**CHAPTER 1**

**INTRODUCTION**

This chapter involves the introduction on the project which include the background of study, problem statement, objectives, hypothesis, research question and the significance of the study.

**1.1 BACKGROUND**

Heart disease is one of the top diseases in developing country. The term of “heart disease” also can be used as “cardiovascular disease”. Cardiovascular disease also refers to conditions that involved blocked vessel or narrowed that can lead to other symptoms like stroke, heart attack and chest pain (angina). The heart unable to push the required amount of blood to other parts of the body to fulfil the normal functionalities of the body, and due to this, ultimately the heart failure occurs. (Bui et al., 2011). These are the symptoms of heart disease including shortness of breath, weakness of physical body, swollen feet, and fatigue with related signs, for example, elevated jugular venous pressure and peripheral edema caused by functional cardiac or noncardiac abnormalities. (Durairaj & Ramasamy, 2016).

The Department of Statistic Malaysia (DOSM) have press release regarding the statistics on causes of death in Malaysia this year. DOSM states that Ischaemic heart diseases remained as the principle causes of death, 15.6 per cent followed by Pneumonia (11.8%). Cerebrovascular diseases (7.8%). Transport accidents (3.7%) and Chronic lower respiratory diseases (2.6%).

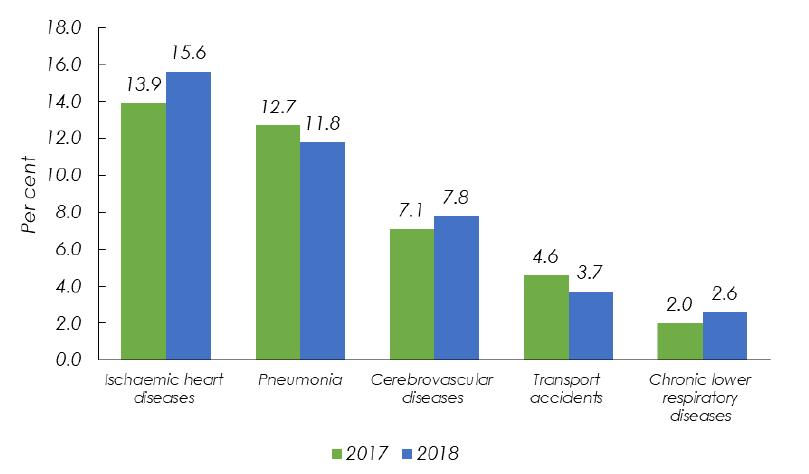


Figure 1.1: The main causes of death in 2017 and 2018.

According to Wei and Gersh in 1987, cardiovascular disease is the most frequent single cause of death in persons over 65 years of age. Only 17% of people with heart failure are less than 65 years of age (Mair et al., 1996). Cardiovascular disease also can lead to other disease that some of the patients infected by the bacteria after the operation. The operation still can be considered as too pricy for some family from bad financial background. This disease is dreadful especially to patients whom aged 40 years and above. The accurate and proper diagnosis of the heart disease risk in patients is necessary for reducing their associated risks of severe heart issues and improving security of heart. (Al-Shayea, 2011). Thus, this article wants to create awareness towards Malaysians about this disease and the factors that affecting it. In addition, this article also analysed the data that mostly been used-Cleveland Heart Disease dataset from the UCI Repository by using multiple techniques of analysis.

**1.2 PROBLEM STATEMENT**

Heart disease still Malaysian biggest killer in 2019. According to New Straits Time, the deputy health of minister Dr Lee Boon Chye said that heart disease has remained the leading cause of death among Malaysians for 13 years from 2005 and 2017. He also states that it was a non-communicable disease, such as diabetes and hypertension, as well as obesity, unhealthy diet and lifestyle, which had contributed to the increase in heart disease.

There are some problems that arise during the study. First, people are not aware about the factors that can lead to heart disease. For example, exercising, healthy diet and smoking. These factors are the factors that can be controlled by human itself. They can reduce the risk heart disease by taking care of this factors.

Next, healthcare industry such as hospitals and clinics have many data of the patients over the years that have been kept in their own databases. These data can provide the analysis of the risk factors of diseases. But they are not properly mined to extract the relationships among the diagnosis data and clinical.

Another problem regarding heart disease is the cost of health services. It will not be a problem to do a simple check-up at government hospitals or clinics, but for the operation cost, it will be a burden to them especially for the people that have heart disease or their family have it, it will be difficult for the moderate- and low-income family. This will affect themselves because they tend to not get the treatment at the hospital.

**1.3 RESEARCH OBJECTIVE**

These are several objectives for this study including:

1. To describe the factors that contributes for heart disease.
2. To identify the association between the factors and heart disease.
3. To predict the heart disease within its factors using logistic regression.

**1.4 RESEARCH HYPOTHESIS**

These are the hypothesis of the study:

1. The factors described for heart disease.
2. There exist association between the factors and heart disease.
3. The heart disease predicted within its factors using logistic regression.

**1.5 RESEARCH QUESTION**

These are the question of the study:

1. How to describe the factors that contributes for heart disease?
2. Is there any association between the factors and heart disease?
3. Does the heart disease can be predicted within its factors by using logistic regression?

**1.6 SIGNIFICANCE OF THE STUDY**

There are 13 factors that are focused in this article, which are age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels and Thallium scan.

From the findings of this study, researchers will know the factors that contributes significantly towards the risk of heart disease. Thus, this will help the researchers to conduct the research based on the significant factors and to help them from choosing the factors that not significant. It will improve the research regarding the heart disease.

Next, this article also can create awareness among the people in Malaysia or other country regarding this issue. This article can increase the information among the people of the factors that affecting heart disease. So, they will be more aware of their health because heart disease can lead to death. The more knowledge health people have about heart disease, the more they will appreciate what a health threat it is. People will have their own initiatives on preventing it and will be more motivated to reduce of becoming a heart disease morbidity or mortality statistic.

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 INTRODUCTION**

This chapter describes the review of literature for some keywords and terminologies related to the research including the discussion of similar, and past research. All data collected are collected from past research comprising published and unpublished materials such as journals, articles, reports, books, and web articles.

**2.2 DEFINITION OF TERMS**

**2.2.1 AGE**

According to Shubhankar Rawat(2019) said that Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life and coronary fatty streaks can begin to form in adolescence. American Heart Association News(2018) indicate that the people having heart attacks were increasingly young, from 27 percent at the start of the study to 32 percent at the end. Asia Pacific Cohort Studies Collaboration(2016) found that  there were 2915 deaths from cardio heart disease(CHD) and the risk of CHD increased substantially with age. Shubhankar Rawat(2019), estimated that 82 percent of people who die of coronary heart disease are 65 and older. The European Society of Cardiology (ESC) reported that 26 million adults worldwide were diagnosed with heart disease and 3.6 million were diagnosed every year(Amin Ul Haq, Muhammad Hammad Memon, Salah Ud Din, Ruinan Sun, Jianping Li, Jalaluddin Khan, Ijaz Ahad, and Zhilong Lai). Jousilhati et al reported that increases in total cholesterol, blood pressure, and diabetes were the risk factors to the heart disease were most strongly associated with the age. From the previous study, it can conclude that age is significant to heart disease.

**2.2.2 GENDER**

National Health and Nutrition Examination Surveys (NHANES) have shown that over the past two decades the prevalence of myocardial infarctions has increased in midlife (35 to 54 years) women, while declining in similarly aged men(A.H.E.M. Maas & Y.E.A Appelman) .European Heart Survey on stable angina pectoris it was found that women are less likely to be referred for functional testing for ischaemia and that a lower rate of diagnostic angiograms and interventional procedures are performed compared with men. According to Shubhankar Rawat(2019) indicates that Men are at greater risk of heart disease than pre-menopausal women but once past menopause, it has been argued that a woman’s risk is similar to a man’s. Meanwhile Harvard Health Publishing said that woman's symptoms are often different from a man's, and she's much more likely than a man to die within a year of having a heart attack. According to Shubhankar Rawat(2019) said that if a female has diabetes, she is more likely to develop heart disease than a male with diabetes.

**2.2.3 Angina (Chest Pain)**

**Angina (Chest Pain) also known as** Angina Pectoris, Acute Coronary Syndrome, Microvascular Angina, Prinzmetal’s Angina, stable angina, Unstable Angina, Variant Angina, Vasospastic Angina, Cardiac Syndrome X stated by National Heart, Lung and Blood Institute. According to Shubhankar Rawat(2019) define that angina is chest pain or discomfort caused when your heart muscle doesn’t get enough oxygen-rich blood and it may feel like pressure or squeezing in your chest. Harvard Health Publishing also defined that Angina is pain in the chest that comes on with exercise, stress, or other things that make the heart work harder and it is an extremely common symptom of coronary artery disease, which is caused by cholesterol-clogged coronary arteries and also known as the network of arteries that nourish the heart muscle. WebMD also defined that [Angina](https://www.webmd.com/heart-disease/heart-disease-angina) is chest pain that happens because there isn't enough [blood](https://www.webmd.com/heart/anatomy-picture-of-blood) going to part of your [heart](https://www.webmd.com/heart/picture-of-the-heart) and it can feel like a [heart attack](https://www.webmd.com/heart-disease/guide/heart-disease-heart-attacks), with pressure or squeezing in your chest. From the previous study, individuals who have chest pain have to change their lifestyles and the ways they take medicine.

**2.2.4 Resting Blood Pressure**

Blood pressure (BP) is the [pressure](https://en.wikipedia.org/wiki/Pressure) of circulating [blood](https://en.wikipedia.org/wiki/Blood) on the walls of [blood vessels](https://en.wikipedia.org/wiki/Blood_vessel) and most of this pressure is due to work done by the heart by pumping blood through the [circulatory system](https://en.wikipedia.org/wiki/Circulatory_system) , this definition stated by Wikipedia. Blood pressure also known as the pressure in the large arteries and systematic circulation and it is measured in mercury(mmHg), under atmospheric pressure. ([George A. Kelley](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kelley%20GA%5BAuthor%5D&cauthor=true&cauthor_uid=11828203), [Kristi A. Kelley](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kelley%20KA%5BAuthor%5D&cauthor=true&cauthor_uid=11828203) & [Zung Vu Tran](https://www.ncbi.nlm.nih.gov/pubmed/?term=Tran%20ZV%5BAuthor%5D&cauthor=true&cauthor_uid=11828203)) founded that  aerobic exercise reduces resting systolic and diastolic blood pressure in adults. Shubhankar Rawat(2019) also indicate that Over time, high blood pressure can damage arteries that feed your heart and Shubhankar Rawat(10 August) also founded that, high blood pressure that occurs with other conditions such as obesity, high cholesterol or diabetes can increases your risk even more. From the previous study, ([George A. Kelley](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kelley%20GA%5BAuthor%5D&cauthor=true&cauthor_uid=11828203), [Kristi A. Kelley](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kelley%20KA%5BAuthor%5D&cauthor=true&cauthor_uid=11828203) & [Zung Vu Tran](https://www.ncbi.nlm.nih.gov/pubmed/?term=Tran%20ZV%5BAuthor%5D&cauthor=true&cauthor_uid=11828203)) founded that  aerobic exercise reduces resting systolic and diastolic blood pressure in adults.

