FACULTY OF COMPUTING AND INFORMATICS

KD24203 DATA MINING AND WAREHOUSING

Semester 2 Session 2021/2022

Project Title: Heart Disease Prediction using Classification

Prepared for: Dr. Florence Sia Fui Sze

Prepare By,

No.	Name	Matric No.
1.	CHAN KAI KHEE	BI20110118
2.	LEE YI FENG	BI20110003
3.	TEW ZHI CHYI	BI20110110
4.	TOH SIN TONG	BI20110062

Table of Contents

1.0 Introduction

The heart is one of the important parts of the human body. It is the generator of the human body to transfer nutrients or power through the human body. Nowadays, the heart disease death rates have been rising annually which has become the top killer in the world as in World Health Organization reports [1]. According to the Centers for Disease Control and Prevention as provided by the United States government [2], it shows the 2020 annual Centers for Disease Control and Prevention survey data of 400k adults and their health status. There are a lot of factors that will cause heart diseases to become the leading cause of death globally. For example,lack of physical activity, unhealthy diet, smoking and alcohol drinking. The most common symptoms for heart disease are chest pain, shortness of breath, feeling sweaty, etc. These symptoms are to warn people that their physical health might be having problems currently. However, most people choose to ignore these small pains in the body. But these small pains might show the signals to the people as a reminder to take a rest and carry out a regular body check up.

From time to time, much research and real-time data are available. Due to the technology generation, many open sources can access the patient's data and research found on the internet. Thus, if the information has been predicted in advance, it will help doctors to detect this disease in a short time. Additionally, it can provide doctors the correct diagnosis which may allow them to arrange the treatment for the patients. Currently, machine learning and artificial intelligence play an important role in all industries, especially medical. There are many different machine learning models used to diagnose the disease and predict future trends or results. Hence, in order to solve this problem, the researchers want to investigate the data mining algorithm for predicting heart diseases on the personal key risks by using classification methods. There are several classification techniques to predict heart diseases. For instance, decision trees, logistic regression and support vector machines.

There are numerous classification methods to solve this real-time problem. Hence, the classification method proposed to use the Artificial Neural Network to predict heart disease. Artificial Neural Network also known as Neural Network which is used to model intricate patterns and prediction obstacles. It likes the human brain to observe from data and make decisions through commenting on similar events [3]. ANN has the potential to learn and model non-linear and complex relationships. Thus, it can be said that it plays an important role as many of the relationships between inputs and outputs are non-linear and complicated. In

addition, ANN is different from other prediction models because it does not impose any restrictions on the input variables like how they should be allocated. Also, after learning from the initial inputs and their relationships, it can infer unknown relationships on unknown data as well, thus making the model generalize and predict on unknown data.

2.0 State of Arts/Literature Review

Author	Classification Method	Domain	Description
Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). [4]	Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Artificial Neural Network, Naive Bayer, Decision Tree	This research is to design a machine-learning-based medical intelligent decision support system for the diagnosis of heart disease.	The classifiers logistic regression with 10-fold cross-validation showed best accuracy 89% when selected by FS algorithm Relief.
Mohan, S., Thirumalai, C., & Srivastava, G. (2019). [5]	Naive Bayes, Logistic Regression, Hybrid Random Forest with Linear Model, Generalized Linear Model, Support Vector Machine, Decision Tree, DL, Neural Networks, Gradient Boosted Trees, Language Model, K-Nearest Neighbor	Prediction of cardiovascular disease.	Enhanced performance of the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).
Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). [6]	Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, and Ensemble Model	Predicting cardiovascular diseases or heart related diseases.	Every algorithm has outperformed in some cases but some will perform poorly in some other cases.
Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan,	Logistic Regression, K-Nearest Neighbor, Support Vector Machine (Radial Basis Function), Linear	Identification of heart diseases with	The use of feature selection algorithms to choose the suitable

A., & Saboor, A. (2020). [7]	Support Vector Machine, Naive Bayes, Decision Tree, Artificial Neural Network	feature selection in E-Healthcare.	characteristics, which improved classification accuracy and reduced diagnosis system processing time.
Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). [8]	Artificial Neural Network	The proposed solution, which is based on Artificial Neural Networks, provides a decision support system to identify three main heart diseases: mitral stenosis, aortic stenosis and ventricular septal defect.	This algorithm performs well and accuracy is excellent, with a heart disease classification accuracy of 92%.
Patel, J., TejalUpadhyay, D., & Patel, S. (2015). [9]	Logistic Model Tree, J48 with ReducedErrorpruning, Random Forest	The main purpose is to use the different algorithm of Decision Trees to enhance the Decision Tree accuracy in identifying heart disease patients.	The best algorithm J48 based on UCI data has the highest accuracy among the three models.
Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). [10]	Decision Tree, Naive Bayes, and Artificial Neural Network	Early heart disease prediction using data mining techniques.	Decision Tree is the best algorithm as it has outperformed with the accuracy of 99.62%.

Table 1

There are different types of classification methods that have been used in the research papers, such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes,

Artificial Neural Network (ANN), and etc. Through all the research papers in Table 1, the researchers can conclude that the most frequent classification method that is used is Artificial Neural Network (ANN). Furthermore, each of the research papers have a different result on deciding the best classification method. As a result, since ANN does not achieve the highest accuracy among all the datasets, the researchers decided to use ANN to discover more in this project.

In many ways, the diagnosis of heart disease based on standard medical history has been considered inaccurate. Noninvasive techniques such as machine learning are reliable and effective for classifying healthy persons and people with heart disease. According to Haq et al., 2018 [4], people with heart disease can be easily identified and classified using the proposed system. There are seven types of classification methods, three feature selection algorithms, the cross-validation method, and seven classifiers performance evaluation metrics are used in this paper. The classification methods are Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Tree and Artificial Neural Network. As a result, the classifiers logistic regression with 10-fold cross-validation showed best accuracy 89% when selected by FS algorithm Relief [4].

In healthcare of cardiology, improved decision making detection of heart disease is critical Li et al., 2020 [7]. In this paper, classification algorithms, feature selection, and cross-validation methods are used to increase the performance of the system. The classification methods are Logistic Regression, K-Nearest Neighbor, Support Vector Machine (Radial Basis Function), Linear Support Vector Machine, Naive Bayes, Decision Tree and Artificial Neural Network. Thus, the ANN is the best predictive system based on the specificity Li et al., 2020 [7].

According to Ghwanmeh et al., 2013 [8], the purpose of the research is to classify the type of heart diseases which are mitral stenosis, aortic stenosis and ventricular septal defect. Thus, Artificial Neural Network (ANN) has been used by Ghwanmeh et al., 2013 [8]. It also stated that the series of experiments have been conducted by using real medical data to examine the performance and accuracy of ANN. As a result, the accuracy for ANN is 92% which proves that performance and accuracy of the system are acceptable [8].

3.0 Data Preparation

The dataset that is used is "Personal Key Indicators of Heart Disease" which downloaded from Kaggle. The dataset includes 319795 observations and 18 variables. The variables are HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease and SkinCancer. Figure 1 shows the first 27 rows of the raw dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
2	No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
3	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
4	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
5	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
6	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
7	Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
8	No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes
9	No	31.64	Yes	No	No	5	0	Yes	Female	80 or older	White	Yes	No	Good	9	Yes	No	No
10	No	26.45	No	No	No	0	0	No	Female	80 or older	White	No, borde	No	Fair	5	No	Yes	No
11	No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No
12	Yes	34.3	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No	Poor	15	Yes	No	No
13	No	28.71	Yes	No	No	0	0	No	Female	55-59	White	No	Yes	Very good	5	No	No	No
14	No	28.37	Yes	No	No	0	0	Yes	Male	75-79	White	Yes	Yes	Very good	8	No	No	No
15	No	28.15	No	No	No	7	0	Yes	Female	80 or older	White	No	No	Good	7	No	No	No
16	No	29.29	Yes	No	No	0	30	Yes	Female	60-64	White	No	No	Good	5	No	No	No
17	No	29.18	No	No	No	1	0	No	Female	50-54	White	No	Yes	Very good	6	No	No	No
18	No	26.26	No	No	No	5	2	No	Female	70-74	White	No	No	Very good	10	No	No	No
19	No	22.59	Yes	No	No	0	30	Yes	Male	70-74	White	No, borde	Yes	Good	8	No	No	No
20	No	29.86	Yes	No	No	0	0	Yes	Female	75-79	Black	Yes	No	Fair	5	No	Yes	No
21	No	18.13	No	No	No	0	0	No	Male	80 or older	White	No	Yes	Excellent	8	No	No	Yes
22	No	21.16	No	No	No	0	0	No	Female	80 or older	Black	No, borde	No	Good	8	No	No	No
23	No	28.9	No	No	No	2	5	No	Female	70-74	White	Yes	No	Very good	7	No	No	No
24	No	26.17	Yes	No	No	0	15	No	Female	45-49	White	No	Yes	Very good	6	No	No	No
25	No	25.82	Yes	No	No	0	30	No	Male	80 or older	White	Yes	Yes	Fair	8	No	No	No
26	No	25.75	No	No	No	0	0	No	Female	80 or older	White	No	Yes	Very good	6	No	No	Yes
27	No	29.18	Yes	No	No	30	30	Yes	Female	60-64	White	No	No	Poor	6	Yes	No	No

Figure 1: First 27 rows of the dataset "Personal Key Indicators of Heart Disease"

The dataset is to make prediction of heart disease, relevant organizations for example hospital can utilize the data to better predict patients with heart disease. From the figure 1 above, there might contain junk data or invalid data. In this case, data preparation should be carried out to ensure the data is clean and suitable for analyzation. Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data. The common main processes of data preparation are normalization, conversion, missing value imputation and resampling. The processes of data preparation will be showed after this.

The process of data preparation will start from the data collection. The dataset "Personal Key Indicators of Heart Disease" is downloaded from Kaggle. After that, data cleaning is carried out to deal with missing data, invalid data and remove duplicates data. After a clean data is produced, the format of data could be changed to ease the data mining and data evaluation process. Last, the data which has been processed could be stored or channelled into a third party application such as excel.

3.1 Feature Extraction with R

Firstly, the data is imported with the function "read.csv()". Next, "str()" function is used to show the internal structure of the variables. "summary()" function is used to produce result summaries of the results of various model fitting functions and "head()" function is used to get the first 6 rows of data.

```
> #read data
> hd<-read.csv("C:\\Users\\kheec\\Downloads\\KD24203 Data mining&w\\project\\heart_2020_cleaned.csv")
> str(hd[1,])
'data.frame':   1 obs. of  18 variables:
 $ HeartDisease    : chr "No"
 $ BMI             : num 16.6
 $ Smoking         : chr "Yes"
 $ AlcoholDrinking : chr "No"
 $ Stroke          : chr "No"
 $ PhysicalHealth  : num 3
 $ MentalHealth    : num 30
 $ DiffWalking     : chr "No"
 $ Sex             : chr "Female"
 $ AgeCategory     : chr "55-59"
 $ Race            : chr "White"
 $ Diabetic        : chr "Yes"
 $ PhysicalActivity: chr "Yes"
 $ GenHealth       : chr "Very good"
 $ SleepTime       : num 5
 $ Asthma          : chr "Yes"
 $ KidneyDisease   : chr "No"
 $ SkinCancer      : chr "Yes"
```

Figure 2: Importing dataset in R

```
> summary(hd)
 HeartDisease            BMI           Smoking          AlcoholDrinking      Stroke          PhysicalHealth    MentalHealth
 Length:319795      Min.   :12.02   Length:319795      Length:319795      Length:319795      Min.   : 0.000   Min.   : 0.000
 Class :character   1st Qu.:24.03   Class :character   Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.000
 Mode  :character   Median :27.34   Mode  :character   Mode  :character   Mode  :character   Median : 0.000   Median : 0.000
                    Mean   :28.33                                                           Mean   : 3.372   Mean   : 3.898
                    3rd Qu.:31.42                                                           3rd Qu.: 2.000   3rd Qu.: 3.000
                    Max.   :94.85                                                           Max.   :30.000   Max.   :30.000
 DiffWalking            Sex             AgeCategory           Race             Diabetic         PhysicalActivity    GenHealth
 Length:319795      Length:319795      Length:319795      Length:319795      Length:319795      Length:319795      Length:319795
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


   SleepTime         Asthma          KidneyDisease       SkinCancer
 Min.   : 1.000   Length:319795      Length:319795      Length:319795
 1st Qu.: 6.000   Class :character   Class :character   Class :character
 Median : 7.000   Mode  :character   Mode  :character   Mode  :character
 Mean   : 7.097
 3rd Qu.: 8.000
 Max.   :24.000
```

Figure 3: Summary of the dataset

```
> head(hd)
  HeartDisease   BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth DiffWalking    Sex AgeCategory  Race Diabetic
1           No 16.60     Yes              No     No              3           30          No Female       55-59 White      Yes
2           No 20.34      No              No    Yes              0            0          No Female 80 or older White       No
3           No 26.58     Yes              No     No             20           30          No   Male       65-69 White      Yes
4           No 24.21      No              No     No              0            0          No Female       75-79 White       No
5           No 23.71      No              No     No             28            0         Yes Female       40-44 White       No
6          Yes 28.87     Yes              No     No              6            0         Yes Female       75-79 Black       No
  PhysicalActivity GenHealth SleepTime Asthma KidneyDisease SkinCancer
1              Yes Very good         5    Yes            No        Yes
2              Yes Very good         7     No            No         No
3              Yes      Fair         8    Yes            No         No
4               No      Good         6     No            No        Yes
5              Yes Very good         8     No            No         No
6               No      Fair        12     No            No         No
```

Figure 4: First 6 rows of the dataset

3.2 Data Cleaning

3.2.1 Handling Missing Values

```
> #Check for missing values
> sapply(hd, function(x) sum(is.na(x)))
    HeartDisease              BMI          Smoking  AlcoholDrinking           Stroke   PhysicalHealth     MentalHealth
               0                0                0                0                0                0                0
     DiffWalking              Sex      AgeCategory             Race         Diabetic PhysicalActivity        GenHealth
               0                0                0                0                0                0                0
       SleepTime           Asthma     KidneyDisease       SkinCancer
               0                0                0                0
```

Figure 5: Check missing value in the dataset

After the data is imported, the "sapply()" function is used to check for missing values in the dataset. From the figure 5 above, it shows that there is no missing value in this dataset. Thus, the data is cleaned and is ready for the next step.

