values (creating a new function to be reused in next attributes)
# 3. One value is using a comma (,) rather than a dot (.) to show decimal point.

# Removing units
survey['BodyWeight'] = survey['BodyWeight'].str.replace(r'[a-zA-Z]', '')

def specialCharCheck(value):
    if isinstance(value, str):
        if ',' not in value and '-' not in value and '~' not in value and '/' not in value:
            try:
                return float(value)
            except ValueError:
                pass

        if '-' in value or '~' in value:
            if "~" in value:
                separator = "~"
            else:
                separator = "-"
            try:
                number1, number2 = map(float, value.split(separator))
                # Calculate the average
                average = (number1 + number2) / 2
                return average
            except ValueError:
                pass
        elif ',' in value:
            value = value.replace(",", ".")
            try:
                return float(value)
            except ValueError:
                pass
        elif '/' in value:
            value = value.split('/')[1]
            try:
                return float(value)
            except ValueError:
                pass

    return None

# Apply the function to remove all inconsistency
survey['BodyWeight'] = survey['BodyWeight'].apply(specialCharCheck)
```

Figure 70: Function Transform BodyWeight

```
# BodyHeight
# Since BMI calculation will be using Meter Squared (m2), then All the values will be changed to Meter first
# Now, there are 3 types of units used in the survey dataset, Feet, Meter, and Centimeter
# 1 Foot is 0.3048 Meter, while 1 centimeter is 0.01 meter

#1. Need to create an if-else to catch all measuring unit
#2. inside each if-else, remove the unit, use the specialCharCheck to remove all inconsistency, then convert the value into meters

def heightCheck(value):
    if isinstance(value, float):
        return float(value)
    elif "cm" in value:
        value = re.sub('[a-zA-Z]', '', value)
        floatValue = specialCharCheck(value)
        floatValue = floatValue * 0.01
        return round(floatValue, 2)
    elif "m" in value or "meter" in value or "meters" in value:
        value = re.sub('[a-zA-Z]', '', value)
        floatValue = specialCharCheck(value)
        # To double check if someone entered 'm' instead of 'cm'
        if floatValue > 100:
            floatValue = floatValue * 0.01
        return round(floatValue, 2)
    elif "feet" in value:
        value = re.sub('[a-zA-Z]', '', value)
        floatValue = specialCharCheck(value)
        floatValue = floatValue * 0.3048
        return round(floatValue, 2)
    else:
        value = re.sub('[a-zA-Z]', '', value)
        floatValue = specialCharCheck(value)
        floatValue = floatValue * 0.01
        return round(floatValue, 2)

survey['BodyHeight'] = survey['BodyHeight'].apply(heightCheck)
```

Figure 71: Function Transform BodyHeight

```
# Creating the "BMI" Attribute
survey['BMI'] = np.where(
    (survey['BodyWeight'].isnull()) | (survey['BodyHeight'].isnull()),
    np.nan,
    (survey['BodyWeight'] / (survey['BodyHeight'] ** 2)).round(2)
)
```

Figure 72: BMI Creation

All the code snippets above show the pre-processing for "BodyWeight" and "BodyHeight" to create a new attribute "BMI". First, both attributes' data is turned into lowercase, so the measurement unit checking can be done easier. Next, both attributes are changed into a standard measurement unit, where BodyWeight is standardized into Kilogram (kg), while BodyHeight is standardized into Meters (m). Finally, BMI can be created using the formula of:

$$BMI = \frac{BodyWeight}{BodyHeight^2} / BMI = \frac{kg}{m^2}$$

The BMI will be saved as float, with 2 numbers after decimal point saved.

```
def healthCheck(value):
    if isinstance(value, float):
        return float(value)
    elif isinstance(value, str):
        value = re.sub('[a-zA-Z]', '', value)
        floatValue = specialCharCheck(value)
        if floatValue is not None:
            return round(floatValue, 2)
        else:
            return 0.0

survey['PhysicalHealth'] = survey['PhysicalHealth'].apply(healthCheck)
survey['MentalHealth'] = survey['MentalHealth'].apply(healthCheck)
```

Figure 73: Function Transform Health Data

Next attributes are "PhysicalHealth" and "MentalHealth", where some values contain strings values. The pre-processing step cover removing all alphabetical and special characters from both attributes, then replacing all remaining numeric values into a float, so both attributes become numerical.

```
survey['Diabetic'] = survey['Diabetic'].replace('No, but borderline Diabetes', 'No, borderline diabetes')
```

Figure 74: Diabetic Data Standardization

For "Diabetic" attribute, answer standardization is applied, so the value "No, but borderline Diabetes" will be changed into "No, borderline diabetes".

```
survey['SleepTime'] = survey['SleepTime'].str.replace(r'[a-zA-Z]', '')
survey['SleepTime'] = survey['SleepTime'].apply(specialCharCheck)
```

Figure 75: SleepTime Data Standardization

Finally, for "SleepTime", the same function for "BodyWeight" is applied to remove all inconsistency, where all alphabetical and special characters are removed, then the numeric value is changed into float value. After all of this, the survey data will be checked again to make sure no inconsistency remains.

```
# Drop BodyWeight and BodyHeight Column
survey.drop(['BodyWeight', 'BodyHeight'], axis=1, inplace=True)
✓ 0.0s

# Rearrange columns in a specific order
survey = survey[['HeartDisease', 'BMI', 'Smoking', 'AlcoholDrinking', 'Stroke',
                 'PhysicalHealth', 'MentalHealth', 'DiffWalking', 'Sex', 'AgeCategory', 'Race',
                 'Diabetic', 'PhysicalActivity', 'GenHealth', 'SleepTime', 'Asthma', 'KidneyDisease', 'SkinCancer']]
✓ 0.0s

# Dataframe Concatenation
fullData = pd.concat([survey, data])
✓ 0.0s
```

Figure 76: Survey Data Merging to Main Data

Before the merging of the survey data into the main data, there are multiple steps must be taken. First, "BodyWeight" and "BodyHeight" will be dropped since it is not needed anymore. Next, the columns need to be rearranged so it follows the main data arrangement. Finally, both data is concatenated into one, named "fullData". Now that the data has been combined into one, the rest of the pre-processing steps can be done.

7.3.3 Main Data – Data Pre-processing

```
fullData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 319892 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      319892 non-null  object
 1   BMI               319815 non-null  float64
 2   Smoking           319892 non-null  object
 3   AlcoholDrinking   319892 non-null  object
 4   Stroke            319892 non-null  object
 5   PhysicalHealth    319889 non-null  float64
 6   MentalHealth      319889 non-null  float64
 7   DiffWalking       319891 non-null  object
 8   Sex               319892 non-null  object
 9   AgeCategory       319892 non-null  object
 10  Race              319892 non-null  object
 11  Diabetic          319892 non-null  object
 12  PhysicalActivity  319892 non-null  object
 13  GenHealth         319892 non-null  object
 14  SleepTime         319871 non-null  float64
 15  Asthma            319892 non-null  object
 16  KidneyDisease     319890 non-null  object
 17  SkinCancer        319892 non-null  object
dtypes: float64(4), object(14)
memory usage: 46.4+ MB
```

Figure 77: Data Viewing

After the data merging, all data checking must be done to decide on what pre-processing steps need to be taken. From the result of the info above, there are some missing data that can be imputed.

Check for Duplicated Data

```
fullData.duplicated().sum()
```

```
18080
```

Dropping Duplicates

```
fullData.drop_duplicates(inplace=True)
```

Figure 78: Duplicates Dropping

Next, part of removing unused data is to drop all duplicate data inside the dataframe. This is to prevent data leakage, where the same identical data is found in both training and testing dataset,

which can cause biased performance of the model, thus resulting in poor performance (Chorev, 2023). From the above code snippet, all 18.080 duplicate code is removed.

```
# Removing the spacing for AgeCategory
def spacingRemover(string):
    if '-' in string:
        string = string.replace(' ', '')
    return string

fullData['AgeCategory'] = fullData['AgeCategory'].apply(spacingRemover)


# Replacing 'Caucassian' into 'White', 'African' into 'Other', and 'Very Good' becomes 'Very good'
fullData['Race'] = fullData['Race'].replace('Caucassian', 'White')
fullData['Race'] = fullData['Race'].replace('African', 'Other')
fullData['GenHealth'] = fullData['GenHealth'].replace('Very Good', 'Very good')
fullData['GenHealth'] = fullData['GenHealth'].replace('Normal', 'Fair')
fullData['AgeCategory'] = fullData['AgeCategory'].replace('19 and Under', '18-24')
fullData['AgeCategory'] = fullData['AgeCategory'].replace('20-24', '18-24')
```

Figure 79: Inconsistencies Fixing

The next step covers the data standardization, where certain data has different wording being used, even though it covers the same meaning. First, all unique values of each attribute are printed to investigate all inconsistency present in the dataframe. After that, each inconsistency can be dealt with on by one. In the above code snippet, function spacingRemover is used to remove all spacing before and after the hypen (18 – 24 becomes 18-24) so the data become one. In attribute “Race”, “GenHealth”, and “AgeCategory” there are some inconsistencies that can be fixed manually.

```
# Function to calculate missing values by column
def missingValuesTable(df):
    # Total missing values
    mis_val = df.isnull().sum()

    # Percentage of missing values
    mis_val_percent = 100 * df.isnull().sum() / len(df)

    # Make a table with the results
    mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)

    # Rename the columns
    mis_val_table_ren_columns = mis_val_table.rename(
    columns = {0 : 'Missing Values', 1 : '% of Total Values'})

    # Sort the table by percentage of missing descending
    mis_val_table_ren_columns = mis_val_table_ren_columns[
        mis_val_table_ren_columns.iloc[:,1] != 0].sort_values(
    '% of Total Values', ascending=False).round(1)

    # Print some summary information
    print ("Your selected dataframe has " + str(df.shape[1]) + " columns.\n"
           "There are " + str(mis_val_table_ren_columns.shape[0]) +
            " columns that have missing values.")

    # Return the dataframe with missing information
    return mis_val_table_ren_columns

missingValuesTable(fullData)
```

```
Your selected dataframe has 18 columns.
There are 6 columns that have missing values.
```

	Missing Values	% of Total Values
BMI	75	0.0
SleepTime	20	0.0
PhysicalHealth	3	0.0
MentalHealth	3	0.0
KidneyDisease	2	0.0
DiffWalking	1	0.0

Figure 80: Missing Data Overview

The next step is to check for any missing data. Using the function “missingValuesTable”, it will print out all attributes that contains missing value, and the amount of the missing values. From the result above, it shows that six attributes have missing values. To fix it, imputation can be done, using Average or Mean for numerical value, and Most Frequent or Mode for categorical value.

```
# Replacing the missing numerical value with Mean (Average)
mean_bmi = fullData['BMI'].mean()
fullData['BMI'].fillna(float(f'{mean_bmi:.2f}'), inplace=True)

#For MentalHealth, PhysicalHealth, and SleepTime, the value should not have any decimal points
fullData['PhysicalHealth'].fillna(fullData['PhysicalHealth'].mean().round(0), inplace=True)
fullData['MentalHealth'].fillna(fullData['MentalHealth'].mean().round(0), inplace=True)
fullData['SleepTime'].fillna(fullData['SleepTime'].mean().round(0), inplace=True)
missingValuesTable(fullData)
```

Figure 81: Missing Numerical Data Imputation

For “BMI”, “PhysicalHealth”, “MentalHealth”, and “SleepTime”, since all these attributes has numerical value, then missing value can be imputed using Mean or the average of the whole attribute. For “BMI”, the average can have two decimal points, while for the rest, since it is counted in days, then it is set to no decimal point.

```
# Replacing the missing categorical value with Mode (Most Frequent), first, value counts can be used to see the mode
print(fullData["KidneyDisease"].value_counts())
print("\n")
print(fullData["DiffWalking"].value_counts())
print("\n")
```

```
KidneyDisease
No     290031
Yes     11779
Name: count, dtype: int64


DiffWalking
No     257455
Yes     44356
Name: count, dtype: int64
```

```
fullData['KidneyDisease'].fillna('No', inplace=True)
fullData['DiffWalking'].fillna('No', inplace=True)

missingValuesTable(fullData)
```

```
Your selected dataframe has 18 columns.
There are 0 columns that have missing values.
```

	Missing Values	% of Total Values

Figure 82: Missing Categorical Data Imputation

Finally, for the categorical value attributes, the missing values can be imputed using Most Frequent or Mode. First, the code above prints out all unique values for each attribute. Now from the output, the value “No” is the most frequent for both “KidneyDisease” and “DiffWalking”. With that, both attributes’ missing values are imputed with “No”.

The next step is to prepare the data for Model Building process. Since most the data in the dataframe are in categorical form, Data encoding must be utilized. Data Encoding itself refers to converting the categorical variables into numerical variables, so it can be fitted into the machine learning models (Khan, 2022). First, to determine which encoding techniques to be used, checking the unique values in each variable is important.

```
# Print all Columns that is Object, to show the unique values
for column in fullData.columns:
    if fullData[column].dtype == 'object':
        unique_values = fullData[column].unique()
        print(f"Unique values for column '{column}': {unique_values}")
```

✓ 0.1s

```
Unique values for column 'HeartDisease': ['No' 'Yes']
Unique values for column 'Smoking': ['No' 'Yes']
Unique values for column 'AlcoholDrinking': ['No' 'Yes']
Unique values for column 'Stroke': ['No' 'Yes']
Unique values for column 'DiffWalking': ['No' 'Yes']
Unique values for column 'Sex': ['Female' 'Male']
Unique values for column 'AgeCategory': ['18-24' '55-59' '45-49' '50-54' '40-44' '35-39' '30-34' '25-29'
 '80 or older' '65-69' '75-79' '70-74' '60-64']
Unique values for column 'Race': ['Asian' 'White' 'Other' 'Black' 'American Indian/Alaskan Native'
 'Hispanic']
Unique values for column 'Diabetic': ['No' 'Yes' 'No, borderline diabetes' 'Yes (during pregnancy)']
Unique values for column 'PhysicalActivity': ['No' 'Yes']
Unique values for column 'GenHealth': ['Fair' 'Good' 'Very good' 'Poor' 'Excellent']
Unique values for column 'Asthma': ['No' 'Yes']
Unique values for column 'KidneyDisease': ['No' 'Yes']
Unique values for column 'SkinCancer': ['No' 'Yes']
```

Figure 83: Initial Encoding Viewing

From the snippet above, most of the categorical variables only consist of two choices. For these variables, picking label encoding is sufficient. Meanwhile, for those variables that contains more than 2 unique values, one hot encoding is utilized.

```
label=LabelEncoder()
fullData['HeartDisease']=label.fit_transform(fullData['HeartDisease']).astype('uint8')
fullData['Smoking']=label.fit_transform(fullData['Smoking']).astype('uint8')
fullData['AlcoholDrinking']=label.fit_transform(fullData['AlcoholDrinking']).astype('uint8')
fullData['Stroke']=label.fit_transform(fullData['Stroke']).astype('uint8')
fullData['DiffWalking']=label.fit_transform(fullData['DiffWalking']).astype('uint8')
fullData['Sex']=label.fit_transform(fullData['Sex']).astype('uint8')
fullData['PhysicalActivity']=label.fit_transform(fullData['PhysicalActivity']).astype('uint8')
fullData['Asthma']=label.fit_transform(fullData['Asthma']).astype('uint8')
fullData['KidneyDisease']=label.fit_transform(fullData['KidneyDisease']).astype('uint8')
fullData['SkinCancer']=label.fit_transform(fullData['SkinCancer']).astype('uint8')
```

Figure 84: Label Encoding

The code snippet above shows the process of encoding all two-choices variables using label encoding. The means that one of the unique values will be assigned into zero (0), while the other one will be one (1). In this case, No is assigned as zero, while Yes is assigned as one. In the case of variable “Sex”, female is assigned as zero, while male as one.

```
# The Old data will be dropped after Data Visualization

dummy=pd.get_dummies(fullData['AgeCategory'],prefix='AgeCategory').astype(int)
fullData=pd.concat([fullData,dummy],axis=1)

dummy=pd.get_dummies(fullData['Race'],prefix='Race').astype(int)
fullData=pd.concat([fullData,dummy],axis=1)

dummy=pd.get_dummies(fullData['Diabetic'],prefix='Diabetic').astype(int)
fullData=pd.concat([fullData,dummy],axis=1)

dummy=pd.get_dummies(fullData['GenHealth'],prefix='GenHealth').astype(int)
fullData=pd.concat([fullData,dummy],axis=1)
```

✓ 0.1s

Figure 85: One Hot Encoding

The next encoding step is to use one hot encoding to transform all categorical variables that contains more than two unique values. One hot encoding will create a new column in the dataframe for each unique value for each variable. After that, it will assign the value one (1) into the column that represents the data, while assigning the rest with zero (0). For example, in 'AgeCategory' variable, if the original data is '25-29', then after one hot encoding, for that specific data, the column 'AgeCategory_25-29' will contain one, while the rest of "AgeCategory" will be zero.

After One hot encoding has been conducted, the old attributes can be deleted from the dataframe, but the removal will be done after the data visualization steps, since using the old attributes are simpler than using the one hot encoded one.

7.4 Data Visualization

Data Visualization refers to a visualization of the data using graphical representation, such as charts, graphs, and maps. Data visualization provides an easier readability and understanding towards the patterns inside the data (Tableau, 2023). Data visualization can also be used to show the correlation between two attributes and be used in a way to help the developer to decide on the next step of data processing.

Another function that Data visualization has is it can provide the non-technical audiences with insights and information from the graphics created. By this, explaining the important details from the project's outcome can be easier. Due to this, Data Visualization will also become one of the project's outcomes.

7.4.1 Visualization for Data Analysis and Deployment

In this visualization part, there are two (2) mains visualizations results that are going to be used to do data analysis from the dataset. All the visualizations will be Bivariate, meaning that it will only shows the correlation between two variables, one of which will be on the target variable, which is "HeartDisease". The goal these visualizations is to get a graphical representation of the correlation between the target variable and one of the feature variables.

```
import numpy as np

heart_disease_counts = fullData.groupby(["AgeCategory", "HeartDisease"]).size().unstack()

desired_order = ["18-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-59", "60-64", "65-69", "70-74", "75-79", "80 or older"]

heart_disease_counts_sorted = heart_disease_counts.reindex(desired_order)

fig, ax = plt.subplots(figsize=(24, 12))

bar_width = 0.35
index = np.arange(len(desired_order))

yes_bars = ax.bar(index, heart_disease_counts_sorted[1], bar_width, label='Heart Disease', color='#E67373')
no_bars = ax.bar(index + bar_width, heart_disease_counts_sorted[0], bar_width, label='No Heart Disease', color='#7373E6')

ax.set_xlabel("Age Category\n(Age Range)")
ax.set_ylabel("Count")
ax.set_title("Grouped Bar Chart of Heart Disease by Age Category")
ax.set_xticks(index + bar_width / 2)
ax.set_xticklabels(desired_order)

ax.legend()

for rect in yes_bars + no_bars:
    height = rect.get_height()
    ax.annotate(f"{height}", xy=(rect.get_x() + rect.get_width() / 2, height),
                xytext=(0, 3), textcoords="offset points",
                ha='center', va='bottom')

plt.show()
```

Figure 86: Bar Chart Visualization Code

The code snippet above shows the code to create a grouped bar chart visualization. This code is used for the correlation between the target variable and a categorical variable.

```
sns.set(style="darkgrid")
sns.set(rc={'figure.figsize': (16, 8)})

# Creating a figure composed of 3 matplotlib.Axes objects
f, (ax_box1, ax_box2, ax_hist) = plt.subplots(3, sharex=True, gridspec_kw={"height_ratios": (.15, .15, .85)})

# Extract 'BMI' values for heart disease and non-heart disease
heartDiseaseBMI = fullData.loc[fullData['HeartDisease'] == 1, "BMI"]
noHeartDiseaseBMI = fullData.loc[fullData['HeartDisease'] == 0, "BMI"]

# Assigning a graph to each ax
sns.boxplot(x=heartDiseaseBMI, ax=ax_box1, color="#E67373")
ax_hist.hist(heartDiseaseBMI, color="#E67373", alpha=0.5, label="Heart Disease", density=True)
sns.boxplot(x=noHeartDiseaseBMI, ax=ax_box2, color='#7373E6')
ax_hist.hist(noHeartDiseaseBMI, color='#7373E6', alpha=0.5, label="No Heart Disease", density=True)