**2.2.5 Serum Cholestrol**

**According to healthline it is founded that cholesterol is significant with heart disease it is** because low-density lipoproteins (LDL) can build up in your arteries and restrict or block blood flow but your body still needs a little cholesterol for healthy digestion and to make vitamin D and certain hormones. According to Shubhankar Rawat(10 August), A high level of low-density lipoprotein (LDL) cholesterol (the “bad” cholesterol) is most likely to narrow arteries and high level of triglycerides, a type of blood fat related to your diet, also ups your risk of heart attack, however a high level of high-density lipoprotein (HDL) cholesterol (the “good” cholesterol) lowers your risk of heart attack. The older you are, the more frequently you have to check for serum cholesterol for example you have to checked your serum cholesterol every year and for adult is is suggested to check four to six years.

**2.2.6 Fasting Blood Sugar**

 A blood sample will be taken after an overnight fast, fasting blood sugar level less than 100 mg/dL (5.6 mmol/L) is normal, fasting blood sugar level from 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes and if it's 126 mg/dL (7 mmol/L) or higher on two separate tests, you have diabetes. Research from healthessential showed that fasting can help lower blood pressure, reduce cholesterol, control diabetes and reduce weight. Another research founded that the reducing of sugar can decrease the risk of cancer, diabetes and obesity but it is suggested to refer from dietician if we want to fasting blood sugar because fasting blood sugar also can give a negative impact for example if individual not fasting with correct way, it will cause to gastric .

**2.2.7 Resting ECG**

Based on physimed, the resting electrocardiogram is a test that measures the electrical activity of the heart, the resting ECG can detect certain heart conditions such as hypertrophy of heart, ischemia, myocardial infarction, sequelae of myocardial infarction, cardiac arrhythmias, etc and test takes about 5 minutes and no preparation is necessary. Shubhankar Rawat(2019),  people at low risk of cardiovascular disease, the US Preventive Services Task Force  (USPSTF) concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits and for people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.

**2.2.8 Maximum heart rate achieved**

MedicalNewsToday defined that heart rate is one of the 'vital signs,' or the important indicators of health in the human body and it is measuring the number of times per minute that the heart contracts or beats. According to Mayoclinic, normal resting heart rate for adult’s ranges from 60 to 100 beats per minute, however a lower heart rate at rest implies more efficient heart function and better cardiovascular fitness and for example, a well-trained athlete might have a normal resting heart rate closer to 40 beats per minute. From the research, it is founded that if heart rate is too fast and exceed 100 beats per minute it is dangerous and called as tachycardia but it is also dangerous when heart rate less than 60 beats per minute excluding the athlete who always exercise and when heartbeat less than 60 to them, it is consider good. When you have symptom of max heart rate such as short breath, fainting and difficult to breath, you should see the doctor.

**2.2.9 Exercise induced angina**

According to mayoclinic, Angina which may also be called angina pectoris and is often described as squeezing, pressure, heaviness, tightness or pain in your chest. Mayoclinic also stated that some people who have angina symptoms describe angina as feeling like a vise is squeezing their chest or feeling like a heavy weight has been placed on their chest and angina may be a new pain that needs evaluation by a doctor, or recurring pain that goes away with treatment. Shubhankar Rawat (2019) describe that Angina is usually felt in the centre of your chest, but may spread to either or both of your shoulders, or your back, neck, jaw or arm and it can even be felt in your hands. It is founded that tracking stable angina can reduced angina symptom.

**2.2.10 Peak exercise ST segment**

Patients with have coronary artery disease always failed to achieved workload or called as heart beat and the ST depression will not reach 1 mm. individuals with healthy condition with normal heart rate will induce to normal ST depression which not caused by ischemia and 20% of healthy subjects will appear during exercise test. Shubhankar Rawat (2019), treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Shubhankar Rawat(2019) also indicate that exercise ECGs with up-sloping ST-segment depressions are typically reported as an ‘equivocal’ test and. Shubhankar Rawat(2019) also said in general the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease, it is stated that duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test.

**2.2.11 ST depression induced by exercise relative to rest**

Stress test was measured by electrocardiogram. Electrocardiogram can determine if the individual have high potential in coronary artery disease. The presence of STsegment changes, either depression or elevation, on the ECG during the treadmill test often suggests presence of CAD and warrants further management([Yoke Ching Lim](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lim%20YC%5BAuthor%5D&cauthor=true&cauthor_uid=27440279), [Swee-Guan Teo](https://www.ncbi.nlm.nih.gov/pubmed/?term=Teo%20SG%5BAuthor%5D&cauthor=true&cauthor_uid=27440279)& [Kian-Keong Poh](https://www.ncbi.nlm.nih.gov/pubmed/?term=Poh%20KK%5BAuthor%5D&cauthor=true&cauthor_uid=27440279)). Electrocardiogram will show highest heart rate when the individuals in a stress or depression but electrocardiogram will shows a normal rhythm when the individuals in rest condition.  The treadmill ECG stress test is also often used to evaluate for chronotropic incompetency, in which the patient is unable to mount an adequate heart rate response to exercise, leading to symptoms such as effort-related dyspnoea and lethargy([Yoke Ching Lim](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lim%20YC%5BAuthor%5D&cauthor=true&cauthor_uid=27440279), [Swee-Guan Teo](https://www.ncbi.nlm.nih.gov/pubmed/?term=Teo%20SG%5BAuthor%5D&cauthor=true&cauthor_uid=27440279)& [Kian-Keong Poh](https://www.ncbi.nlm.nih.gov/pubmed/?term=Poh%20KK%5BAuthor%5D&cauthor=true&cauthor_uid=27440279)).

**2.2.12 Number of major vessels (0-3) colored by fluoroscopy (vca)**

Vessels are functioned to bring blood to and from heart. Vessels are including superior vena cava, inferior vena cava, pulmonary arteries, pulmonary veins and aorta. Coronary heart disease also known as coronary artery disease (CAD) is a chronic disease in which the coronary arteries, responsible for transporting oxygenated blood to heart muscles, get narrowed and are not able to convey enough fresh blood to this blood-pumping organ([Hamid Reza Marateb](https://www.ncbi.nlm.nih.gov/pubmed/?term=Marateb%20HR%5BAuthor%5D&cauthor=true&cauthor_uid=26109965) & [Sobhan Goudarzi](https://www.ncbi.nlm.nih.gov/pubmed/?term=Goudarzi%20S%5BAuthor%5D&cauthor=true&cauthor_uid=26109965)). According to MayoClinic defines that coronary artery disease develops when the major blood vessels that supply your heart with blood, oxygen and nutrients (coronary arteries) become damaged or diseased where Cholesterol-containing deposits (plaque) in your arteries and inflammation are usually to blame for coronary artery disease

**2.2.13 Thallium scan (Thal)**

β-Thalassemia is an inherited haemoglobin disorder resulting in chronic hemolytic anemia that typically requires life-long transfusion therapy([Dimitrios T. Kremastinos](https://www.ahajournals.org/doi/full/10.1161/circheartfailure.109.913863) et al..). Thalassemia is inherited blood disorder and will cause the haemoglobin will less in your body which haemoglobin is function as to carry oxygen and this will cause a critical disease which is called as anaemia. According to mayoclinic indicates that there are a few symptoms in thalassemia such as fatigue, weakness, pale, facial bone deformities, slow growth, abdominal swelling and dark urine. If you have these symptoms it is important to get a check-up from the doctor.

**CHAPTER 3**

**METHODOLOGY**

**3.1 INTRODUCTION**

This chapter will discuss about methodology including source of dataset, description of dataset and method that have been used to achieve our objective.

**3.2 SOURCE OF DATASET**

The dataset that have been used in this article was named as Heart disease dataset whereby the dataset were taken from the UCI machine learning respiratory. This dataset contains 303 observations and 14 columns of dataset. This dataset dates were taken from 1998 and consist database from Cleveland (303 observations).

**3.3 DESCRIPTION OF DATASET**

The dataset contains 14 variables (Age gender CPT RBP SCH FBS RES MHR EIA OPK PES VCA THA STATUS). We wish to identify the significance between dependent variable (status=predicted attribute) and all independent variables. This dataset was focusing on individuals where their age is between 29-77 years old which high potential to get heart disease according to their age.

Table 3.1: The Data Description

|  |  |  |
| --- | --- | --- |
| Variables | Data type | Description |
| Age | Numeric | Age of patient |
| Gender | Character | 1=male, 0=female |
| CPT | Character | Chest pain type: 1=typical angina, 2=atypical angine, 3=non-anginal pain, 4=asymptomatic |
| RBP | Numeric | Resting blood pressure (mm Hg) |
| SCH | Numeric | Serum cholesterol (mg/dl) |
| FBS | Character | Fasting blood sugar: 1 if >120 mg/dl; 0 otherwise |
| RES | Character | Resting electrocardiographic results: 0=normal; 1=having ST-T Wave abnormality, 2=showing probable or definite left ventricular hypertrophy |
| MHR | Numeric | Maximum heart rate achieved |
| EIA | Character | Exercise induced angina; 1=yes,0=no |
| OPK | Numeric | ST depression induced by exercise relative to rest |
| PES | Character | The slope of the peak exercise ST segment; 1=up sloping, 2=flat,3=down sloping. |
| VCA | Character | Number of major vessels (0-3) coloured by fluoroscopy |
| THA | Character | Thallium scan; 3=normal,6=fixed defect, 7=reversable defect |
| STATUS | Character | Predicted attribute, 1=heart disease, 0=no heart disease. |

**3.4 DESCRIPTIVE STATISTICS**

Descriptive statistic is descriptive coefficients that summarize the analysis of a given dataset which can be either sample of the population or can be presentation to interpret for the whole entire population.

**SYNTAX ERROR FOR PROC LOGISTIC**

|  |
| --- |
| The syntax for Proc logistic,  PROC LOGISTIC <statement>;  OUTEST=<statement> COVOUT;  MODEL <statement>  /SELECTION=stepwise  SLENTRY=**0.3**  SLSTAY=**0.35**  LACKFIT  EXPB;  OUTPUT OUT=<statement> |

**3.5 INFERENTIAL STATISTIC**

Inferential statistic will make a prediction by using data that were given in descriptive statistic. Inferential statistics will be estimating parameters and mean are taken from the sample data and inferential statistic will estimate about population parameter. Then inferential statistic also useful for hypothesis testing.