3.2.2 Handling Duplicate Values

```
> hd<-hd[!duplicated(hd),]
> hd
```

hd	319795 obs. of 18 variables

hd	301717 obs. of 18 variables

Figure 6: Removing duplicate data

Next, the "!duplicated()" function is used to remove all the duplicates data to ensure all the data are unique. The new data is stored as "hd" to replace the old dataset. As the result shown from the figure 6 above, the observations of the dataset decreased, meaning that there are duplicate values that have been removed.

3.2.3 Encoding categorical features

```
> #Encoding categorical labels
> hd$HeartDisease<-ifelse(hd$HeartDisease=="Yes",1,0)
> hd$Smoking<-ifelse(hd$Smoking=="Yes",1,0)
> hd$AlcoholDrinking<-ifelse(hd$AlcoholDrinking=="Yes",1,0)
> hd$Stroke<-ifelse(hd$Stroke=="Yes",1,0)
> hd$DiffWalking<-ifelse(hd$DiffWalking=="Yes",1,0)
> hd$Sex<-ifelse(hd$Sex=="Male",1,0)
> unique(hd$AgeCategory)
 [1] "55-59"       "80 or older" "65-69"       "75-79"       "40-44"       "70-74"       "60-64"       "50-54"       "45-49"
[10] "18-24"       "35-39"       "30-34"       "25-29"
> hd$AgeCategory = factor(hd$AgeCategory,
+                         levels = c('18-24','25-29','30-34','35-39','40-44','45-49',
+                                    '50-54','55-59','60-64','65-69','70-74','75-79','80 or older'),
+                         labels = c(1,2,3,4,5,6,7,8,9,10,11,12,13))
> unique(hd$Race)
[1] "White"                            "Black"                            "Asian"
[4] "American Indian/Alaskan Native" "Other"                            "Hispanic"
> hd$Race = factor(hd$Race,
+                  levels = c('White','Black','Asian','American Indian/Alaskan Native','Other','Hispanic'),
+                  labels = c(1, 2, 3, 4, 5, 6))
> unique(hd$Diabetic)
[1] "Yes"                "No"                 "No, borderline diabetes" "Yes (during pregnancy)"
> hd$Diabetic = factor(hd$Diabetic,
+                      levels = c('Yes','No','No, borderline diabetes','Yes (during pregnancy)'),
+                      labels = c(1, 2, 3, 4))
> hd$PhysicalActivity<-ifelse(hd$PhysicalActivity=="Yes",1,0)
> unique(hd$GenHealth)
[1] "Very good" "Fair"      "Good"      "Poor"      "Excellent"
> hd$GenHealth = factor(hd$GenHealth,
+                       levels = c('Very good','Fair','Good','Poor','Excellent'),
+                       labels = c(1, 2, 3, 4, 5))
> hd$Asthma<-ifelse(hd$Asthma=="Yes",1,0)
> hd$KidneyDisease<-ifelse(hd$KidneyDisease=="Yes",1,0)
> hd$SkinCancer<-ifelse(hd$SkinCancer=="Yes",1,0)
```

Figure 7: Encoding categorical features

Next, the “ifelse()” and “factor()” functions are used to convert categorical data into integer format so that the data with converted categorical values can be provided to the different models. Besides, "unique()" function is used to get the unique values in each variable. There are total of 18 variables in this dataset and 14 of them are categorical attributes and 4 of them are numerical attributes. Thus, the 14 of the categorical attributes have been converted into integer format.

```
> #Convert factor to numeric
> hd$AgeCategory=as.numeric(hd$AgeCategory)
> hd$Race=as.numeric(hd$Race)
> hd$Diabetic=as.numeric(hd$Diabetic)
> hd$GenHealth=as.numeric(hd$GenHealth)
```

Figure 8: Convert factor value to numeric value

“as.numeric()” function has also been used to convert factor value to numeric value to facilitate the following process.

```
> #Check again the data types of variables
> str(hd)
'data.frame':   319795 obs. of  18 variables:
 $ HeartDisease    : num  0 0 0 0 0 1 0 0 0 0 ...
 $ BMI             : num  16.6 20.3 26.6 24.2 23.7 ...
 $ Smoking         : num  1 0 1 0 0 1 0 1 0 0 ...
 $ AlcoholDrinking : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Stroke          : num  0 1 0 0 0 0 0 0 0 0 ...
 $ PhysicalHealth  : num  3 0 20 0 28 6 15 5 0 0 ...
 $ MentalHealth    : num  30 0 30 0 0 0 0 0 0 0 ...
 $ DiffWalking     : num  0 0 0 0 1 1 0 1 0 1 ...
 $ Sex             : num  0 0 1 0 0 0 0 0 0 1 ...
 $ AgeCategory     : num  8 13 10 12 5 12 11 13 13 10 ...
 $ Race            : num  1 1 1 1 1 2 1 1 1 1 ...
 $ Diabetic        : num  1 2 1 2 2 2 2 1 3 2 ...
 $ PhysicalActivity: num  1 1 1 0 1 0 1 0 0 1 ...
 $ GenHealth       : num  1 1 2 3 1 2 2 3 2 3 ...
 $ SleepTime       : num  5 7 8 6 8 12 4 9 5 10 ...
 $ Asthma          : num  1 0 1 0 0 0 1 1 0 0 ...
 $ KidneyDisease   : num  0 0 0 0 0 0 0 0 1 0 ...
 $ SkinCancer      : num  1 0 0 1 0 0 1 0 0 0 ...
```

Figure 9: Checking the data types of variables

“str()” function is used again to check the data types of variables to ensure all the data have converted into numeric values.

3.3 Data Sampling

```
> set.seed(500)
> hd_new<-sample(1:nrow(hd),500)
> hd_new<-hd[hd_new,]
```

Figure 10: Sampling the data

Due to the massive dataset which contains 300k of observations, data sampling without replacement is performed. The data is crunched to derive useful information out of it. Data

sampling is simply meant that a subset of data. From the figure 10 show above, “set.seed()” function is used to get the same rows of data when executing the code multiples time. After that, “sample()” function is used to generate 500 random rows as a sample from the “Personal Key Indicators of Heart Disease” dataset. The sampled data has been stored in “hd_new”.

3.4 Feature Selection

```
> #One way anova
> BMI.aov<-aov(HeartDisease~BMI, data=hd_new)
> summary(BMI.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
BMI           1    0.3 0.30090   4.106 0.0433 *
Residuals   498   36.5 0.07329
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Sex.aov<-aov(HeartDisease~Sex, data=hd_new)
> summary(Sex.aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
Sex           1   0.85  0.8522   11.81 0.00064 ***
Residuals   498  35.95  0.0722
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #Two way anova
> two.way<-aov(HeartDisease~BMI+AgeCategory,data=hd_new)
> summary(two.way)
             Df Sum Sq Mean Sq F value  Pr(>F)
BMI           1   0.30  0.3009   4.378  0.0369 *
AgeCategory   1   2.34  2.3402  34.050 9.7e-09 ***
Residuals   497  34.16  0.0687
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: One way ANOVA and Two way ANOVA

ANOVA tests whether any of the group means are different from the overall mean of the data by checking the variance of each individual group against the overall variance of the data. ANOVA is performed by using “aov()” function, this will calculate the test statistic for ANOVA and determine whether the levels of the independent variable generate significant variation among the groups. After running the model, “summary()” function is used to print the summary of the model.

The Pr(>F) column is the p-value of the F-statistic. In One way ANOVA, the p-value of the BMI is 0.433($p>0.001$) whereas the p-value of the Sex is 0.00064($P<0.001$). Thus it appears that the sex has a real impact on the heart disease than BMI while applying One Way ANOVA. In Two way Anova, the p-value of Age Category is lower than BMI, thus it also shows that the Age Category has a real impact on the heart disease than BMI. The variables will be statistically significant when the p-value is lower than 0.001.

```
> #Multiple variables
> n<-names(hd_new[,2:18])
> form<-as.formula(paste("HeartDisease~",paste(n[!n %in% "use"],collapse="+")))
> anova<-aov(form,data=hd)
> summary(anova)
                    Df Sum Sq Mean Sq   F value  Pr(>F)
BMI                  1     55    55.4   772.826 < 2e-16 ***
Smoking              1    267   267.1  3727.099 < 2e-16 ***
AlcoholDrinking      1     52    52.4   730.937 < 2e-16 ***
Stroke               1    866   866.4 12089.107 < 2e-16 ***
PhysicalHealth       1    399   399.3  5571.948 < 2e-16 ***
MentalHealth         1     22    22.5   313.767 < 2e-16 ***
DiffWalking          1    296   296.4  4135.753 < 2e-16 ***
Sex                  1    148   147.9  2064.369 < 2e-16 ***
AgeCategory          1    774   773.6 10793.686 < 2e-16 ***
Race                 1      0     0.1     1.022  0.3121
Diabetic             1    119   119.2  1662.687 < 2e-16 ***
PhysicalActivity     1      4     3.6    50.663 1.1e-12 ***
GenHealth            1      1     0.6     7.809  0.0052 **
SleepTime            1      0     0.0     0.046  0.8298
Asthma               1     18    18.4   256.050 < 2e-16 ***
KidneyDisease        1    140   140.1  1955.153 < 2e-16 ***
SkinCancer           1     13    13.1   183.271 < 2e-16 ***
Residuals       301699  21622     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova1<-aov(HeartDisease~BMI+Smoking+AlcoholDrinking+Stroke+PhysicalHealth+MentalHealth+
+               DiffWalking+Sex+AgeCategory+Diabetic+PhysicalActivity+Asthma+KidneyDisease+
+               SkinCancer,data=hd_new)
> summary(anova1)
                   Df Sum Sq Mean Sq F value   Pr(>F)
BMI                 1  0.301  0.3009   5.034 0.025305 *
Smoking             1  0.636  0.6360  10.640 0.001185 **
AlcoholDrinking     1  0.007  0.0074   0.124 0.724494
Stroke              1  1.731  1.7312  28.963 1.15e-07 ***
PhysicalHealth      1  0.879  0.8786  14.699 0.000143 ***
MentalHealth        1  0.006  0.0058   0.097 0.755463
DiffWalking         1  1.446  1.4462  24.195 1.19e-06 ***
Sex                 1  0.858  0.8582  14.358 0.000170 ***
AgeCategory         1  0.678  0.6778  11.339 0.000819 ***
Diabetic            1  0.443  0.4432   7.414 0.006704 **
PhysicalActivity    1  0.376  0.3756   6.284 0.012512 *
Asthma              1  0.059  0.0591   0.989 0.320521
KidneyDisease       1  0.113  0.1131   1.892 0.169592
SkinCancer          1  0.277  0.2767   4.629 0.031932 *
Residuals         485 28.990  0.0598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: ANOVA with multiple variables

From the figure 12 above, multiples variables are adding into the model to get the variables which are more important in predicting heart disease. From the summary of “anova” and “anova1” and by observing the p-value, it is clear to see that BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWaling, Sex, AgeCategory, Diabetic, PhysicalActivity, Asthma, KidneyDisease and SkinCancer are the important variables in predicting heart disease. Other variables probably not adding much information to the model.

```
> #Find the best-fit model
> model.set<-list(BMI.aov,Sex.aov,two.way,anova,anova1)
> model.names<-c("BMI.aov","Sex.aov","two.way","anova","anova1")
> aictab(model.set, modnames = model.names)

Model selection based on AICc:

         K     AICc Delta_AICc AICcWt Cum.Wt        LL
anova1  16    28.24       0.00      1      1      2.44
two.way  4    85.23      56.99      0      1    -38.57
Sex.aov  3   108.72      80.48      0      1    -51.33
BMI.aov  3   116.33      88.09      0      1    -55.14
anova   19 61012.99   60984.75      0      1 -30487.49
```

Figure 13: Best-fit model

To validate BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWaling, Sex, AgeCategory, Diabetic, PhysicalActivity, Asthma, KidneyDisease and SkinCancer are the important variables in predicting heart disease, “aictab()” function is used to perform model selection based on AICc(Akaike information criterion). AIC calculates the information value of each model by balancing the variation explained against the number of parameters used. The lowest AIC value means more information explained in the model. The model with the lowest AIC score is the best fit for the data. Based on the result from figure 13, it appears anova1 is the best fit. The anova1 model has the lowest AIC value, and 100% of the AIC weight, which means that it explains 100% of the total variation in the dependent variable that can be explained by the full set of models. Thus, it proved that BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWaling, Sex, AgeCategory, Diabetic, PhysicalActivity, Asthma, KidneyDisease and SkinCancer are the important variables in predicting heart disease from the dataset.