# Remove x-axis name for the boxplots
ax_box1.set(xlabel='')
ax_box2.set(xlabel='')

plt.legend(title='', loc=2, labels=['Heart Disease', 'No Heart Disease'], bbox_to_anchor=(1.02, 1), borderaxespad=0.)
plt.show()
```

Figure 87: Boxplot Histogram Visualization Code

The code snippet above shows the code to create a boxplot and histogram visualization. This code is used to show the correlation between the target variable and a numerical variable.

7.4.1.1 Correlation between ‘HeartDisease’ and ‘AgeCategory’ variable

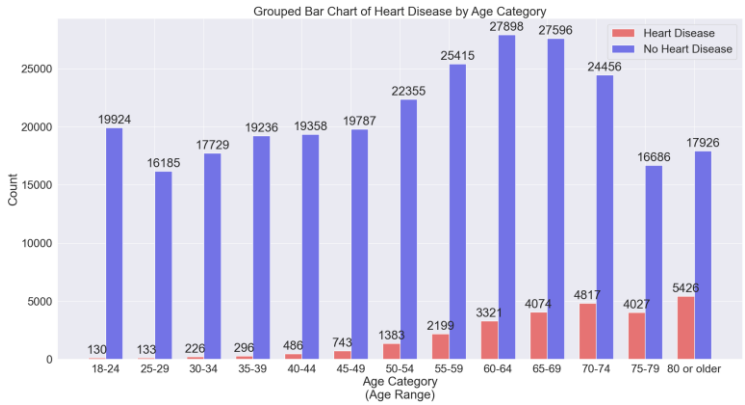


Figure 88: Data Visualization - HeartDisease and AgeCategory

The above visualization shows the correlation between “HeartDisease” and “AgeCategory”. From the generated visualization, it shows that Heart Disease become more common as a person ages, shown by the increase in heart disease case amount, even though the decreased number of samples in of older people. The above graph also shows that even younger audiences still have a risk to suffer from a Heart Disease.

Age Category	**The risk of Heart Disease in Percentage (%)**
18-24	0.64%
25-29	0.81%
30-34	1.25%
35-39	1.51%
40-44	2.44%
45-49	3.61%
50-54	5.82%
55-59	7.96%
60-64	10.63%
65-69	12.86%
70-74	16.45%
75-79	19.44%
80 or older	23.23%

Table 7: HeartDisease and Agecategory Correlation

The percentage table above shows an upward trend in the risk of heart disease, as a person ages. It shows that age is one of the main factors in the risk of heart disease.

7.4.1.2 Correlation between ‘HeartDisease’ and ‘Diabetic variable

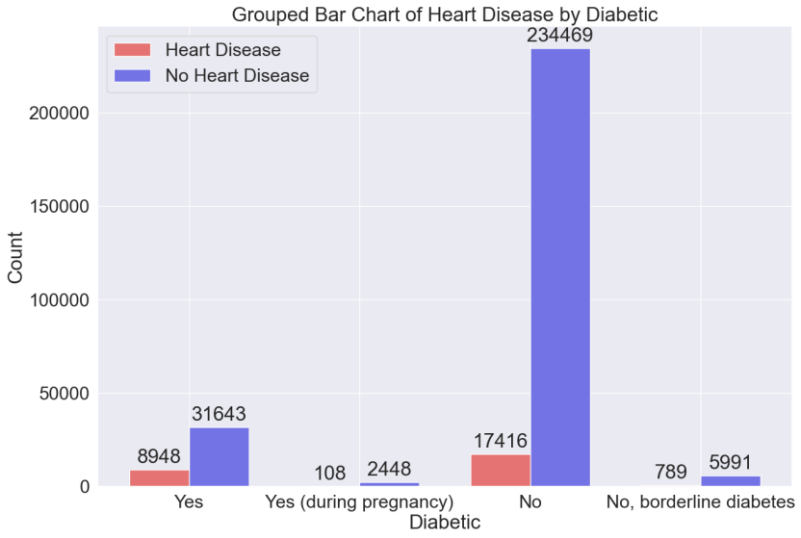


Figure 89: Data Visualization - HeartDisease and Diabetic

The above visualization shows the correlation between “HeartDisease” and “Diabetic” variable. It shows that people with people that has diabetes is prone to suffer from a heart disease.

Diabetic	The risk of Heart Disease in Percentage (%)
Yes	22.04%
Yes (during pregnancy)	4.22%
No	6.91%
No, borderline diabetes	11.63%

Table 8: HeartDisease and Diabetic Correlation

This table shows that through the number of the “Yes” cases relative to the sample amount for each answer, where 22.04% of people with diabetes also suffer from heart disease, compared to 6.91% from people without diabetes. It also shows that people with prediabetes or borderline diabetes have a higher percentage of risk of heart disease.

7.4.1.3 Correlation between ‘HeartDisease’ and ‘Race’ variable

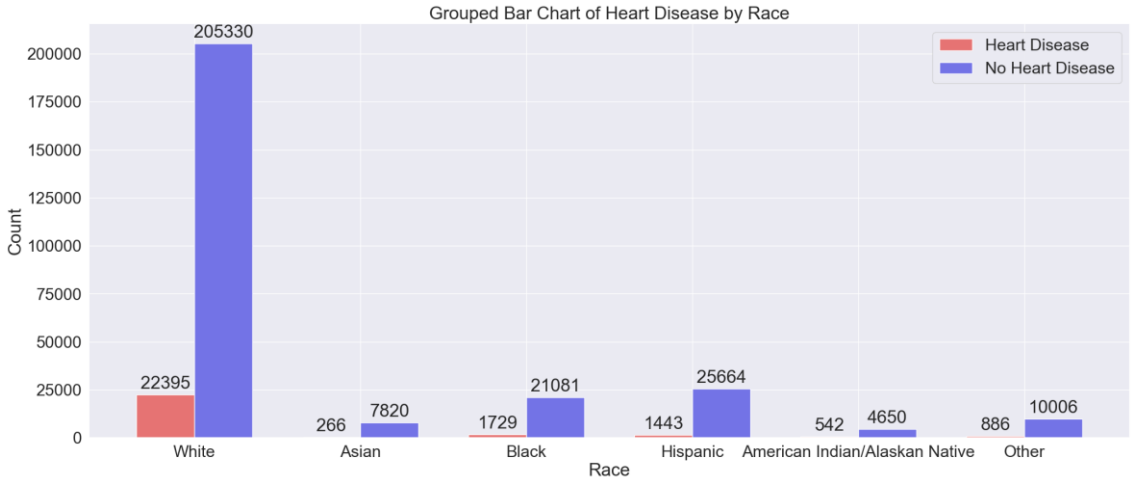


Figure 90: Data Visualization - HeartDisease and Race

The above grouped bar chart shows the correlation between “HeartDisease” with “Race” variable. It shows that majority of the data is from White or Caucasian race. Even though the distribution is heavily imbalanced, a conclusion still can be pulled:

Race	The risk of Heart Disease in Percentage (%)
White	9.83%
Asian	3.28%
Black	7.58%
Hispanic	5.32%
American Indian/Alaskan Native	10.43%
Other	8.13%

Table 9: HeartDisease and Race Correlation

Race can have an effect towards the risk of heart disease. According to studies conducted by Javed et al. (2022), Black adults have a higher chance of heart disease risk factors, such as hypertension and obesity relative to white adults, while American Indian adults are 1.5 times more likely to be diagnosed with heart disease. Since the studies is done in America, it can also have an effect due to the racial/ethnic problem that may have been established. On the other hand, a study by Sasayama (2008) shows that Asian have a lower risk factors of heart disease compared to Western people. This shows that Race can have an effect towards the risk of Heart Disease, especially due to culture, lifestyle choices, ancestry and social problems.

7.4.1.4 Correlation between ‘HeartDisease’ and ‘GenHealth’ variable

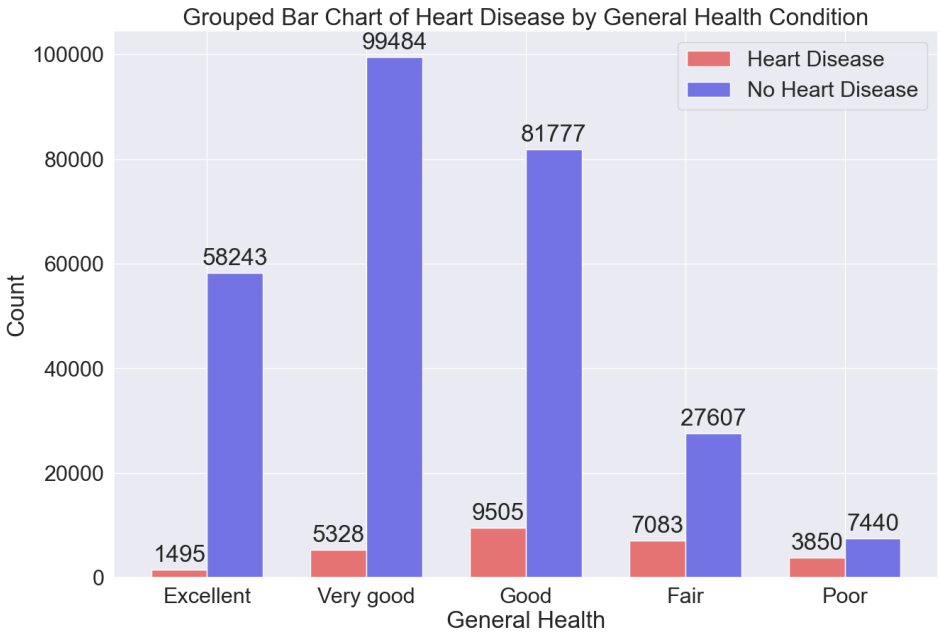


Figure 91: Data Visualization - HeartDisease and GenHealth

The above group bar chart shows the correlation between “HeartDisease” and “GenHealth” variable. From the graph above, it shows a pattern as the general health of a person gets lower, the higher their chance of suffering from a heart disease.

General Health Status	The risk of Heart Disease in Percentage (%)
Excellent	2.5%
Very Good	5.08%
Good	10.41%
Fair	20.41%
Poor	34.1%

Table 10: HeartDisease and GenHealth Correlation

From the table above, it shows an upward trend of the risk of heart disease, based on a person’s general health. It shows that general health can be one of the main indicators used by a person to measure themselves, especially towards the risk of a heart disease.

7.4.1.5 Correlation between ‘HeartDisease’ and ‘Smoking’ variable

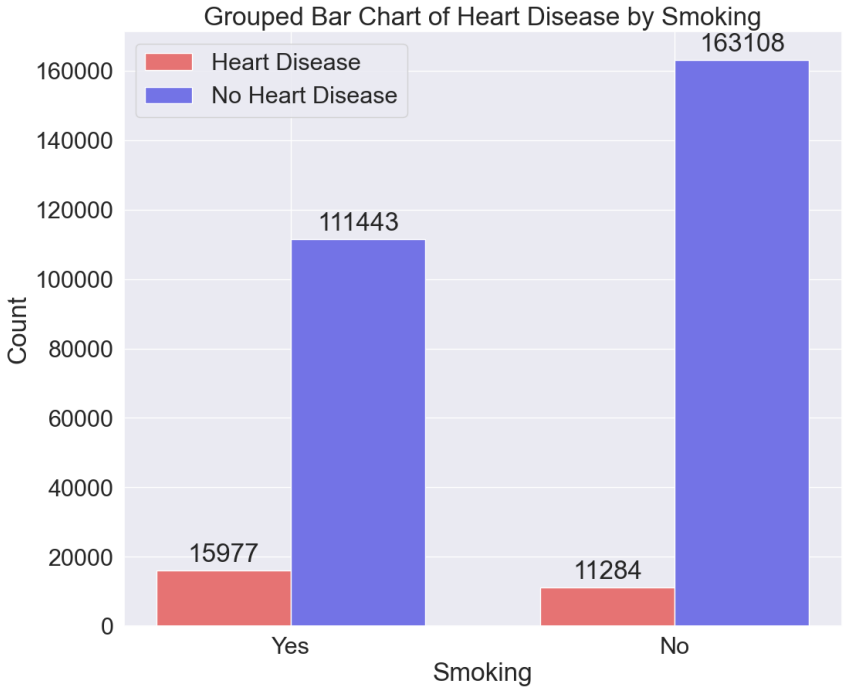


Figure 92: Data Visualization - HeartDisease and Smoking

Smoking	The risk of Heart Disease in Percentage (%)
Yes	12.53%
No	6.47%

Table 11: HeartDisease and Smoking Correlation

The above bar chart and table shows the correlation between “HeartDisease” and “Smoking” variable. It shows that people that smokes have a higher chance of heart disease, with the risk percentage almost doubles from 6.47% becoming 12.53%. This is due to smoking causes plaque build-up in the arteries, thus becoming one of the main risk factors of Heart Disease (NIH, 2022).

7.4.1.6 Correlation between ‘HeartDisease’ and ‘KidneyDisease’ variable

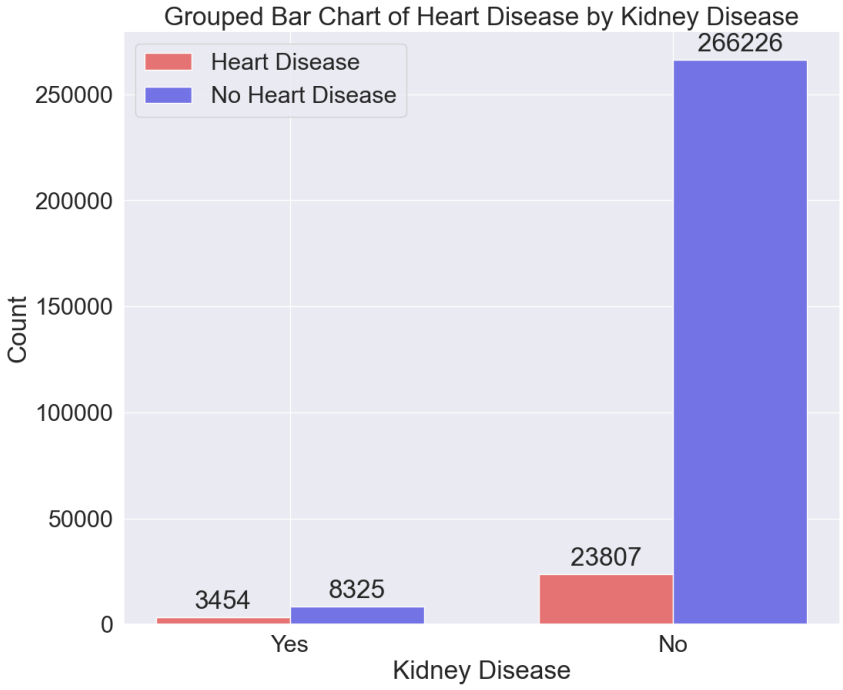


Figure 93: Data Visualization - HeartDisease and KidneyDisease

Kidney Disease	The risk of Heart Disease in Percentage (%)
Yes	29.32%
No	8.2%

Table 12: HeartDisease and KidneyDisease Correlation

The above bar chart and table shows the correlation between “HeartDisease” with “KidneyDisease” Variable. From the results shown above, it can be concluded that people with kidney disease may have a higher risk of suffering from a heart disease. This can be due to Chronic Kidney Disease also boost the risk factors of Heart Disease, such as diabetes, hypertension, and inflammation (Sarnak et al., 2019). Due keep in mind that the risk percentage and the visualization can be biased due to the lack of sample and very imbalanced distribution.

7.4.1.7 Correlation between ‘HeartDisease’ and ‘AlcoholDrinking’ variable

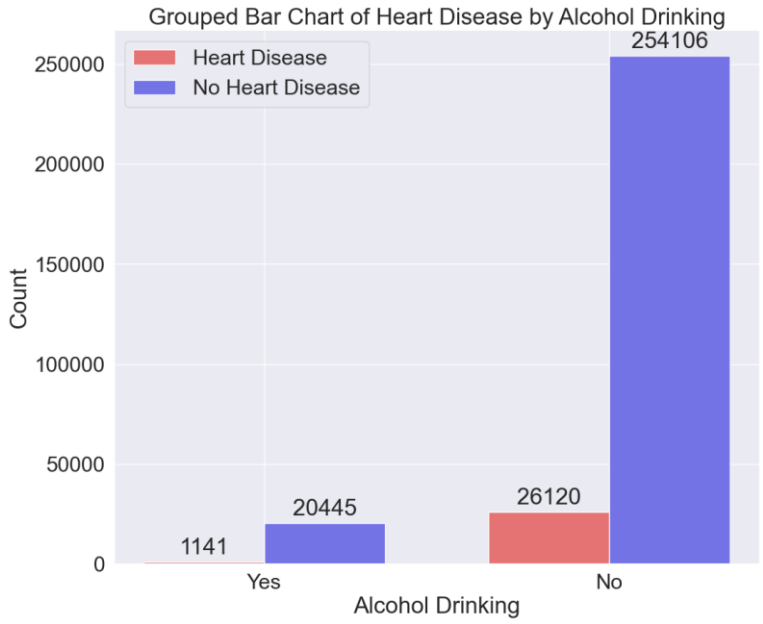


Figure 94: Data Visualization - HeartDisease and AlcoholDrinking

Alcohol Drinking	The risk of Heart Disease in Percentage (%)
Yes	5.28%
No	9.32%

Table 13: HeartDisease and AlcoholDrinking Correlation

The above bar chart and table shows the correlation between “HeartDisease” and “AlcoholDrinking” variable. Here shows that a heavy alcohol consumer may have a lower risk percentage compared to others. But due keep in mind that the data is highly imbalanced, the distribution of the class is very much biased towards no alcohol drinker, thus it can be concluded that not enough data is collected to said that the result is accurate. But Heavy Alcohol Consumption can cause a long-term effect on the kidney and liver, which can indirectly become a risk factor towards heart disease (Alcohol Think Again, 2023).

7.4.1.8 Correlation between ‘HeartDisease’ and ‘Sex’ variable

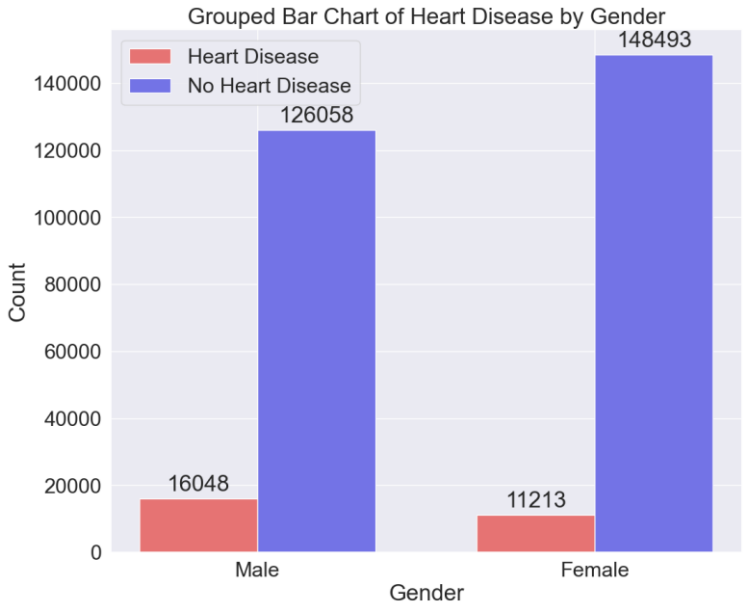


Figure 95: Data Visualization - HeartDisease and Sex

Sex	The risk of Heart Disease in Percentage (%)
Male	11.29%
Female	7.02%

Table 14: HeartDisease and Sex Correlation

The bar chart and the table above show the correlation between “HeartDisease” and “Sex” variable. It shows that Male have a higher risk of heart disease compared to Female. This can be due to hormones factors, where female can produce Estrogen and Progesterone that can boost blood vessel health, while men only relying on Testosterone, where the production starts to decrease after age 40. Other factors such as cholesterol buildup, differences in body fat location, and stress can also have an affect towards the risk (Aduli, 2023).

7.4.1.9 Correlation between ‘HeartDisease’ and ‘Stroke’ variable

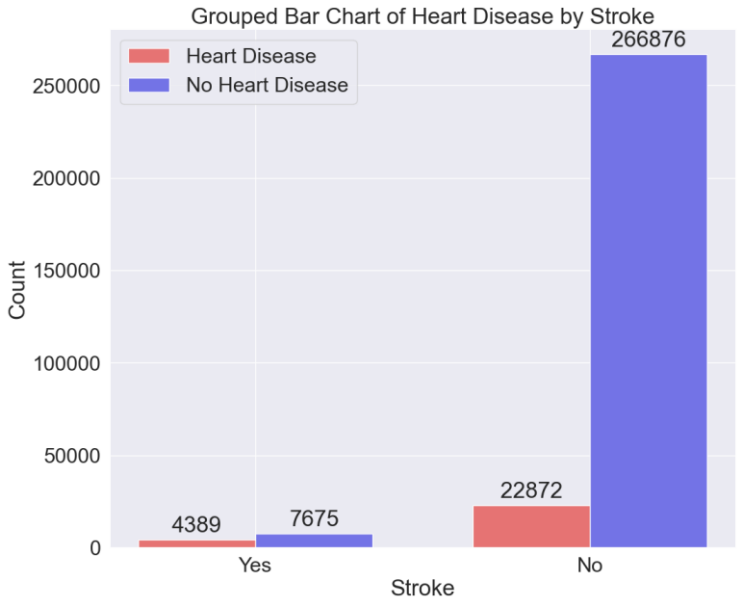


Figure 96: Data Visualization - HeartDisease and Stroke

Stroke	The risk of Heart Disease in Percentage (%)
Yes	36.38%
No	7.89%

Table 15: HeartDisease and Stroke Correlation

The bar chart and the table above show the correlation between “HeartDisease” and “Stroke” variable. Even though the data is highly imbalanced with bias towards the people with no stroke, it still can be concluded that people with stroke also have a higher risk of heart disease, almost 5 times higher based on the data visualization. This can be due to the similarity of cause between stroke and heart disease, where a blockage of blood vessels and arteries causes these diseases. Thus, people with stroke can lead to heart disease, and vice versa (Fogoros, 2022).

7.4.1.10 Correlation between ‘HeartDisease’ and ‘Asthma’ variable

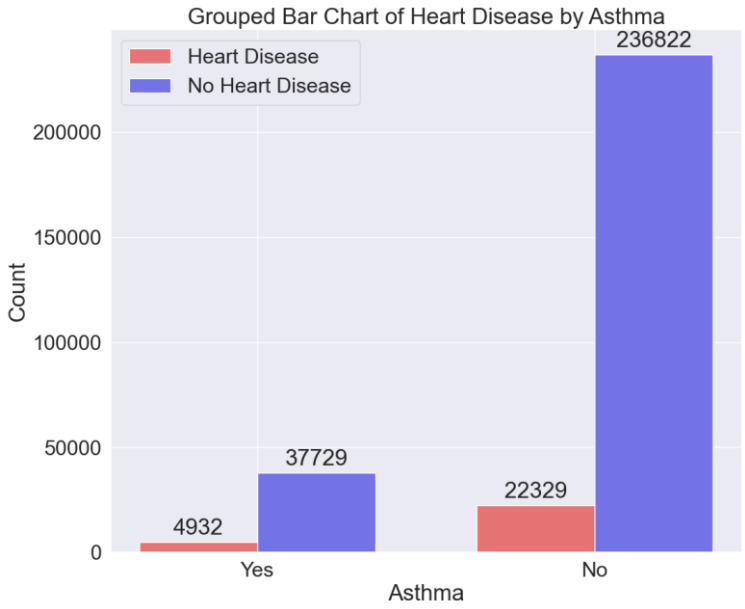


Figure 97: Data Visualization - HeartDisease and Asthma

Asthma	The risk of Heart Disease in Percentage (%)
Yes	11.56%
No	8.61%

Table 16: HeartDisease and Asthma Correlation

The bar chart and the table above show the correlation between “HeartDisease” and “Asthma” variable. Even though the data bias towards the people without Asthma, it can be concluded that people with Asthma has a higher risk of heart disease compared to people without asthma. This can be due to the inflammation caused by long-term asthma can also cause damages towards blood vessels, thus increasing the risk of high blood pressure (Taylor, 2023).

7.4.1.11 Correlation between ‘HeartDisease’ and ‘SkinCancer’ variable

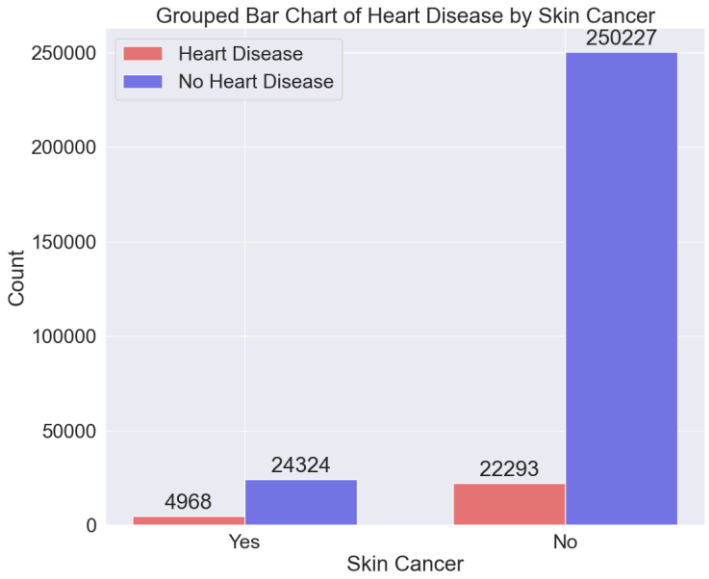


Figure 98: Data Visualization - HeartDisease and SkinCancer

Skin Cancer	The risk of Heart Disease in Percentage (%)
Yes	16.96%
No	8.18%

Table 17: HeartDisease and SkinCancer Correlation

The bar chart and the table above show the correlation between “HeartDisease” and “SkinCancer” variable. Though the data is very imbalanced and biased, an insight still can be pulled that people with Skin Cancer has a higher risk of Heart Disease. This can be caused due to Melanoma, the malignant tumour associated with skin cancer, can spread through metastatis to all internal organ, including the heart, which can cause heart diseases, such as arrhythmia (irregular heartbeat) and heart failure (Babar et al., 2020).

7.4.1.12 Correlation between ‘HeartDisease’ and ‘PhysicalActivity’ variable

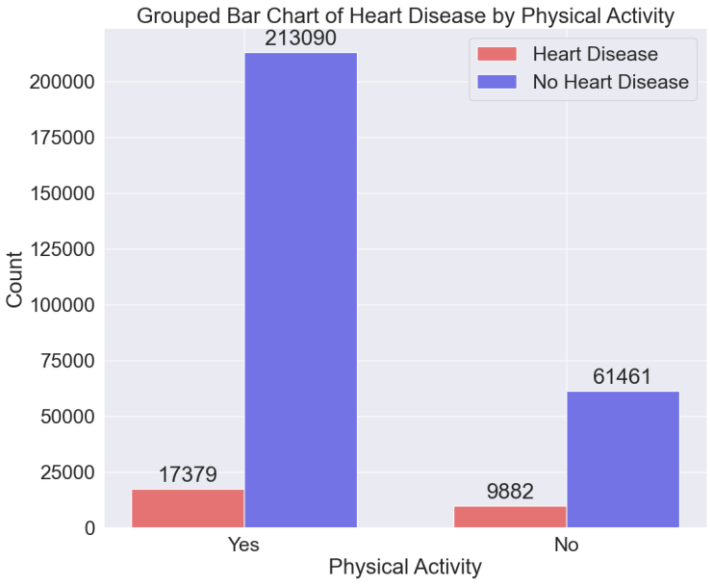


Figure 99: Data Visualization - HeartDisease and PhysicalActivity

Physical Activity	The risk of Heart Disease in Percentage (%)
Yes	7.54%
No	13.85%

Table 18: HeartDisease and PhysicalActivity Correlation

The bar chart and the table above show the correlation between “HeartDisease” and “PhysicalActivity” variable. Even though the data is imbalanced, it shows that people that do Physical Activity, such as sports or exercise has a lower chance of suffering from heart disease, since being physically active has been proven to maintain a healthy weight, thus reducing the chance of high blood pressure (NHS, 2021).

7.4.1.13 Correlation between ‘HeartDisease’ and ‘DiffWalking’ variable

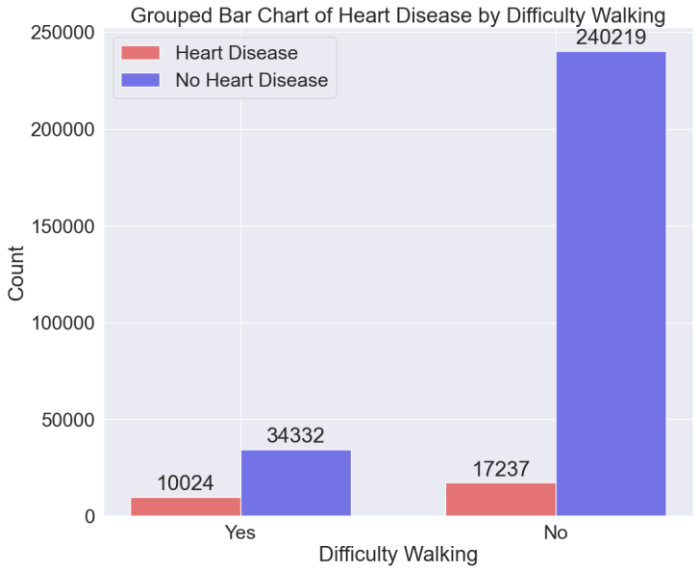


Figure 100: Data Visualization - HeartDisease and DiffWalking

Difficulty Walking	The risk of Heart Disease in Percentage (%)
Yes	22.59%
No	6.69%

Table 18: HeartDisease and DiffWalking Correlation

The bar chart and the table above show the correlation between “HeartDisease” and “DiffWalking” variable. Even though the data distribution is imbalanced, it still shows that people with difficulty walking or climbing stairs has a higher risk of heart disease. This can be correlated with the previous visualization between “HeartDisease” and “PhysicalActivity”, since people with difficulty walking will have a harder time to do physical activity such as exercising and sports.

7.4.1.14 Correlation between ‘HeartDisease’ and ‘BMI’ variable

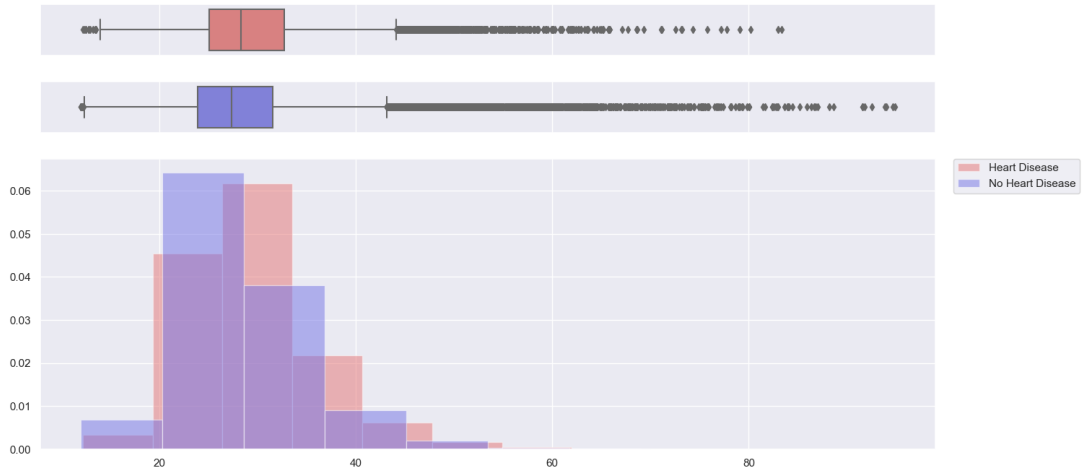


Figure 101: Data Visualization - HeartDisease and BMI

The boxplot and the histogram above show the correlation between “HeartDisease” and “BMI” variable. From both visualizations, it shows that people with higher BMI have a higher risk of heart disease overall. It shows from the distribution of the histogram, as the BMI increases, the amount of heart disease also increases. From the boxplot, it also shows that the average BMI for people with heart disease is much higher compared to the people without heart disease. This is due to higher BMI increases the risk of high blood pressure, which can lead to a heart disease.

7.4.1.15 Correlation between ‘HeartDisease’ and ‘SleepTime’ variable

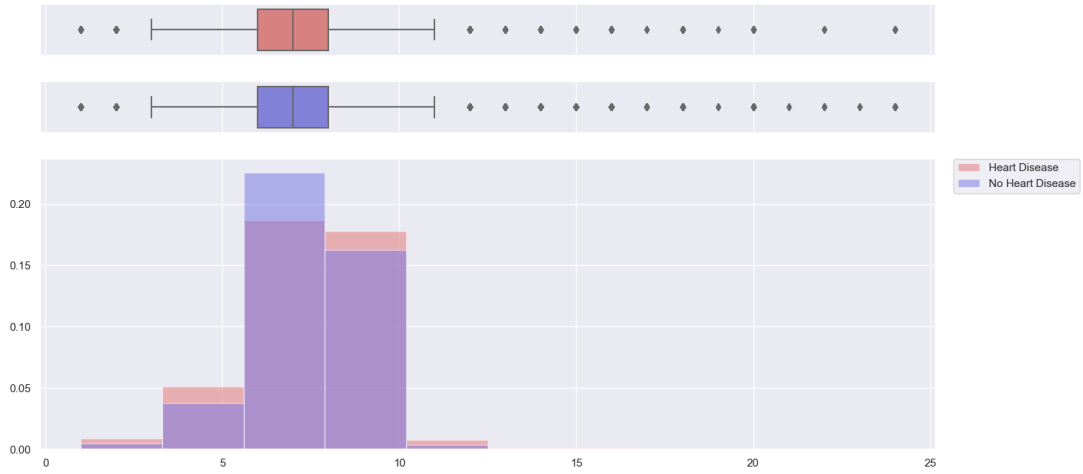


Figure 102: Data Visualization - HeartDisease and SleepTime

The boxplot and the histogram above show the correlation between “HeartDisease” and “SleepTime” variable. From both visualizations, it shows that sleep time has a low effect towards heart disease, since the data distribution is similar for both people with heart disease and people without heart disease. But, sleeping disorder such as insomnia is linked to high blood pressure, high stress level, and unhealthier lifestyle over time, thus can increase the risk of heart disease also in the long-term (Nagai et al., 2010).

7.4.1.16 Correlation between ‘HeartDisease’ and ‘PhysicalHealth’ variable

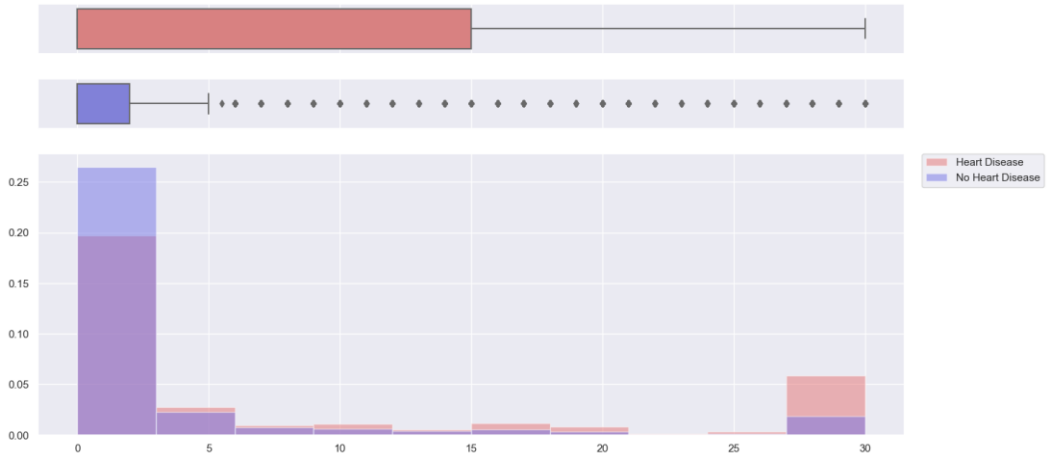


Figure 103: Data Visualization - HeartDisease and PhysicalHealth

The boxplot and the histogram above show the correlation between “HeartDisease” and “PhysicalHealth” variable. From the above visualizations, it shows that physical health has a high effect towards the risk of heart disease, as the physical health of a person worsen, their risk of heart disease also increases. This is shown through the amount of physical health problems a person has in 30 days. As the number of days increase, the risk of heart disease also increases.

7.4.1.17 Correlation between ‘HeartDisease’ and ‘MentalHealth’ variable

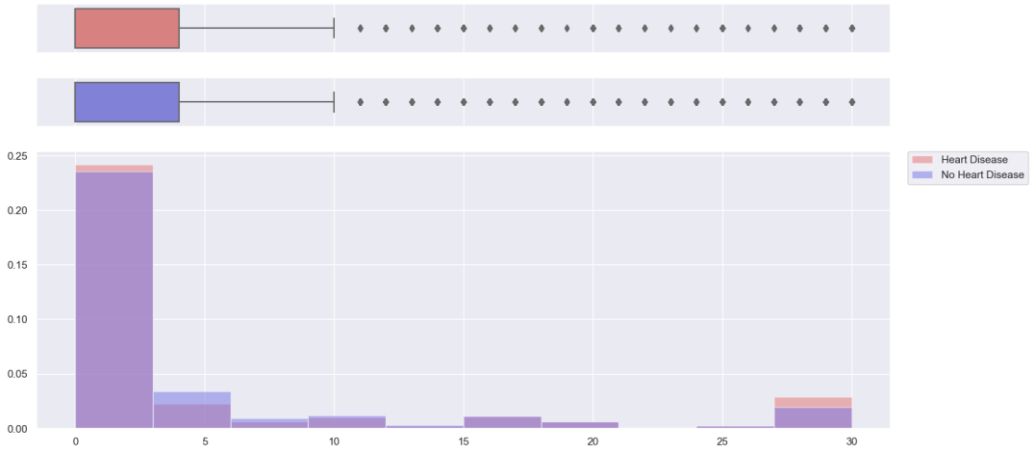


Figure 104: Data Visualization - HeartDisease and MentalHealth

The boxplot and the histogram above show the correlation between “HeartDisease” and “MentalHealth” variable. From the above visualizations, it shows that mental health has an effect towards the risk of heart disease, but not as high as “PhysicalHealth” variable, since it is more distributed equally compared to the “PhysicalHealth” variable.

7.4.2 Visualization for Project Development

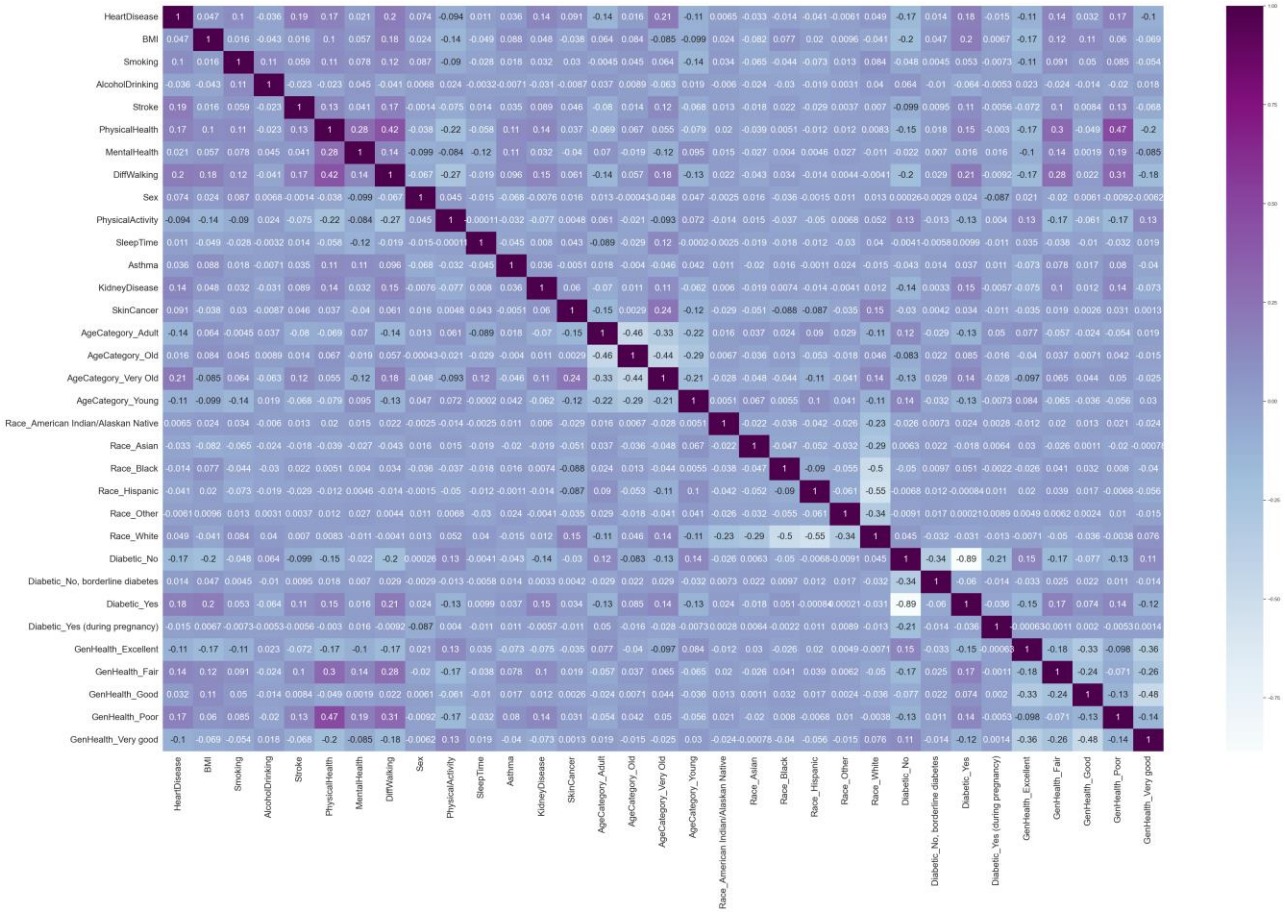


Figure 105: Correlation Heatmap

The above picture shows the correlation heatmap figure of the project. Correlation heatmap can be utilized to gain an understanding of how each variable correlate with each other. The correlation can be both positive and negative. Positive correlation means that the variable has a positive impact towards the other variable, shown with warmer colour. On the other hand, negative correlation means that the variable has a negative impact towards the other variable, shown with cooler colour (Chip, 2023).

Some examples that can be observed from the above heatmap is between GenHealth_Poor and PhysicalHealth, meaning if the GenHealth_Poor variable is 1 (Indicating the person have a poor general health condition), means it has a 47% positive impact towards PhysicalHealth (The person physical Health value will also go up).

Another example is between HeartDisease and Diabetic_No, where if the Diabetic_No variable is 1 (Indicating the person do not have diabetes), then it has a 11% negative impact towards HeartDisease (the person is 11% more unlikely to have a heart disease)

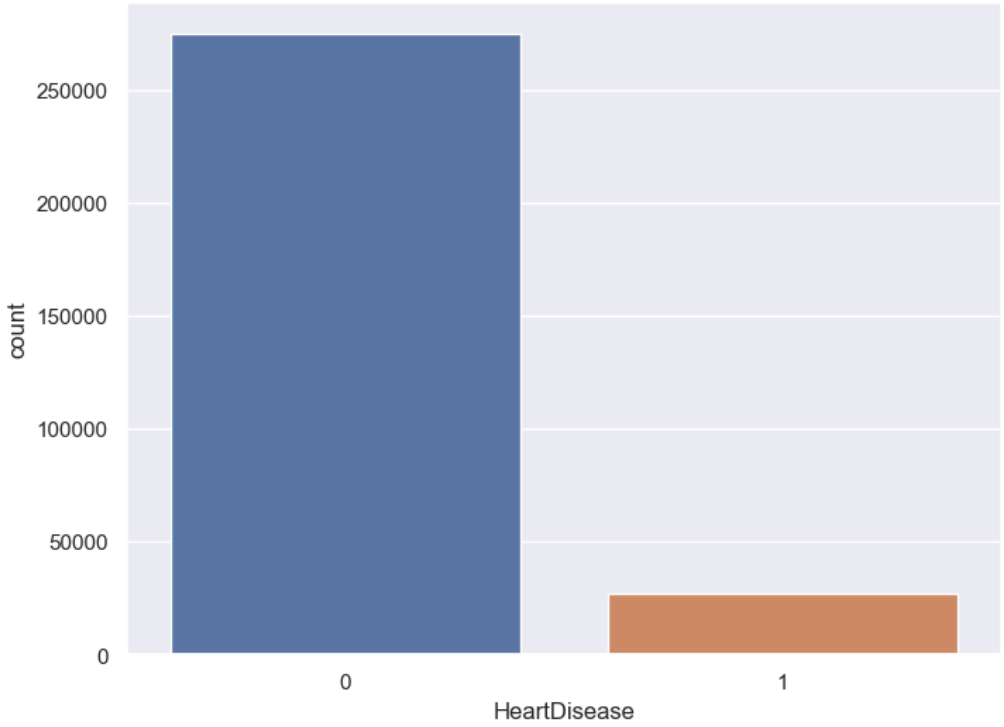


Figure 106: Target Variable Class Distribution

Target Variable class distribution is also a very important visualization. From this visualization, it is determined that the target variable, in this case HeartDisease, is very imbalanced and must be balanced before model building starts. If not, then the model performance will suffer, due to the lower sample amount of a specific class. In this case, the Machine Learning model will have a worse performance to predict HeartDisease value 1, or Yes, due to the lack of data provided. The distribution balancing can be done through Data Sampling Techniques, which can be implemented before Model Building (Brownlee, 2020).

7.5 Model Building

Model Building step refers to the activity to build or create a Machine Learning Model using the pre-processed data that has been prepared during the previous step. The ML model is built using a set of training data, and then the newly created model is validated using a set of testing data (Towards AI, 2021). Both the training data and the testing data comes from the pre-processed data, split into two, usually in 70:30 or 80:20 split, with the dominant amount is used for training (GeeksforGeeks, 2023). In creating the Machine Learning model, picking the correct model is also important, since different models has different purposes and target. In the case of this dataset, where the Machine Learning model's objective is to classify the new observation between one of two classes, then Binary Classification is the most compatible model type. In this project, the ML model will decide based on the data inputted, if the person is having a risk of heart disease or not (Karabiber, 2023). In Model Building, there are some steps that needs to be taken before Model Building can be initiated.

7.5.1 Setting Target and Feature Dataset

Before model building, the dataset must be separated into two datasets, first consisting of the Feature or Input variables, usually referred as "x". The other dataset consists of the Target variable, usually referred as "y".