**3.5 LOGISTIC REGRESSION**

[PROC LOGISTIC](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax01.htm) <options>;

[BY](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax04.htm) variables;

[CLASS](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax05.htm) variable <(options)> <variable <(options)> …> </ options>;

[CODE](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax08.htm) <options>;

[CONTRAST](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax09.htm) 'label' effect values<, effect values, …> </ options>;

[EFFECT](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax12.htm) name=effect-type(variables </ options>);

[EFFECTPLOT](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax13.htm) <plot-type <(plot-definition-options)>> </ options>;

[ESTIMATE](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax14.htm) <'label'> estimate-specification </ options>;

[EXACT](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax15.htm) <'label'> <INTERCEPT> <effects> </ options>;

[EXACTOPTIONS](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax17.htm) options;

[FREQ](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax18.htm) variable;

[ID](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax19.htm) variables;

[LSMEANS](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax20.htm) <model-effects> </ options>;

[LSMESTIMATE](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax21.htm) model-effect lsmestimate-specification </ options>;

[<label:> MODEL](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax22.htm) variable <(variable\_options)> = <effects> </ options>;

[<label:> MODEL](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax22.htm) events/trials = <effects> </ options>;

[NLOPTIONS](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax25.htm) options;

[ODDSRATIO](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax26.htm) <'label'> variable </ options>;

[OUTPUT](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax27.htm) <OUT=SAS-data-set> <keyword=name <keyword=name …>> </ option>;

[ROC](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax29.htm) <'label'> <specification> </ options>;

[ROCCONTRAST](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax30.htm) <'label'> <contrast> </ options>;

[SCORE](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax31.htm) <options>;

[SLICE](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax32.htm) model-effect </ options>;

[STORE](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax33.htm) <OUT=>item-store-name </ LABEL='label'>;

[STRATA](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax34.htm) effects </ options>;

[<label:> TEST](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax35.htm) equation1 <,equation2, …> </ option>;

[UNITS](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax36.htm) <independent1=list1 <independent2=list2 …>> </ option>;

[WEIGHT](http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/statug_logistic_syntax37.htm) variable </ option>;

**3.5.1 Assumptions of Logistic Regression**

1. Observations to be independent of each other
2. No multicollinearity among the independent variables.
3. Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
4. Model should be fitted correctly
5. Error term must be independent.

**3.5.2 Data checking and data cleaning**

The data checking is conducted to identify the missing value and invalid values in the data sets. Since there are invalid data found in the data sets, data cleaning was conducted. The data with invalid values will be omitted. For example, there are -100000 in Thallium scan (THA) which is not logical as THA is a categorical variable with only three choice of answers (3=normal,6=fixed defect, 7=reversable defect).

**3.5.3 Model fit statistic**

**Akaike Information Criterion (AIC)**

Akaike Information Criterion (AIC) used to check whether the model is over fit or not. AIC work to balance the trade-off between complexity of model and goodness of fit which AIC will describe the fit of model and penalize them when there are over fitting in the model or when model is overly complex. The basic formula of AIC is **AIC = -2(log-likelihood) + 2K** where k is number of parameters, n is number of sample and log is likelihood measure of fit. The lowest value of AIC is considered good as it is indicating that there is superior balance trade-off between complexity of model and goodness of fit.

**3.5.4 Logistic Regression analysis**

**R-squared and Adjusted R-square**

R-squared and adjusted r-square are used to measure proportion of dependent variables that is explaining independent variable but adjusted R-square measure proportion of dependent variable that really help to explain independent variable. The higher the r-square the model is considered to be good and adjusted r-square is more accurate rather than r-square because r-square doesn’t provide any ways to stop adding more independents variable.

**Multicollinearity**

Before start the research, it is assuming that the variance is constant mean there is no multicollinearity in the model. Variance of inflation (VIF) and tolerance have been used to test whether there is multicollinearity or not. There is multicollinearity if VIF more than 10 and tolerance equal than 0.1. data transform is one way to solve the multicollinearity issues.

**Analysis of Variance (ANOVA)**

ANOVA or Analysis of Variance used to test the difference between two or more means, it is mostly known as Analysis of Variance rather than Analysis of Means. ANOVA table also is a collection of statistical method and their association estimation of group means. Analysis of variance (ANOVA) is a statistical test for detecting differences in group means when there is one parametric dependent variable and one or more independent variables ([Steven Sawyer](https://www.researchgate.net/profile/Steven_Sawyer)).  T-test and F-test in ANOVA table were used to check whether the model is significant or not, and the model can be removed if it is not significant because it not contributes to model.

**Wald chi-square**

Wald chi-square also called as Wald chi-square test is a test that can be used to check explanatory variables that are significant in the model. Significant means that there is association between the model and the variables that is significant will affecting the model. The formula of Wald chi-square is (Beta/standard error) ^2 where beta is the coefficient which we test against on the null hypothesis that is 0. If the Wald chi-square showed the value of variables is zero, it is indicated that Hnull are failed to been rejected and the variable can be removed from the model as it is not affecting the model or not contribute to the model. Meanwhile, when the value of variables not equal to zero, Hnull will be rejected and the variable cannot be removed from the model as it will affect the model.

**Maximum likelihood estimation**

Maximum likelihood estimation (MLE) is a method for estimates the parameter of probability distribution. This technique will maximise the likelihood function, and the data that were produced from the process actually were being observed. The goal of maximum likelihood estimation is the observed data produced most probable through process in estimate the parameter by maximizing likelihood function. If the likelihood function is differentiable the derivative test can be used for determining maximum.

**Odds ratio estimation**

Odd ratio estimation is to measure the strength of associations between two variables. Odd ratio is define as the presence or absence of odd B in A or vice versa. There is positive association when odd ratio is more than 1 while if the odd ratio is less than 1, it is showed that there is negative association which means one variable will reduces another variable for example A and B is negative correlation so when A presence it will reduce B. When odd ratio is equal with 1 so it showed that the variables is independent that means the variables are not giving effect to each other’s.

**GOODNESS OF FIT**

Goodness of fit used to measure how well observation to model. Goodness of fit can measure statistical hypothesis testing to test the normality of residual, to test whether to sample are drawn from distribution or to check whether the output follow the distributions. The formula for goodness of fit used Pearson chi-square test.

**Contrast estimation**

Contrast is a linear combination of variables such as all the coefficients are added up to zero. Contrast will allow comparisons between different treatments. There are 3 type of contrast set for example orthogonal contrast, polynomial contrast and orthonormal contrast. Orthogonal contrast is a set of contrast which is the distict number and the coefficient of cross product are zero and the sample size assume to be equal. Polynomial contrast are used to test polynomial pattern with more mean for example linear, quadratic and etc while orthonormal contrast or known as orthogonal contrasts that satisfy additional condition for example to each contrast and sum square coefficient are add up to one.

**3.6 Testing Accuracy**

**Kolmogorov Smirnov**

The Kolmogorov-Smirnov statistic measures a distance between empirical function of two samples or it is also can measure the distance between empirical distribution function of the sample and cumulative distribution function of reference distributions. Kolmogorov-Smirnov also can be used to test whether the model is significant or not. According to Wikipedia,  the Kolmogorov–Smirnov test (K–S test or KS test) is a [nonparametric test](https://en.wikipedia.org/wiki/Nonparametric_statistics) of the equality of continuous (or discontinuous) one-dimensional [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) that can be used to compare a [sample](https://en.wikipedia.org/wiki/Random_sample) with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). The K-S test is very useful for comparing two sample because it is sensitive to location and shape of the empirical distribution’s functions of two samples.

**Classification Table (CONFUSION MATRIX)**

Confusion matrix is a table that is used to describe the performance of dataset model of classification which true values is known. Confusion matrix will allow the visualization performance of algorithm and make it allow to identify the confusion between class. Confusion matrix is a summarized of prediction result as the number of correct and incorrect values have been count and broken by each class. Confusion matrix will give us insight what types of error that always being made and can identify which error that contain incorrect values.

**3.7 Model selection**

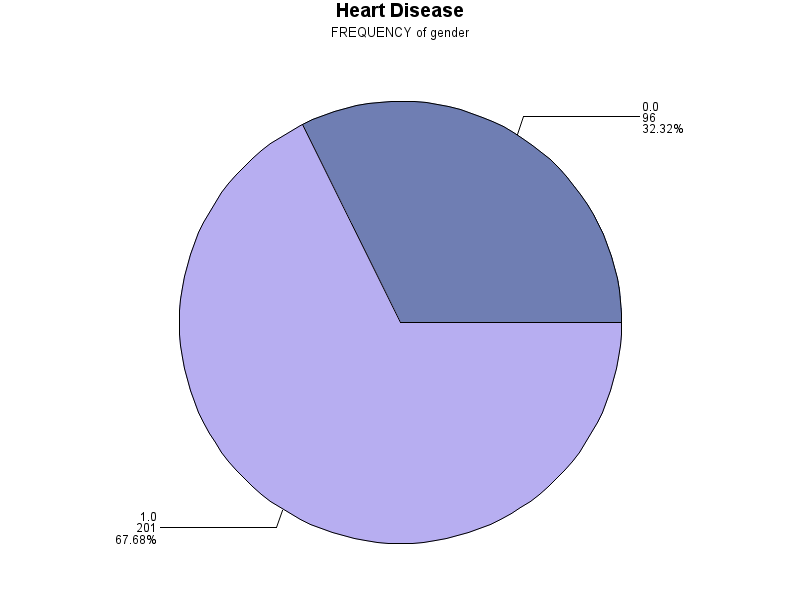
Model selection is process of combining data and prior information to select a statistical model. Model selection is the task of selecting a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) from a set of candidate models, given data state by Wikipedia. Model selection also involved design of experiment which can be suite with the problem and reduce the problem that can interrupt data. Some algorithms will select the variable that they want to be included in the model and return the best or variable that is significant to be included in the model for example stepwise regression, backward. There are main approaches for model selection including optimize some selection criterion, test of hypotheses and hoc methods.

**CHAPTER 4**

**RESULTS AND ANALYSIS**

This chapter consist of the analysis of the study by using the appropriate method and multiple technique of analysis.

**4.1 DESCRIPTIVE STATISTICS**



1=Male

0=Female

Figure 4.1: The distribution of gender

Figure 4.1 show that 67.68% of respondents are males while 32.32% are females.