4.0 Data Mining

Data mining is the process to identify trends and patterns and establish relationships by sorting the large data sets. It can help to generate new opportunities or solve business problems through the analysis of the data. Enterprises can predict future trends and make more-informed business decisions by using data mining techniques and tools. Data mining uses advanced analytics techniques to find useful information in large data sets, which is the key part of data analytics overall and one of the core task in data science.

4.1 Data Visualization

Data visualization helps to see and understand trends and patterns in data by using the charts and graphs. Graphs are plotted to show the insight of each variable in this project. The plotted graphs are separated into 2 categories which are Numerical Variable and Categorical Variable as shown below.

4.1.1 Numerical

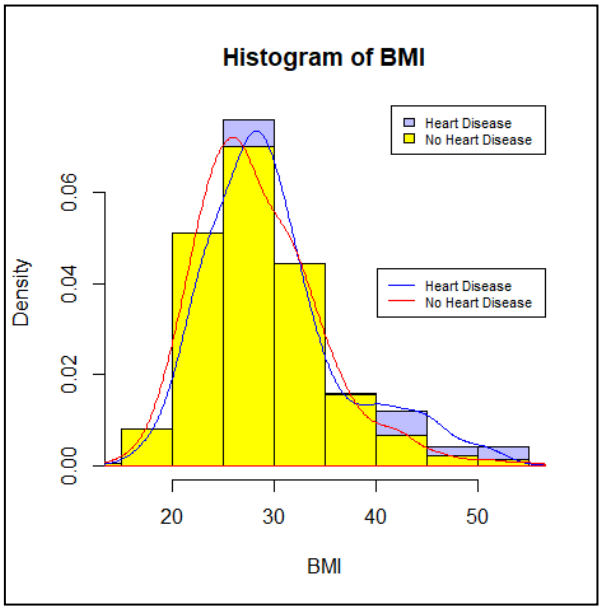

Figure 14: Heart Disease by BMI

Body Mass Index (BMI) is a value derived from the height and the mass of a person. In the figure above, the top point of the blue line (has heart disease) has a higher BMI than the red line (no heart disease). This shows that the higher the BMI, the higher the probability of getting heart disease.

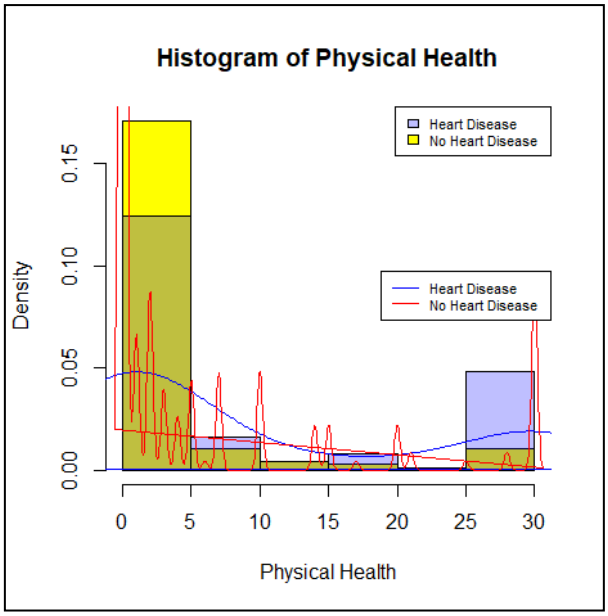

Figure 15: Heart Disease by Physical Health

Physical Health variable in this data set shows the days during the past 30 days was the personal physical health not good. From the figure above, the density of the blue bar (has heart disease) is almost 4 times higher than the yellow bar (no heart disease). This shows that the more days of the physical health is not good, the more chance to get heart disease.

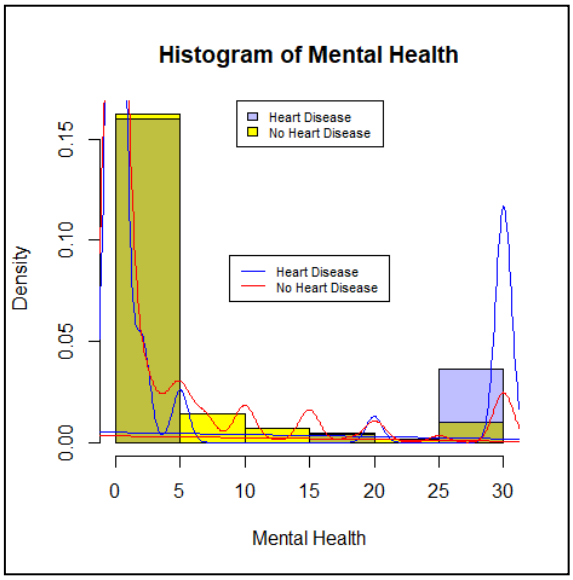

Figure 16: Heart Disease by Mental Health

Mental Health variable in this data set shows the days during the past 30 days was the personal mental health not good. From the figure above, the density of the blue bar (has heart disease) is almost 3 times higher than the yellow bar (no heart disease). This shows that the more day of the mental health is not good, the higher chance to get heart disease.

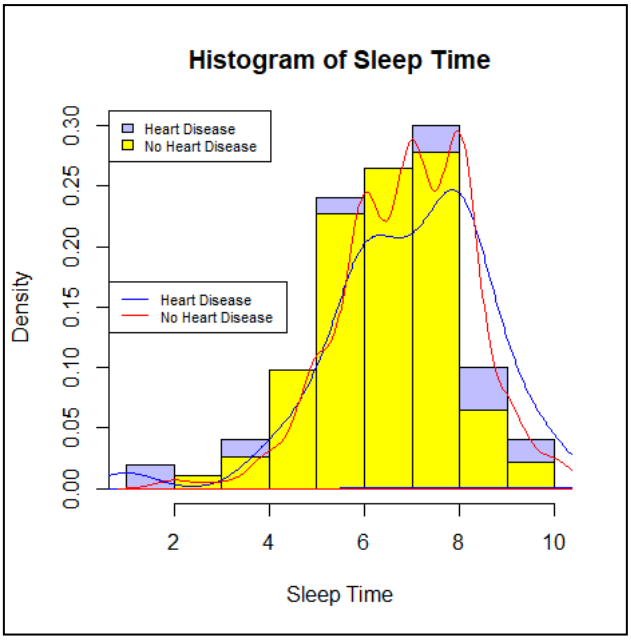

Figure 17: Heart Disease by Sleep Time

Sleep Time variable is the hours of sleeping in 24-hour period. From the figure above, the sleep time of heart disease person and the sleep time of no heart disease person are almost the same. This shows that the chance of getting heart disease is less affected by sleep time.

4.1.2 Categorical

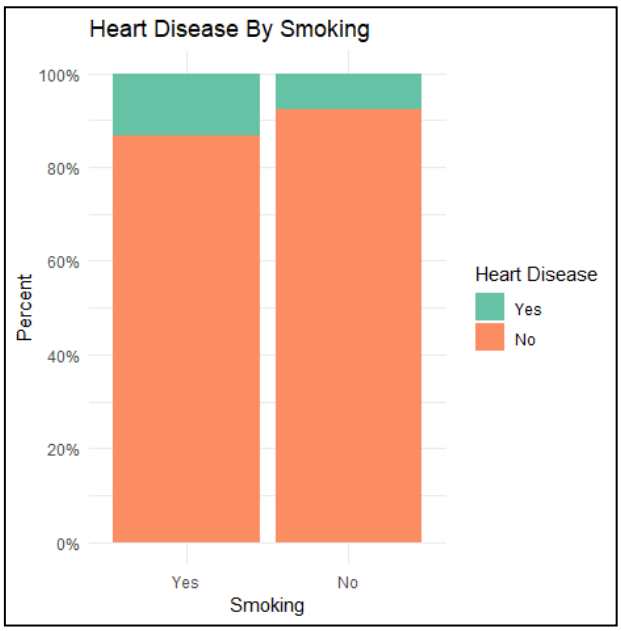

Figure 18: Heart Disease by Smoking

From the figure above, the percentage of getting heart disease with smoking is slightly higher than the percentage of getting heart disease with no smoking. This shows that a smoking person will has a higher risk to get heart disease compared to a non-smoking person.

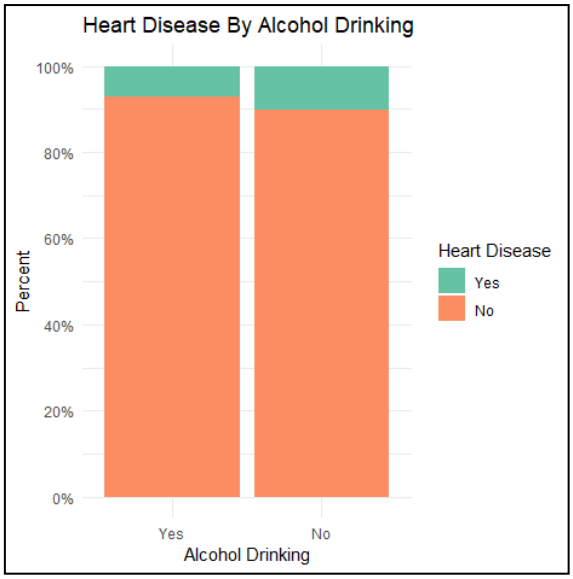

Figure 19: Heart Disease by Alcohol Drinking

From the figure above, the percentage of getting heart disease with alcohol drinking and the percentage of getting heart disease with no alcohol drinking are almost the same. This shows that a person getting heart disease has not much relation.

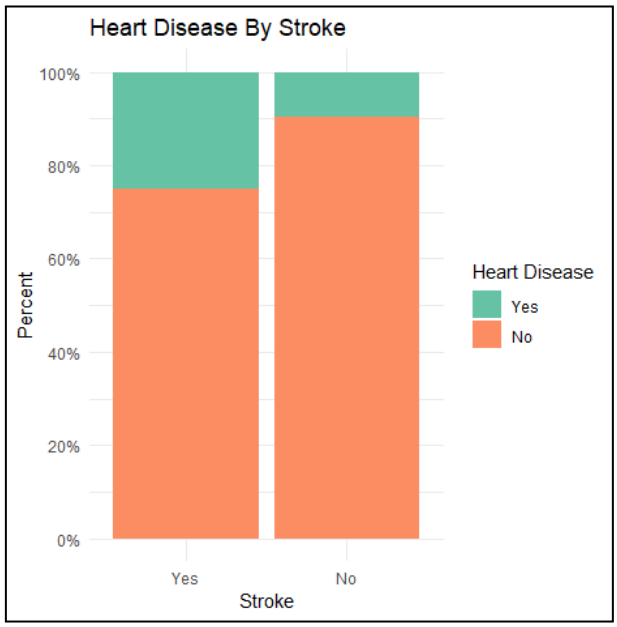

Figure 20: Heart Disease by Stroke

From the figure above, a person with stroke history has 2 times higher of percentage of getting heart disease compared to the person with no stroke history. This shows that the stroke has much impact on heart disease.

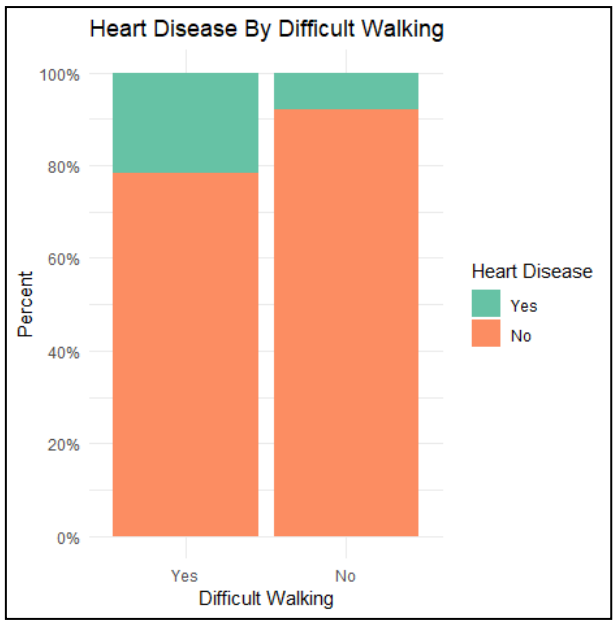

Figure 21: Heart Disease by Difficult Walking

From the figure above, a person with difficulty on walking has approximately 3 times higher of percentage of getting heart disease compared to the person with no difficulty on walking. This shows that the percentage of getting heart disease will be higher on the person who has difficulty on walking.