```
# Separating Input (x) with Target (y)
x = fullData.drop(['HeartDisease'], axis=1)
y = fullData['HeartDisease']
```

Figure 107: Data Feature-Target Split

The code snippet above shows the process of separating the Feature or the Input variable and the Target variable.

7.5.2 Data Normalization

First, Data Normalization can be applied to the numerical value attributes inside the dataset. Normalization refers to the technique to transform the features of the numerical value attributes to use a common scale (Javatpoint, 2023b). Normalization can improve the performance and training quality of the Machine Learning Model (Google, 2023).

```
# FOR OVERSAMPLING

scaler = MinMaxScaler()
x_scaled = scaler.fit_transform(x)
```

Figure 108: Data Normalization

The code snippet above shows the normalization being done towards the feature variables. The normalization method used for this data is MinMaxScaler.

7.5.2 Data Train-Test Split

Next, the data is separated into two sets, where the first set is for training the machine learning model, while the second set is for the testing and assessing the trained model's performance (GeeksforGeeks, 2023). Usually for Train-Test split, the dataset is split following 70:30 or 80:20 percentage, depending on the amount of data available. Since the data amount is sufficient, 70:30 split is eligible for this project.

```
# Splitting the Train-Test Data
x_train, x_test, y_train, y_test = train_test_split(x_scaled,y,test_size=0.3,random_state=42,stratify=y)

print(len(x_train))
print(len(x_test))
print(len(y_train))
print(len(y_test))
```

```
211268
90544
211268
90544
```

Figure 109: Data Train-Test Split

The code snippet above shows the process of Train-Test split, with 70:30 percentage split. The below output shows the result and the amount of data for training and testing, where training dataset has 211.628 data, while the test has 90.544 data.

7.5.3 Data Sampling

Data sampling refers to techniques that can be used in case of imbalanced data distribution (Brownlee, 2020). The problem with imbalanced data distribution is that the minority class (the class with less data) can be ignored by the machine learning models' algorithm during training phase to achieve a better performance, causing a higher False Positive or False Negative prediction. To counteract this, Data Sampling will try to balance the class distribution, thus allow the ML algorithms to train with both classes equally.

There are two main data sampling methods, called Oversampling and Undersampling. Oversampling refers adding new data for the minority class, while Undersampling refers to removing data from the majority class. Another method of data sampling can be through combined sampling, where Oversampling and Undersampling is used concurrently to create a balanced data (Brownlee, 2020). For this project, all three methods are used to be compared during model building step. All sampling techniques are applied only to the training dataset. This is to prevent data leakage, which means that the test data has been used during data training, thus provides a good performance during training, but poor performance during deployment. Data leakage can happen if data sampling, such as oversampling is applied to the dataset before the train-test split (Javatpoint, 2023a).

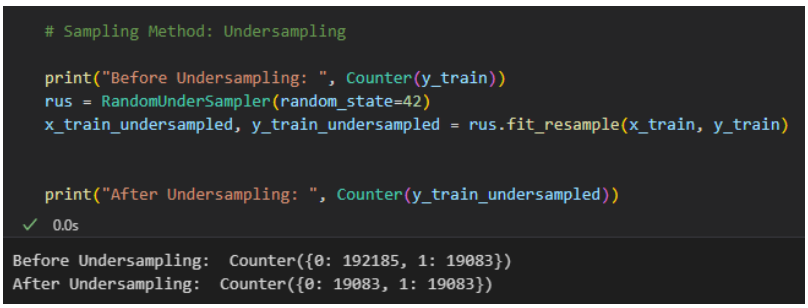


Figure 110: Data Undersampling

The code snippet above shows the undersampling technique usage, where using RandomUnderSampler, the train dataset majority class data amount is reduced to be equal to the minority class data amount. From the result above, the majority class data is reduced from 192.185 to 19.083.

```
# Sampling Method: Oversampling

print("Before SMOTE sampling: ", Counter(y_train))
# Apply SMOTE only to the training set
sm = SMOTE()
x_train_oversampled, y_train_oversampled = sm.fit_resample(x_train, y_train)

print("After SMOTE sampling: ", Counter(y_train_oversampled))
```

✓ 1.2s

```
Before SMOTE sampling:  Counter({0: 192185, 1: 19083})
After SMOTE sampling:  Counter({0: 192185, 1: 192185})
```

Figure 111: Data Oversampling

The code snippet above shows the oversampling technique usage, where using SMOTE (Synthetic Minority Oversampling Technique) is used, where it synthesizes new examples for the minority class, to increase the data amount to be equal to the majority class. From the result above, the minority class data amount is increased from 19.083 to 192.185.