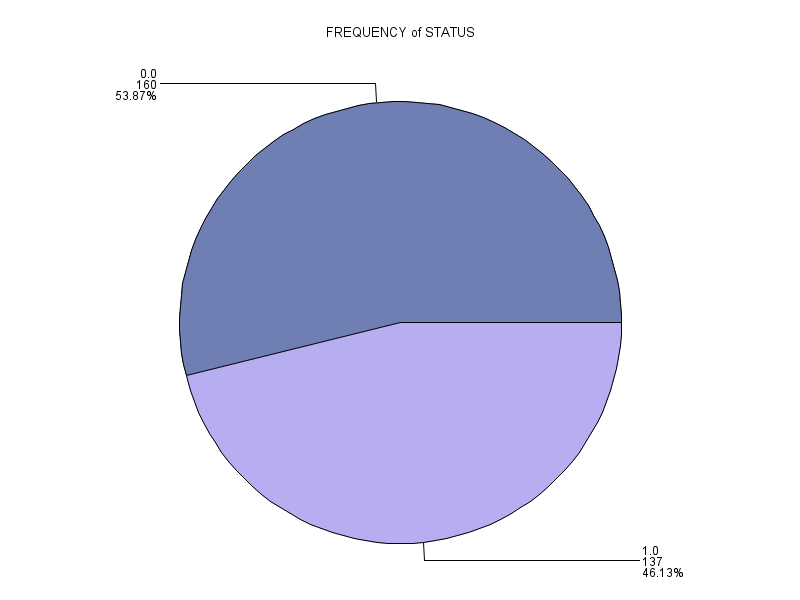
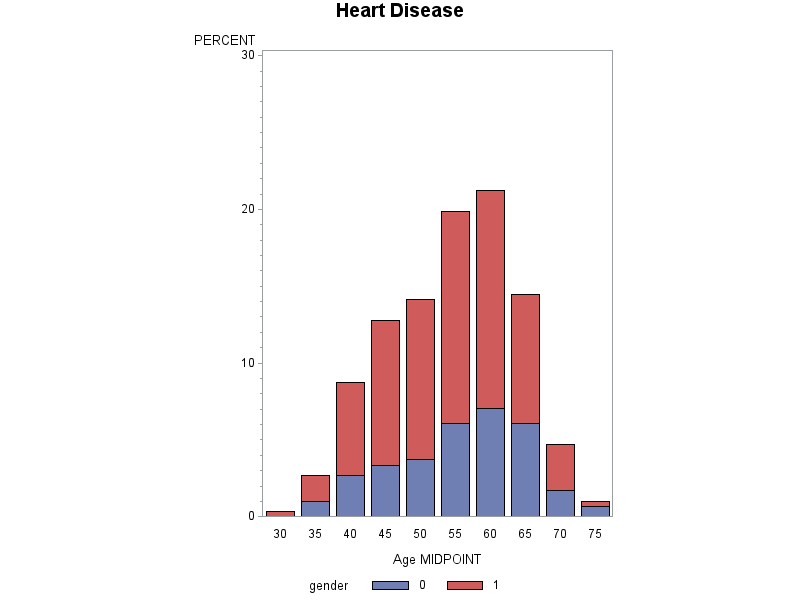


Figure 4.2: The distribution of heart disease

46.13% of respondents have heart disease while 53.87% do not have heart disease.

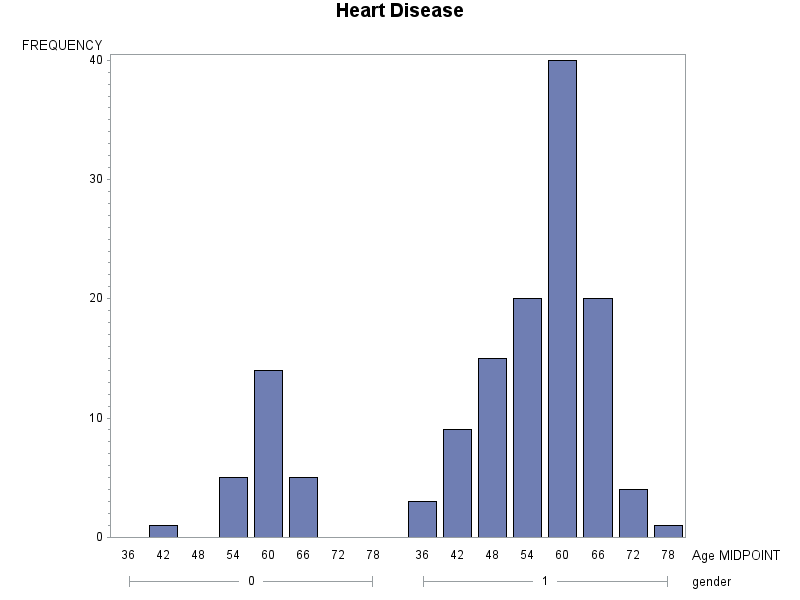


1=Male

0=Female

Figure 4.3: The percentages of age based on gender

The midpoint of age at 60 years old give the highest contribution with 7% are females and 14% are males. Meanwhile midpoint of age at of 30 only has male patients with 0.5% of overall percentages.

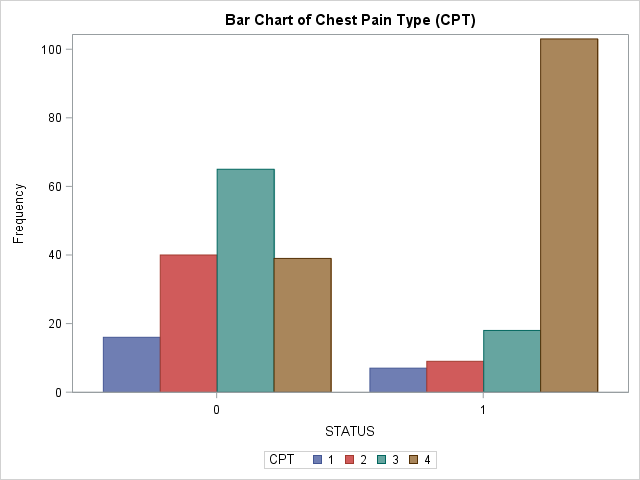


1=Male

0=Female

Figure 4.4: The frequency of patients with heart disease based on age and classified by gender.

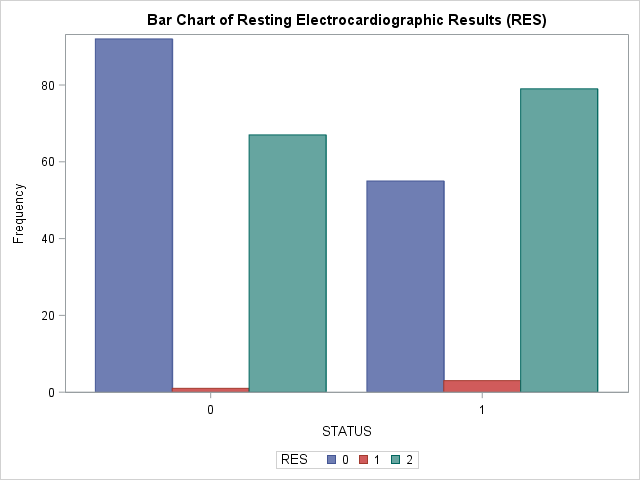
Figure above shows that for both male and female, the numbers of patients are highest at age midpoint of 60 years old with frequency of 14 and 39 respectively. The lowest numbers of respondents with heart disease for males are at the age midpoint of 78 years old. For females, there are none of respondents are from age midpoint of 36, 48, 72.



1=Typical Angina 2=Atypical Angina 3=Non-Anginal Pain 4= Asymptomatic

Figure 4.5: Bar chart of Chest pain type (CPT)

Figure 4.5 shows the chest pain type for all patients without heart disease compared to patients with heart disease. The highest number of patients with heart disease is 103 patients come from patients having asymptotic chest pain while for patients without heart disease, 65 of them from group of non-anginal pain. Typical number has the lowest frequency of patients for both with and without heart diseases which indicate that typical angina type is not major causes of the heart disease.



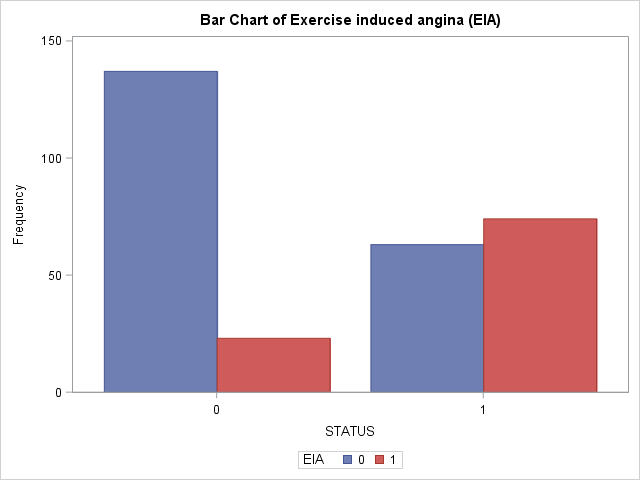
0=normal

1=having ST-T Wave abnormality

2=showing probable or definite left ventricular hypertrophy

Figure 4.6: Bar chart of Resting Electrocardiographic Result (RES)

The bar chart above, it can be observed that 79 patients with heart disease problem show that their Resting Electrocardiographic Result (RES) at probable or definite left ventricular hypertrophy which indicate the higher probability of getting cardiac diseases. 92 patients without heart disease have normal RES while only 4 patients out of 297 patients having ST-T Wave abnormality.

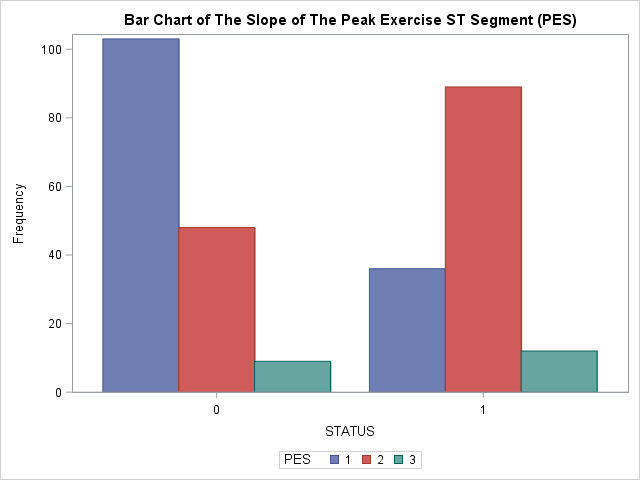


0=No

1=Yes

Figure 4.7: Bar chart of Exercise induced angina (EIA)

Figure 4.7 shows does exercise induced angina pain release the pain or not. For patients with heart disease, 74 patients agree that EIA can reduce the pain while 63 patients are disagreed. For patients without heart disease, 137 disagree and 23 agree that EIA can reduce the pain. In total, 97 patients believe EIA useful while 200 patients don’t.