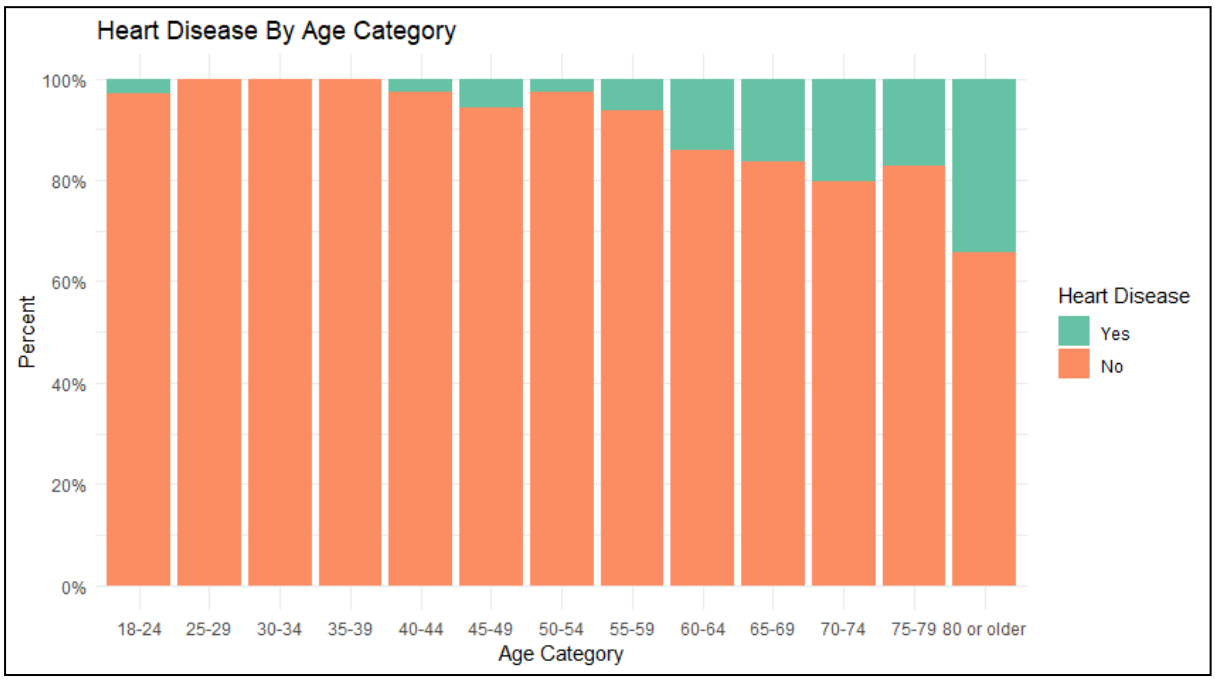

Figure 22: Heart Disease by Age Category

From the figure above, the percentage of getting heart disease becomes higher as the age category becomes higher. This shows that an older person will have a higher chance of getting heart disease.

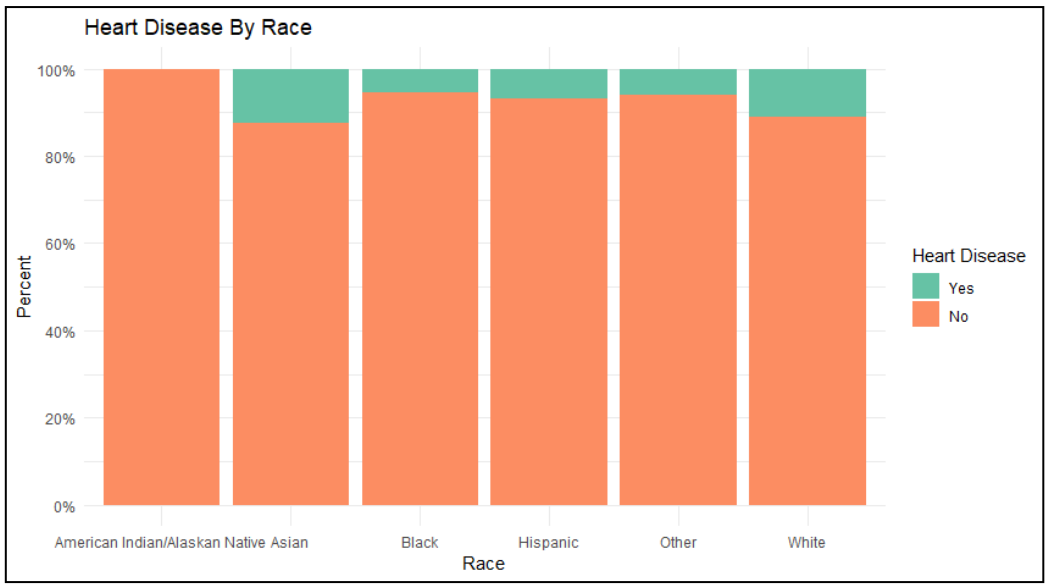

Figure 23: Heart Disease by Race

From the figure above, the percentage of getting heart disease of all races are not much different. This shows that no matter the race are, there will be a chance of getting heart disease.

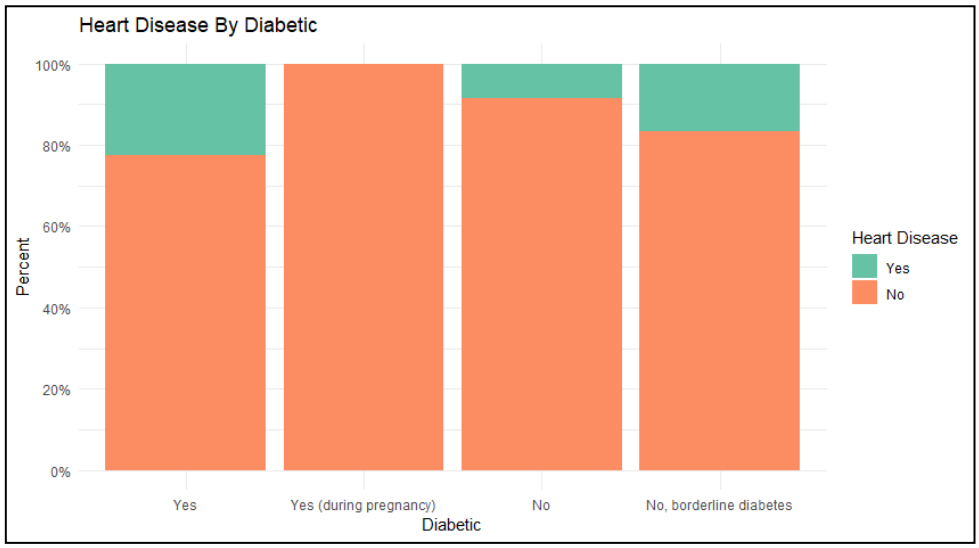

Figure 24: Heart Disease by Diabetic

From the figure above, the percentage of a non-pregnancy diabetic patient getting heart disease is highest compared to the other categories. For the category with no diabetic, the percentage of getting heart disease is higher for a person who is borderline diabetes. This shows that diabetic will increase the risk of getting heart disease.

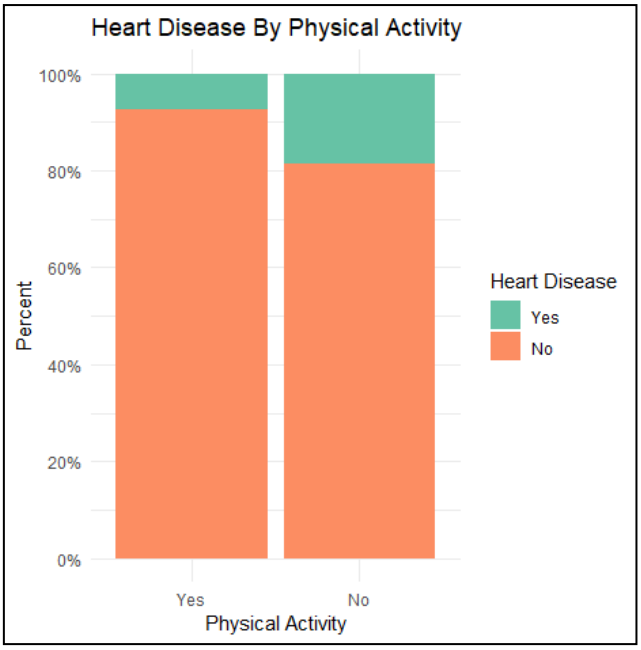

Figure 25: Heart Disease by Physical Activity

Physical Activity variable is showing whether a person doing physical activity or exercise during the past 30 days other than their regular job or not. From the figure above, the percentage of getting heart disease is 2 times higher for the person without doing the physical activity. This shows that doing the physical activity like running or sporting will decrease the risk of getting heart disease.

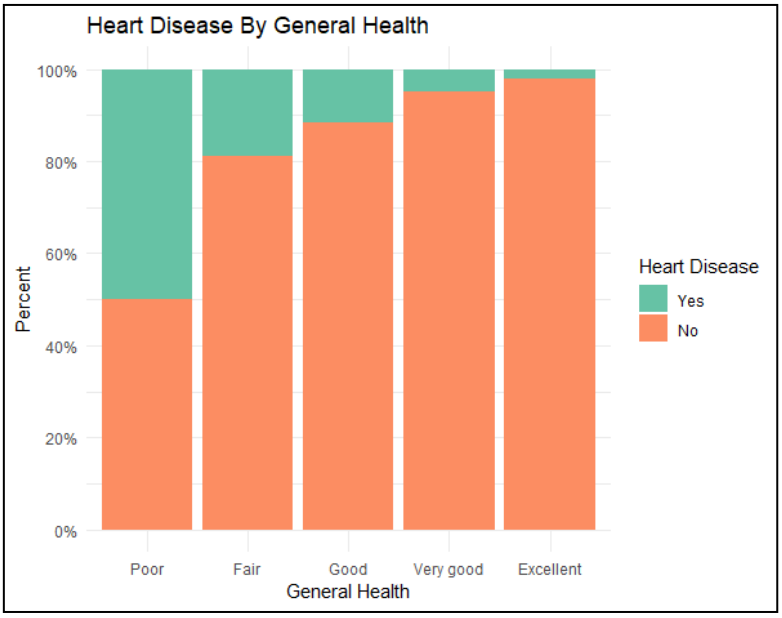

Figure 26: Heart Disease by General Health

From the figure above, the percentage of getting heart disease is almost 50% with the poor general health. From the poor category to excellent category, the percentage of getting heart disease is decreasing significantly. This shows that maintain an excellent general health is good to get away from heart disease.

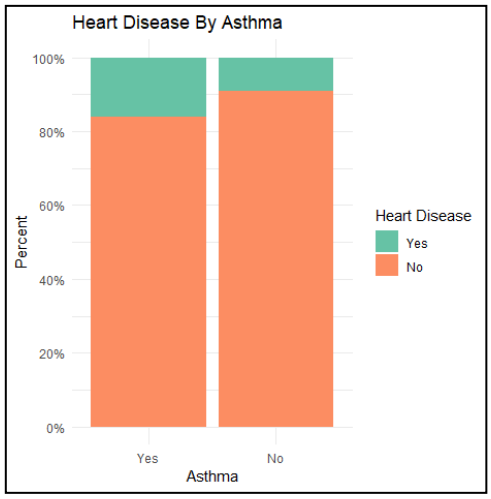

Figure 27: Heart Disease by Asthma

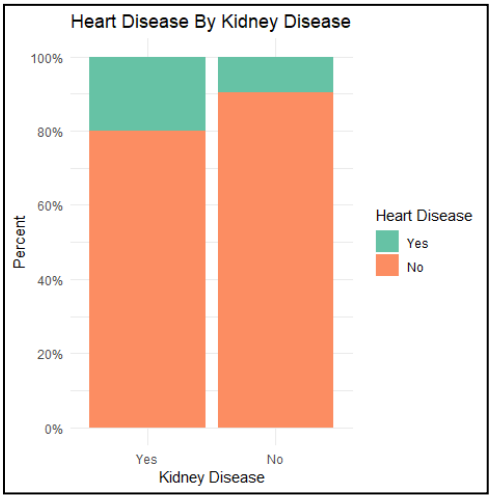

Figure 28: Heart Disease by Kidney Disease

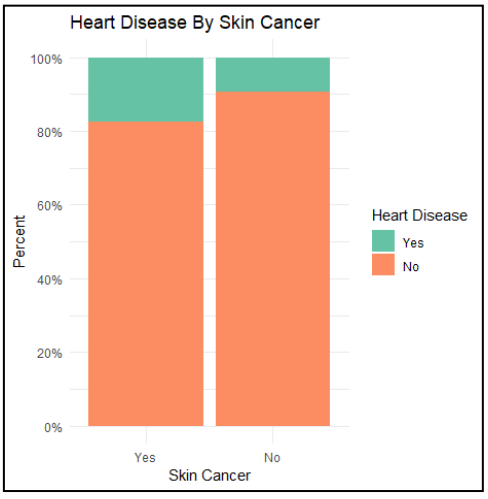

Figure 29: Heart Disease by Skin Cancer

From the figures above, the percentage of an asthma patient getting heart disease is slightly higher than the non-asthma person. The percentage of a kidney disease patient getting heart disease is 2 times higher than the person without kidney disease. Besides, the percentage of a skin cancer patient getting heart disease is higher than the person with no skin cancer. These 3 graphs show that the other disease will affect the chance of getting heart disease.