```
# Sampling Method: Combined Sampling (SMOTE & RandomUnderSampler)

print("Before sampling: ", Counter(y_train))

# define pipeline
over = SMOTE(sampling_strategy=0.3)
under = RandomUnderSampler(sampling_strategy=1)
steps = [('o', over), ('u', under)]
pipeline = Pipeline(steps=steps)

# fit and apply the pipeline
X_resampled, y_resampled = pipeline.fit_resample(x_train, y_train)

print("After sampling: ", Counter(y_resampled))
```

✓ 0.7s

```
Before sampling:  Counter({0: 192185, 1: 19083})
After sampling:  Counter({0: 57655, 1: 57655})
```

Figure 112: Data Combined Sampling

The code snippet above shows the combined sampling technique usage, where using both SMOTE and RandomUnderSampler together through Pipeline usage, it allows to apply both techniques at the same time. This will reduce the majority class amount and increase the minority class amount at the same time. From the result above, it shows that both majority and minority classes data amount become 57.655.

7.5.4 Initial Model Building

In this step, the dataset is used to train and validate all machine learning models that are considered for the project. After that, the performance for all models is compared with each other to determine on which model is the best for the dataset, and to be used for the deployment. For this project, there are eight (8) Machine Learning models used for the initial model building. Those are MLP(Multi-Layer Perceptron), Linear SVC (Support Vector Classification), Random Forest, XGBoost, Decision Tree, Ada Boost, K-Nearest Neighbors,

and Gradient Boost. All these models are tested with all three data sampling techniques, so a comparison on what data sampling technique is best for the project, and what machine learning model is best for the project can be decided on the same time.

```
# Model Performance Metrics Function
def cross_val(model):
    accuracies=cross_val_score(estimator=model,X=x_train_undersampled,y=y_train_undersampled,cv=5, verbose=2)
    return accuracies.mean()*100

def fit_evaluate(model):
    name=model.__class__.__name__
    model.fit(x_train_undersampled,y_train_undersampled)
    y_pred=model.predict(x_test)
    cross=cross_val(model)
    print(classification_report(y_test,y_pred))
    print (confusion_matrix(y_test,y_pred))
    a_s=accuracy_score(y_test,y_pred)*100
    pre_sc=precision_score(y_test,y_pred)*100
    rec_sc=recall_score(y_test,y_pred)*100
    f1_sc=f1_score(y_test,y_pred)*100
    roc_sc=roc_auc_score(y_test,y_pred)*100
    result=pd.DataFrame([[name,cross,a_s,pre_sc,rec_sc,f1_sc,roc_sc]],columns=['model','accuracy_train_cv','accuracy_test','precision_score','recall_score','f1_score','roc_auc_score'])
    return result
```

Figure 113: Function for Model Building

```
# Models Performance Metrics Comparison
# models=[AdaBoostClassifier()]
models=[MLPClassifier(),LinearSVC(),RandomForestClassifier(),XGBClassifier(),DecisionTreeClassifier(),KNeighborsClassifier(),AdaBoostClassifier(),GradientBoostingClassifier()]
# models = [svm.SVC(kernel='linear')]
result_models=pd.DataFrame(columns=['model','accuracy_train_cv','accuracy_test','precision_score','recall_score','f1_score','roc_auc_score'])
for model in models:
  print(model.__class__.__name__)
  results = fit_evaluate(model)
  result_models=pd.concat([result_models,results])
result_models.sort_values(by='recall_score',ascending=False)
```

Figure 114: Model Building

The code snippets above show the functions to create, train, and test all models. The function above allows all the building to be done for all models back-to-back. The confusion matrix, classification report, and a summary table are also shown for each model.

7.5.5 Model Performance Comparison

The model performance is compared between all models, with all different data sampling techniques. The goal is to choose the best configuration so the model's performance is best suited for the project's objective.

First, a brief explanation of the performance metrics is required. In Machine Learning, the model's performance can be measured using a confusion matrix. A confusion matrix is 2x2 matrix that consist of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), and used to visualize the results of the prediction (Simplilearn, 2023b). In the case of this project, the target variable of "HeartDisease" has only two unique values, zero (0) as Negative, means that a person has a low risk of Coronary Artery Disease (CAD), and one (1) as Positive, means that a person has a high risk or already diagnoses with Coronary Artery Disease. Confusion Matrix checks the model's results based on the amount of:

- True Positive (TP): The prediction outcome is 1, the actual data is 1.
- False Positive (FP): The prediction outcome is 0, the actual data is 1.
- False Negative (FN): The prediction outcome is 1, the actual data is 0.
- True Negative (TN): The prediction outcome is 0, the actual data is 0.

Based on these four results, four performance metrics can be created (Harikrishnan, 2021):

- Accuracy: The number of correct predictions over the total data amount.
- Precision: Positive predictive value, which covers the amount of true positive over the true positive and false positive. Thus, precision value become higher when False Positive decreases.
- Recall: Sensitivity, which covers the amount of true positive over true positive and False Negative. Thus, Recall value become higher when False Negative decreases.
- F-1 Score: Metrics that covers both precision and recall.

All these performance metrics have the formula as below (Harikrishnan, 2021):

- $Accuracy = \frac{TN+TP}{TN+FN+FP+TP}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F-1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$

Since this project's main objective is to be a warning detection tool for detecting Coronary Artery Disease (CAD), Recall is the important metric to be measured. Recall covers the False Negative (FN), where the person is predicted to be CAD-free, but they have a high risk or already diagnosed with CAD. A high recall score means that the False Negative number is lower.

The next performance metric to be measured is accuracy of the prediction. Accuracy means that the Machine Learning model can correctly predict the data. Meanwhile, Precision is not the most important metrics in these case since a high false positive means that the person also has a high risk of CAD based on the Machine Learning model, thus the person can start taking preventive measure and made changes to reduce the risk. From all the considerations above, Both Recall and Accuracy is the main metrics to determine the Machine Learning Model to be used for the project.

7.5.4.1 Model Results for Undersampling

	model	accuracy_train_cv	accuracy_test	precision_score	recall_score	f1_score	roc_auc_score
0	XGBClassifier	75.336702	72.729281	21.863286	78.454390	34.196781	75.307616
0	GradientBoostingClassifier	76.154174	73.801688	22.570854	78.197603	35.030539	75.781413
0	LinearSVC	76.143699	73.817150	22.539346	77.928589	34.965572	75.668760
0	AdaBoostClassifier	76.033656	74.623388	22.924366	76.607973	35.288816	75.517157
0	RandomForestClassifier	73.701733	70.849532	20.188531	75.421864	31.851278	72.908708
0	MLPClassifier	74.715721	74.077796	22.215763	74.761555	34.253060	74.385731
0	KNeighborsClassifier	72.352366	71.895432	20.404456	72.792859	31.874281	72.299593
0	DecisionTreeClassifier	66.333930	66.507996	16.457153	66.434336	26.379549	66.474823

Figure 115: Undersampling Results

The results above show the results for all models' performances using the training dataset that has been undersampled using RandomUnderSampler. From the results above, GradientBoost seems to be the best balance between Test Accuracy and Recall score.

7.5.4.2 Model Results for Oversampling

	model	accuracy_train_cv	accuracy_test	precision_score	recall_score	f1_score	roc_auc_score
0	LinearSVC	76.276504	73.640440	22.404756	77.879677	34.798525	75.549605
0	AdaBoostClassifier	82.973697	80.523281	25.759984	61.445341	36.301246	71.931422
0	MLPClassifier	82.723678	79.390131	23.486266	56.774272	33.227180	69.204949
0	KNeighborsClassifier	86.470588	78.747349	21.687733	51.821961	30.578304	66.621347
0	GradientBoostingClassifier	86.283269	85.462317	30.715667	48.544876	37.624982	68.836336
0	DecisionTreeClassifier	88.219164	83.245715	21.011609	30.985571	25.042000	59.710060
0	RandomForestClassifier	92.199964	86.571170	26.667448	27.818538	27.230834	60.111585
0	XGBClassifier	90.992533	88.854038	33.823529	24.468085	28.395062	59.857455

Figure 116: Oversampling Results

The results above show the results for all models' performances using the training dataset that has been oversampled using SMOTE. From the result above, LinearSVC has the best balance between Recall and Accuracy.

7.5.4.3 Model Results for Combined Sampling

	model	accuracy_train_cv	accuracy_test	precision_score	recall_score	f1_score	roc_auc_score
0	LinearSVC	76.108750	73.607307	22.400534	78.001956	34.805620	75.586463
0	MLPClassifier	78.520510	72.733699	20.803565	71.924676	32.272578	72.369351
0	AdaBoostClassifier	79.591536	78.192923	24.785281	69.515774	36.541861	74.285119
0	KNeighborsClassifier	79.787529	73.583009	20.368209	66.153094	31.146550	70.236904
0	GradientBoostingClassifier	81.257480	80.496775	26.537986	65.566153	37.783180	73.772684
0	RandomForestClassifier	84.467956	80.868970	24.881881	55.380289	34.336619	69.389996
0	XGBClassifier	82.727430	84.145830	29.427829	54.022989	38.100987	70.579836
0	DecisionTreeClassifier	78.119851	75.811760	18.578559	49.608706	27.033150	64.011065

Figure 117: Combined Sampling Results

The results above show the results for all models' performances using the training dataset that has been sampled using both SMOTE and RandomUnderSampler. Based on the result above, LinearSVC has the best balance between Recall and Accuracy.

7.5.4.4 Evaluation of the results

Based on the results above, all results have low precision score, means that the False Positive case is rampant from all Machine Learning models. But as discussed previously, low precision still allows the project's objective to be met, where the Machine Learning model is used as an early warning detection tool for the users. With this, False Positive case allow the users to be aware of their risk of Heart Disease, thus they can start their preventive measures to reduce the risk.

Considering the accuracy, recall, and the time needed for the model fitting, Gradient Boosting algorithm with Undersampled data will be the machine learning model used for the project's deployment.

7.5.6 Hyperparameter Tuning

After the best model has been chosen, a further optimization of the model can be done through Hyperparameter Tuning or shorten as HP Tuning. Hyperparameter tuning refer to tweaking the model's parameter to get the best results out of the Machine Learning model (AWS, 2023). Hyperparameter Tuning can take a long time, especially if doing a manual HP Tuning, where tweaking the model's parameter one by one to find the optimal result can be time consuming. Instead, using available HP Tuning algorithms from libraries such as Keras and Scikit-learn allow automating HP Tuning, which can take less time, and can be done without active monitoring by the developer.

For this project, the developer has decided to use KerasTuner, RandomizedSearchCV, and GridSearchCV for the automatic HP Tuning. KerasTuner is based on the Keras library, which allows the users to put in the range for each parameter, then KerasTuner will print out the result of the best combination of the parameters (Keras Team, 2023). Next, RandomizedSearchCV is based on Scikit-learn library, Which randomly use a set of hyperparameters, get the score for each set, and will provide the best set of hyperparameters as an output (scikit-learn, 2023b). Finally, GridSearchCV is also from Scikit-learn, but all combinations of hyperparameters are tested, then the best set of hyperparameters will be provided as an output. This means that GridSearchCV is more exhaustive but will give out the best results (scikit-learn, 2023a). During the HP Tuning step, the performance will be measured using both Classification Report and Confusion Matrix.

7.5.6.1 Gradient Boosting before Tuning

```
# Performance metrics for Best Model

best_model_gradient=GradientBoostingClassifier()
best_model_gradient.fit(x_train_undersampled,y_train_undersampled)
best_model_gradient_prediction=best_model_gradient.predict(x_test)
print("Performance Metrics before Tuning for GradientBoostingClassifier")
print(classification_report(y_test,best_model_gradient_prediction))
print (confusion_matrix(y_test,best_model_gradient_prediction))
```

✓ 5.4s

```
Performance Metrics before Tuning for GradientBoostingClassifier
              precision    recall  f1-score   support

           0       0.97      0.73      0.84     82366
           1       0.23      0.78      0.35      8178

    accuracy                           0.74     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.74      0.79     90544

[[60428 21938]
 [ 1783  6395]]
```

Figure 118: GradientBoost Base Result

The code snippet and the result above show the result of the Gradient Boost algorithm before any tuning is applied. The result of this version will be used as the baseline comparison for HP Tuning step.

7.5.6.2 Gradient Boosting using KerasTuner

```
def buildModel(hp):
  model = GradientBoostingClassifier(
    learning_rate = hp.Float('learning_rate', min_value = 0.01, max_value = 0.5),
    random_state= hp.Int('random_state', min_value = 0, max_value = 256),
    n_estimators= hp.Int('n_estimators', min_value = 1, max_value = 500),
    loss= hp.Choice('loss', ['deviance', 'exponential', 'log_loss'])
  )
  return model

tuner = kt.SklearnTuner(oracle = kt.oracles.BayesianOptimizationOracle(objective = kt.Objective('score', 'max'),
        max_trials = 7), hypermodel = buildModel, cv = StratifiedKFold(5), overwrite = True)

tuner.search(x_train_df, y_train_df)

print()
best_model = tuner.get_best_models(num_models=1)[0]
best_model
```

✓ 5m 36.8s

```
Trial 7 Complete [00h 00m 44s]
score: 0.7538910459427357

Best score So Far: 0.761725198635422
Total elapsed time: 00h 05m 37s
INFO:tensorflow:Oracle triggered exit
```

```
GradientBoostingClassifier(learning_rate=0.04850701521306425,
                           loss='exponential', n_estimators=333,
                           random_state=164)
```

Figure 119: KerasTuner HP Tuning

```
best_model_Gradient_keras=GradientBoostingClassifier(learning_rate = 0.04850701521306425, loss = 'exponential', n_estimators = 333, random_state=164)
best_model_Gradient_keras.fit(x_train_undersampled,y_train_undersampled)
best_model_Gradient_prediction=best_model_Gradient_keras.predict(x_test)
print("Performance Metrics after Keras Tuning for GradientBoostingClassifier")
print(classification_report(y_test,best_model_Gradient_prediction))
print (confusion_matrix(y_test,best_model_Gradient_prediction))
```

✓ 17.8s

```
Performance Metrics after Keras Tuning for GradientBoostingClassifier
              precision    recall  f1-score   support

           0       0.97      0.73      0.84     82366
           1       0.23      0.78      0.35      8178

    accuracy                           0.74     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.74      0.79     90544

[[60325 22041]
 [ 1759  6419]]
```

Figure 120: GradientBoost with KerasTuner's Settings

The first code snippet above shows the process of KerasTuner HP Tuning. The best set of parameters from KerasTuner is used in the Gradient Boosting algorithm to get the performance results. Comparing the Classification Report, the results are identical with the baseline performance. Meanwhile, comparing the confusion matrix shows an improvement in lower False Positive and higher True Positive. This means that the model is more accurate on detecting if the person has a high risk of CAD, and not mistaken them into the low-risk category. But the performance results worsen on both the True Negative and False Negative. This means that the model makes more mistake in predicting if the person has a low risk of CAD.

7.5.6.3 Gradient Boosting using RandomizedSearchCV – Accuracy

```
params={'n_estimators':[100,200,300,400,500],"learning_rate":[0.01,0.05,0.1,0.2,0.3,0.5], 'random_state': [0, 42, 128, 1]}
randSearchCV_gradient_accuracy=RandomizedSearchCV(best_model_gradient,params,scoring='accuracy',cv=5)
randSearchCV_gradient_accuracy.fit(x_train_undersampled,y_train_undersampled)
prediction=randSearchCV_gradient_accuracy.predict(x_test)
print("Performance Metrics after Tuning for GradientBoostingClassifier for accuracy")
print(classification_report(y_test,prediction))
print (confusion_matrix(y_test,prediction))
```

✓ 9m 41.5s

```
Performance Metrics after Tuning for GradientBoostingClassifier for accuracy
              precision    recall  f1-score   support

           0       0.97      0.73      0.84     82366
           1       0.23      0.78      0.35      8178

    accuracy                           0.74     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.74      0.79     90544

[[60319 22047]
 [ 1776  6402]]
```

Figure 121: GradientBoost with RandomizedSearchCV - Accuracy

The above code snippet shows the usage and result of RandomizedSearchCV on the Gradient Boosting algorithm, with the scoring target set on Accuracy. Comparing the Classification Report with the baseline model, it is also identical on every metrics. Next, comparing the Confusion Matrix results, the True Positive and the False Positive result is a little bit better compared to the baseline performance, but still inferior compared to the KerasTuner's result. Same result also on the True Negative and False Negative, where the performance worsens, but also inferior compared to KerasTuner's result.

7.5.6.4 Gradient Boosting using RandomizedSearchCV – Recall

```
params={'n_estimators':[100,200,300,400,500],"learning_rate":[0.01,0.05,0.1,0.2,0.3], 'random_state': [0, 42, 128, 1]}
randSearchCV_gradient_recall=RandomizedSearchCV(best_model_gradient,params,scoring='recall',cv=5)
randSearchCV_gradient_recall.fit(x_train_undersampled,y_train_undersampled)
prediction=randSearchCV_gradient_recall.predict(x_test)
print("Performance Metrics after Tuning for GradientBoostingClassifier for Recall")
print(classification_report(y_test,prediction))
print (confusion_matrix(y_test,prediction))
```

✓ 6m 47.2s

```
Performance Metrics after Tuning for GradientBoostingClassifier for Recall
              precision    recall  f1-score   support

           0       0.97      0.73      0.83     82366
           1       0.22      0.79      0.35      8178

    accuracy                           0.73     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.73      0.79     90544

[[60072 22294]
 [ 1737  6441]]
```

Figure 122: GradientBoost with RandomizedSearchCV - Recall

The above code snippet shows the usage and result of RandomizedSearchCV on the Gradient Boosting algorithm ,with the scoring target set on Recall. Comparing the Classification Report with the baseline model, the accuracy and Positive Precision has dropped by 0.01, while the recall for Positive has increased by 0.01. This means that the Recall result is much better compared to the baseline, with both False Positive and True Positive results getting better. On the other hand, The True Negative and False Negative results are worse compared on the baseline model. This is all confirmed when comparing the Confusion Matrix of this result compared with the baseline performance.

7.5.6.5 Gradient Boosting using GridSearchCV – Accuracy

```
params={'n_estimators':[100,200,300,400,500],"learning_rate":[0.01,0.05,0.1,0.2,0.3], 'random_state': [0, 42, 128, 1]}
gridSearch_gradient_acc=GridSearchCV(best_model_gradient,params,scoring='accuracy',cv=5)
gridSearch_gradient_acc.fit(x_train_undersampled,y_train_undersampled)
prediction=gridSearch_gradient_acc.predict(x_test)
print("Performance Metrics after Tuning for GradientBoostingClassifier for Recall")
print(classification_report(y_test,prediction))
print (confusion_matrix(y_test,prediction))
```

✓ 71m 18.2s

```
Performance Metrics after Tuning for GradientBoostingClassifier for Recall
              precision    recall  f1-score   support

           0       0.97      0.73      0.84     82366
           1       0.23      0.79      0.35      8178

    accuracy                           0.74     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.74      0.79     90544

[[60278 22088]
 [ 1733  6445]]
```

Figure 123: GradientBoost with GridSearchCV - Accuracy

The above code snippet shows the usage and result of GridSearchCV on the Gradient Boosting algorithm, with the scoring target set on Accuracy. Comparing the Classification Report result with the baseline model, The Positive Recall has increased by 0.01, while the rest are still the same. Next, comparing the Confusion Matrix, the value of False Positive has decreased, while the value of True Positive has increased, thus meaning that the model has a better performance detecting if the person has high risk of CAD. Meanwhile, The True Negative value has decreased, and the False Negative value has also increased, which means that the model has a worse performance detecting if the person has a low risk of CAD, compared to the baseline model.

7.5.6.6 Gradient Boosting using GridSearchCV – Recall

```
params={'n_estimators':[100,200,300,400,500],"learning_rate":[0.01,0.05,0.1,0.2,0.3], 'random_state': [0, 42, 128, 1]}
gridSearch_gradient_recall=GridSearchCV(best_model_gradient,params,scoring='recall',cv=5)
gridSearch_gradient_recall.fit(x_train_undersampled,y_train_undersampled)
prediction=gridSearch_gradient_recall.predict(x_test)
print("Performance Metrics after Tuning for GradientBoostingClassifier for Recall")
print(classification_report(y_test,prediction))
print (confusion_matrix(y_test,prediction))
```

✓ 68m 23.4s

```
Performance Metrics after Tuning for GradientBoostingClassifier for Recall
              precision    recall  f1-score   support

           0       0.97      0.73      0.83     82366
           1       0.22      0.79      0.35      8178

    accuracy                           0.73     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.73      0.79     90544