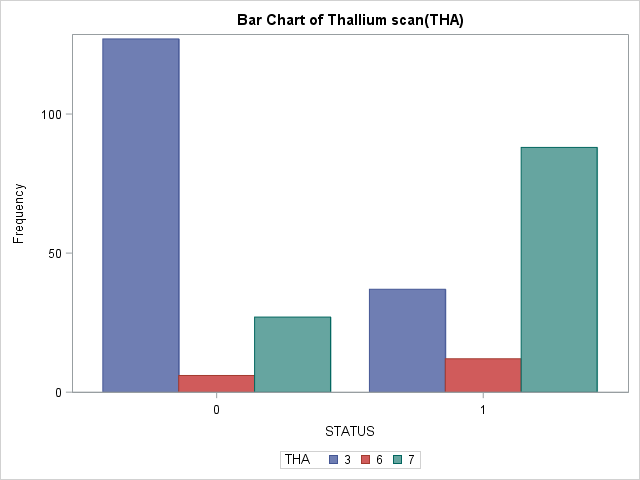
1=Upsloping

2=Flat

3=Down sloping

Figure 4.8: Bar Chart of the Slope of the Peak Exercise ST Segment (PES)

Based on figure above, the upsloping has the highest frequency of patients with 103 patients out of 160 patients without heart disease which indicates is considered normal. For patients have heart disease, 89 patients out of 137 patients are suffered from flat PES and 12 patients have down sloping PES. Having flat and down sloping PES are considered abnormal.

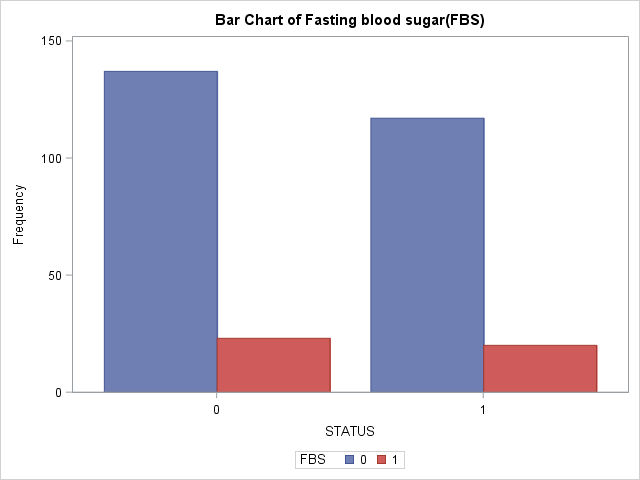


3=Normal

6=Fixed Defect 7=Reversable Defect

Figure 4.9: Bar Chart of the Thallium scan (THA)

Bar chart above shows that 127 patients out of 160 patients that have no heart disease have normal THA while 100 patients out of 137 patients that have heart disease have abnormal THA which either fixed or reversable defect.



1= if >120 mg/dl

0 =otherwise

Figure 4.10: Bar Chart of Fasting blood sugar (FBS)

Figure above shows the fasting blood pressure of patients, either less than or greater than 120mg/dl. It is observed that the majority for both with and without heart disease have fasting blood pressure less than or equal to 120mg/dl which is 117 and 137 respectively. Thus, FBS do not have major influence in predicting heart disease. Besides, only 20 patients that have FBS greater than120mg/dl also have heart disease.

Table 4.1: The frequency statistic of other categorical variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Category | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| Gender | 0  1 | 96  201 | 32.32  67.68 | 96  297 | 32.32  100.00 |
| CPT | 1  2  3  4 | 23  49  83  142 | 7.74  16.50  27.95  47.81 | 23  72  155  297 | 7.74  24.24  52.19  100.00 |
| RES | 0  1  2 | 147  4  146 | 49.49  1.35  49.16 | 147  151  297 | 49.49  50.84  100.00 |
| FBS | 0  1 | 254  43 | 85.52  14.48 | 254  297 | 85.52  100.00 |
| EIA | 0  1 | 200  97 | 67.34  32.66 | 200  297 | 67.34  100.00 |
| PES | 1  2  3 | 139  137  21 | 46.80  46.13  7.07 | 139  276  297 | 46.80  92.93  100.00 |
| VCA | 0  1  2  3 | 174  65  38  20 | 58.59  21.89  12.79  6.73 | 174  239  277  297 | 58.59  80.47  93.27  100.00 |
| THA | 3  6  7 | 164  18  115 | 55.22  6.06  38.72 | 164  182  297 | 55.22  61.28  100.00 |

Table 4.1 shows the frequency of each categorical variable. Each of variables must have total frequency 297 out of 303 observation as some of observation that have invalid values are dropped from the data. Variables FBS and EIA only have two rank of responses while others have more than two rank of responses.

Table 4.2: The summary statistics of all variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | N | N miss | Mode | Mean | Standard deviation | Minimum | Maximum | Sum |
| Age | 297 | 0 | 58.00 | 54.54 | 9.05 | 29.00 | 77.00 | 16199.00 |
| RBP | 297 | 0 | 120.00 | 131.69 | 17.76 | 94.00 | 200.00 | |  | | --- | | 39113.00 | |
| SCH | 297 | 0 | 197.00 | 247.35 | 52.00 | 126.00 | |  | | --- | | 564.00 | | 73463.00 |
| MHR | 297 | 0 | 162.00 | 149.60 | 22.94 | 71.00 | 202.00 | 44431.00 |
| OPK | 297 | 0 | 0.00 | 1.06 | 1.17 | 0.00 | 6.20 | 313.50 |

Table 4.2 shows the summary for n observation, n missing observation, mode, mean, standard deviation, minimum, maximum and sum total for each of the variables. The most frequent age of respondent is 58 years old where it has mode=58.00 with mean age of 54.54 years old. The SCH has the highest mean with value of 247.35 compared to others numeric variables.

**4.2 INFERENTIAL STATISTIC**

The data inserted by using SAS command. There are six null observations that already been discarded. The remaining of the observations is 297 observations.

**4.2.1 Chi- square test of independent**

Chi-square test of independent are used to identify the association between categorical variables.

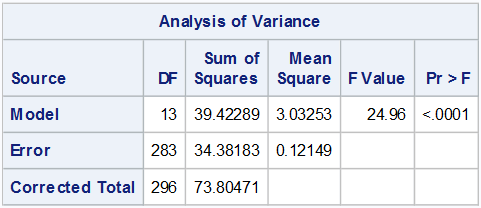
Table 4.3: The chi-square test of categorical variables

|  |  |  |
| --- | --- | --- |
| **Variables** | **Chi-square** | **Prob** |
| Gender | 23.0305 | <.0001 |
| CPT | 77.2758 | <.0001 |
| FBS | 0.0030 | 0.9565 |
| RES | 9.5755 | 0.0083 |
| EIA | 52.7295 | <.0001 |
| PES | 43.4732 | <.0001 |
| VCA | 72.3005 | <.0001 |
| THA | 82.4601 | <.0001 |

Table above shows the test for association on independents and dependent categorical variables. Since FBS has p-value=0.9565 which is greater than 0.05, thus indicate that there is association between FBS and status of heart disease. Besides, the others variables have no association with status of heart disease since the p-value are less than 0.05.

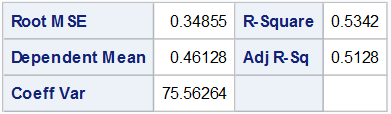
**4.2.2 Logistic Regression**

Table 4.4: The Analysis of Variance (ANOVA) table of the full model



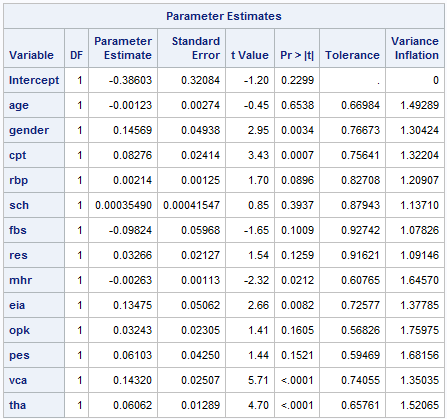
Based on the table above, the p-value is less than 0.05, the default value of alpha. Thus, it can be concluded that the model is significant. The value of mean square error is 0.12149 which is low. If the value of F compared with the statistical table, it also can be concluded that the model is significant.

Table 4.5: R-square and Adjusted R-Square.



Based on the table above, the value for R-squared is 0.5342. 53.42 percent of the total variation of heart disease explained by age, gender, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels and Thallium scan. The rest explained by other factors.

Table 4.6: Parameter Estimates for Tolerance and Variance Inflation.



The table above shows that all the value for tolerance greater than 0.1 and all the value for Variance Inflation (VIF) for all variables below 10. This can be concluded that the multicollinearity does not exist. The last analysis used to obtain the best model is by conducting stepwise selection method.

Table 4.7: Analysis of Maximum Likelihood Estimates



The estimated model for the full model is =-2.5836 + 0.0235 AGE + 1.6700 (GENDER=0) + 2.3730 (CPT=1) + 0.9249 (CPT=2) + 1.9797 (CPT=3) - 0.0277 RBP - 0.00444 SCH - 0.5741 FBS +0.4864 (RES=0) - 0.5140 (RES=1) + 0.0197MHR + 0.6533 (EIA=0) - 0.3906 OPK + 0.6067 (PES=1) - 0.6954 (PES=2) + 2.1886 (VCA=0) -0.0487 (VCA=1) - 1.0829 (VCA=2) + 1.4332 (THA=3) + 1.6016 (THA=6). The p-value for variable gender, CPT=1, CPT=3, RBP, VCA=0, THA=3 AND THA=6 are less than 0.05. Hence, it can be concluded that the variables are significant. But, for the other variables that are not mentioned, not significant because the p-value is greater than 0.05.

Table 4.8: Type 3 analysis of effects of full model

|  |  |  |  |
| --- | --- | --- | --- |
| **Type 3 Analysis of Effects** | | | |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Age** | 1 | 0.8755 | 0.3494 |
| **Gender** | 1 | 9.1375 | 0.0025 |
| **CPT** | 3 | 18.5592 | 0.0003 |
| **RBP** | 1 | 5.5666 | 0.0183 |
| **SCH** | 1 | 1.1804 | 0.2773 |
| **FBS** | 1 | 0.9387 | 0.3326 |
| **RES** | 2 | 1.5875 | 0.4521 |
| **MHR** | 1 | 2.8251 | 0.0928 |
| **EIA** | 1 | 2.1316 | 0.1443 |
| **OPK** | 1 | 2.6678 | 0.1024 |
| **PES** | 2 | 7.3642 | 0.0252 |
| **VCA** | 3 | 29.4157 | <.0001 |
| **THA** | 2 | 12.1873 | 0.0023 |

Based on the p-value, variable gender, CPT, RBP, PES, VCA and THA have p-value less than 0.05, it means that those variables are significant while others are not significant.

**4.2.3 Stepwise Model Selection of Logistic Regression**

Table 4.9: Stepwise Selection.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of Stepwise Selection** | | | | | | | |
| **Step** | **Effect** | | **DF** | **Number In** | **Score Chi-Square** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Entered** | **Removed** |
| **1** | **THA** |  | 2 | 1 | 82.4601 |  | <.0001 |
| **2** | **VCA** |  | 3 | 2 | 54.9320 |  | <.0001 |
| **3** | **CPT** |  | 3 | 3 | 39.2233 |  | <.0001 |
| **4** | **OPK** |  | 1 | 4 | 15.7603 |  | <.0001 |
| **5** | **PES** |  | 2 | 5 | 8.6634 |  | 0.0131 |
| **6** | **GENDER** |  | 1 | 6 | 6.5413 |  | 0.0105 |
| **7** | **RBP** |  | 1 | 7 | 5.4749 |  | 0.0193 |

Table 4.9 depicts that the variable remained after selection of stepwise chosen. The p-value for all variables is significance as it less than 0.05.