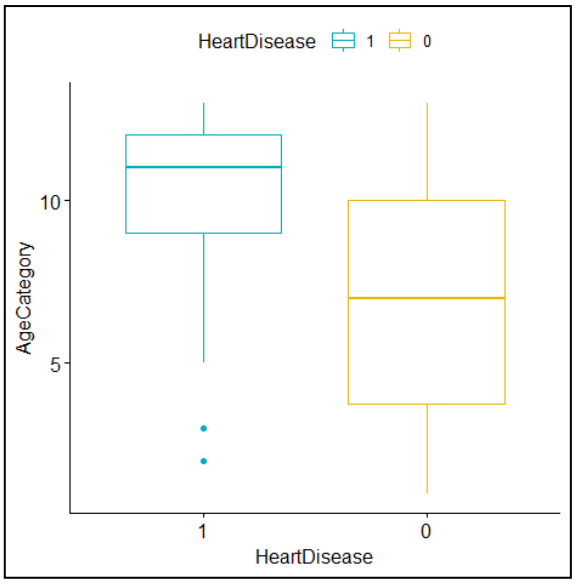

Figure 30: Boxplot of Heart Disease by Age Category

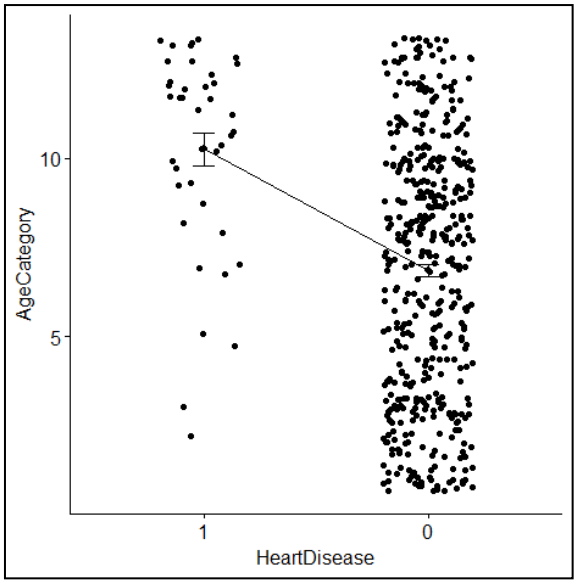

Figure 31: Mean Plot of Heart Disease by Age Category

From the figures above, the mean of the person who has heart disease is higher than the person with no heart disease. This can be said that the mean age of getting heart disease is higher and the mean age of no heart disease is lower. Therefore, the higher the age, the higher the risk of getting heart disease.

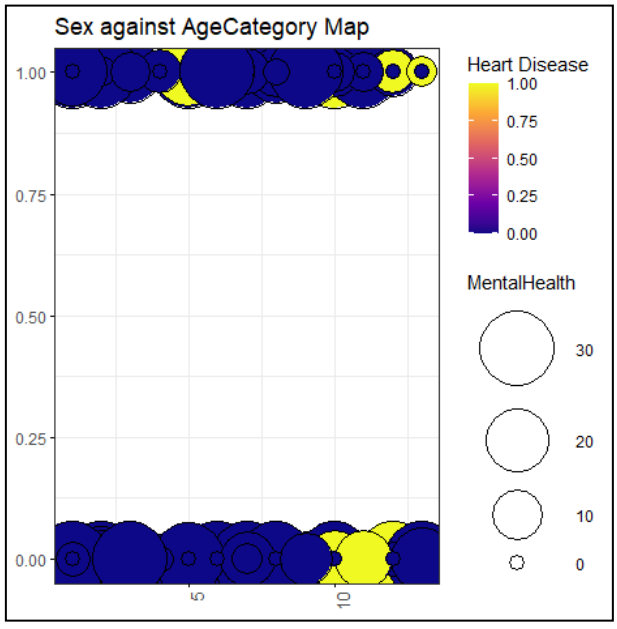

Figure 32: Plot Sex against Age Category map with the condition of heart disease and Mental Health

From the figure above, the upper circles are male, and the lower circles are female. The yellow circles are the person with heart disease and blue circles are without heart disease. The x-axis is Age Category. By observing the pattern of this map, the female category with higher age, higher mental health will have the higher chance to get heart disease.

4.2 Artificial Neural Network

Artificial Neural Network (ANN) is inspired by the human brain, it is acted like the biological neurons send signals to each other. Artificial neural networks have many layers with many interconnected nodes (neuron). It has 3 main layers which are input layer, hidden layer, and output layer. The neural network can be single layer neural network or multi-layer neural network, based on the number of neurons in the hidden layer. The artificial neural network algorithm is used in the heart disease prediction in heart disease data set in this project.

```
> hd_new[2:18]<-scale(hd_new[2:18]) #Set the input variables into same scale
> summary(hd_new) #All variables have a mean equal to 0
  HeartDisease          BMI              Smoking        AlcoholDrinking       Stroke        PhysicalHealth      MentalHealth
 Min.   :0.00   Min.   :-1.8235   Min.   :-0.8536   Min.   :-0.2783   Min.   :-0.1874   Min.   :-0.4743   Min.   :-0.51427
 1st Qu.:0.00   1st Qu.:-0.6946   1st Qu.:-0.8536   1st Qu.:-0.2783   1st Qu.:-0.1874   1st Qu.:-0.4743   1st Qu.:-0.51427
 Median :0.00   Median :-0.1594   Median :-0.8536   Median :-0.2783   Median :-0.1874   Median :-0.4743   Median :-0.51427
 Mean   :0.08   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.:0.00   3rd Qu.: 0.4541   3rd Qu.: 1.1692   3rd Qu.:-0.2783   3rd Qu.:-0.1874   3rd Qu.:-0.1397   3rd Qu.: 0.09304
 Max.   :1.00   Max.   : 6.2732   Max.   : 1.1692   Max.   : 3.5865   Max.   : 5.3249   Max.   : 2.8724   Max.   : 3.12959
  DiffWalking          Sex            AgeCategory          Race            Diabetic       PhysicalActivity    GenHealth
 Min.   :-0.423   Min.   :-0.9831   Min.   :-1.6808   Min.   :-0.4867   Min.   :-2.4139   Min.   :-1.6426   Min.   :-1.0671
 1st Qu.:-0.423   1st Qu.:-0.9831   1st Qu.:-0.8574   1st Qu.:-0.4867   1st Qu.: 0.1986   1st Qu.:-1.6426   1st Qu.:-1.0671
 Median :-0.423   Median :-0.9831   Median : 0.2404   Median :-0.4867   Median : 0.1986   Median : 0.6076   Median : 0.2803
 Mean   : 0.000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.:-0.423   3rd Qu.: 1.0151   3rd Qu.: 0.7893   3rd Qu.:-0.4867   3rd Qu.: 0.1986   3rd Qu.: 0.6076   3rd Qu.: 0.2803
 Max.   : 2.360   Max.   : 1.0151   Max.   : 1.6127   Max.   : 2.6174   Max.   : 5.4235   Max.   : 0.6076   Max.   : 1.6276
   SleepTime           Asthma         KidneyDisease       SkinCancer
 Min.   :-4.10624   Min.   :-0.3997   Min.   :-0.2243   Min.   :-0.3293
 1st Qu.:-0.72272   1st Qu.:-0.3997   1st Qu.:-0.2243   1st Qu.:-0.3293
 Median :-0.04602   Median :-0.3997   Median :-0.2243   Median :-0.3293
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.63069   3rd Qu.:-0.3997   3rd Qu.:-0.2243   3rd Qu.:-0.3293
 Max.   : 6.04433   Max.   : 2.4968   Max.   : 4.4490   Max.   : 3.0308
>
> #Divide data into train and test datasets
> set.seed(12345)
> data<-sample(2,nrow(hd_new),replace=TRUE,prob=c(0.7,0.3))
> train.data<-hd_new[data==1,]
> test.data<-hd_new[data==2,]
```

Figure 33: Train dataset and Test dataset

From the figure above, the input variables from the dataset are set into same scale. Scaling input variables is a critical step in using neural network models to make it simple for a model to learn and understand the problem. Next, the dataset is split into train dataset and test dataset using “sample” function. Train dataset contains 70% of the data and test dataset contains 30% of the data.

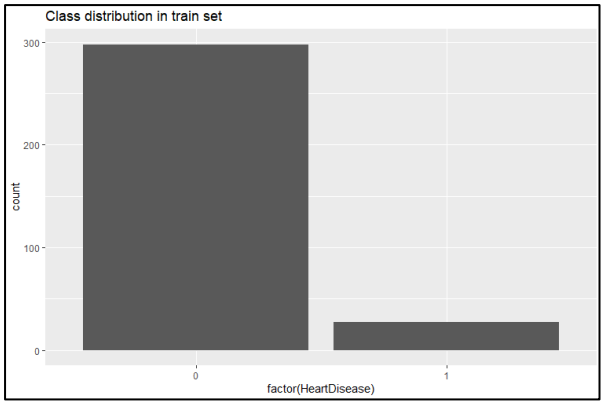

Figure 34: Class distribution in train dataset

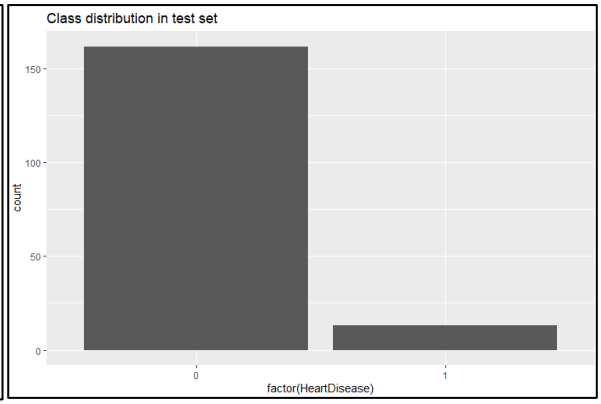

Figure 35: Class distribution in test dataset

From the figure above, the bar charts have shown the distribution of the dataset. They showed that the train dataset and test dataset have the same pattern in the bar chart. Hence, the distribution of the data is average and there won't be much difference in the number of heart disease patient that will cause a bad result after this.

4.2.1 Artificial Neural Network Model

The library package "neuralnet" is used in this model to call the function inside the package. The neuralnet does not accommodate the use of y~. to specify all the other variables in the data as inputs. The as.formula() function is used to solve this problem as shown in below.

```
###################Create first neural network model(1 hidden layer, 4 nodes)###
set.seed(12345)
#Implements neural network on training data
n<-names(hd_new[,2:18])
form<-as.formula(paste("HeartDisease~",paste(n[!n %in% "use"],collapse="+")))
nn1<-neuralnet(form,
               data=train.data,
               hidden=4, err.fct="ce", linear.output=FALSE)
```

Figure 36: Neural Network Algorithm in R

From the figure above, the neural network model is trained by 5 arguments which are "*form*" (used to input the variables name), "*data=train.data*" (the train data that used to train the model), "*hidden=4*" (4 nodes in 1 hidden layer is set), "*err.fct=ce*" (to deal with binary outcomes), "linear.output=FALSE" (not ignore act.fct).

```
> summary(nn1)
                   Length Class      Mode
call                  6   -none-     call
response            325   -none-     numeric
covariate          5525   -none-     numeric
model.list            2   -none-     list
err.fct               1   -none-     function
act.fct               1   -none-     function
linear.output         1   -none-     logical
data                 18   data.frame list
exclude               0   -none-     NULL
net.result            1   -none-     list
weights               1   -none-     list
generalized.weights   1   -none-     list
startweights          1   -none-     list
result.matrix        80   -none-     numeric
> nn1$response[1:20]
 [1] 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
> nn1$net.result[[1]][1:20]
 [1] 1.742829e-06 2.665011e-13 3.281190e-02 3.458589e-04 7.693233e-04 7.673956e-04 7.866909e-11 3.475480e-09 7.997085e-11 7.888747e-11 3.324050e-02
[12] 9.994156e-01 1.807258e-02 1.595893e-13 1.152704e-03 3.475491e-09 3.292459e-02 1.571330e-13 1.513041e-07 7.957941e-11
> nn1$result.matrix
                                    [,1]
error                       8.346733e+00
reached.threshold           9.099999e-03
steps                       1.804000e+03
Intercept.to.1layhid1      -9.965099e+00
BMI.to.1layhid1             2.699846e+01
Smoking.to.1layhid1         2.101047e+01
AlcoholDrinking.to.1layhid1 -5.186208e+00
Stroke.to.1layhid1          4.799915e+01
PhysicalHealth.to.1layhid1  5.187575e+00
MentalHealth.to.1layhid1   -1.153877e+00
DiffWalking.to.1layhid1     2.474626e+01
Sex.to.1layhid1             4.165384e+01
AgeCategory.to.1layhid1    -8.033175e+01
Race.to.1layhid1           -4.937325e+01
Diabetic.to.1layhid1        1.118536e+01
```

Figure 37: Result of Neural Network Algorithm in R

The summary of the result of neural network model is shown in the figure above.