[[60004 22362]
 [ 1724  6454]]
```

Figure 124: GradientBoost with GridSearchCV - Recall

Finally, the above code snippet shows the usage and result of GridSearchCV on Gradient Boosting algorithm, with the scoring target set on Recall. Comparing the Classification Report, both Positive Precision and Accuracy has decreased by 0.01, while the Positive Recall has increased by 0.01 also. This is proven when during the Confusion Matrix comparison, where the False Negative and True Negative value are the best compared to all other tuning settings, while the True Positive and False Positive are the worst compared to all other tuning settings.

7.5.6.7 Evaluation of the Results

With the following HP Tuning results, there is a consistent pattern happening on all HP Tuning settings results, Where the False Negative and True Negative results are working opposite with False Positive and True Positive. In this case, Getting a better Recall score means sacrificing a on Precision, or even Accuracy. In this case, a compromise where the Recall score improvement did not have a very huge effect on both the Precision and the Accuracy of the model. On all settings, the difference in value on all performance metrics are very small, thus a calculation using the Confusion Matrix can be used to get the finer detail.

HP Tuning Setting	Positive Improvement in Value Count	Negative Degradation in Value Count	Results
Baseline Model	0	0	0
KerasTuner	24	-103	-79
RandomizedSearchCV – Accuracy	7	-109	-102
RandomizedSearchCV – Recall	46	-356	-310
GridSearchCV – Accuracy	50	-150	-100
GridSearchCV – Recall	59	-424	-365

Table 19: HP Tuning Results Comparison

The above table shows the comparison results from each HP Tuning setting using the Confusion Matrix. Using baseline model as the main benchmark, then the improvement on the Positive value (1) and the degradation on the Negative value (0) is then measured. The Positive Improvement is the margin between the HP Tuning's True Positive with the baseline model's True Positive, while the Negative Degradation is the margin between the HP Tuning's True Negative with the baseline model's True Negative. Based on the result above, the most optimal HP Tuning comes from KerasTuner. Thus, KerasTuner Parameter's set will be applied to the Gradient Boosting Algorithm that is used for the deployment process.

7.6 Summary

In conclusion, All steps of Data Analysis for the project have been done. Initial Data Exploration has for both Survey Data and Main Dataset has been conducted, with notes on the appropriate data pre-processing steps that are needed for the dataset. Next, Data Cleaning also has been done, with the first step being Survey Data Inconsistency Fixing and Data Transformation, thus allowing the Survey Data to be concatenated into the Main Dataset. After that, pre-processing such as Removing Duplicates, More Inconsistency Fixing, Missing Data Imputation, and Data Encoding has been done to prepare the dataset for the Model Building and Data Visualization. Next, Data Visualization has been done for each feature variables, with the results showing the correlation between each feature variable with the target variable. After that, Model Building Preparation has also been conducted, such as Dataset Splitting, Train-Test Splitting, Data Normalization, and finally Data Sampling. For the initial model building, eight (8) Machine Learning algorithms are used to decide the most optimal Machine Learning model for this project. A comparison between different Data Sampling Techniques also has been conducted, with Gradient Boosting Algorithm using Undersampled Data has been chosen as the most optimal combination after an in-depth discussion regarding the performance metrics and consideration. Finally, using Hyperparameter Tuning, The Gradient Boosting algorithm also has been tuned to produce a better score of the model's prediction.

Chapter 8: Results and Discussion

8.1 Introduction

In this chapter, the results of the project are going to be the main topic of discussion. First, a discussion regarding the Machine Learning model result is going to be conducted, including the satisfactory level of the result achieved, also with comparison of other developers' works of the same dataset. With that comparison, benchmarking can be done to validate the results that has been achieved by the project. After that, discussion regarding the deployment method will also be done, how the deployment works, and what is the result of the deployment process.

8.2 Model Building Results

A comparison between the Project's Machine Learning model result and other developers' works can be done to add validity of the results. Comparison also can show that this project's result has fulfilled the project's objective, and eligible to be used to for the target users.

8.2.1 Model Building Results – This Project

The first discussion will be covering this project's work. Due to the usage of survey data, some pre-processing steps are added. First, data pre-processing is done on the survey data, to allow it to be merged with the main data. This covers the data transformation creating "BMI" variable from "BodyWeight" and "BodyHeight". Next, data standardization to where inconsistencies are resolved, and language translation from Bahasa Indonesia into English Language are done. After that, the survey data can be merged into the main data. After that, another round of inconsistencies fixing is done to ensure data standardization. Finally, Data Encoding is done using both One Hot Encoding and Label Encoding to allow the data to be fitted into the Machine Learning Models. After that, the data is splitted into x (feature variables) and y (target variables) before data normalization through Min Max Scaling. Next, the data is divided into train-test with 70:30 data split. Finally, the finalized model used RandomUnderSampler on the train datasets to undersample the data to avoid imbalanced class distribution, using Gradient Boosting algorithm, and HP Tuning using Keras Tuner.

Model Name	Accuracy	Precision	Recall	F-1
Gradient Boosting	0.74	0.23	0.78	0.35

Table 20: The Project's Final Model Result

The above table shows the performance of the chosen model built for this project. It shows that the model has adequate Accuracy and Recall but fell short in Precision. This means that the model is more likely to produce False Positive results, where people that has low risk of Heart Disease is predicted to have a high risk of Heart Disease. Despite this shortcoming, the model still fits the project's objective. As discussed in the previous chapter, the focus of the model chosen would be in Recall, then Accuracy. Thus, having a model with the highest Recall with adequate Accuracy is sufficient for the project. The performance of the model can be improved by gathering more data, especially on the minority class, to increase the model's learning variety.

For the benchmark, three (3) other developers' works are going to be used for the benchmarking. These developers post their works and results on the website Kaggle, in the Code section of the main data page. The benchmark will review the choice of model, the model's performance, and the pre-processing and preparation steps that has been taken by the other developers to achieve their results. By this, it allows a retrospective review on why the developer has chosen these specific steps during the previous chapter.

8.2.2 Model Building Results – Developer 1

The first developer is Elsayed (2022). For Pre-processing, Elsayed only performed data transformation for diabetic, where Elsayed remove "No, borderline diabetes" and "Yes (during pregnancy)", making it into "No" and "Yes" respectively. Next, Elsayed did the data encoding using both One Hot Encoding for attributes with more than two unique values and perform manual encoding to change the two unique values variables into zero and one, which is identical to Label Encoding. Finally, Elsayed split the dataset into train-test with 80:20 data split and applied Standard Scaler normalization technique on both the train feature and the test feature datasets. For model building, Elsayed try building the model using K-Nearest Neighbors and Decision Tree.

Model Name	Accuracy	Precision	Recall	F-1
K-Nearest Neighbors	0.90	0.35	0.14	0.20
Decision Tree	0.86	0.23	0.25	0.24

Table 21: Elsayed's Final Model Results

The above table shows the results of the model building done by Elsayed. It shows that both model's has a very high Accuracy, both fell short on both Precision and Recall. This means that there are a lot of False Positive and False Negative cases in the models built. This can be due to no data sampling techniques applied towards the dataset before model building, thus causing the model to be built using very imbalanced data. Elsayed also did not remove any duplicates data from the dataset.

8.2.3 Model Building Results – Developer 2

The second developer is Mohaimin (2022). The first Pre-processing that Mohaimin done is manual normalization, using Min-Max scaler logic. After that, Mohaimin encode both using Label Encoding for two unique values variables and One Hot Encoding for the variables with more than two unique values. After that, Mohaimin goes straight to Feature-Target splitting, and then Train-Test split, with 80:20 data split. Next, Mohaimin did a checking on the class

balancing for both Training and Testing datasets. Since the data is unbalanced, Mohaimin oversample both the training dataset and the testing dataset using Random Oversampling with Replacement. After that, Mohaimin start model building by using Random Forest.

Model Name	Accuracy	Precision	Recall	F-1
Random Forest	0.59	0.83	0.22	0.35

Table 22: Mohaimin's Final Model Result

The table above shows the results of the model building done by Mohaimin. It shows that the accuracy and recall is very abysmal, with 0.59 and 0.22 value respectively. This means that the accuracy of the overall prediction is not very good, while the recall means that the False Negative result of the prediction is also very high. The only upside is the Precision score, means that the False Positive result is very low. This low score can be explained due to oversampling also applied to the testing dataset, which can cause Unrealistic Evaluation results. This also means that if the prediction already struggling during the testing period, showing the same data multiple times during testing will also make the performance worsen. Mohaimin also did not remove any duplicates data, which may reduce the performance of the model even more.

8.2.4 Model Building Results – Developer 3

Finally, the last benchmark is using the results from J. Hossen (2023). First, Hossen removed all the duplicated data from the dataset, then, he goes straight for Encoding, where he used Label Encoder on all attributes of the dataset. After that, Hossen did feature-target split and train-test split back-to-back, with 70:30 Split of Train-Test. Hossen then test the performance of the dataset using Decision Tree. After evaluating the result, Hossen decided to do oversampling using RandomOverSampler and applied it to the base x (feature variables) and y (target variables) dataset. After that, Hossen redid the train-test split with the same 70:30 ratio, then evaluate the result again using Decision Tree again. He then decided to use the oversampled data for all the next model buildings. Hossen did multiple experiments using different types of models, but in the end, He decided to finalize his result using Extra Tree.

Model Name	Accuracy	Precision	Recall	F-1
Extra Tree	0.97	0.94	0.99	0.97

Table 23: Hossen's Final Model Result

The table above shows the result of the final model building done by Hossen. From the result above, it shows that the performance of his model is almost perfect, with a very good Accuracy and Precision, and an almost perfect Recall score. This result can be explained by the usage of

oversampling to the dataset before Train-Test split has been done. Due to this, the data from the oversampling technique also get mixed in into the testing dataset. This is conflicting with the main objective of data sampling, where artificially creating an equal distribution, the Machine Learning will learn through the training dataset with less bias, and more focused on the attributes of the dataset (Brownlee, 2020). Due to this method, Hossen accidentally allowed data leakage during the Machine Learning model building. Data Leakage refers to the scenario where the test data also exist in the training data. Using RandomOverSampler, where the amount of minority class is artificially added by picking samples at random replacement (ImbalancedLearn, 2023). This means that there are high chances that during the train-test split of Hossen's project, some artificial data included in both training and testing dataset, causing data leakage.This causes an illusion where the performance during ML training and testing are very high, while during deployment, the accuracy will become very low (Javatpoint, 2023a). With that, the testing will not be representative of the real-world cases.

8.3 Evaluation on Model Building Results

Project	Model Name	Accuracy	Precision	Recall	F-1
This Project	Gradient Boosting	0.74	0.23	0.78	0.35
Elsayed (2022)	K-Nearest Neighbors	0.90	0.35	0.14	0.20
	Decision Tree	0.86	0.23	0.25	0.24
Mohaimin (2022)	Random Forest	0.59	0.83	0.22	0.35
Hossen (2023)	Extra tree	0.97	0.94	0.99	0.97

Table 24: Comparison between All Developers' Final Results

The above table shows the final comparison between this project's result with other developers' results. Compared to Elsayed, the accuracy of this project may not be the best, but the Recall of this project completely outclassed Elsayed's score. Compared to Mohaimin's, both Accuracy and Recall score of this project is much superior. Finally, This project's results are inferior compared to Hossen's, but Hossen may have accidentally caused data leakage towards the ML Model, thus the results can be considered unrealistic. This means that this project's results are sufficient and valid, especially compared to other developers' works. Even though that the result achieved is not the best, the result still can be considered as viable, especially due to the nature of the project, the Personal Key Indicators (PKIs) chosen, and the data gathered. Finally, the deployment step also can be conducted to deploy the Machine Learning model so it can be used by the target users.

8.4 Deployment Results

After the model has been created and validated, it is proven that the results are more than sufficient to be used for the project deployment. For this project, the deployment will be in a form of Web Application, utilizing the library Streamlit that is available in Python. Streamlit allows easier deployment for Python codes, especially for data scientist and analyst, since all the popular libraries for both areas are also compatible with Streamlit. Moreover, Streamlit development can be done entirely through Python, without any frontend development experiences required. Due to all these supporting factors, Streamlit has been chosen for this project's deployment.

Figure 125: Deployment - Introdution Page - Part 1

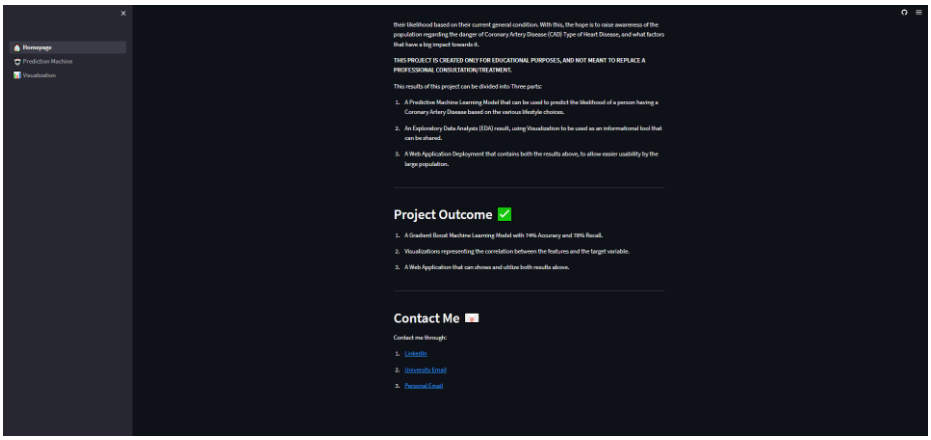


Figure 126: Deployment - Introdution Page - Part 2

The deployment Web Application contains three (3) pages. The first and default page contains an introduction of the project for visitors. It explained the main problem that the project is trying to tackle, the deliverables, and the outcome of the projects. The page also contains the contact info of the developer through email and LinkedIn.

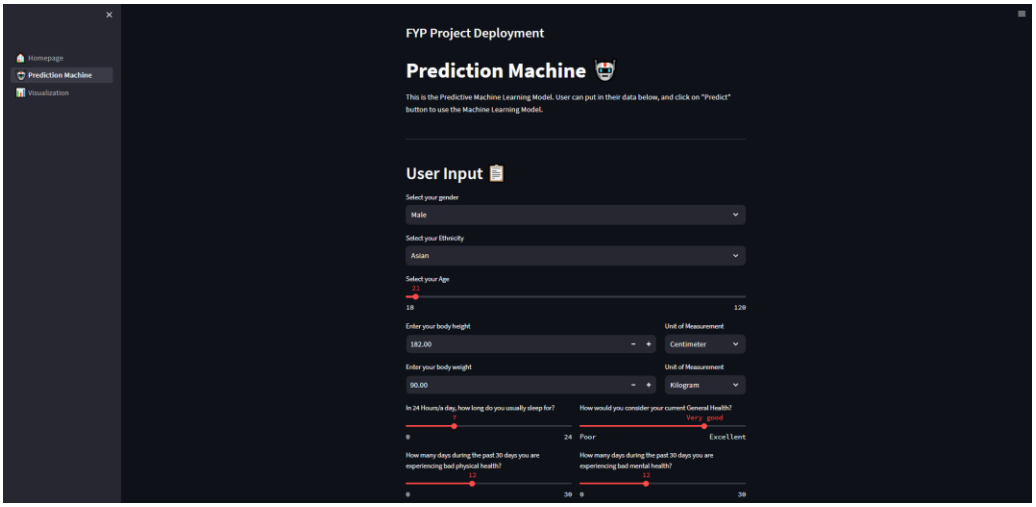


Figure 127: Deployment - Prediction Page - Part 1

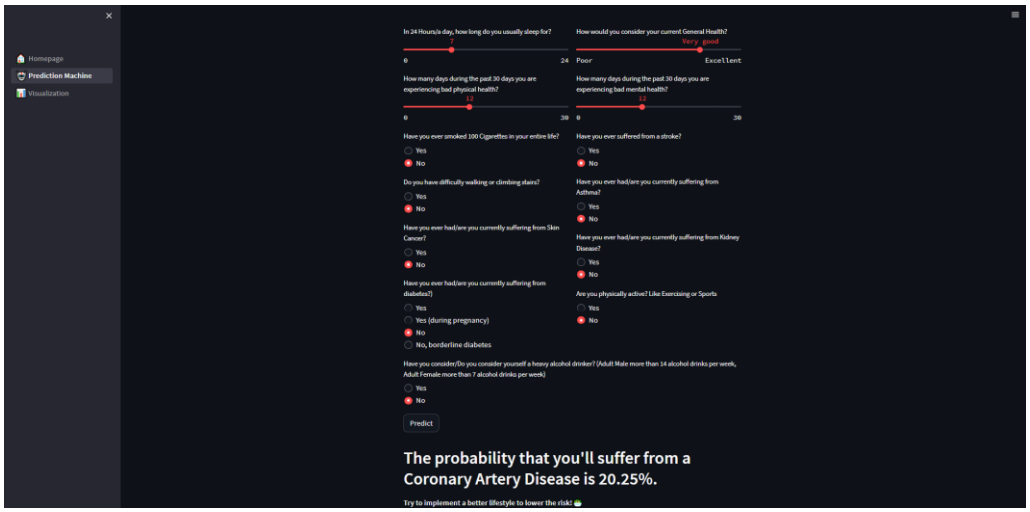


Figure 128: Deployment - Prediction Page - Part 2

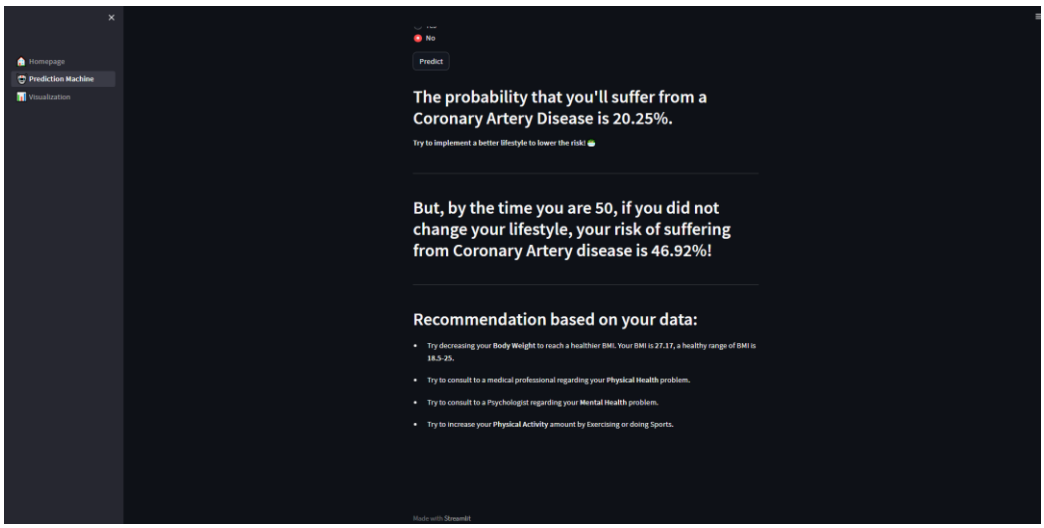


Figure 129: Deployment - Prediction Page - Part 3

The second page is called the Prediction Machine Page, where the user can use the created Machine Learning model to predict their own risk of Coronary Artery Disease (CAD) using their own information. When the user submitted their data, a prediction of the current age will be displayed, with the probability percentage of the risk. For younger users, the results will also show the risk prediction when they are older if they keep their current habits. Finally, based on the data submitted, a recommendation of lifestyle changes will also be shown for the user, which can educate them on what aspect of their lifestyle that they need to improve upon.

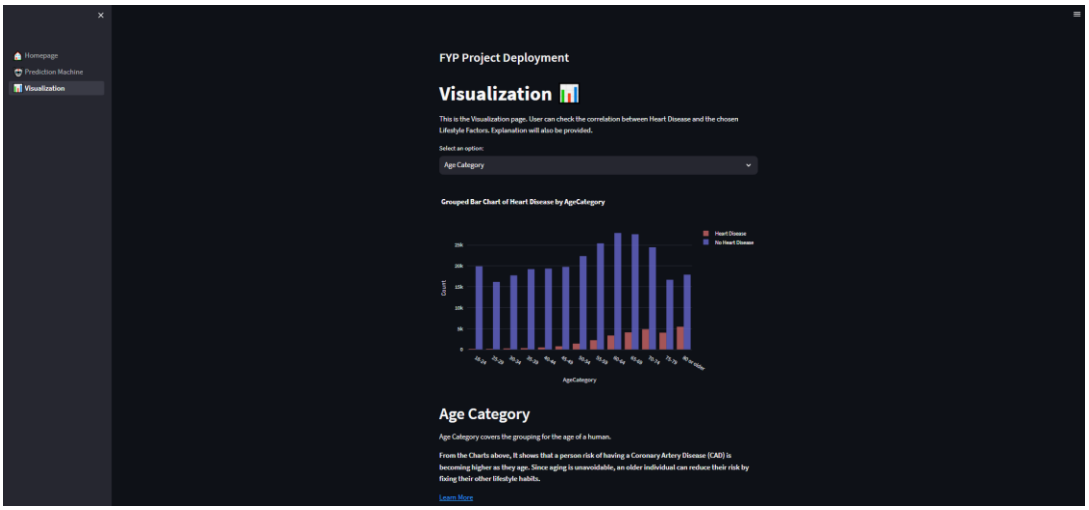


Figure 130: Deployment - Visualization Page

The third page is called the Visualization page, where the user can see and learn about the correlation between a feature attribute with the target attribute based on their selection. The same visualizations used in the Data Visualization phase are used in this page, with enhancement made with the usage of interactive chart, where the user can zoom in/out the chart, and hover over the data to get a more detailed information. Below each visualization, a short explanation regarding the correlation between the selected feature attribute with the target attribute. A link will also be shown in case the users want to learn more regarding the topic.

Chapter 9: Conclusions and Reflections

In conclusion, this project's main objective is to increase public awareness regarding the risk factors that can influence Coronary Artery Disease (CAD) based on lifestyle factors. The main outcome of this project is an early warning detection tool utilizing Machine Learning to predict a user's risk of CAD based on their current lifestyle. The problems, benefits, and aims of the project also has been discussed thoroughly in the respective chapter. A detailed literature review that includes an in-depth knowledge learning of the area of interest, such as Coronary Artery Disease, the currently available diagnosis tools, and Machine Learning usage in the medical field has been conducted. A comparison study covering all previous similar studies done by other researchers also has been conducted. Next, research of the technical part of this project, such as the Programming language, Integrated Development Environment (IDE), and Operating System (OS) of the project has been conducted, with Python, Visual Studio Code, and Windows 10 chosen for respective categories. Comparison regarding the methodologies that can be suitable for the project also has been done, with CRISP-DM being chosen as the methodology of this project. Next, research methodology of the project also has been discussed, with using both the available dataset and survey data collected through questionnaire for the project's dataset. An analysis from the results of the validation part of the questionnaire survey also has been done, which validates the need of the project and its outcome. Finally, Data Analysis has been done thoroughly and accounting for all required actions needed based on the project's needs. From the Data Analysis steps, Data Visualization of the dataset is created, where insights regarding the dataset, especially the correlations between the target variable and the feature variables are shown in-depth. Finally, Model Building to create the Machine Learning model of the project also has been done, with appropriate steps taken and further optimization also has been done. Finally, the outcomes of the project also have been compared with others' works, and the outcome also has been deployed in form of Web Application which allows for ease-of-access for general population.