Table 4.10: Model Fit Statistics

|  |  |  |
| --- | --- | --- |
| **Model Fit Statistics** | | |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 411.946 | 240.071 |
| **SC** | 415.640 | 269.621 |
| **-2 Log L** | 409.946 | 224.071 |

The table 4.10 shows the model fit statistics. The value for intercept and covariates should be less than the value of the intercept only.

Table 4.11: Testing Global Null Hypothesis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Testing Global Null Hypothesis: BETA=0** | | | |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 185.8753 | 7 | <.0001 |
| **Score** | 147.4065 | 7 | <.0001 |
| **Wald** | 83.9106 | 7 | <.0001 |

Based on the p-value for all tests above, H0 will be rejected because the p-value for each test are less than alpha value which is 0.05.

Table 4.12: Analysis of Effects of selected variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Type 3 Analysis of Effects** | | | |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Gender** | 1 | 7.8935 | 0.0050 |
| **CPT** | 3 | 31.6367 | <.0001 |
| **RBP** | 1 | 5.3116 | 0.0212 |
| **OPK** | 1 | 4.9429 | 0.0262 |
| **PES** | 2 | 11.7249 | 0.0028 |
| **VCA** | 3 | 33.5895 | <.0001 |
| **THA** | 2 | 14.5719 | 0.0007 |

Table 4.12 shows the hypothesis tests for each of variables selected in the model individually. All independents variables above are significance to the model since the p-values are less than 0.05.

Table 4.13: Association of predicted probabilities and observed responses

|  |  |  |  |
| --- | --- | --- | --- |
| **Association of Predicted Probabilities and Observed Responses** | | | |
| **Percent Concordant** | 90.9 | **Somers' D** | 0.818 |
| **Percent Discordant** | 9.1 | **Gamma** | 0.818 |
| **Percent Tied** | 0.0 | **Tau-a** | 0.408 |
| **Pairs** | 21920 | **c** | 0.909 |

Percent concordant measure the percentage of pairs where the observation with the desired outcome (event) has a higher predicted probability than the observation without the outcome (non-event). So, the percent concordant for this model is 90.9% which is greater than 80%. The model is considered as a good model. Percent discordant for this model is 9.1% which measure of the percentage of pairs where the observation with the desired outcome (event) has the lower predicted probability than the observation without the outcome (non-event).

The area under curve (c-statistics) is 0.909. It will range from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response. Since the value is greater than 0.90 it can be concluded that this model is excellent. But the model may be faced with problem of over-fitting. The value for Somer’s D is 0.818. It is a common measure for assessing predictive power of a credit risk model. Since the value is greater than 0.4, the model has better discrimination power than the model with random scores.

Table 4.14: The Analysis of Maximum Likelihood Estimates

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr >**  **ChiSq** | **Standardized Estimate** | **Exp (Est)** |
| **Intercept** |  | 1 | 0.3400 | 1.8290 | 0.0345 | 0.8525 |  | 1.405 |
| **gender** | **0** | 1 | 1.3994 | 0.4981 | 7.8935 | 0.0050 | 0.3615 | 4.053 |
| **CPT** | **1** | 1 | 2.7545 | 0.6883 | 16.0148 | <.0001 | 0.4066 | 15.713 |
| **CPT** | **2** | 1 | 1.2968 | 0.5403 | 5.7602 | 0.0164 | 0.2658 | 3.658 |
| **CPT** | **3** | 1 | 2.4426 | 0.4923 | 24.6136 | <.0001 | 0.6053 | 11.503 |
| **RBP** |  | 1 | -0.0247 | 0.0107 | 5.3116 | 0.0212 | -0.2418 | 0.976 |
| **OPK** |  | 1 | -0.4980 | 0.2240 | 4.9429 | 0.0262 | -0.3202 | 0.608 |
| **PES** | **1** | 1 | 0.6527 | 0.8624 | 0.5727 | 0.4492 | 0.1799 | 1.921 |
| **PES** | **2** | 1 | -0.8871 | 0.7968 | 1.2393 | 0.2656 | -0.2442 | 0.412 |
| **VCA** | **0** | 1 | 2.1953 | 0.9044 | 5.8928 | 0.0152 | 0.5972 | 8.983 |
| **VCA** | **1** | 1 | -0.1127 | 0.9575 | 0.0138 | 0.9063 | -0.0257 | 0.893 |
| **VCA** | **2** | 1 | -0.6359 | 1.1102 | 0.3281 | 0.5668 | -0.1173 | 0.529 |
| **THA** | **3** | 1 | 1.5237 | 0.4208 | 13.1107 | 0.0003 | 0.4184 | 4.589 |
| **THA** | **6** | 1 | 1.6071 | 0.7414 | 4.6982 | 0.0302 | 0.2118 | 4.988 |

Table 4.14 shows that the coefficients of PES (1 versus 3 and 2 versus 3) and VCA (1 versus 3 and 2 versus 3) are not significance while others are significance. Having gender of females compared to males will increase the log odds heart disease (status) of by 1.3994. For CPT, having CPT of 1 or 2 or 3 compared to CPT of 4 will increase the log odds of heart disease by 2.75545, 1.2968 and 2.4426 respectively. Besides, for every one-unit change in RPB or OPK, the log odds of heart disease decrease by 0.0247 and 0.4980 respectively. Next, having THA of 3 or 6 compared to 7 will increase the log odds of heart disease by 1.5237 and 1.6071.

Table 4.15: The Odds Ratio Estimates

|  |  |  |  |
| --- | --- | --- | --- |
| **Odds Ratio Estimates** | | | |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **Gender 0 vs 1** | 4.053 | 1.527 | 10.758 |
| **CPT 1 vs 4** | 15.713 | 4.077 | 60.552 |
| **CPT 2 vs 4** | 3.658 | 1.268 | 10.546 |
| **CPT 3 vs 4** | 11.503 | 4.382 | 30.190 |
| **RBP** | 0.976 | 0.955 | 0.996 |
| **OPK** | 0.608 | 0.392 | 0.943 |
| **PES 1 vs 3** | 1.921 | 0.354 | 10.413 |
| **PES 2 vs 3** | 0.412 | 0.086 | 1.963 |
| **VCA 0 vs 3** | 8.983 | 1.526 | 52.870 |
| **VCA 1 vs 3** | 0.893 | 0.137 | 5.836 |
| **VCA 2 vs 3** | 0.529 | 0.060 | 4.665 |
| **THA 3 vs 7** | 4.589 | 2.012 | 10.469 |
| **THA 6 vs 7** | 4.988 | 1.166 | 21.332 |

An odds ratio is the exponential coefficient and can be interpret as the multiplicative change in odds of one-unit change in predictor variables. Based on table 4.14, for one unit increase of RBP and OPK, the odds of heart disease (versus no heart disease) will increase by factors of 0.976 and 0.608 respectively. Besides, having CPT of 1, 2 or 3 compared to CPT of 4 will increase the odds of heart disease by a factor of 15.713, 3.658 and 11.503 individually. For PES, having PES 1o 2 compared to 3 rise the odds of heart disease by .921 or 0.412. For VCA, having VCA of 0, 1 and 2 compared to VCA of 4, the odds of heart disease increase by 8.983, 0.893 or 0.529 respectively. Lastly, having THA of 3 or 6 compared to 7 will rise the odds ratio by 4.589 and 4.988.

Table 4.16: The Goodness of Fit

|  |  |  |
| --- | --- | --- |
| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 7.7368 | 8 | 0.4596 |

Table 4.16 shows the p-value =0.4596 is greater than 0.05 which indicates the model fits data well. Based on appendix 1 of Hosmer and lemeshow goodness of fit test for full model, the p-value are increase from 0.0724 to 0.4596 thus indicate that reduced model of stepwise method is better than full model. This rule might be tough to achieve if you are working on large sample and small event rate.

**4.3 TESTING ACCURACY**

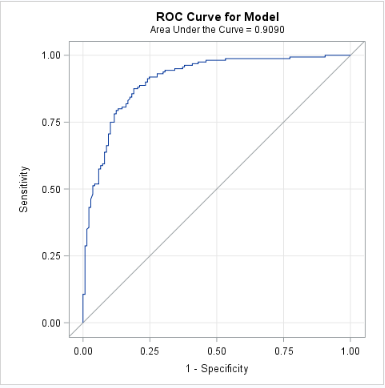


Figure 4.11: Area under curve.

The sensitivity, or true positive rate of the model, is shown on the y-axis while the false positive rate, or 1 minus the specificity, is given on the x-axis. The line shows how these two outcome measures vary with different threshold values. The ROC curve always starts at the point (0,0) i.e. threshold of value 1. This means at this threshold; it will be no sensitivity of 0 but will correctly label of all the good care cases (FP = 0).

The area under curve should be more than 0.7 in both training and validation samples. Since the area under the curve is 0.909 which is greater than 0.8, it is considered as an excellent score. The ROC curve captures all thresholds simultaneously. The higher the threshold, or closer to (0,0), the higher the specificity and the lower the sensitivity. The lower the threshold, or closer to (1,1), the higher the sensitivity and lower the specificity.

The value of area under curve (AUC) can be obtained by:

= 0.909 + 0.5 \*0.0 =0.909.

Can be referred at table 4.12.

Table 4.17: Classification table (Confusion Matrix).

|  |  |  |
| --- | --- | --- |
|  | **Into: status** | |
| **0** | **1** |
| **N** | **N** |
| **From: status** | 36 | 5 |
| **0** |
| **1** | 7 | 41 |

From the table 4.17, the value for TN is 36, FP is 5, FN is 7 and TP is 41. There are two possible predicted classes: “0” for surviving heart disease and “1” for having heart disease. The classifier made a total of 89 predictions. Out of those 89 predicted, “0” 43 times, and “1” 46 times. In reality, 41 people in the sample surviving heart disease while 48 people having the heart disease.

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

* **Accuracy**: (36+41) / 89 = 0.86
* **Misclassification Rate**: (5+7) / 89 = 0.13
* **Sensitivity / True Positive Rate**: 41 / 48 = 0.85
* **False Positive rate**: 5 / 41 = 0.12
* **Specificity / True Negative Rate**: 36 / 41 = 0.88
* **Precision**: 41 / 46 = 0.89
* **Prevalence**: 48 / 89 = 0.54

Based on the calculated value above, the accuracy is 0.86 which is greater than 0.5. it describes on how often the classifier correct. The percentage of people of heart disease who correctly identified as having the condition is 85%. The percentage of people of not having heart disease who correctly identified as having the condition is 88%. The percentage of people of not having heart disease but identified as having heart disease is 12%. The percentage when the predicted is true and it is correct is 89%.