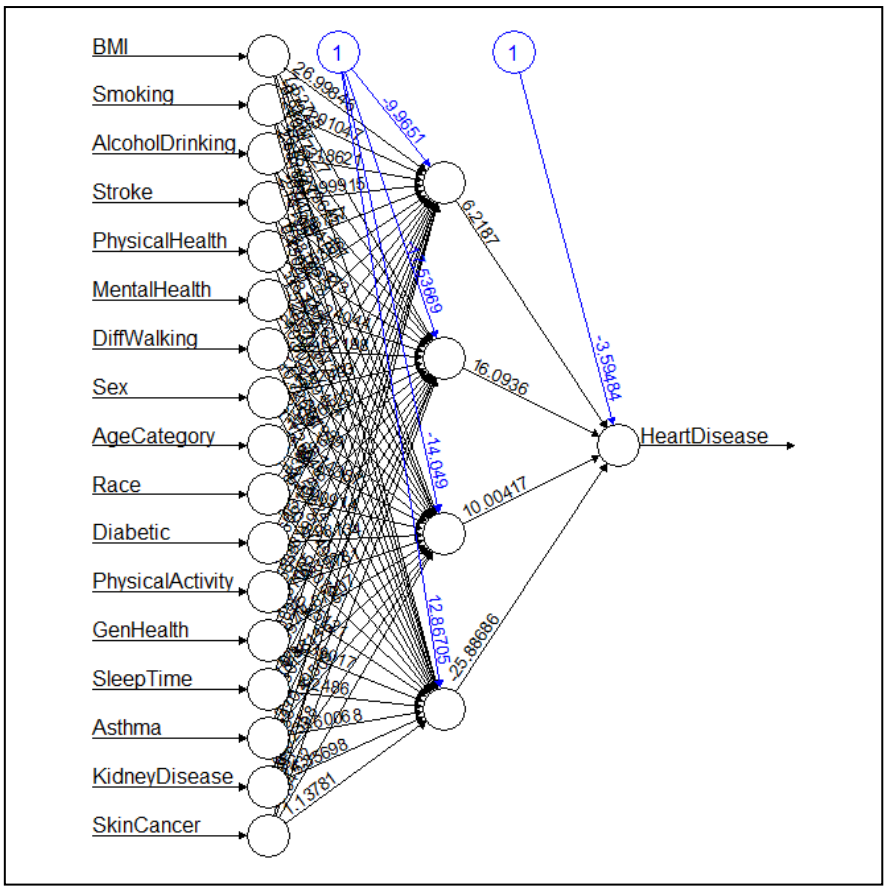

Figure 38: Plot of Neural Network Model

From the figure above, the plot of the neural network model with 1 hidden layer and 4 hidden nodes is shown. The second node in the hidden layer has the highest weight with 16.09. The last node in the hidden layer has the negative value of weight which is -25.88. This shows that these two node will affect the output most but the other two node will also have some effect to the output because it is not near to zero.

```
> #Compute the predicted values for training set
> trainPred1<-neuralnet::compute(nn1,nn1$covariate)$net.result
> trainPred1<-apply(trainPred1,c(1),round)
> #Create confusion matrix on train data
> confusionMatrix(table(trainPred1,train.data$HeartDisease,dnn=c("Predicted","Actual")))
Confusion Matrix and Statistics

         Actual
Predicted   0   1
        0 297   1
        1   1  26

               Accuracy : 0.9938
                 95% CI : (0.9779, 0.9993)
    No Information Rate : 0.9169
    P-Value [Acc > NIR] : 2.651e-10

                  Kappa : 0.9596

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9966
            Specificity : 0.9630
         Pos Pred Value : 0.9966
         Neg Pred Value : 0.9630
             Prevalence : 0.9169
         Detection Rate : 0.9138
   Detection Prevalence : 0.9169
      Balanced Accuracy : 0.9798

       'Positive' Class : 0
```

Figure 39: Result of Training Set Prediction

By using the compute() and apply() functions, the prediction is carried out for the training data set. Next, confusionMatrix() function is used to calculate the accuracy of the model which is 99.38% for the training set.

```
> #Compute the predicted values for testing set
> testPred1<-neuralnet::compute(nn1,test.data[,1:18])$net.result
> testPred1<-apply(testPred1,c(1),round)
> #Create confusion matrix on test data
> confusionMatrix(table(testPred1,test.data$HeartDisease,dnn=c("Predicted","Actual")))
Confusion Matrix and Statistics

         Actual
Predicted   0   1
        0 149   8
        1  13   5

               Accuracy : 0.88
                 95% CI : (0.8224, 0.9242)
    No Information Rate : 0.9257
    P-Value [Acc > NIR] : 0.9891

                  Kappa : 0.2586

 Mcnemar's Test P-Value : 0.3827

            Sensitivity : 0.9198
            Specificity : 0.3846
         Pos Pred Value : 0.9490
         Neg Pred Value : 0.2778
             Prevalence : 0.9257
         Detection Rate : 0.8514
   Detection Prevalence : 0.8971
      Balanced Accuracy : 0.6522

       'Positive' Class : 0
```

Figure 40: Result of Testing Set Prediction

After completing the prediction of the test set, the confusion matrix function is used again to calculate the accuracy of the test set. The accuracy of the test set is 88% as shown in figure above.

5.0 Evaluation

In any model construction, evaluation and validation are critical aspect to validate results after data mining process had been done. Data mining method that is used in this project is classification and the algorithm used is neural network algorithm. In this session, the parameter setting of neural network is being fine tuned in order to achieve the high quality of the model. The process will be showed in the following part.

Besides, neural network model validation is usually based on some specified network performance measure of data that was not utilised in model development (a "test set"), despite the fact that there is no well-formulated or theoretical methodology for doing so. This performance measure is frequently used to assess the superiority of network architecture, learning algorithms, or neural network applications, in addition to trained network validation.

There are three frequently used performance measures which are Mean Absolute Error (MAE), Mean Squared Error (MSE), and percent good classification.

In this project, confusion matrix and Mean Squared Error (MSE) are used to perform validation by using R programming. The evaluation and validation result will be discussed in the following part.

5.1 Fine tune process

Fine-tuning is a technique for putting transfer learning into practise. Fine-tuning is a procedure that takes a model that has already been trained for one task and tunes or tweaks it to perform a second task that is similar to the first. The reason needed to perform fine tuning is because the neural network that has already been designed and trained is allowed to take advantage of what the model has already learned without having to develop it from scratch.

When building a model from scratch, approaches through trial-and-error must be tried. There are a few ways to fine tune the parameter in neural network by changing the number of layers, types of layers, number of nodes in each layer, etc. In this project, number of layers and number of nodes in each layer are being fine-tuned to achieve high quality of the model.

Remark: model1 -> neural network model with 1 hidden layer and 4 nodes
model2 -> neural network model with 2 hidden layer and 6 nodes
model3 -> neural network model with 3 hidden layer and 6 nodes

As mentioned before, a neural network model is performed with 1 hidden layer and 4 nodes. The coding is used to redo the neural network process by different hidden layer and nodes.

```
###################Create first neural network model(1 hidden layer, 4 nodes)#############################
set.seed(12345)
#Implements neural network on training data
n<-names(hd_new[,2:18])
form<-as.formula(paste("HeartDisease~",paste(n[!n %in% "use"],collapse="+")))
nn1<-neuralnet(form,
               data=train.data,
               hidden=4, err.fct="ce", linear.output=FALSE)
summary(nn1)
nn1$response[1:20]
nn1$net.result[[1]][1:20]
nn1$result.matrix
plot(nn1)
```

Figure 41: Neural Network Algorithm in R for model1

Here, a neural network model with 2 hidden layer and 6 nodes and a neural network model with 3 hidden layer and 6 nodes are built in order to carry out the fine tuning process.

```
###########Create second neural network model(2 hidden layer, 6 nodes)###################################
set.seed(12345)
#Implements neural network on training data
n<-names(hd_new[,2:18])
form<-as.formula(paste("HeartDisease~",paste(n[!n %in% "use"],collapse="+")))
nn2<-neuralnet(form,
               data=train.data,
               hidden=c(6,2), err.fct="ce", linear.output=FALSE)
nn2$net.result[[1]][1:20]
nn2$result.matrix
plot(nn2)
```

```
###########Create third neural network model(3 hidden layer, 6 nodes)###################################
set.seed(12345)
#Implements neural network on training data
n<-names(hd_new[,2:18])
form<-as.formula(paste("HeartDisease~",paste(n[!n %in% "use"],collapse="+")))
nn3<-neuralnet(form,
               data=train.data,
               hidden=c(6,3), err.fct="ce", linear.output=FALSE)
nn3$net.result[[1]][1:20]
nn3$result.matrix
plot(nn3)
```

Figure 42: Neural Network Algorithm in R for model2 and model3

From the figures below, the plots of neural network models are shown. The strength of the connection between units is represented by a weight. If the weight from node 1 to node 2 is larger, this indicates that neuron 1 has a stronger influence on neuron 2. If the weights are close to zero, adjusting this input will have little effect on the output. Negative weights indicate that increasing this input will result in a lower output. A weight determines how much of an impact the input has on the output.