Limitations and problem faced during this project is mainly comes from the dataset and the Machine Learning model results. Even though the result is acceptable and can be used according to the objective, the result still can be improved to increase its accuracy. This problem is caused by the lack of variance and very imbalance class distribution from the dataset. If the dataset has better variance and properly balanced, then the produced results could be higher.

As the next step of the project, the developer can improve upon the Machine Learning Models by doing a bigger scale of questionnaire, with more complex questions that can correlate more with Coronary Artery Disease. The developer can also enhance the features available in the deployment since the features and customization are mainly limited by the capability of the Streamlit library. Creating a web application from scratch using proper front-end language can make the deployment results to be more personalized, with more features. But the deployment outcome of the project is already satisfactory for the current scope of the project.

In the end, the developer feels that satisfactory results have been achieved through the process of the project. The developer feels that all steps has been taken carefully to make sure that the results of the project can be used according to the objective of the project, and can be useful for general population, especially to increase the general population's awareness regarding the risk of Coronary Artery Disease (CAD) based on lifestyle factors.

References

ActiveState. (2022a, July 11). *What Is Numpy Used For In Python?* Retrieved February 25, 2023, from https://www.activestate.com/resources/quick-reads/what-is-numpy-used-for-in-python/

ActiveState. (2022b, July 12). *What Is Matplotlib In Python? How to use it for plotting?* Retrieved February 25, 2023, from https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/

ActiveState. (2022c, August 9). *What Is Pandas in Python? Everything You Need to Know*. Retrieved February 25, 2023, from https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/

Aduli, F. (2023). *Do men have a higher risk for heart disease?* Louisiana Heart and Vascular. Retrieved July 16, 2023, from https://www.louisianaheart.org/blog/do-men-have-a-higher-risk-for-heart-disease#:~:text=That%20said%2C%20men%20have%20larger,found%20in%20smaller%20blood%20vessels.

Alcohol Think Again. (2023, July 11). *How does alcohol cause cardiovascular disease?* Retrieved July 16, 2023, from https://alcoholthinkagain.com.au/alcohol-and-your-health/long-term-health-effects/cardiovascular-disease

American Heart Association. (2021, November 1). *Saturated Fat*. Retrieved February 16, 2023, from https://www.heart.org/en/healthy-living/healthy-eating/eat-smart/fats/saturated-fats

AWS. (2023). *What is Hyperparameter Tuning? - Hyperparameter Tuning Methods Explained*. Amazon Web Services, Inc. Retrieved July 18, 2023, from https://aws.amazon.com/what-is/hyperparameter-tuning/#:~:text=Hyperparameter%20tuning%20allows%20data%20scientists,the%20model%20as%20a%20hyperparameter.

Babar, A., Lak, H. M., Chawla, S., Mahalwar, G., & Maroo, A. (2020). Metastatic melanoma presenting as a ventricular arrhythmia. *Cureus*. https://doi.org/10.7759/cureus.7634

Bhat, A. (2023, February 9). *Questionnaires: The ultimate guide, advantages & examples*. QuestionPro. Retrieved March 2, 2023, from https://www.questionpro.com/blog/what-is-a-questionnaire/

Bigelow, S. J. (2020, May 18). *Operating System (OS)*. WhatIs.com. Retrieved February 17, 2023, from https://www.techtarget.com/whatis/definition/operating-system-OS

British Heart Foundation. (2023). Global Heart & Circulatory Diseases [Fact sheet]. In *bhf.org.uk*. Retrieved February 8, 2023, from https://www.bhf.org.uk/-/media/files/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf

Brownlee, J. (2020, January 14). *Tour of Data Sampling Methods for Imbalanced Classification*. MachineLearningMastery.com. Retrieved July 10, 2023, from https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/

Burns, E. (2021, March 30). *Machine Learning*. TechTarget. Retrieved February 8, 2023, from https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

Calculator. (n.d.). *BMI Calculator*. Retrieved February 16, 2023, from https://www.calculator.net/bmi-calculator.html

Cambridge Dictionary. (2023). Methodology. In *Cambridge Dictionary*. Retrieved February 26, 2023, from https://dictionary.cambridge.org/dictionary/english/methodology

CDC. (2017, August 29). *NHIS - Adult Tobacco Use - Glossary*. Centers for Disease Control and Prevention. Retrieved March 9, 2023, from https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm#:~:text=Every%20day%20smoker%3A%20An%20adult,at%20the%20time%20of%20interview.

CDC. (2020, April 28). *Health Effects of Smoking and Tobacco Use*. Centers for Disease Control and Prevention. Retrieved February 10, 2023, from https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm

CDC. (2021a, January 26). *Heart Disease: It Can Happen at Any Age | CDC*. Centers for Disease Control and Prevention. Retrieved February 10, 2023, from https://www.cdc.gov/heartdisease/any_age.htm

CDC. (2021b, July 19). *Coronary Artery Disease | cdc.gov*. Centers for Disease Control and Prevention. Retrieved February 9, 2023, from https://www.cdc.gov/heartdisease/coronary_ad.htm

CDC. (2022a, April 22). *Drinking too much alcohol can harm your health. Learn the facts | CDC*. Centers for Disease Control and Prevention. Retrieved February 9, 2023, from https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm

CDC. (2022b, July 12). *Heart Disease Resources | CDC*. Centers for Disease Control and Prevention. Retrieved February 8, 2023, from https://www.cdc.gov/heartdisease/about.htm

CDC. (2022c, September 24). *Effects of Overweight and Obesity*. Centers for Disease Control and Prevention. Retrieved February 10, 2023, from https://www.cdc.gov/healthyweight/effects/index.html

CDC. (2022d, October 14). *Heart Disease Facts | cdc.gov*. Centers for Disease Control and Prevention. Retrieved February 16, 2023, from https://www.cdc.gov/heartdisease/facts.htm

CDC. (2022e, October 24). *LDL and HDL Cholesterol and Triglycerides*. Centers for Disease Control and Prevention. Retrieved February 16, 2023, from https://www.cdc.gov/cholesterol/ldl_hdl.htm

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, *2*, 100016. https://doi.org/10.1016/j.health.2022.100016

Chia, J. (2023, January 3). *5 Best R IDE & Editors (Ranked & Reviewed for 2023!)*. Justjooz. Retrieved February 23, 2023, from https://justjooz.com/best-r-ides/

Chip. (2023, May 3). *How to read a correlation heatmap?* QuantHub. Retrieved July 10, 2023, from https://www.quanthub.com/how-to-read-a-correlation-heatmap/#:~:text=A%20correlation%20heatmap%20is%20a,closely%20related%20different%20variables%20are.

Chorev, S. (2023, March 24). *A practical guide to data cleaning*. Deepchecks. Retrieved July 18, 2023, from https://deepchecks.com/what-is-data-cleaning/#:~:text=Duplicate%20entries%20can%20ruin%20the,disappointing%20the%20model%20in%20production.

Choudhury, A. (2020, June 9). *Top 12 R Packages For Machine Learning In 2020*. Analytics India Magazine. Retrieved February 24, 2023, from https://analyticsindiamag.com/top-12-r-packages-for-machine-learning-in-2020/

Clear, J. (2020, February 4). *How Long Does it Take to Form a Habit? Backed by Science.* James Clear. Retrieved February 9, 2023, from https://jamesclear.com/new-habit

Cleveland Clinic. (2022, August 31). *What Blood Tests Detect Heart Problems?* Retrieved February 10, 2023, from https://my.clevelandclinic.org/health/diagnostics/16792-blood-tests-to-determine-risk-of-coronary-artery-disease

Codecademy Team. (n.d.). *What Is an IDE?* Codecademy. Retrieved February 25, 2023, from https://www.codecademy.com/article/what-is-an-ide

Cordeiro, M. (2022, January 4). *Why Data Scientists Should use Jupyter Notebooks with Moderation ? | Towards Data Science*. Medium. Retrieved February 25, 2023, from https://towardsdatascience.com/why-data-scientists-should-use-jupyter-notebooks-with-moderation-808900a69eff

Coursera. (2022a, June 27). *9 Best Python Libraries for Machine Learning*. Retrieved February 24, 2023, from https://www.coursera.org/articles/python-machine-learning-library

Coursera. (2022b, November 14). *What Is Python Used For? A Beginner's Guide*. Retrieved February 23, 2023, from https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python

DataRobot. (2023, May 3). *Data Preparation for Machine Learning | DataRobot Artificial Intelligence Wiki*. DataRobot AI Platform. Retrieved July 5, 2023, from https://www.datarobot.com/wiki/data-preparation/#:~:text=What%20is%20Data%20Preparation%20for,uncover%20insights%20or%20make%20predictions.&text=Improperly%20formatted%20%2F%20structured%20data.

Dhruv, S. (2019, December 15). *Factors to Consider When Choosing a Programming Language*. Aalpha. Retrieved February 23, 2023, from https://www.aalpha.net/blog/factors-to-consider-when-choosing-a-programming-language/

Duggal, N. (2023, January 6). *Top 10 Python Libraries for Data Science for 2023*. Simplilearn.com. Retrieved February 25, 2023, from https://www.simplilearn.com/top-python-libraries-for-data-science-article

Elsayed, M. (2022, June 3). *Heart disease prediction*. Kaggle. Retrieved July 18, 2023, from https://www.kaggle.com/code/andls555/heart-disease-prediction

Faccioni, J. L. (2022, November 23). *Python Scripts vs. Jupyter Notebooks: Pros and Cons*. LearnPython. Retrieved July 6, 2023, from https://learnpython.com/blog/python-scripts-vs-jupyter-notebooks/

Fogoros, R. N., MD. (2022, May 5). *Heart problems that go hand in hand with strokes*. Verywell Health. Retrieved July 16, 2023, from https://www.verywellhealth.com/heart-problems-that-occur-with-strokes-1746119#:~:text=Heart%20problems%20associated%20with%20strokes,an%20embolus%20to%20the%20brain.

Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012046. https://doi.org/10.1088/1757-899x/1022/1/012046

GeeksforGeeks. (2023, May 4). *Splitting Data for Machine Learning Models*. Retrieved July 8, 2023, from https://www.geeksforgeeks.org/splitting-data-for-machine-learning-models/

Google. (2023). *Normalization*. Google for Developers. Retrieved July 9, 2023, from https://developers.google.com/machine-learning/data-prep/transform/normalization#:~:text=The%20goal%20of%20normalization%20is,training%20stability%20of%20the%20model.

Harikrishnan, B. (2021, December 12). Confusion matrix, accuracy, precision, recall, F1 score. *Medium*. Retrieved July 18, 2023, from https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Hatch, C. (2022, June 28). *Adult Cancer Survivors Have Higher Risk of Cardiovascular Disease Than Those Without Cancer, Study Shows*. Johns Hopkins Medicine Newsroom. Retrieved March 9, 2023, from https://www.hopkinsmedicine.org/news/newsroom/news-releases/adult-cancer-survivors-have-higher-risk-of-cardiovascular-disease-than-those-without-cancer-study-shows#:~:text=The%20authors%20found%20that%20survivors,those%20with%20and%20without%20cancer.

Health Effects of Smoking and Tobacco Use. (2020, April 28). Centers for Disease Control and Prevention. Retrieved February 15, 2023, from https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm

Hossen, J. (2023, June 4). *Heart disease prediction (Accuracy: 0.96%)*. Kaggle. Retrieved July 18, 2023, from https://www.kaggle.com/code/jahidhossen/heart-disease-prediction-accuracy-0-96/notebook

Hossen, M. (2022). Heart Disease Prediction Using Machine Learning Techniques. *American Journal of Computer Science and Technology, 5*. https://doi.org/10.11648/j.ajcst.20220503.11

Hotz, N. (2023a, January 19). *KDD and Data Mining*. Data Science Process Alliance. Retrieved February 27, 2023, from https://www.datascience-pm.com/kdd-and-data-mining/

Hotz, N. (2023b, January 19). *What is CRISP DM?* Data Science Process Alliance. Retrieved February 27, 2023, from https://www.datascience-pm.com/crisp-dm-2/

Hotz, N. (2023c, January 31). *What is SEMMA?* Data Science Process Alliance. Retrieved February 27, 2023, from https://www.datascience-pm.com/semma/

Hu, Z., Lin, X., Kaminga, A. C., & Xu, H. (2020, August 25). *Impact of the COVID-19 Epidemic on Lifestyle Behaviors and Their Association With Subjective Well-Being Among the General Population in Mainland China: Cross-Sectional Study*. National Library of Medicine. Retrieved February 10, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7458584/

IBM. (n.d.). *What is Machine Learning? | IBM*. Retrieved February 8, 2023, from https://www.ibm.com/my-en/topics/machine-learning

ImbalancedLearn. (2023). *RandomOverSampler*. Retrieved July 18, 2023, from https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

Indeed Editorial Team. (2021, December 15). *What Is Research Methodology? (Why It's Important and Types)*. Indeed Career Guide. Retrieved March 2, 2023, from https://www.indeed.com/career-advice/career-development/research-methodology

Institute for Academic Development. (2022, August 29). *Literature review*. The University of Edinburgh. Retrieved February 11, 2023, from https://www.ed.ac.uk/institute-academic-development/study-hub/learning-resources/literature-review

Iterators. (2021, September 15). *Data Collection: Best Methods + Practical Examples | Iterators*. Retrieved March 2, 2023, from https://www.iteratorshq.com/blog/data-collection-best-methods-practical-examples/

Jain, K. (2017, September 12). *Python vs. R vs. SAS – which tool should I learn for Data Science?* Analytics Vidhya. Retrieved February 24, 2023, from https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/

Jain, K. (2020, June 25). *Scikit-learn(sklearn) in Python – the most important Machine Learning tool I learnt last year!* Analytics Vidhya. Retrieved February 17, 2023, from https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/

Javatpoint. (2023a). *Data leakage in machine Learning*. Retrieved July 17, 2023, from https://www.javatpoint.com/data-leakage-in-machine-learning#:~:text=In%20simple%20words%2C%20data%20leakage,poorly%20in%20deployment%20or%20production.%22

Javatpoint. (2023b). *Normalization in Machine Learning*. Retrieved July 9, 2023, from https://www.javatpoint.com/normalization-in-machine-learning

Javed, Z., Maqsood, M., Yahya, T., Amin, Z., Acquah, I., Valero-Elizondo, J., Andrieni, J. D., Dubey, P., Jackson, R. K., Daffin, M. A., Cainzos-Achirica, M., Hyder, A. A., & Nasir, K. (2022). Race, Racism, and Cardiovascular Health: Applying a Social Determinants of Health Framework to Racial/Ethnic Disparities in Cardiovascular Disease. *Circulation: Cardiovascular Quality and Outcomes Logo*, *15*(1). https://doi.org/10.1161/circoutcomes.121.007917

Jeba, E. (2023, June 18). *.ipynb vs .py - MLearning.ai - Medium*. Medium. Retrieved July 6, 2023, from https://medium.com/mlearning-ai/ipynb-vs-py-9d17fbce6669#:~:text=ipynb%20files%20are%20a%20great,py%20files%20are%20better%20suited.

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012072. https://doi.org/10.1088/1757-899x/1022/1/012072

Johns Hopkins Medicine. (2021, November 1). *It's Never Too Late: Five Healthy Steps at Any Age*. Retrieved February 8, 2023, from https://www.hopkinsmedicine.org/health/wellness-and-prevention/its-never-too-late-five-healthy-steps-at-any-age

Kanade, V. (2022, August 30). *What Is Machine Learning? Definition, Types, Applications, and Trends for 2022*. Spiceworks. Retrieved February 8, 2023, from https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/

Karabiber, F. (2023). *Binary Classification*. LearnDataSci. Retrieved July 8, 2023, from https://www.learndatasci.com/glossary/binary-classification/

Katari, K. (2021, December 15). *Seaborn: Python - Towards Data Science*. Medium. Retrieved February 25, 2023, from https://towardsdatascience.com/seaborn-python-8563c3d0ad41

Kelley, K. (2023, June 9). *What is Data Analysis? Methods, Process and Types Explained*. Simplilearn.com. Retrieved July 5, 2023, from https://www.simplilearn.com/data-analysis-methods-process-types-article#what_is_data_analysis

Keras Team. (2023). *KerasTuner*. Keras Documentation. Retrieved July 18, 2023, from https://keras.io/keras_tuner/

Khabaza, T. (2010). *CRISP-DM*. Khabaza. Retrieved February 27, 2023, from http://khabaza.codimension.net/index_files/crispdm.htm

Khan, T. (2022, January 2). *Different types of Encoding - AI ML Analytics*. AI ML Analytics. Retrieved July 15, 2023, from https://ai-ml-analytics.com/encoding/

Khete, T. (2022, May 21). *R vs Python for Data Science and visualization: The language debate*. Medium. Retrieved February 24, 2023, from https://medium.com/@tusharkhete118/r-vs-python-for-data-science-and-visualization-the-language-debate-1aac453e7e29

Kumar, M. (2021, December 15). *R Overview and History*. Medium. Retrieved February 23, 2023, from https://medium.com/@ArtisOne/r-overview-and-history-75ecb036d0df

Lawson, C. (2021, November 15). *How SAS analytics uses machine learning to power data analysis*. Selerity. Retrieved February 24, 2023, from https://seleritysas.com/blog/2021/02/05/how-sas-analytics-uses-machine-learning-to-power-data-analysis/

Lifewire. (2022, August 5). *Windows 10: Everything You Need to Know*. Retrieved February 17, 2023, from https://www.lifewire.com/windows-10-2626217