Table 4.18: Kolmogorov Smirnov (KS) Statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Kolmogorov-Smirnov Two-Sample Test (Asymptotic)** | | | |
| **KS** | 0.304828 | **D** | 0.618503 |
| **KSa** | 2.875739 | **Pr > KSa** | <.0001 |

The D statistic is the metrics used to report Kolmogorov Smirnov score. “KS” showing in the output table ‘K-S Two-Sample Test (Asymptotic)’. The D statistic is the maximum difference between the cumulative distributions between events (Y=1) and non-events (Y=0). For this model, D=0.618503. the higher value of D the better the model distinguishes between events and non-events.

Table 4.19: A lift table for the model

|  |  |  |  |
| --- | --- | --- | --- |
| **Decile** | **Status** | | |
| **Count** | **Number of Responses** | **% of events** |
| **0** | 8 | 7.00 | 18.92 |
| **1** | 9 | 8.00 | 21.62 |
| **2** | 9 | 5.00 | 13.51 |
| **3** | 9 | 5.00 | 13.51 |
| **4** | 9 | 5.00 | 13.51 |
| **5** | 9 | 3.00 | 8.11 |
| **6** | 9 | 1.00 | 2.70 |
| **7** | 9 | 1.00 | 2.70 |
| **8** | 9 | 2.00 | 5.41 |
| **9** | 9 | 0.00 | 0.00 |

The table above shows that the percent of targets (events) covered at a given decile level after the data split into two datasets; 70% for training and 30% for validation.

Table 4.20: The frequency of the target for validation data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **status** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **0** | 52 | 58.43 | 52 | 58.43 |
| **1** | 37 | 41.57 | 89 | 100.00 |

From the table 4.20, it shows that the frequency for the validation data of having heart disease is 37 people while 52 people are surviving heart disease.

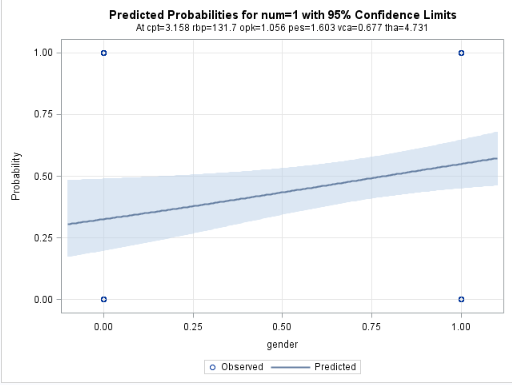


Figure 4.12: Fit plot for a regression.

Figure 4.12 illustrates the predicted probabilities when the target is equal to 1 with 95 percent confidence limits. The class gender is used. There is an increasing pattern of the predicted line and observed line.

**CHAPTER 5**

**CONCLUSION AND RECOMMENDATIONS**

**5.1 INTRODUCTION**

This chapter covers the conclusion of this research. Furthermore, recommendations for further research and the conclusion that can be drawn also be discussed.

**5.2 CONCLUSION**

In conclusion, the overall data sets composed of 5 numerical and 8 categorical independent variables with one categorical dependent variable. Majority of the patients are male aged between 29-77 years old. From the chi-square test, only one variable which is FBS has the association with status of heart disease while the others categorical independent variables are non-correlated with status. In this study, logistic regression analyses were used to analyse the heart disease data. From the result of reducing the model from 13 independent variables to 7 independent variables using stepwise selection method, it can be concluded that only variablesTHA- Thallium scan, VCA- Number of major vessels, CPT- Chest pain type, OPK- ST depression induced by exercise relative to rest, PES- The slope of the peak exercise ST segment, GENDER, and RBP- Resting blood pressure are significant for predicting the heart disease. Using the Cleveland Heart Disease database, this project provides guidelines to train and test the system. Furthermore, this project also includes a study of the testing accuracy results which include sensitivity, specificity, and accuracy.

**5.3 RECOMMENDATION**

There is a need for combinational and more complex models to increase the accuracy of predicting the early onset of cardiovascular diseases. For further researcher, instead of only focusing on logistic regression, other model should be used such as support vector machines, decision tree, neural networks, ensemble algorithm, etc .Rani et. al (2018) suggest researcher should involve the development of tool to predict the risk of having a heart disease of a prospective patient.

**REFERENCES**

Al-Shayea Q.K., “Artificial neural networks in medical diagnosis”, International

Journal of Computer Science Issues, vol. 8, no. 2, pp. 150-154, 2011.

# American Heart Association News (2018, July 18). Heart health at any age – 40,

# 50, 60 and beyond. Retrieved From

# [https://www.heart.org/en/news/2018/07/20/heart-health-at-any-age-40-50-60-and- beyond](https://www.heart.org/en/news/2018/07/20/heart-health-at-any-age-40-50-60-and-%20%20%20%20%20%20beyond)

Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and

Ruinan Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,” Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. Retrieved from

<https://www.hindawi.com/journals/misy/2018/3860146/cta/>

Bui A.L., Horwich T.B., and Fonarow G.C., “Epidemiology and risk profile of

heart failure”, Nature Reviews Cardiology, vol. 8, no. 1, pp. 30-41, 2011.

Beatrice N. J., (2019, January 23). Cardiovascular diseases have almost doubled in

13 years: Deputy Health Minister. New Straits Time. Retrieved from

<https://www.nst.com.my/news/nation/2019/01/453741/cardiovascular-diseases-have-almost-doubled-13-years-deputy-health>

Durairaj M. and Ramasamy N., “A comparison of the perceptive approaches for

pre-processing the dataset for predicting fertility success rate”, International Journal of Control Theory and Applications, vol. 9, pp. 256-260, 2016.

Department of Statistic Malaysia, Official Portal (2019, October 30). Statistic on

Causes of Death, Malaysia, 2019. Retrieved from

<https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=RUxlSDNkcnRVazJnakNCNVN2VGgrdz09>

Kelley, George & Kelley, Kristi & Tran, Zung. (2000). Exercise and bone mineral

density in men: A meta-analysis. Journal of applied physiology (Bethesda, Md.

: 1985). 88. 1730-6. 10.1152/jappl.2000.88.5.1730.

<https://www.researchgate.net/publication/12520531_Exercise_and_bone_mineral_density_in_men_A_meta-analysis/citation/download>

Lim, Y. C., Teo, S. G., & Poh, K. K. (2016). ST-segment changes with exercise

stress. *Singapore medical journal*, *57*(7), 347–353.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4958709/>

Maas, Angela & Appelman, Yolande. (2010). Gender differences in coronary

heart disease. Netherlands heart journal: monthly journal of the Netherlands

Society of Cardiology and the Netherlands Heart Foundation. 18. 598-602.

10.1007/s12471-010-0841-y. Retrieved From

<https://www.researchgate.net/publication/49817773_Gender_differences_in_coronary_heart_disease/citation/download>

Marateb, H. R., & Goudarzi, S. (2015). A noninvasive method for coronary artery

diseases diagnosis using a clinically-interpretable fuzzy rule-based system. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, *20*(3), 214–223.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4468223/>

Mair FS, Crowley TS, (1996). Bundred PE.Prevalence, aetiology and

management of heart failure in general practice. Br J Gen Pract; 46:77–9.

Rani K.S, Chaitanya M.S and Kiran G.S (2018). A Heart Disease Prediction

Model Using Logistic Regression by Cleveland Database. International

Journal of Trend in Scientific Research and Development (IJTSRD) Volume

2 Issues 3. Retrieved from

<http://Www.Ijtsrd.Com/Computer-Science/Data-Miining/11401/A-Heart-Disease-Prediction-Model-Using-Logistic-Regression/K-Sandhya-Rani>

Sawyer, Steven. (2009). Analysis of Variance: The Fundamental Concepts.

Journal of Manual & Manipulative Therapy. 17. 27E-38E.

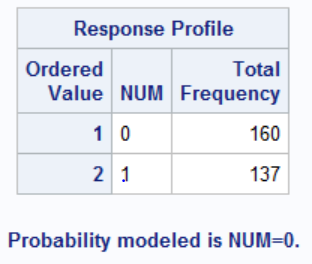
10.1179/jmt.2009.17.2.27E. Retrieved from

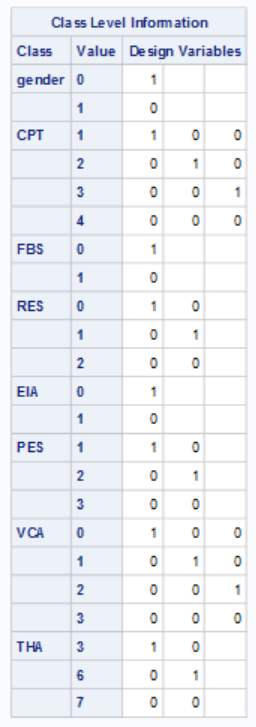
<https://www.researchgate.net/publication/272311020_Analysis_of_Variance_The_Fundamental_Concepts>

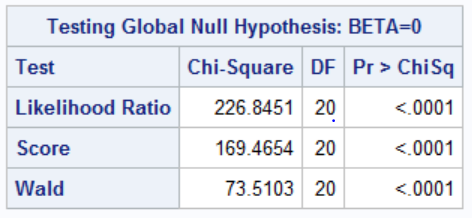
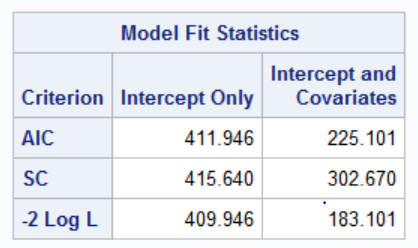
Wei JY, Gersh BJ., (1987).Heart disease in the elderly. Curr Probl Cardiol; 12:1

65.