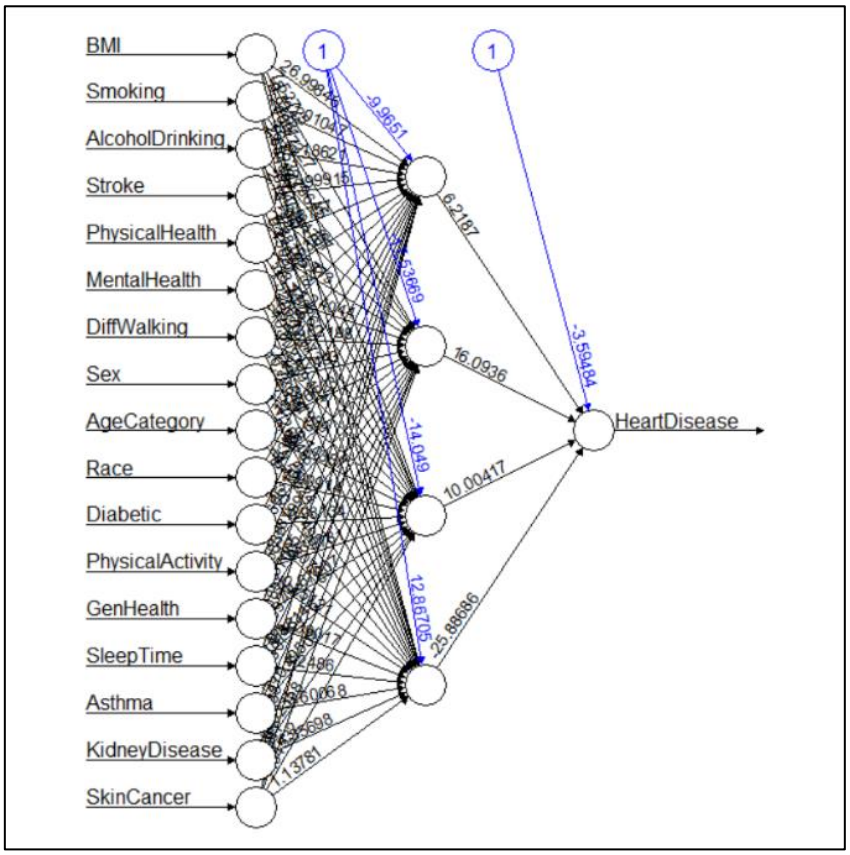

Figure 43: Plot of Neural Network Model 1

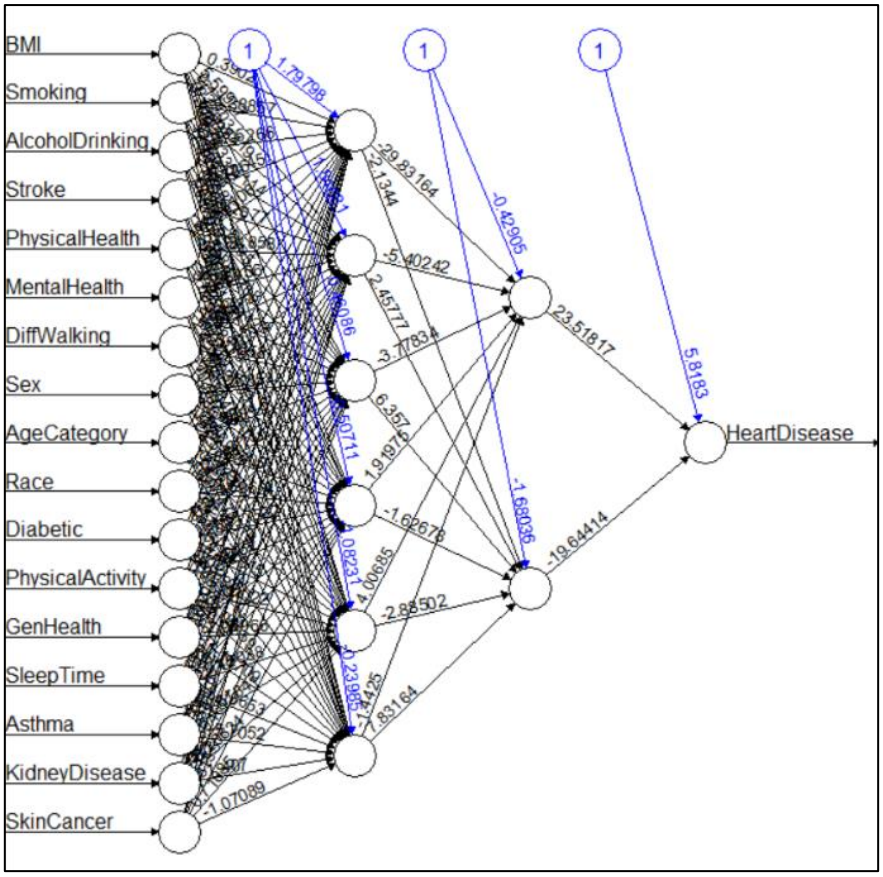

Figure 44: Plot of Neural Network Model 2

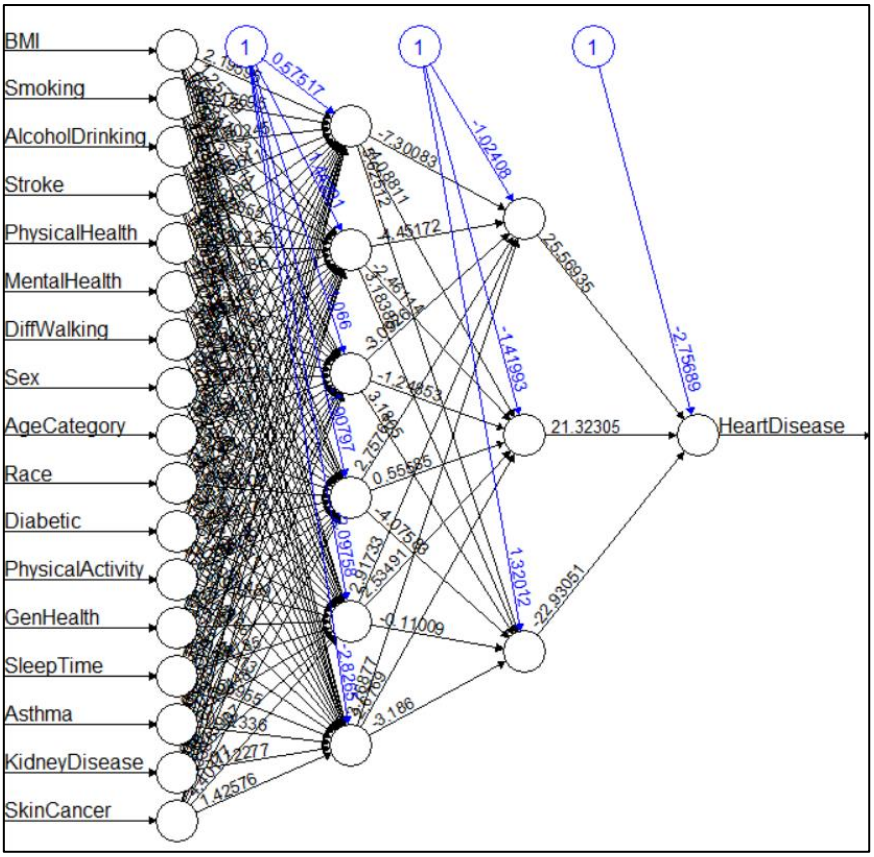

Figure 45: Plot of Neural Network Model 3

5.1.1 Confusion Matrix

```
Confusion Matrix and Statistics

          Actual
Predicted   0   1
        0 297   1
        1   1  26

               Accuracy : 0.9938
                 95% CI : (0.9779, 0.9993)
    No Information Rate : 0.9169
    P-Value [Acc > NIR] : 2.651e-10

                  Kappa : 0.9596

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9966
            Specificity : 0.9630
         Pos Pred Value : 0.9966
         Neg Pred Value : 0.9630
             Prevalence : 0.9169
         Detection Rate : 0.9138
   Detection Prevalence : 0.9169
      Balanced Accuracy : 0.9798

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Actual
Predicted   0   1
        0 149   8
        1  13   5

               Accuracy : 0.88
                 95% CI : (0.8224, 0.9242)
    No Information Rate : 0.9257
    P-Value [Acc > NIR] : 0.9891

                  Kappa : 0.2586

 Mcnemar's Test P-Value : 0.3827

            Sensitivity : 0.9198
            Specificity : 0.3846
         Pos Pred Value : 0.9490
         Neg Pred Value : 0.2778
             Prevalence : 0.9257
         Detection Rate : 0.8514
   Detection Prevalence : 0.8971
      Balanced Accuracy : 0.6522

       'Positive' Class : 0
```

Figure 46: Confusion matrix of model1(left(training set), right(testing set))

```
Confusion Matrix and Statistics

          Actual
Predicted   0   1
        0 298   0
        1   0  27

               Accuracy : 1
                 95% CI : (0.9887, 1)
    No Information Rate : 0.9169
    P-Value [Acc > NIR] : 5.731e-13

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.9169
         Detection Rate : 0.9169
   Detection Prevalence : 0.9169
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Actual
Predicted   0   1
        0 148   4
        1  14   9

               Accuracy : 0.8971
                 95% CI : (0.8423, 0.9379)
    No Information Rate : 0.9257
    P-Value [Acc > NIR] : 0.93773

                  Kappa : 0.4476

 Mcnemar's Test P-Value : 0.03389

            Sensitivity : 0.9136
            Specificity : 0.6923
         Pos Pred Value : 0.9737
         Neg Pred Value : 0.3913
             Prevalence : 0.9257
         Detection Rate : 0.8457
   Detection Prevalence : 0.8686
      Balanced Accuracy : 0.8029

       'Positive' Class : 0
```

Figure 47: Confusion matrix of model2(left(training set), right(testing set))

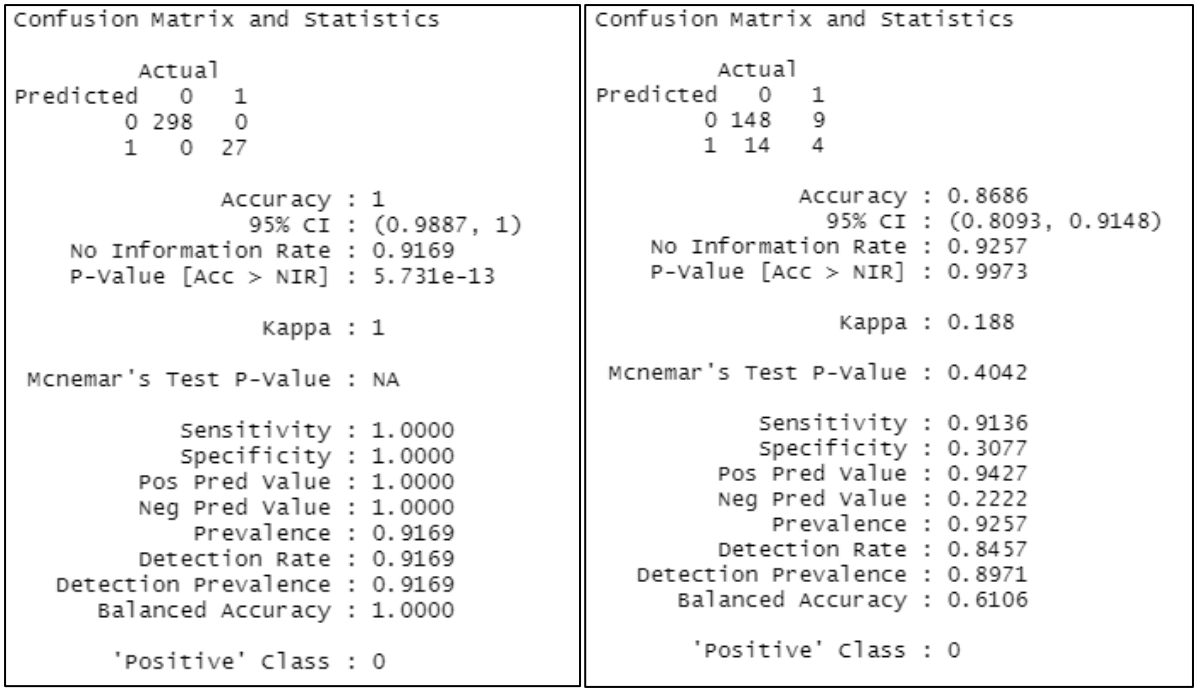

Figure 48: Confusion matrix of model3(left(training set), right(testing set))

Metrics such as precision, sensitivity and specificity are commonly used to evaluate the performance of a predictive model. These mentioned metrics can be easily computed using the function “confusionMatrix()”.

From the figure showed above, it is clear to notice that model2 with 2 hidden layer and 6 nodes has the highest accuracy among the three models. The accuracy for training set is 100% and the accuracy for testing set is 89.71%. In fact, each neuron/unit in a network learns a weighted sum of its inputs. The weights for each of these connections, also known as model parameters, must be learned during the learning process. As a result, the number of connections in the network or the parameters that must be learned is a good way to think about model capacity. The main point to remember is that the higher the model capacity, the more complex the functions that the network can learn. If there is enough data to learn from, the ability to learn more complex functions could lead to improved performance. This is why fine tune the hidden layer and nodes will affect the accuracy of the models. But in this project, the model with 2 hidden layers performed better than the model with 3 hidden layers. It is because accuracy is not only depending on the number of hidden layer and nodes, but also depend on the quality of the model and the quality of the training data.

5.2 Mean Squared Error (MSE)

```
> pr.nn1 <- compute(nn1,test.data[,1:18])
> pr.nn1_ <- pr.nn1$net.result*(max(hd_new$HeartDisease)-min(hd_new$HeartDisease))+min(hd_new$HeartDisease)
> test.r1 <- (test.data$HeartDisease)*(max(hd_new$HeartDisease)-min(hd_new$HeartDisease))+min(hd_new$HeartDisease)
> MSE.nn1 <- sum((test.r1 - pr.nn1_)^2)/nrow(test.data)
>
> pr.nn2 <- compute(nn2,test.data[,1:18])
> pr.nn2_ <- pr.nn2$net.result*(max(hd_new$HeartDisease)-min(hd_new$HeartDisease))+min(hd_new$HeartDisease)
> test.r2 <- (test.data$HeartDisease)*(max(hd_new$HeartDisease)-min(hd_new$HeartDisease))+min(hd_new$HeartDisease)
> MSE.nn2 <- sum((test.r2 - pr.nn2_)^2)/nrow(test.data)
>
> pr.nn3 <- compute(nn3,test.data[,1:18])
> pr.nn3_ <- pr.nn3$net.result*(max(hd_new$HeartDisease)-min(hd_new$HeartDisease))+min(hd_new$HeartDisease)
> test.r3 <- (test.data$HeartDisease)*(max(hd_new$HeartDisease)-min(hd_new$HeartDisease))+min(hd_new$HeartDisease)
> MSE.nn3 <- sum((test.r3 - pr.nn3_)^2)/nrow(test.data)
```

Figure 49: Mean Square Error of the three models

MSE is a network performance function. It measures the network's performance according to the mean of squared errors. The larger the number, the larger the error. MSE would be considered as an extra validation in this project to verify the result of confusion matrix.

```
> print(paste(MSE.nn1,MSE.nn2,MSE.nn3))
[1] "0.114217158912996 0.104459205736553 0.131097045501141"
```

Figure 50: Result of the Mean Square Error

In conclusion, the result has showed that the model2 with 2 hidden layer and 6 nodes has the lowest MSE. Thus, it indicates and validates that model2 is the best performing model among the three models. In this project, the most suitable hidden layer and nodes that could be used are 2 and 6 respectively for this dataset in order to get the best accuracy and performance.

6.0 Data Warehouse

6.1 Business scenario

The type of business used as a model for the project is a medical organization from the United States which is the Centers for Disease Control and Prevention. This organization is set up to collect data overall in 50 states as well as the District of Columbia and three United States territories. There are many different factors that influence heart disease like smoking, alcohol, BMI, Age, physical health, mental health, age, etc.

6.2 Data warehouse conceptual modelling

A data warehouse is a process of collecting and managing data by using data warehouse elements. The elements of a data warehouse are data cleaning, data integration, and data consolidation. Usually, it is used to connect and analyze data from heterogeneous sources. The

data warehouse is the centralized data repository that can be analyzed to make better-informed decisions. Moreover, the data warehouse is concerned with historical data. The data warehouse uses Online Analytical Processing (OLAP) which serves users or knowledge workers for the purpose of data analysis and decision-making. These systems can organize and present data in a variety of formats to satisfy the needs of different users. In today's business environment, an organization needs to be able to report and analyse vast amounts of data with confidence. From customer service to partner integration to top-level executive business decisions, businesses want their data to be consolidated and integrated at various degrees of aggregation. Thus, a data warehouse plays an important role as it facilitates the reporting and analysis process of organizations. As a result of the increase in data, data warehouses are being used more frequently to handle business data. Hence, the data warehouse is suitable for the healthcare industry where it is used to strategize and predict outcomes, create patient treatment reports, etc. Advanced machine learning and big data enable data warehouse systems to predict ailments.

The important data that has to be retrieved must be properly arranged in order to establish a data warehouse system. In order to do this, the researchers created the information package seen in Table 2. To that purpose, the information package represented in Table 2 has been produced. The information package makes it easier to write out the needs for the dimension tables, their hierarchies, and the facts to be modified in the data warehouse's design process. The information package was then used to create the dimension tables shown in Figure 51. The Patient, Habit, and HealthCondition dimensions are each used to create dimension tables. The fact table shown in Figure 52 used for this paper was based on heart disease information. The table contains 500 randomly generated heart disease from 320k observations.