Luna, J. C. (2022, December). *Python vs R for Data Science: Which Should You Learn?* DataCamp. Retrieved February 23, 2023, from https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Mayo Clinic. (2022a, January 15). *Blood tests for heart disease*. Retrieved February 16, 2023, from https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease/art-20049357

Mayo Clinic. (2022b, May 25). *Coronary artery disease - Symptoms and causes*. Retrieved February 9, 2023, from https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613

Mayo Clinic. (2022c, August 25). *Heart disease - Symptoms and causes*. Retrieved February 9, 2023, from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

Mcleod, S., PhD. (2023, February 20). *Qualitative vs Quantitative Research: Differences, Examples & Methods*. Study Guides for Psychology Students - Simply Psychology. Retrieved March 2, 2023, from https://simplypsychology.org/qualitative-quantitative.html

Merriam-Webster. (2023a). anteroposterior. In *The Merriam-Webster Dictionary*. Retrieved February 16, 2023, from https://www.merriam-webster.com/medical/anteroposterior

Merriam-Webster. (2023b). posteroanterior. In *The Merriam-Webster Dictionary*. Retrieved February 16, 2023, from https://www.merriam-webster.com/medical/posteroanterior

Mhadhbi, N. (2021, December 21). *Python Tutorial: Streamlit*. DataCamp. Retrieved July 6, 2023, from https://www.datacamp.com/tutorial/streamlit

Miah, E. (2017, November 7). *Key Factors in The Successful Use of Machine Learning*. Data Science Central. Retrieved February 8, 2023, from https://www.datasciencecentral.com/key-factors-in-the-successful-use-of-machine-learning/

Microsoft. (n.d.). *Windows 10 Home and Pro - Microsoft Lifecycle*. Microsoft Learn. Retrieved February 17, 2023, from https://learn.microsoft.com/en-us/lifecycle/products/windows-10-home-and-pro

Mock, J. (2019, December 26). *Is It Ever Too Late to Start Being Healthy?* Discover Magazine. Retrieved February 8, 2023, from https://www.discovermagazine.com/health/is-it-ever-too-late-to-start-being-healthy

Moffitt, C. (2021, November 15). *16 Reasons to Use VS Code for Developing Jupyter Notebooks*. Practical Business Python. Retrieved July 7, 2023, from https://pbpython.com/vscode-notebooks.html

Mohaimin, M., MD. (2022, April 18). *<unk>HeartDisease EDA <unk> + Prediction*. Kaggle. Retrieved July 18, 2023, from https://www.kaggle.com/code/mushfirat/heartdisease-eda-prediction

Nagai, M., Kario, K., & Kario, K. (2010). Sleep duration as a risk factor for cardiovascular disease- A review of the recent literature. *Current Cardiology Reviews*, *6*(1), 54–61. https://doi.org/10.2174/157340310790231635

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of Healthcare Engineering*, *2022*, 1–9. https://doi.org/10.1155/2022/7351061

Nandal, N., Goel, L., & Tanwar, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. *F1000Research*, *11*, 1126. https://doi.org/10.12688/f1000research.123776.1

Narang, M. (2023, February 6). *Top 11 Programming Languages for Data Science*. KnowledgeHut. Retrieved February 23, 2023, from https://www.knowledgehut.com/blog/data-science/programming-languages-for-data-science

National Cancer Institute. (n.d.). *NCI Dictionary of Cancer Terms | Heart Disease*. Retrieved February 8, 2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/heart-disease

National Institute of Health. (2023). NCI Dictionary of Cancer Terms. In *National Cancer Institute*. Retrieved February 16, 2023, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/complication

NHS. (2021, November 30). *Prevention - Coronary Artery Disease*. Retrieved July 17, 2023, from https://www.nhs.uk/conditions/coronary-heart-disease/prevention/

NIAAA. (2023). *Drinking Levels Defined*. National Institute on Alcohol Abuse and Alcoholism. Retrieved March 9, 2023, from https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking

NIH. (2022, March 24). *How smoking affects the heart and blood vessels*. Retrieved July 16, 2023, from https://www.nhlbi.nih.gov/health/heart/smoking

Nunez, K. (2021, October 20). *How Is Smoking Linked to Heart Disease and Other Heart Issues?* Healthline. Retrieved February 16, 2023, from https://www.healthline.com/health/smoking/how-does-smoking-affect-your-heart

NVIDIA. (n.d.). *What is PyTorch?* NVIDIA Data Science Glossary. Retrieved February 17, 2023, from https://www.nvidia.com/en-us/glossary/data-science/pytorch/

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100–1113. https://doi.org/10.1515/med-2022-0508

Patil, P. (2022, May 30). What is Exploratory Data Analysis? - Towards Data Science. *Medium*. Retrieved July 5, 2023, from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Programiz. (n.d.). *9 Best Python IDEs and Code Editors*. Retrieved February 23, 2023, from https://www.programiz.com/python-programming/ide

Python Institute. (n.d.). *About Python*. Retrieved February 23, 2023, from https://pythoninstitute.org/about-python

Pytlak, K. (2022, February 16). *Personal Key Indicators of Heart Disease*. Kaggle. Retrieved July 5, 2023, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Quantum. (2021, December 11). *Data Science project management methodologies - DataDrivenInvestor*. Medium. Retrieved February 26, 2023, from https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb

Ray, S. (2020, June 26). *8 Proven Ways for improving the "Accuracy" of a Machine Learning Model*. Analytics Vidhya. Retrieved February 16, 2023, from https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

Red Hat. (2019, January 8). *What is an IDE?* Retrieved February 25, 2023, from https://www.redhat.com/en/topics/middleware/what-is-ide

Reuters Staff. (2008, May 26). *Many ignorant of heart attack signs: study*. U.S. Retrieved February 10, 2023, from https://www.reuters.com/article/us-heart-symptoms/many-ignorant-of-heart-attack-signs-study-idUKN2322118520080526

Sarnak, M. J., Amann, K., Bangalore, S., Cavalcante, J. L., Charytan, D. M., Craig, J. C., Gill, J. S., Hlatky, M. A., Jardine, A. G., Landmesser, U., Newby, L. K., Herzog, C. A., Cheung, M., Wheeler, D. A., Winkelmayer, W. C., & Marwick, T. H. (2019). Chronic kidney disease and coronary artery disease. *Journal of the American College of Cardiology*, *74*(14), 1823–1838. https://doi.org/10.1016/j.jacc.2019.08.1017

SAS. (2021a). *SAS History*. Retrieved February 23, 2023, from https://www.sas.com/en_ph/company-information/history.html

SAS. (2021b). *SAS Studio*. Retrieved February 23, 2023, from https://www.sas.com/en_us/software/studio.html

SAS. (2021c). *Why SAS?* Retrieved February 23, 2023, from https://www.sas.com/en_ph/company-information/why-sas.html

Sasayama, S. (2008). Heart Disease in Asia. *Circulation*, *118*(25), 2669–2671. https://doi.org/10.1161/circulationaha.108.837054

Saturn Cloud. (2023, May 25). *What is the ipynb Jupyter Notebook File Extension and How to Open It?* Retrieved July 6, 2023, from https://saturncloud.io/blog/what-is-the-ipynb-jupyter-notebook-file-extension-and-how-to-open-it/#:~:text=The%20ipynb%20file%20extension%20stands,%2C%20visualizations%2C%20and%20narrative%20text.

scikit-learn. (2023a). *GridSearchCV*. Scikit-learn. Retrieved July 18, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

scikit-learn. (2023b). *RandomizedSearchCV*. Scikit-learn. Retrieved July 18, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Shahid. (2021, June 18). *Data Science Tools Comparison -SAS, R, Python*. CodeForGeek. Retrieved February 24, 2023, from https://codeforgeek.com/data-science-tools-comparison-sas-r-python/

Shahjehan, R., & Bhutta, B. (2022). *Coronary Artery Disease* [Internet]. StatPearls. https://pubmed.ncbi.nlm.nih.gov/33231974/

Sidey-Gibbons, J. a. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0681-4

Simplilearn. (2022, November 23). *What is R: Overview, its Applications and what is R used for?* Simplilearn.com. Retrieved February 23, 2023, from https://www.simplilearn.com/what-is-r-article

Simplilearn. (2023a, January 24). *What is PyTorch, and How Does It Work: All You Need to Know*. Retrieved February 17, 2023, from https://www.simplilearn.com/what-is-pytorch-article

Simplilearn. (2023b, February 16). *What is a Confusion Matrix in Machine Learning?* Simplilearn.com. Retrieved July 20, 2023, from https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning#:~:text=A%20confusion%20matrix%20presents%20a,actual%20values%20of%20a%20classifier.

Smith, J. (2023, January 10). *Is red wine good for you?* Retrieved February 16, 2023, from https://www.medicalnewstoday.com/articles/265635

Stinchcombe, C. (2022, April 20). *Aerobic vs. Anaerobic Exercise: Which Benefits You More?* GoodRx. Retrieved February 16, 2023, from https://www.goodrx.com/well-being/movement-exercise/aerobic-vs-anaerobic-exercise

Streamlit. (n.d.). *Cloud*. Retrieved July 6, 2023, from https://streamlit.io/cloud

Tableau. (2023). *What Is Data Visualization? Definition, Examples, And Learning Resources*. Retrieved July 11, 2023, from https://www.tableau.com/learn/articles/data-visualization

Taylor, M. (2023, February 13). How Asthma and Heart Disease Are Connected—and What to Do About It. *Health Central*. Retrieved July 16, 2023, from https://www.healthcentral.com/condition/asthma/how-asthma-and-heart-disease-are-connected

The Free Dictionary. (2023). Diagnostic tool. In *The Free Dictionary*. Retrieved February 16, 2023, from https://medical-dictionary.thefreedictionary.com/Diagnostic+tool

Towards AI. (2021, October 13). What is an ML model? *Towards AI*. Retrieved July 8, 2023, from https://towardsai.net/p/machine-learning/what-is-an-ml-model#:~:text=What%20is%20building%20a%20model,make%20predictions%20and%20obtain%20results.

UCSF. (n.d.). *Coronary Artery Disease Diagnosis*. UCSF Health. Retrieved February 10, 2023, from https://www.ucsfhealth.org/conditions/coronary-artery-disease/diagnosis

UNC-Chapel Hill Writing Center. (2021, September 22). *Literature Reviews*. The Writing Center • University of North Carolina at Chapel Hill. Retrieved February 11, 2023, from https://writingcenter.unc.edu/tips-and-tools/literature-reviews/

Valadkhani, M. (2018, March 9). *Knowledge Discovery in Data (KDD) Process*. LinkedIn. Retrieved February 27, 2023, from https://www.linkedin.com/pulse/knowledge-discovery-data-kdd-process-mohammad-valadkhani/

Verma, S. (2022, January 29). *Why switch to JupyterLab from jupyter-notebook? - Analytics Vidhya*. Medium. Retrieved February 25, 2023, from https://medium.com/analytics-vidhya/why-switch-to-jupyterlab-from-jupyter-notebook-c6d98362945b

Vu, T. (2022, January 25). *7 Reasons Why You Should Use Jupyterlab for Data Science*. Medium. Retrieved February 17, 2023, from https://towardsdatascience.com/7-reasons-why-you-should-use-jupyterlab-for-data-science-7c2a3db8755a

Watson, S. (2020, June 24). *Understanding Coronary Artery Disease and How to Prevent It*. Healthline. Retrieved February 16, 2023, from https://www.healthline.com/health/high-cholesterol/preventing-CAD

Wawro, A. (2023, January 20). *Windows 11 problems and fixes — everything we know so far*. Tom's Guide. Retrieved February 17, 2023, from https://www.tomsguide.com/news/windows-11-problems-and-fixes-everything-we-know-so-far

WebMD Editorial Contributors. (2022, November 1). *Your Heart Rate* (J. Beckerman, Ed.). WebMD. Retrieved February 16, 2023, from https://www.webmd.com/heart-disease/heart-failure/watching-rate-monitor

World Health Organization. (2019, June 11). *Cardiovascular diseases*. Retrieved February 16, 2023, from https://www.who.int/health-topics/cardiovascular-diseases

Yetman, D. (2022, November 9). *How Much Does a Coronary Calcium Scan Cost?* Healthline. Retrieved February 9, 2023, from https://www.healthline.com/health/heart/coronary-calcium-scan-cost

Zhou, S. (2021, June 21). *Knowing the Importance of Being Healthy in your Early 20s*. Flexispot. Retrieved February 8, 2023, from https://www.flexispot.ca/spine-care-center/knowing-the-importance-of-being-healthy-in-your-early-20s/

Appendices

Project Proposal Form (PPF)

Office Record	Receipt
Date Received:	Student name:
Received by whom:	Student number:
	Received by:
	Date:

DRAFT PROJECT PROPOSAL FORM

Proposal ID :

Supervisor : **Assoc. Prof. Dr. Imran Medi**
Dr. Preethi Subramanian
Mr. Raheem Mafas
Dr. Vazeerudeen Hameed
Dr. Murugananthan Velayutham

Student Name : **EDWARD LEONARDO**

Student No : **TP058284**

Email Address : **tp058284@mail.apu.edu.my & edwardleonardo14@gmail.com**

Programme Name : **Computer Science with Specialism in Data Analytics (CSDA)**

Title of project : **Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle**

Please record which module(s) your topic is related to:

Text Analytics and Sentiment Analysis (CT107-3-3-TXSA)
Programming for Data Analysis (CT127-3-2-PDFA)
Data Mining and Predictive Modelling (CT119-3-2-DMPM)

Figure 131: PPF - Page 1

1. Introduction

Currently, the amount of people suffering from heart disease has been on a steady rise. Heart disease is one of the leading causes of death in the world, especially in America, as it has been come the number one cause of death in the country (CDC, 2022c), and the country itself also has become the fourth highest number of heart disease death (Preidt, 2020).

The main cause of this phenomena is High Blood Pressure and High Blood Cholestrol (CDC, 2022b). Several other factors also can cause a higher risk of heart disease, such as diabetes, overweight/obesity, smoking, and excessive alcohol use. Unhealthy Diet and Physical Inactivity lifestyle also can be a major factor in causing heart disease.

One of the major causes why heart disease is still very common, especially in some countries such as America, India, and Indonesia are the general population still not informed or educated on the subject, and even ignorant on the problem. Many of them only start learning about heart disease starting when they have been diagnosed or have suffered from one. This may have been too late for some people, since the best way to deal with heart disease is to have a healthier lifestyle from the start, thus preventing or lowering the .

To combat this issue, educating people early on the cause and the symptoms are the best way on how to suppress the increasing trend of heart disease. This can be done by educating people using information that they can observe daily.

In this project, the analysis will go deep dive on the general information of heart disease, what is the main cause of heart disease, the early symptoms of heart disease, and personal key indicator that have the most effect on causing heart disease.

Figure 132: PPF - Page 2

2. Problem Statement

1. **Lack of Problem Acknowledgement from the Public**

   The main cause of why heart disease is still very fatal and affecting many people is people is still uneducated and even ignorant of the disease (Reuters Staff, 2008). The reason is the lack of education regarding this topic, and people still associating that heart disease will only happened to specific demographic, such as obese people and old people. The general population also did not know that heart disease is very easy to be avoided, mainly by adopting a healthier lifestyle, the risk of heart disease already decreased dramatically.

2. **No easy self-diagnosis tools available**

   The other reason why heart disease cases is still very high is due to the lack of easy self-diagnosis tools available for the public to use. This problem may result in people that already showing an early signs of heart disease being clueless on the issue. Even though the easiest method to detect heart disease is through doing blood test on a lab, an easier way to test it at home is by checking based on their personal key indicator, to gain a better understanding if the person has risk of having a heart disease in the future, so they prevent it by doing some changes on their lifestyle or seek professional help earlier.

Figure 133: PPF - Page 3

3. Project Aim and Objectives

The aim of the analysis is to educate the readers on what is the main causes of heart disease in an easy and generalized manner. This in turn will increase the general population's knowledge on the topic of heart disease, causing more people to be more aware of it, and ultimately to decrease the amount of people suffering from heart disease.

The objectives of this project are:

1. To do an Exploratory Data Analysis (EDA) on Personal Key Indicators of heart disease.
2. To produce an analysis result paper based on the result of the EDA.
3. To create a predictive model using Machine Learning to help predict if a person has a heart disease based on the Personal Key Indicators.
4. To help people do a self-diagnosis of heart disease using key indicators that they can measure themselves.
5. To educate the public about heart disease, the symptoms, and the causes that can increase the risk of heart disease.
6. To reduce the amount of people suffering from heart disease through a preventive method.

Figure 134: PPF - Page 4

4. Literature Review

The literature will focus on relevant information regarding the topic, such as the general information about heart disease, what is the main causes that increase the risk of having a heart disease, the problem on why heart disease cases number is still so high, and how to prevent heart disease. From the review, a better understanding of the topic can be achieved.

Heart disease, as the name suggests, is a range of conditions that affects the blood flow to the heart (CDC, 2022a). there are multiple classifications of heart disease, such as blood vessel disease, irregular heartbeats or arrhythmias, congenital heart defects, disease of the heart muscle, and heart valve disease (Mayo Clinic, 2022). The most common heart disease that is happening is coronary artery disease (CAD), while some others type of heart disease can be obtained due to family genetical inheritance, virus/infection, or a defect during infancy. Coronary artery disease is the one that most people associated with the word of heart disease; thus, CAD is the chosen heart disease that will be the main topic of this project.

The cause of CAD is a buildup of fatty plaques in the arteries, or more medically known as atherosclerosis (Mayo Clinic, 2022). The main cause of atherosclerosis is poor lifestyle choices, such as poor diet, physical inactivity, obesity, and tobacco smoking. Therefore, CAD become the most common heart disease, since obesity rate has been on the rise, and predicted to double by the next eight years (Johnson, 2022). Another main factor of the increase in unhealthy lifestyle rate is the COVID-19 Pandemic, where research has been conducted in China, showing that due to the pandemic, unhealthy lifestyle rate has been increasing, due to people inactivity in their home, stress level increasing, causing people to smoke, and more unhealthy diet due to more delivery food ordering (Hu et al., 2020).