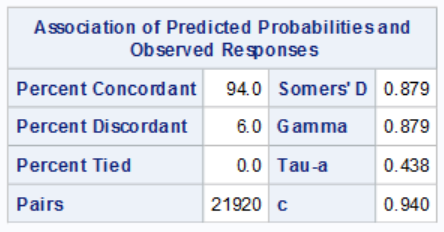
**APPENDIX 1 (FULL MODEL)**

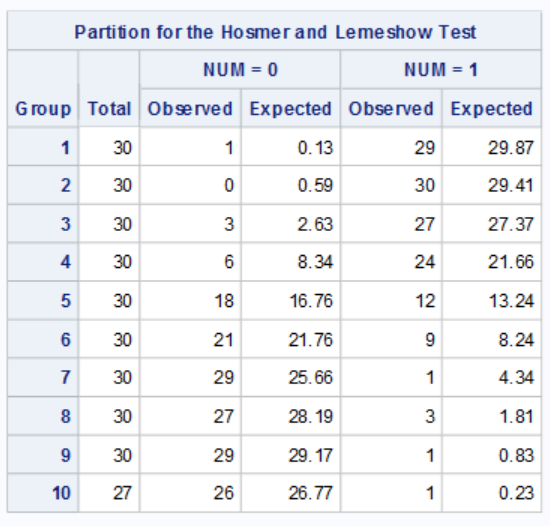


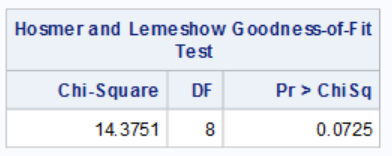


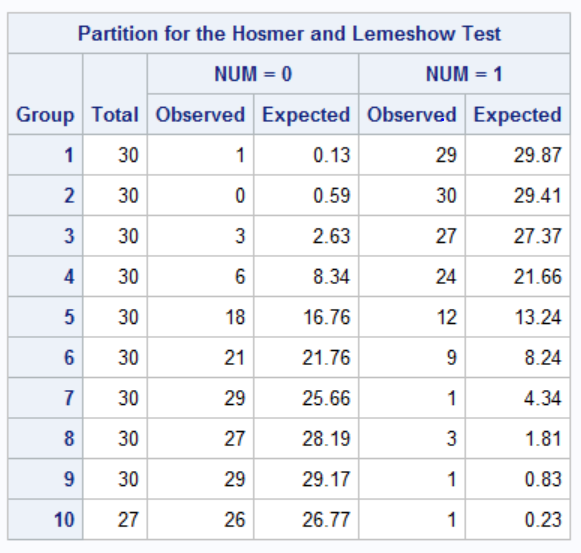


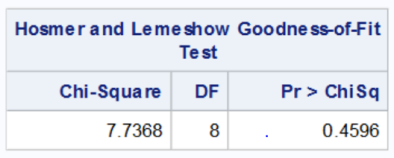




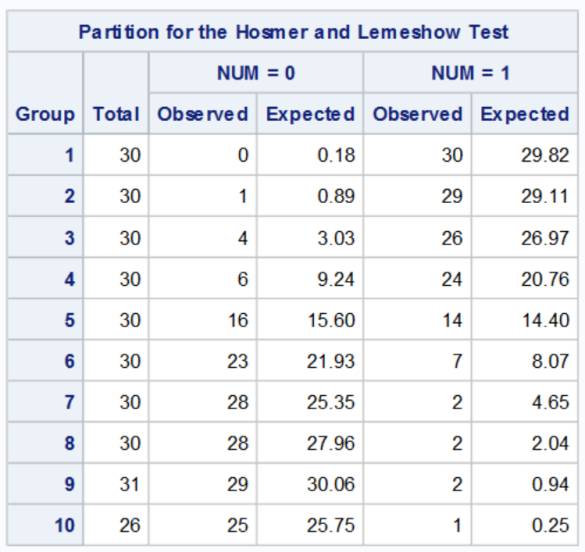


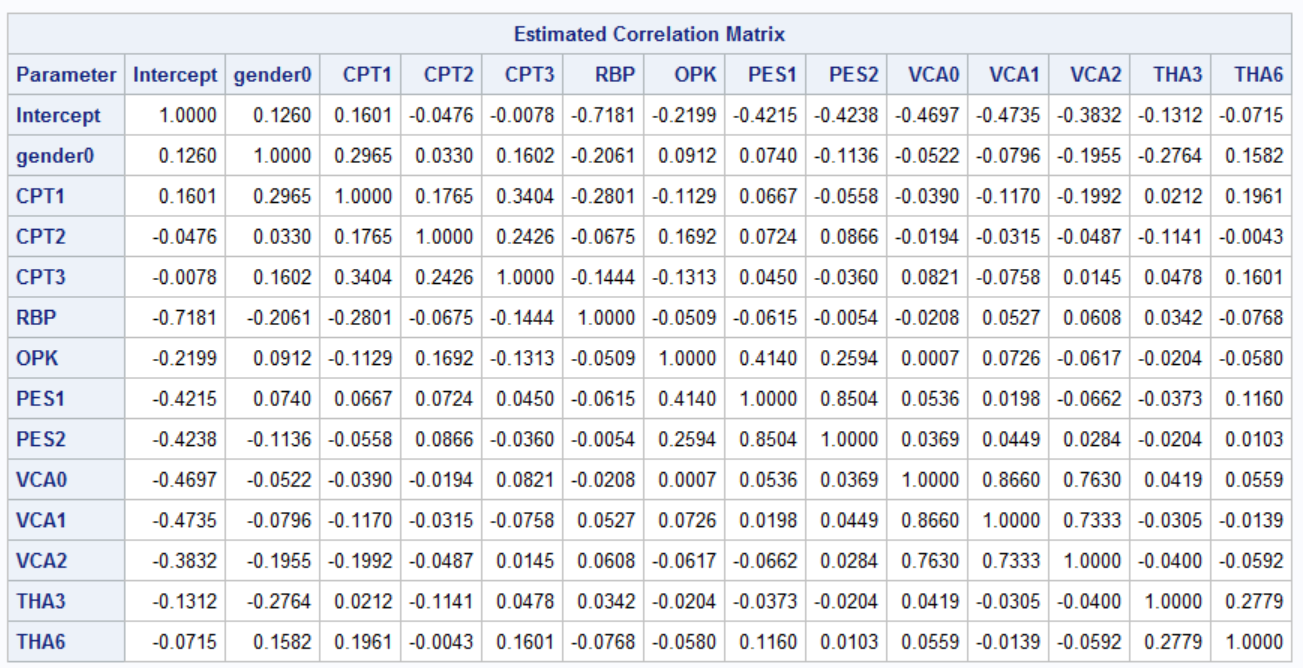






**APPENDIX 2 (REDUCED MODEL)**





**APPENDIX 3**

**Data** work.heart\_disease;

infile "I:\PROJECT 2\heart\_disease.csv" dlm=',';

input Age gender CPT RBP SCH FBS RES MHR EIA OPK PES VCA THA STATUS ;

if vca=-**100000** then delete;

if tha=-**100000** then delete;

**run**;

**proc** **gchart** data=work.heart\_disease;

pie gender / percent=arrow;

**run**;

**proc** **gchart** data=work.heart\_disease;

pie status / percent=arrow;

**run**;

**proc** **gchart** data=work.heart\_disease;

vbar age/ type=percent subgroup=gender;

**run**;

**proc** **gchart** data=work.heart\_disease;

vbar age /group=gender;

where status=**1**;

**run**;

**proc** **sgplot** data=work.heart\_disease;

vbar status/ group=CPT groupdisplay=cluster;

title'Bar Chart of Chest Pain Type (CPT)';

**run**;

**proc** **sgplot** data=work.heart\_disease;

vbar status/ group=RES groupdisplay=cluster;

title'Bar Chart of Resting Electrocardiographic Results (RES)';

**run**;

**proc** **sgplot** data=work.heart\_disease ;

vbar status/ group=EIA groupdisplay=cluster;

title'Bar Chart of Exercise induced angina (EIA)';

**run**;

**proc** **sgplot** data=work.heart\_disease ;

vbar status/ group=PES groupdisplay=cluster;

title'Bar Chart of The Slope of The Peak Exercise ST Segment (PES)';

**run**;

**proc** **sgplot** data=work.heart\_disease ;

vbar status/ group=THA groupdisplay=cluster;

title'Bar Chart of Thallium scan(THA)';

**run**;

**proc** **sgplot** data=work.heart\_disease ;

vbar status/ group=FBS groupdisplay=cluster;

title'Bar Chart of Fasting blood sugar(FBS)';

**run**;

**proc** **freq** data=work.heart\_disease nlevels;

tables \_all\_;

**run**;

**proc** **means** data=work.heart\_disease n nmiss mode mean stddev min max sum maxdec=**2**;

var age rbp sch mhr opk;

**run**;

**proc** **freq** data=heart\_disease;

tables status\*(gender CPT FBS RES EIA PES VCA THA)/ chisq

plots=(freqplot(twoway=grouphorizontal scale=percent));

**run**;

**proc** **logistic** data=work.heart\_disease;

class gender CPT FBS RES EIA PES VCA THA/ param=ref;

model status=Age gender CPT RBP SCH FBS RES MHR EIA OPK PES VCA THA/lackfit;

**run**;

**proc** **logistic** data=work.heart\_disease;

class gender CPT FBS RES EIA PES VCA THA/ param=ref;

model status=Age gender CPT RBP SCH FBS RES MHR EIA OPK PES VCA THA/selection=stepwise exbp stb corrb lackfit;

**run**;

\*70% train 30% valid;

**data** train valid;

infile "C:\Users\Admin\Desktop\SAS\PROJECT SAS 2.0\heart\_disease.xls" dlm='","';

input age gender cpt rbp sch fbs res mhr eia opk pes vca tha num valid train;

if vca=-**100000** then delete;

if tha=-**100000** then delete;

IF \_n\_ LT **89** THEN OUTPUT TRAIN;

ELSE OUTPUT VALID;

**run**;

ods select none;

**PROC LOGISTIC** DATA=TRAIN;

MODEL num = age gender cpt rbp sch fbs res mhr eia opk pes vca tha /CTABLE;

SCORE DATA=VALID OUT=SCORE;

**RUN**;

ods select all;

**PROC** **TABULATE** DATA=SCORE;

CLASS F\_num I\_num;

TABLE F\_num, I\_num;

**RUN**;

**proc** **logistic** data=heart;

MODEL num (event="1") = gender cpt rbp opk pes vca tha;

output out=out2 p=pred;

**run**;

**proc** **npar1way** data=out2 edf;

class num;

var pred;

**run**;

\*split data into two datasets :70%-training 30%-validation;

**proc** **surveyselect** data=heart out=split samprate=**.7** outall;

**run**;

**data** training validation;

set split;

if selected = **1** then output training;

else output validation;

**run**;

ods graphics on;

**proc** **logistic** data=training;

model num = gender cpt rbp opk pes vca tha / ctable pprob=**0.5** outroc=troc;

output out=test p=ppred;

score data=validation out=logit\_file outroc=vroc;

**run**;

**proc** **rank** data=logit\_file descending groups=**10** out=predrank\_Dev;

var P\_1;\*predrank-variable name for ranks;

ranks predrank;

**run**;

**proc** **tabulate** data=predrank\_dev;

class predrank;

var num;

table predrank="" all, num\*(N="Count" SUM="Number of Responses"

COLPCTSUM="% of events")/box="decile";

title Creating a Lift Table for Model;

**run**;

**proc** **freq** data=predrank\_dev;

tables num;

**run**;

**proc** **npar1way** data=logit\_file edf;

class num;

var p\_1;

**run**;

ods graphics on;

**proc** **logistic** data=heart plots=all;

model num= gender cpt rbp opk pes vca tha/ clodds=pl;

**run**;

ods graphics off;

ODS GRAPHICS ON;

**PROC** **LOGISTIC** DATA=HEART PLOTS=EFFECT;

model num (EVENT="1")= gender cpt rbp opk pes vca tha;

OUTPUT OUT=LOGIOUT PREDICTED=PREDPROB;

**RUN**;

ODS GRAPHICS OFF;