Dimension	Patient	Habit	Health Condition
Hierarchies/Category	Sex	Smoking	Heart Disease
	BMI	Alcohol Drinking	Asthma
	Age Category	Physical Activity	Skin Cancer

	Race	Difficult Walking	Kidney Disease
		Sleep Time	Stroke
			Diabetics
			Mental Health
			Physical Health

Table 2: Information Package

Fact (Measures): Days, Truth, Hours

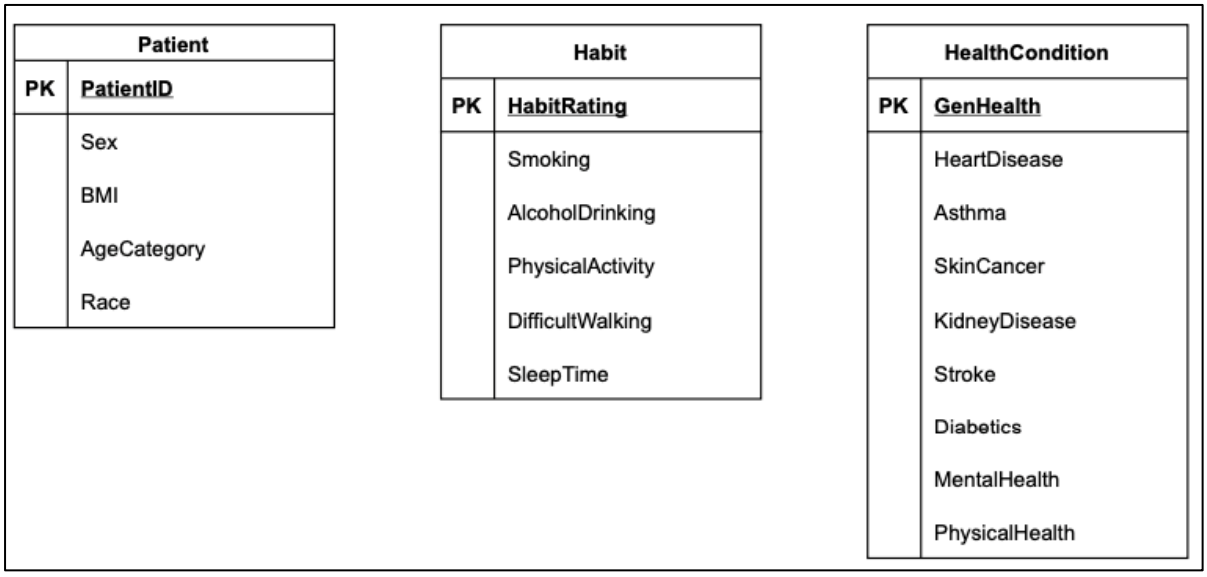

Figure 51: Dimensional tables (Patient, Habit and HealthCondition)

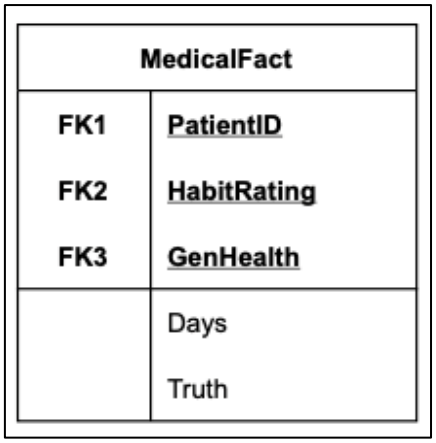

Figure 52: Fact table (MedicalFact)

Sex	BMI	Age Category	Race
Female	Normal weight (18.5 <= BMI < 25.0)	18 - 24	American Indian/Alaskan Native
Male	Obese (30.0 <= BMI < +Inf)	25 - 29	Asian
	Overweight (25.0 <= BMI < 30.0)	30 - 34	Black
	Underweight (BMI < 18.5)	35 - 39	Hispanic
		40 - 44	Other
		45 - 49	White
		50 -54	
		55 - 59	
		60 - 64	
		65 - 69	
		70 - 74	

		75 - 79	
		80 or older	

Table 3: Hierarchies in Patient Dimensions

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Every dimension consists of at least one hierarchy. These hierarchies enable users to analyse data aggregations in simple ways utilising Analysis Services' OLAP functionalities. For example, hierarchies in patient dimensions are shown in Table 3. Patients are included in four hierarchies within the Patients dimension, as shown in Table 3. It is important to note that each patient will be found in each of the four hierarchies. These four different hierarchies provide four different ways to look at summarised customer data. The relationship between the BMI, AgeCategory and Sex can become an important analytical topic. The analysis can be implemented to see which AgeCategory is more normal BMI. Moreover, the race and the BMI also can be analysed to identify the relation between them using data mining techniques. The analytical topic such as the BMI that is mainly used by the customer can be implemented and the organisation can use the data for further uses to detect the heart disease as the preliminary estimate and minimise the burden of the doctors to improve efficiency of the work.

The Star schema was used since the dimension tables were not normalised and the dimensions were not excessively large. Figure 53 depicts the Star schema. The Star schema was also chosen because it has a simple and direct design that is easier for users to understand. Furthermore, there are no foreign keys in the dimension tables of the star schema. It also provides highly optimised capabilities for data warehouse queries. While much of the data was easily translated into the tables needed for the Star schema implementation, some data cleansing was necessary as part of the data cleaning process.

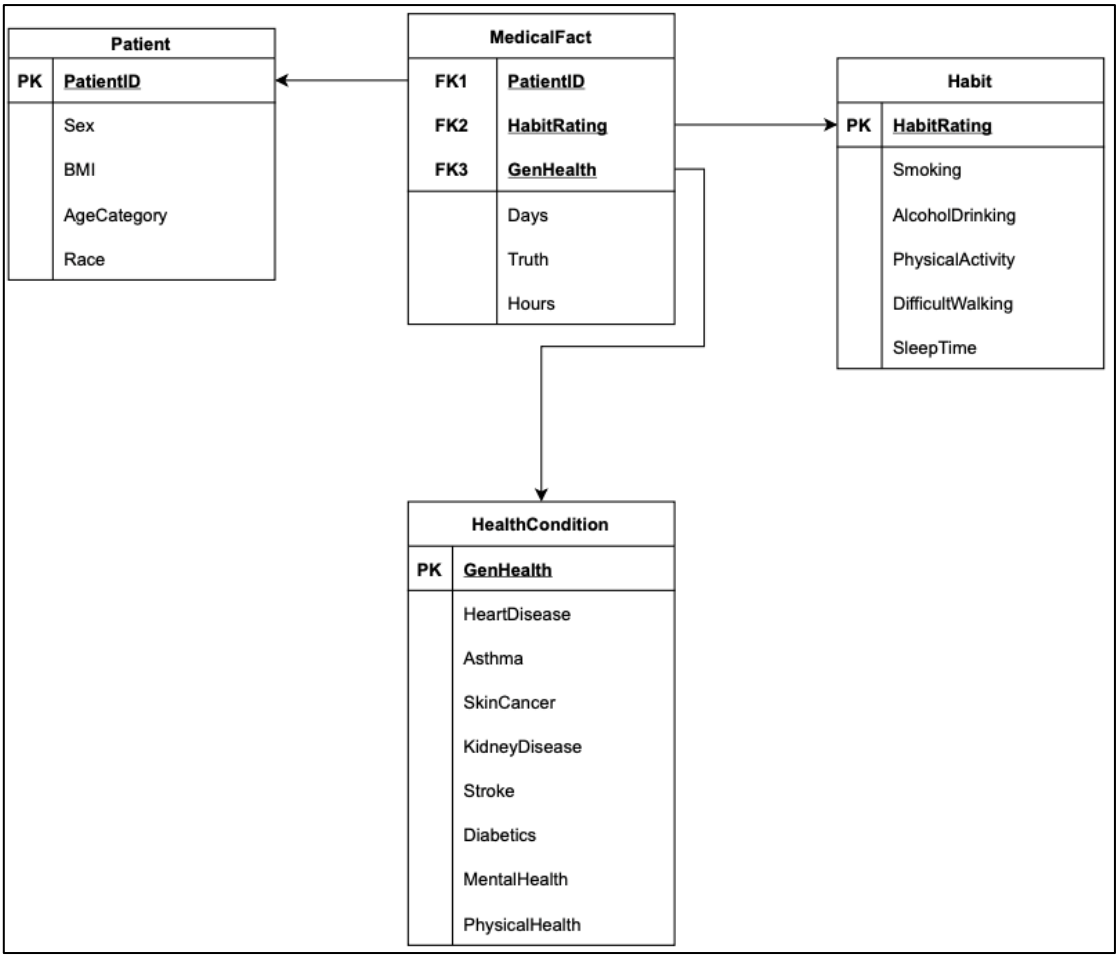

Figure 53: Star Schema

The Medical fact table and 4-dimension table link together to form a star schema as shown in Figure 53. Since the total number of the data columns used in the data set is 18 columns, which is considerable as normal size compared with the other database. The Star schema is suitable to use in this case.

The patient dimension includes the patient ID which unique for each patient from 0001 to 0500, sex, Body Mass Index (BMI), age category, and race. The patient dimension is formed to be able to apply prediction analysis on the heart diseases, and what relationship that hiddern in the variables, the relationship between the race, BMI, age category and sex, their race and many more prediction, reports and information can be generated from this dimension by using different type of Data Mining techniques.

Besides that, the habit dimension stores the habit rating, smoking, alcohol drinking, physical activity, difficult walking and sleep time. The habits such as smoking, alcohol drinking, physical activity and difficult walking are measured in truth which is yes or no while

the sleep time is measured in hours. Through the collection of the data of the patient's habits, it will predict whether the patient has the habits or not. Then, it will result in the habit rating which is measured in truth either yes or no. As a result, the habit dimension will help patients to predict heart disease accurately.

Furthermore, the health condition dimension includes general health, health diseases, skin cancer, kidney disease, stroke, asthma, diabetes, physical health and mental health. The general health is measured into five groups which are poor, fair, good, very good and excellent. All of the diseases are evaluated by truth except mental health and physical health are calculated by the days out of 30 days. The health conditions of the patients are collected to predict the risks of heart disease. This is because the higher the number of the disease gets by the patients, the risks of heart disease increases. So, the diseases are evaluated as yes or no and result in general health based on the disease and health.

Finally, data warehouses enhance strategic decision-making. A data warehouse may become even more powerful with the help of data mining, since it can assist in the analysis of large amounts of data. Using various types of data, an organisation may insight the hidden data, predict the diseases as assistant of the doctor, and more. Techniques for mining enhance work performance through improving accuracy of the model, increasing efficiency, and analysing variables and trends. By using this technique, humans can attempt a golden treatment period on time.

7.0 Conclusion

Heart disease is a critical public health problem. It requires a reliable and cost-effective prediction on a comprehensive dataset to provide experts information to take precautions. In this era of information and technology, machine learning algorithms play an important role in predicting heart disease using classification with high accuracy and reliability. In this paper, the researchers have used the Neural Network algorithm. Thus, there are few aspects that will be summarised which are the data preparation method, data mining method, data evaluation process, and the data warehousing method.

For the data preparation method, it consists of a few processes such as feature extraction, data cleaning, data sampling and feature selection. The feature extraction is a process of transforming raw data into numerical features that may be processed while preserving the information in the original data set where it produces better outcomes than merely applying machine learning to raw data. For the data cleaning process, it begins with dealing with missing values, duplicate values, encoding categorical features. In this paper, the data cleaning process begins with dealing with missing values. Next, the data with duplicate values were removed to ensure that every data was unique. Lastly, the Encoding categorical features converts all categorical data into integer format so that the data with transformed categorical values may be sent to the models to produce and enhance predictions. Furthermore, the data sampling process is a set of approaches for transforming a training dataset in order to balance or improve the class distribution. Thus, Neural Network algorithm can be trained directly on the altered dataset after it has been balanced. Besides that, feature selection process by using Analysis of Variance (ANOVA). ANOVA is used when one variable is numerical and the other is categorical.

In addition, for the data mining process, data visualization for numerical variables and categorical variables are first carried out to extract the useful information from the data. Next, the researchers decided to choose Artificial Neural Network algorithm to perform the heart disease classification in this project. Three neural network models have been built to get the best performance model among them. At the end of the study, the researchers found that out of the three models, the second model with 2 hidden layer and 6 nodes performs the best as it has the highest accuracy which is 100% for train dataset and 89.71% for test dataset. Moreover, in the evaluation process, confusion matrix and result of the Mean Square Error (MSE) have proved the quality of the models.

Furthermore, the data warehousing method is important in the medical field especially for heart disease, as it provides capabilities beyond electronic health records to manage and help patients with heart disease symptoms to manage and reduce mortality. First of all, the researchers created an information package to arrange the requirements of dimensional tables, a fact table and hierarchies. Then, the researchers implemented a star schema because it is easy to understand and provides optimal disk usage.

To summarise, all of these aspects are critical to the data analysis process which can effectively reduce the errors. The research has clearly demonstrated the steps of building a model for heart disease predictions using different models and the results obtained by each model is compared with one another. As a result, the findings of this paper can be utilised as a reference for future scholars conducting research on heart disease prediction using the Neural Network algorithm.

References

1. World Health Organization, *Cardiovascular Diseases*, WHO, Geneva, Switzerland, 2020, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

2. Centers for Disease Control and Prevention. (2022, January 13). FASTSTATS - deaths and mortality. Centers for Disease Control and Prevention. Retrieved May 10, 2022, from https://www.cdc.gov/nchs/fastats/deaths.htm

3. Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

4. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*.

5. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

6. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

7. Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, *8*, 107562-107582.

8. Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). Innovative artificial neural networks-based decision support system for heart diseases diagnosis.

9. Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.

10. Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. *Computer Science & Information Technology Journal*, *24*, 53-59.