Another cause of why CAD is still very common is due to people lack education of the topic, causing them to be oblivious or even ignorant regarding the topic. This cause less people taking preventive action to reduce the risk of them suffering a heart disease.

The best way of reducing the risk of suffering a heart disease, in this case CAD, is to do a preventive measure early in adulthood. This can be done by leading a healthier lifestyle, such as having a healthier diet, become more physically active, keep checking cholesterol and

Figure 135: PPF - Page 5

blood pressure to make sure it is under control, giving up smoking for any smokers, reduce alcohol consumption, and stay in a healthy Body Mass Index (BMI) (NHS, 2021). For people that have suffered from heart disease previously, all the above would still be effective, with the addition of their prescribed medication, to prevent further problems developing.

Figure 136: PPF - Page 6

5. Deliverable

The result of the analysis should produce results that can be easily understood by the general masses. The result should be in a form of an analysis result paper, where the user can be educated and gain the knowledge they need about the subject. Another form of result will be a predictive model using Machine Learning to help public to self-diagnose using personal key indicators. The main user target for this Project is the general population, especially young adults around the age of 20-30s, and adults. Using the result of this project, hopefully the target audience can use the information to make decision regarding to their lifestyle, so they may reduce the risk of suffering from heart disease.

The list of deliverables that should be produced by this project:

1. An Analysis Result Paper based on Exploratory Data Analysis (EDA) using the dataset.
2. A Predictive Model that can be used by the user to self-diagnose if they have a risk to suffer heart disease using personal key indicator.

Figure 137: PPF – Page 7

6. References

CDC. (2022a, July 12). Heart Disease Resources | CDC. Centers for Disease Control and Prevention. Retrieved December 13, 2022, from https://www.cdc.gov/heartdisease/about.htm

CDC. (2022b, September 8). Heart Disease and Stroke | CDC. Retrieved December 14, 2022, from https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm

CDC. (2022c, October 14). Heart Disease Facts | CDC. Centers for Disease Control and Prevention. Retrieved December 14, 2022, from https://www.cdc.gov/heartdisease/facts.htm

Hu, Lin, Kaminga, & Xu. (2020, August 25). Impact of the COVID-19 Epidemic on Lifestyle Behaviors and Their Association With Subjective Well-Being Among the General Population in Mainland China: Cross-Sectional Study. National Library of Medicine. Retrieved December 13, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7458584/

Johnson, S. (2022, October 19). Obesity rates likely to double by 2030 with highest rises in lower-income countries. The Guardian. Retrieved December 13, 2022, from https://www.theguardian.com/global-development/2022/mar/04/obesity-rates-likely-to-double-by-2030-with-highest-rises-in-lower-income-countries

Mayo Clinic. (2022, August 25). Heart disease - Symptoms and causes. Retrieved December 13, 2022, from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

NHS. (2021, November 30). Coronary heart disease - Prevention. NHS UK. Retrieved December 13, 2022, from https://www.nhs.uk/conditions/coronary-heart-disease/prevention/

Preidt. (2020, December 9). Heart Disease Is World's No. 1 Killer. WebMD. Retrieved December 14, 2022, from https://www.webmd.com/heart-disease/news/20201209/heart-disease-is-worlds-no-1-killer

Reuters Staff. (2008, May 26). Many ignorant of heart attack signs: study. U.S. Retrieved December 13, 2022, from https://www.reuters.com/article/us-heart-symptoms/many-ignorant-of-heart-attack-signs-study-idUKN2322118520080526

Figure 138: PPF - Page 8

Ethics Form

Office Record	Receipt – Fast-Track Ethical Approval
Date Received:	Student name: Student number: Received by:
Received by whom:	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: **B.Sc. (Hons) Computer Science Specialism in Data Analytics**

Title of project / assignment: **Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle**

Name of student researcher: **Edward Leonardo**

Name of supervisor / lecturer: **Dr. Murugananthan Velayutham**

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.

Fast-Track Ethical Approval Form ver. 3.0 (Dec 2015) Page 1 of 4

Figure 139: Ethics Form - Page 1

			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 research will be done using a Google Survey Form where the participants can fill out the questions as they find which one is most relatable to their opinion. No coercion will be conducted to any participants, and if there are any questions that the participants may want to pass on, then they are entitled to do so.	
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 *(Student Researcher)* Print Name **Edward Leonardo** Date **28th February 2023**

Fast-Track Ethical Approval Form ver. 3.0 (Dec 2015) Page 2 of 4

Figure 140: Ethics Form - Page 2

Please note that any variation to that contained within this document that in any way affects ethical 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 your 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… … *V. Muruganantham*… Print Name…**Dr. Murugananthan Velayutham** … Date…**3rd March 2023**.
(Supervisor/Lecturer)

E-signature… … … … … … … … … … … … … Print Name… … … … … … … … … … … … … Date… … … …
(Second Academic Signatory)

Figure 141: Ethics Form - Page 3

Office Record	Receipt – Appendix A (Fast-Track Ethics Form)
Date Received:	Student name: Student number:
Received by whom:	Received by: Date:

APPENDIX A
AUTHORISATION FOR USE OF KEY DOCUMENTS

Completion of Appendix A is required when for good reasons key documents are not available when a fast track application is approved by the supervisor/module lecturer and second academic signatory.

I have now checked and approved all the key documents associated with this proposal e.g. consent form, information sheet, questionnaire, interview schedule

Title of project/assignment: **Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle**

Name of student researcher: **Edward Leonardo**

Student ID: **TP058284** Intake: **APU3F2211CS(DA)**

E-signature… … *V. Muruganantham*… Print Name…**Dr. Murugananthan Velayutham** … Date…**3rd March 2023**.
(Supervisor/Lecturer)

Fast-Track Ethical Approval Form ver. 3.0 (Dec 2015) Page 4 of 4

Figure 142: Ethics Form - Page 4

Supervisor Meeting Logs

(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: Edward Leonardo **Date:** 18th January 2023 **Meeting No:** 01

Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle

Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Murugananthan Velayutham

Supervisor's signature: …….

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

1. Introduction to Supervisor
2. Asking Supervisor's opinion for the dataset chosen for the project
3. Asking for Supervisor's preferred Meeting method
4.

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

1. can continue with IR if dataset is suitable for the project's objective
2. Supervisor is fine with either online or on-campus meeting
3.
4.

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

1. Proceed with the IR
2. PPF to be discussed in the next meeting
3.

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.

Project Log Sheet

Figure 143: Supervisor Meeting Log - Meeting 1

(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: Edward Leonardo **Date:** 7th February 2023 **Meeting No:** 02

Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle

Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Murugananthan Velayutham

Supervisor's signature: ……

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

1. Discussing PPF content
2.
3.
4.

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

1. The aim and objective can be revised to be more focused
2.
3.
4.

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

1. Proceed with the IR
2.
3.

[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.

Project Log Sheet

Figure 144: Supervisor Meeting Log - Meeting 2

(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: Edward Leonardo **Date:** 9th March 2023 **Meeting No:** 03

Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle

Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Murugananthan Velayutham

Supervisor's signature: ……

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

1. Discussing current IR progress
2. Discussing Questionnaire's content
3.
4.

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

1. The aim and objective can be revised to be more focused
2. In methodology explanation, explain how the steps will be used during the project's execution
3. The Questionnaire is approved
4.

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

1. Finish the IR
2. Share the questionnaire to gain data
3.

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.

Project Log Sheet

Figure 145: Supervisor Meeting Log - Meeting 3

(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: Edward Leonardo **Date:** 5th July 2023 **Meeting No:** 04

Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle
Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Murugananthan Velayutham
Supervisor's signature: ….V Murugananthan….

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

1. Asking Supervisor opinion regarding Machine Learning Model Result
2. Asking Supervisor regarding changes on the Documentation
3. Asking Supervisor Regarding the next Supervisory Meeting Schedule
4.

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

1. The Model is sufficient, but need to try to improve the Accuracy, focuses on LinearSVC
2. Search other projects to use as benchmarking
3. Research for Web App UI/UX Design, making it user-friendly
4. Next Meeting will be either Wednesday/Thursday, 4PM

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

1. HP Tuning the selected model to get better results
2. Start the documentation process
3.

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.

Project Log Sheet

Figure 146: Supervisor Meeting Log - Meeting 4

(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: Edward Leonardo **Date:** 13th July 2023 **Meeting No:** 05

Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle
Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Murugananthan Velayutham
Supervisor's signature: ... V.Murugananthan.....

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

1. Asking Supervisor opinion regarding Deployment and Final Machine Learning Model Choice
2. Asking Supervisor Regarding the next Supervisory Meeting Schedule
3.
4.

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

1. The results are fine can fine tune
2. Improvement on the deployment can be done
3. Do benchmarking on the model's results, compared to others' works
4. The next meeting is scheduled on 18th of July 2023

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

1. Finish the documentation
2.
3.

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.

Project Log Sheet

Figure 147: Supervisor Meeting Log - Meeting 5

(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: Edward Leonardo **Date:** 18th July 2023 **Meeting No:** 06

Project title: Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle
Intake: APU3F2211CS(DA)

Supervisor's name: Dr. Murugananthan Velayutham
Supervisor's signature: ……

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

1. Asking Supervisor opinion regarding the progress that has been made
2. Recapping the whole project's outcome, and improvement compared to the previous meeting
3. Discussion regarding the GCMT submission
4.

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

1. The results are produced.
2. Demo of the deployment before submission will be good.
3. Recommended to do a meeting with Second Marker, to ask for his opinion
4. The next meeting is scheduled either on the 20th of July or the 24th of July.

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

1. Finish the demo deployment.
2.
3.

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.

Project Log Sheet

Figure 148: Supervisor Meeting Log - Meeting 6

Poster Design

Figure 149: Poster Design

Bahasa Indonesia Version of Questionnaire Design

Survei Mengenai Penyakit Jantung

Kepada Peserta,

Nama saya Edward Leonardo, TP058284, Murid Tahun Terakhir di Asia Pacific University of Technology and Innovation, dengan jurusan Computer Science dengan keahlian dalam Data Analytics. Saya melakukan survei ini dalam rangka penelitian untuk Tugas Akhir Tahun/Skripsi saya, dengan judul "Analisa Data dan Model Prediktif tentang Penyakit Jantung berdasarkan Pola Hidup Masyarakat". Tujuan utama survei ini untuk mendapatkan data dari sudut pandang peserta mengenai penyakit jantung, dan mengambil data mengenai kesehatan umum peserta berdasarkan gaya hidup mereka. Survei ini hanya membutuhkan 10 menit.

Mohon diingat bahwa peserta:

1. Peserta memberikan jawaban survei dengan sukarela, tanpa paksaan dari pihak survei.
2. Peserta diperbolehkan untuk tidak menyelesaikan survei kapan pun dengan alasan apa pun, tanpa dikenakan penalti.
3. Peserta diperbolehkan untuk tidak menjawab pertanyaan yang mereka tidak ingin jawab.
4. Data yang didapat akan disimpan dengan kerahasiaan penuh, dan jika data diterbitkan, maka data identifikasi tidak ada disebarkan.
5. Data hanya digunakan untuk kebutuhan akademik.
6. Kejujuran peserta sangat dihargai.
7. Peserta diperkenankan untuk mendapatkan informasi lebih lanjut mengenai penelitian ini dengan menghubungi pihak survei lewat informasi kontak dibawah

Terima Kasih sebelumnya untuk partisipasinya, Saya harap anda semua memiliki hari yang menyenangkan!

Dengan appresiasi,
Edward Leonardo
TP058284

Informasi Kontak:
Email: TP058284@mail.apu.edu.my / edwardleonardo14@gmail.com
WhatsApp: +601123585769

Figure 150: Questionnaire - Bahasa Indonesia Version - First Page

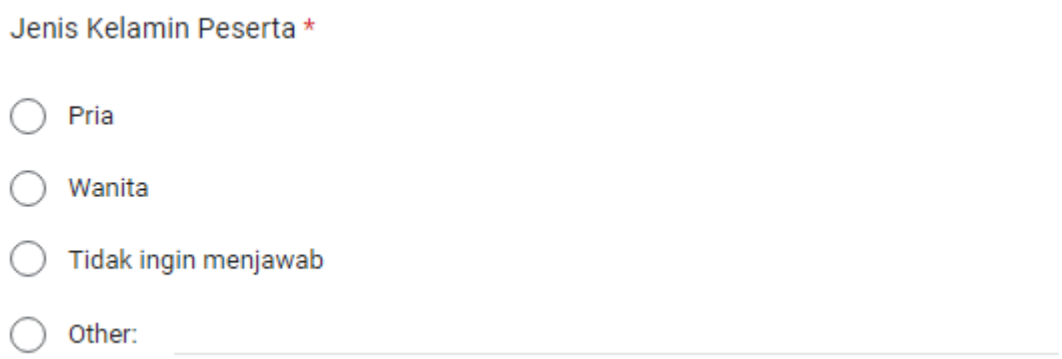


Figure 151: Questionnaire Section A – Bahasa Indonesia Version – Question 1

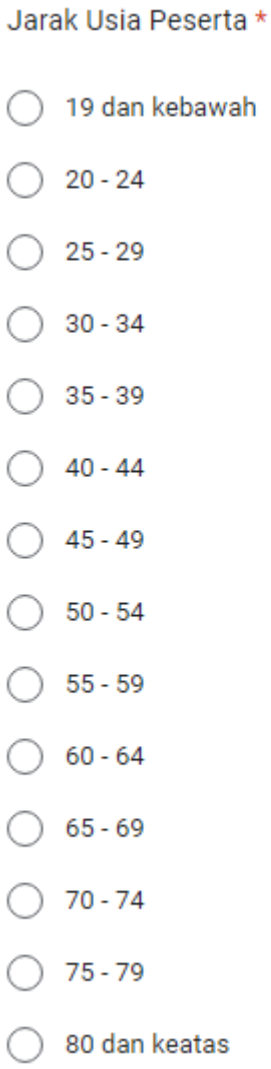


Figure 152: Questionnaire Section A – Bahasa Indonesia Version – Question 2

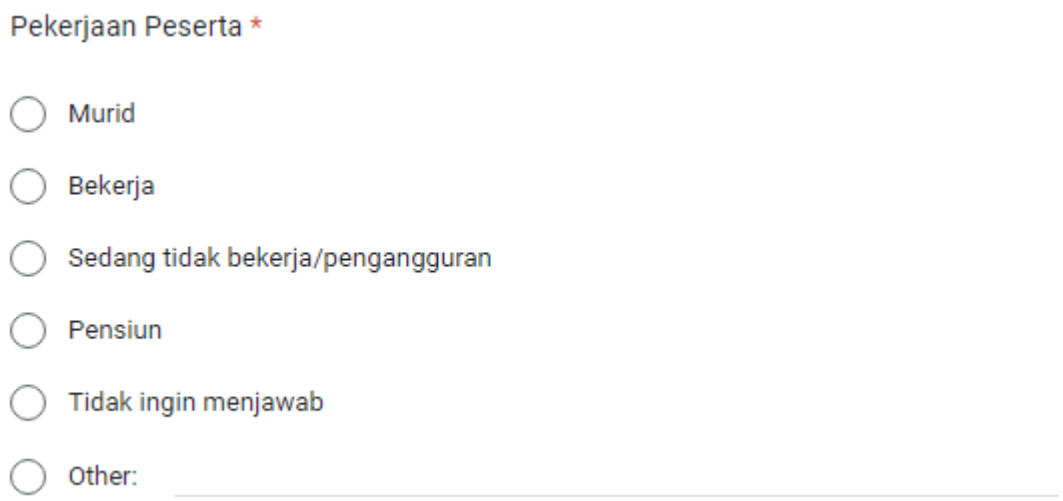


Figure 153: Questionnaire Section A – Bahasa Indonesia Version – Question 3

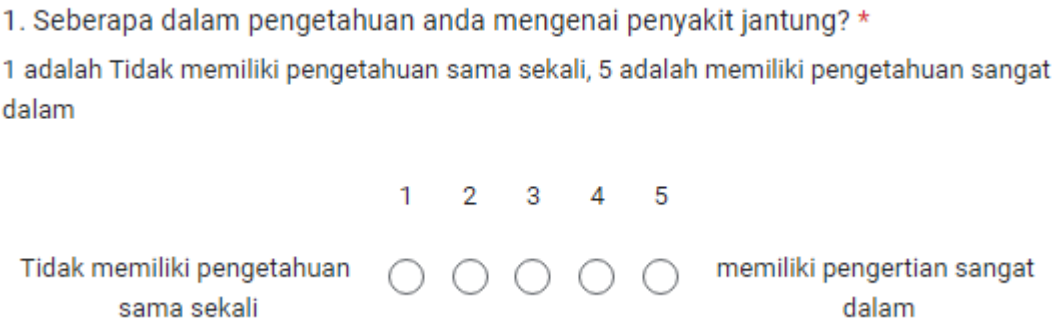


Figure 154: Questionnaire Section B – Bahasa Indonesia Version – Question 1

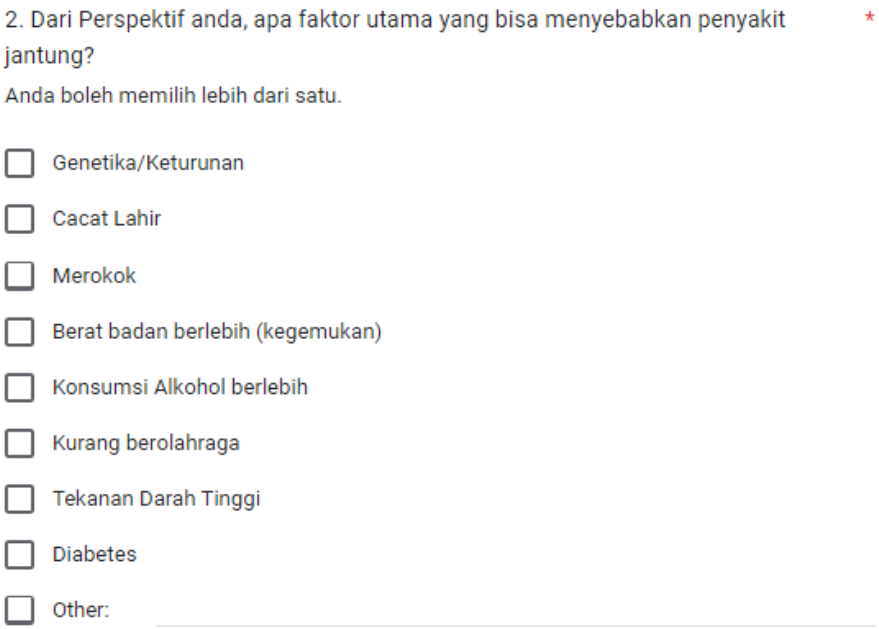


Figure 155: Questionnaire Section B – Bahasa Indonesia Version – Question 2

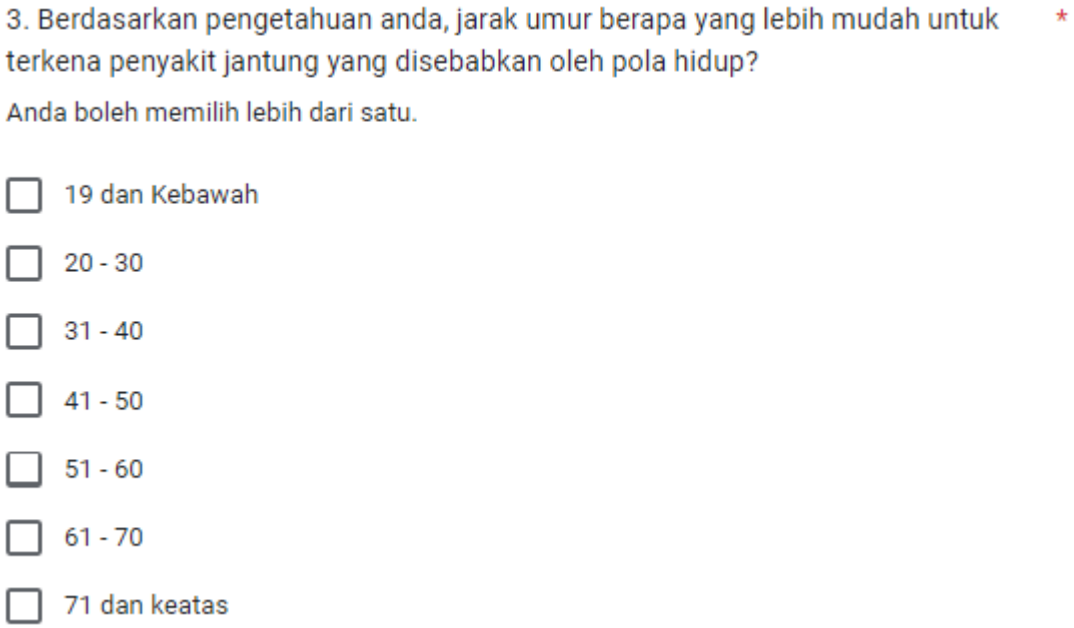


Figure 156: Questionnaire Section B – Bahasa Indonesia Version – Question 3

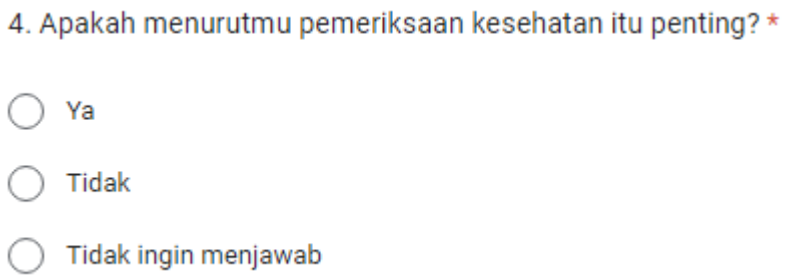


Figure 157: Questionnaire Section B – Bahasa Indonesia Version – Question 4

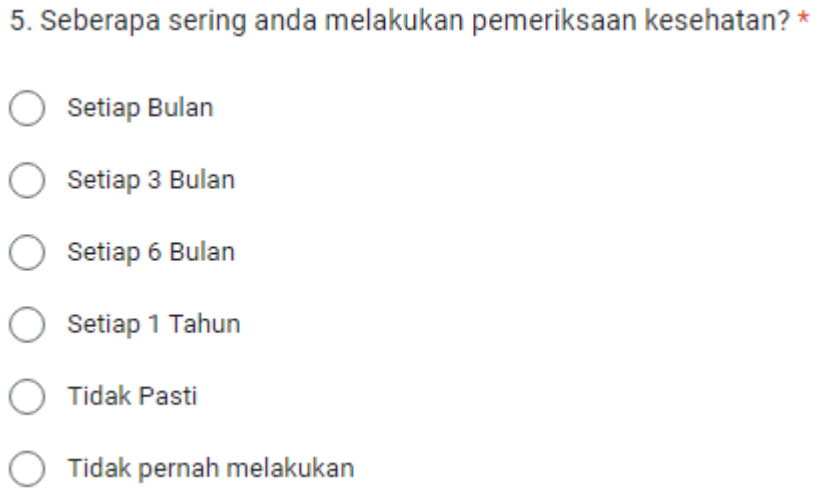


Figure 158: Questionnaire Section B – Bahasa Indonesia Version – Question 5

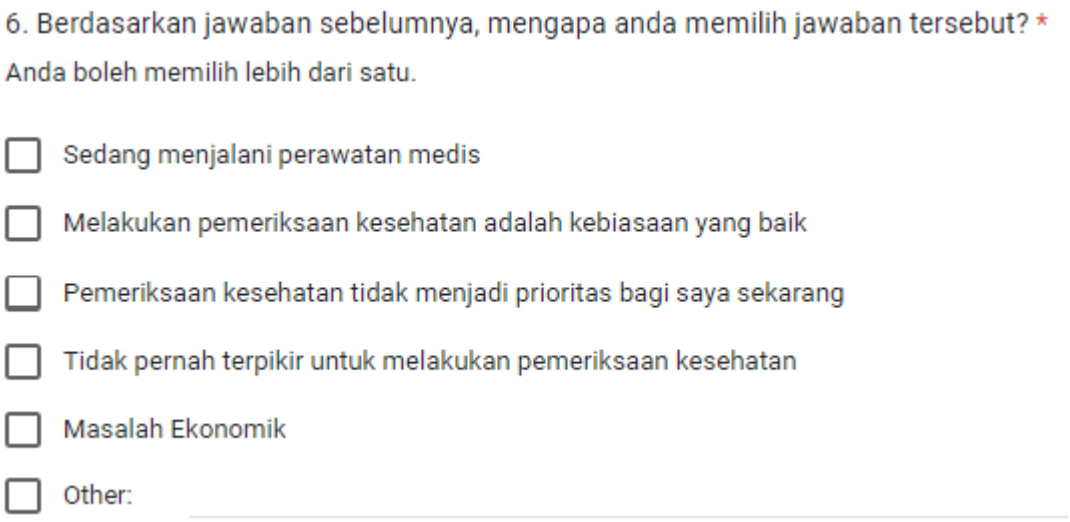


Figure 159: Questionnaire Section B – Bahasa Indonesia Version – Question 6

7. Menurut anda, apakah masyarakat umum mengetahui bahwa Penyakit Jantung yang disebabkan oleh pola hidup bisa dihindari? *

- ◯ Iya, dan semua orang melakukan yang terbaik untuk menghindari itu
- ◯ Iya, tapi hanya sebagian orang mencoba untuk menghindari itu
- ◯ Iya, tapi yang sangat sedikit orang mencoba untuk menghindari itu
- ◯ Tidak
- ◯ Other: _____

Figure 160: Questionnaire Section B – Bahasa Indonesia Version – Question 7

8. Apakah anda mau alat diagnosa mandiri yang bisa membantu anda memeriksa apakah anda memiliki risiko menderita dari penyakit jantung? *

- ◯ Ya
- ◯ Tidak
- ◯ Other: _____

Figure 161: Questionnaire Section B – Bahasa Indonesia Version – Question 8

9. Menurut anda, dalam bentuk media apakah yang terbaik untuk meningkatkan pengetahuan awam mengenai penyakit jantung? *

Anda boleh memilih lebih dari satu.

- ☐ dasbor menunjukkan hasil penelitian
- ☐ Poster
- ☐ Alat Diagnosa mandiri
- ☐ Kampanye oleh medis profesional
- ☐ Aplikasi Web
- ☐ Other: _____

Figure 162: Questionnaire Section B – Bahasa Indonesia Version – Question 9

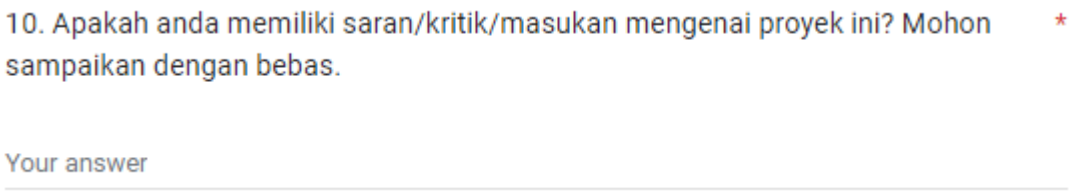


Figure 163: Questionnaire Section B – Bahasa Indonesia Version – Question 10

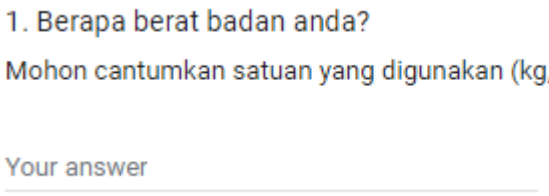


Figure 164: Questionnaire Section C – Bahasa Indonesia Version – Question 1

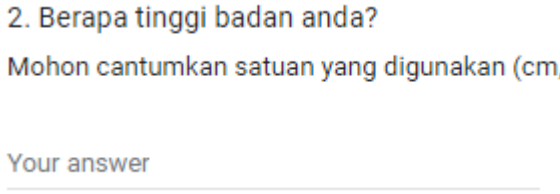


Figure 165: Questionnaire Section C – Bahasa Indonesia Version – Question 2

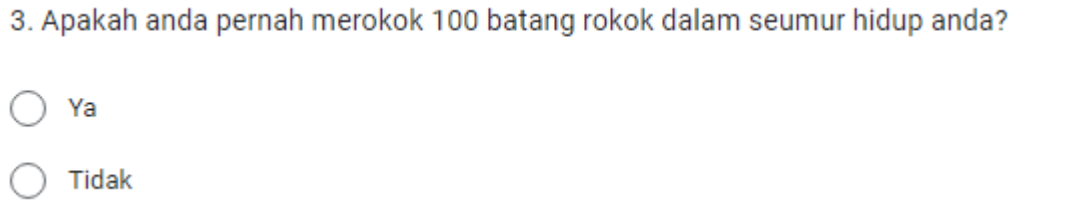


Figure 166: Questionnaire Section C – Bahasa Indonesia Version – Question 3

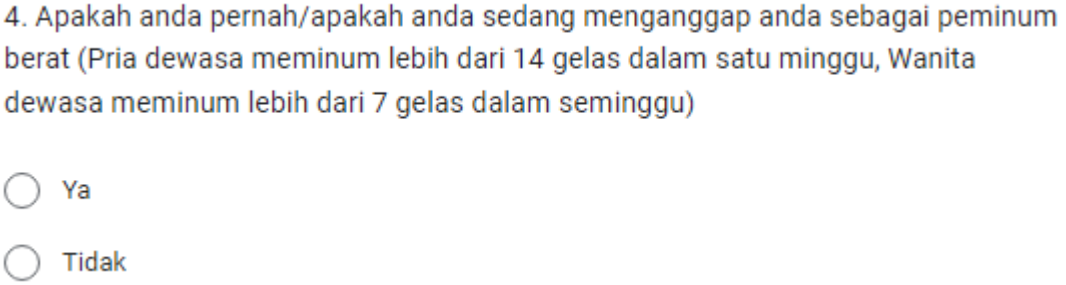


Figure 167: Questionnaire Section C – Bahasa Indonesia Version – Question 4

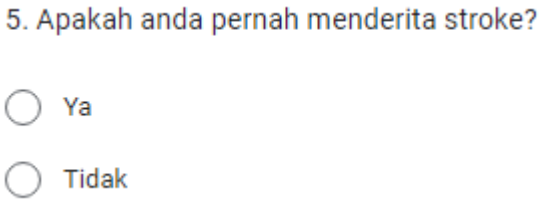


Figure 168: Questionnaire Section C – Bahasa Indonesia Version – Question 5

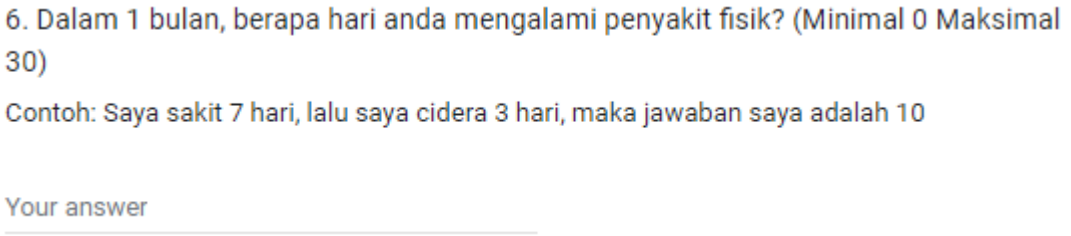


Figure 169: Questionnaire Section C – Bahasa Indonesia Version – Question 6

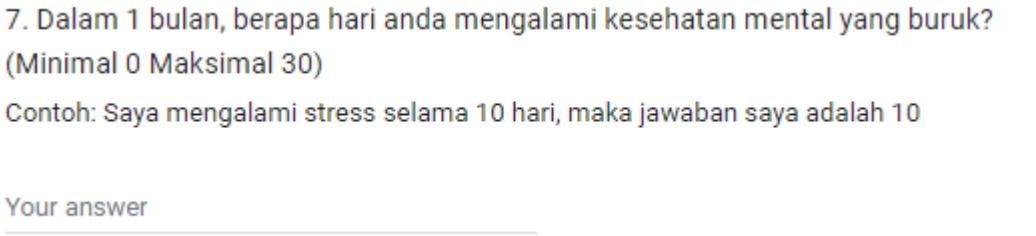


Figure 170: Questionnaire Section C – Bahasa Indonesia Version – Question 7

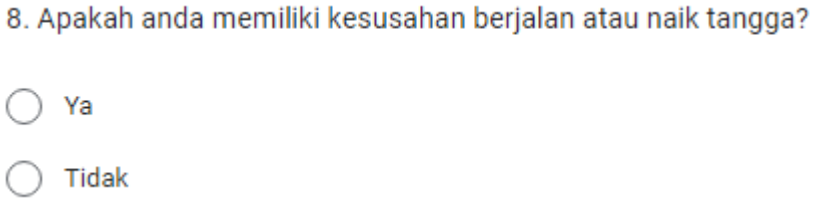


Figure 171: Questionnaire Section C – Bahasa Indonesia Version – Question 8

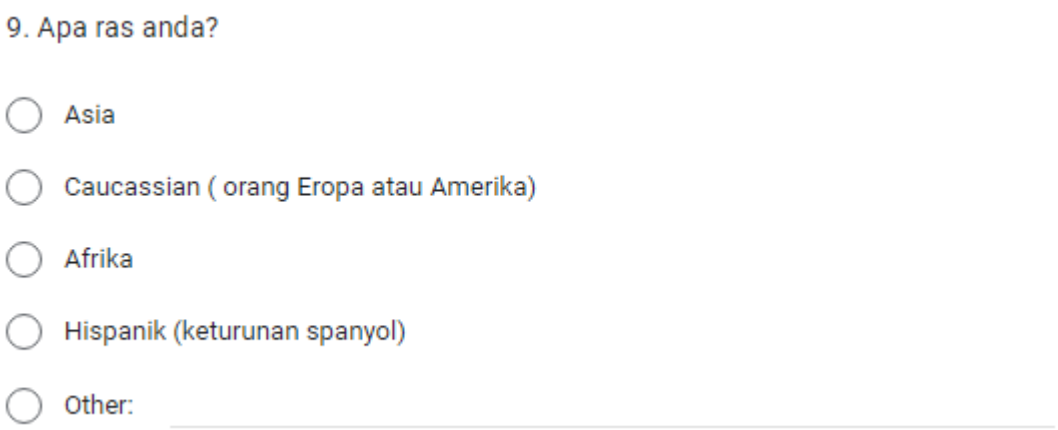


Figure 172: Questionnaire Section C – Bahasa Indonesia Version – Question 9

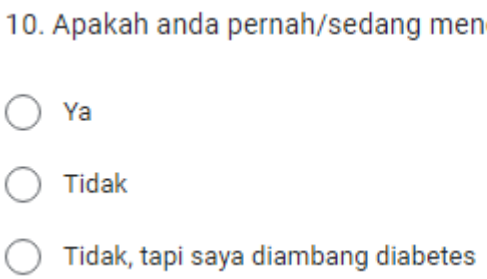


Figure 173: Questionnaire Section C – Bahasa Indonesia Version – Question 10

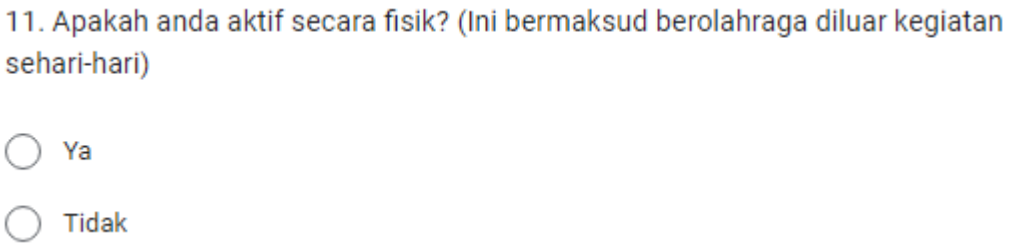


Figure 174: Questionnaire Section C – Bahasa Indonesia Version – Question 11

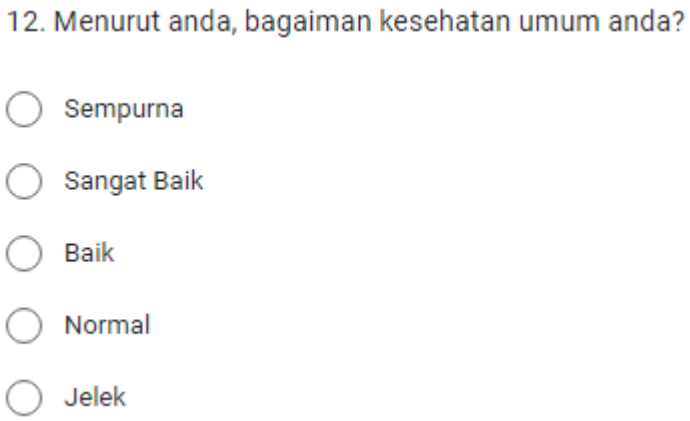


Figure 175: Questionnaire Section C – Bahasa Indonesia Version – Question 12

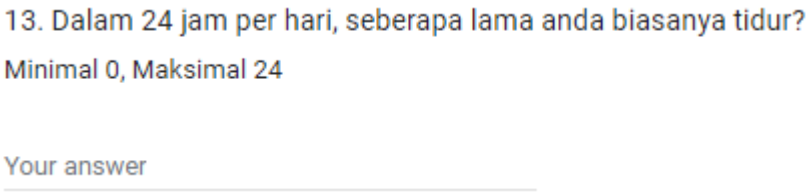


Figure 176: Questionnaire Section C – Bahasa Indonesia Version – Question 13

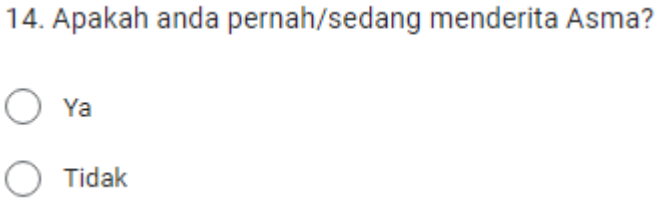


Figure 177: Questionnaire Section C – Bahasa Indonesia Version – Question 14

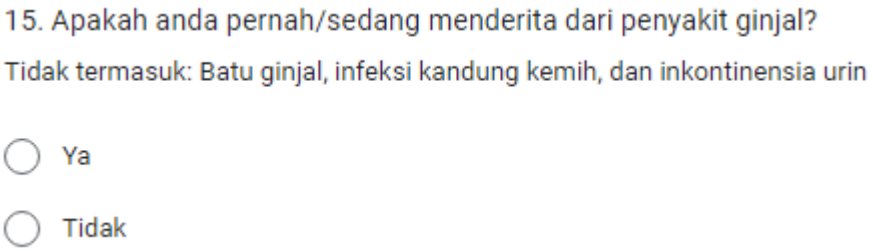


Figure 178: Questionnaire Section C – Bahasa Indonesia Version – Question 15

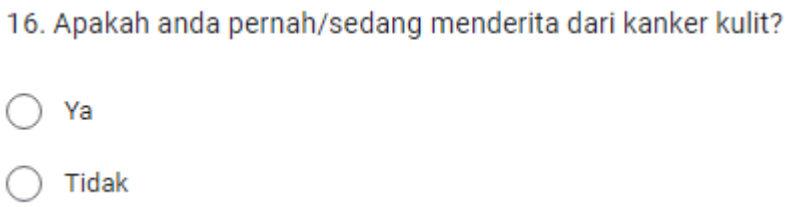


Figure 179: Questionnaire Section C – Bahasa Indonesia Version – Question 16

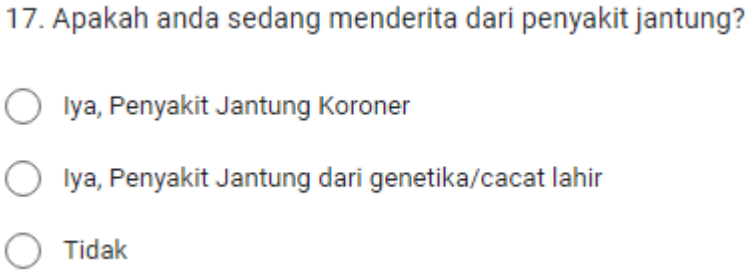


Figure 180: Questionnaire Section C – Bahasa Indonesia Version – Question 17