Ho Chi Minh University of Technology

Office for International Study Programs

Faculty of Applied Science

Course:

PROBABILITY AND STATISTICS

(MT2013)

Group: CC01 - 05

PREDICT HEART DISEASE USING LOGISTIC REGRESSION

Authors:	*Student's ID:*
Nguyen Van Binh	2153223
Dinh Le Minh Quan	2152262
Le Minh Quy	2153758
Dinh Minh Tri	2153059
Le Quoc Tuan	2153944

Supervisor:

Dr. Phan Thi Huong

Ho Chi Minh City, April 2023

Leader: Dinh Minh Tri - tri.dinhminh@hcmut.edu.vn

Authors	**Contribution**	**Assigned Tasks**
Nguyen Van Binh	20%	Select data, code R, drop and re-define data based on its usage, reference solving methods
Dinh Le Minh Quan	20%	Code R, train and test the data, come up with special model for the report
Le Minh Quy	20%	Pre-process data, code R, quantitatively plot data to ease the analyzing process
Dinh Minh Tri	20%	Organize and divide tasks, write report, research, "build", and explain mathematical model for coding
Le Quoc Tuan	20%	Search for topics, Data visualizing, write report, reference data sources, coding methods

Lecturer's assessment

Members	**Score**	**Assessment**
Nguyen Van Binh		
Dinh Le Minh Quan		
Le Minh Quy		
Dinh Minh Tri		
Le Quoc Tuan		

Contents

1 Data introduction

1.1 Dataset description

The dataset that is used in this project is about heart disease diagnosis. Here are some general details of the dataset:

- **Title:** Heart Disease Database
- **Source Information:**

  (a) Creators: V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

  (b) Date: July, 1988
- **Number of Instances:** 303 (with 165 tested positive and 138 tested negative for heart disease)
- **Number of Variables:** 14 *(Described in section 1.2)*

1.2 Variables description

Variable	**Data type** *cont = continuous, dis = discrete	**Unit**	**Description**
age	$\{x \in N \mid 29 \leq x \leq 77\}$, cont	years	Age of a person
sex	$x = 0$ or $x = 1$, dis	none	Sex of a person (0 = female, 1 = male)
cp	$\{x \in N \mid 0 \leq x \leq 3\}$, dis	none	Type of chest pain (0 = Typical angina, 1 = Atypical angina, 2 = Non-anginal pain, 3 = Asymptomatic)
trtbps	$\{x \in N \mid 94 \leq x \leq 200\}$, cont	mmHg	Resting blood pressure
chol	$\{x \in N \mid 126 \leq x \leq 564\}$ cont	mg/dl	Cholesterol measurement fetched via BMI sensor
fbs	$x = 0$ or $x = 1$, dis	none	Fasting blood sugar (> 120 mg/dl, 0 = false, 1 = true)

restecg	$\{x \in N \mid 0 \leq x \leq 2\}$, dis	none	Results of resting electrocardiographic (0 = Normal, 1 = Having ST-T wave abnormality, 2 = Showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalachh	$\{x \in N \mid 71 \leq x \leq 202\}$, cont	bpm	Maximum heart rate achieved
exng	$x = 1$ or $x = 1$, dis	none	Exercise induced angina (0 = no, 1 = yes)
oldpeak	$\{x \in R \mid 0 \leq x \leq 6.2\}$, cont		ST depression induced by exercise relative to rest (More details)
slp	$\{x \in N \mid 0 \leq x \leq 2\}$, dis	none	The slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)
ca	$\{x \in N \mid 0 \leq x \leq 3\}$, dis	unit(s)	Number of major vessels
thal	$\{x \in N \mid 1 \leq x \leq 3\}$, dis	none	A blood disorder called thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect)
output	$x = 0$ or $x = 1$, dis	none	Heart disease (0 = no, 1 = yes)

2 Background

2.1 Logistic Regression

2.1.1 Definition

Logistic regression (or logit regression) is a process of estimating the probability of a discrete outcome, based on a given dataset of independent variables. It is the appropriate regression analysis to conduct when the dependent variable is binary (with value 0 or 1).

Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables. This regression technique is similar to linear regression and can be used to predict the Probabilities for classification problems.

In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formula:

$$P(X) = \frac{1}{1 + e^{-X\beta}}$$

- P: "Success probability" - the probability of the dependent variable equaling a success/case rather than a failure/non-case (the probability of a 1).

- $X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$: The dependent variable.

- k: The number of parameters.

- X_i: The independent variables.

- β_0: The intercept.

- β_i: The coefficient of X_i

With one X variable, the theoretical model for P has an elongated signoidal shape with asymptotes at 0 and 1, although in sample estimates we may not see the mentioned shape if the range of the variable is limited.

After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit, which will be discussed in a later section.

2.1.2 Logistic Regression vs. Linear Regression

Linear regression models are used to identify the relationship between a continuous dependent variable and one or more independent variables. When there is only one independent variable and one dependent variable, it is known as simple linear regression,

but as the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit through a set of data points, which is typically calculated using the least squares method.

Similar to linear regression, logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false. The unit of measure also differs from linear regression as it produces a probability, but the logit function transforms the S-curve into a straight line.

With its usage to solve Classification problems, Logistic regression will be used in this project to determine the probability of heart attacks by determining the relationship between variables such as the weight, exercise, etc., of an individual and use it to predict whether the person will suffer from a heart attack or any other medical complication.

2.2 Hosmer-Lemeshow test

Overall performance of the fitted model can be measured by several different goodness-of-fit tests, one of which is the Hosmer-Lemeshow test. It is a goodness-of-fit test for logistic regression, especially for risk prediction models, which tells how well the data fits the model. Essentially, it is a chi-square goodness of fit test for grouped data, and is conducted by sorting the n records in the dataset by estimate probability of success, dividing the sorted set into g equal-sized group, and evaluating the Hosmor-Lemeshow C statistic:

$$\hat{C}_g = \sum_{i=1}^{g} [\frac{(O_{s,i} - E_{s,i})^2}{E_{s,i}} + \frac{(O_{f,i} - E_{f,i})^2}{E_{f,i}}]$$

- $O_{s,i}$: the observed number of successes

- $O_{f,i}$: the observed number of failures

- $E_{s,i}$: the expected number of successes in the ith group.

- $E_{f,i}$: the expected number of failures in the ith group.

The null and alternative hypothesis of the test are:

$$\begin{cases} H_0 : O_{s,i} = E_{s,i} \\ H_1 : O_{s,i} \neq E_{s,i} \end{cases}$$

The hypothesis can also be written as:

$$\begin{cases} H_0 : \text{The model fits the data} \\ H_1 : \text{The model does not fit the data} \end{cases}$$

Under the null hypothesis that the model fits the data, we show that $\hat{C}_g$ follows a χ^2 distribution with $(g-2)$ degrees of freedom. Thus, the p-value for the Hosmer-Lemeshow test is:

$$p = \int_{\hat{C}_g}^{\infty} \chi^2_{g-2}(x)\, dx$$

where $\chi^2_{g-2}(x)$ is the probability density function of the χ^2 function with $g-2$ degrees of freedom evaluated at x. The value of g is user-defined, but a commonly used value is $g = 10$, this value has been adopted as the default by most statistical packages.

Hosmer-Lemeshow goodness-of-fit test is useful for unreplicated datasets or for datasets that contain just a few replicated observations, whereas other tests such as the Pearson chi-square good-of-fit test and the deviance goodness-of-fit test require replicated data.

This test is usually run using computers, which is appropriate for this project. The output returns a chi-square value, and a p-value. Small p-values mean that the model is a poor fit.

2.3 Pearson correlation coefficient

Also known as Pearson's r, or simply known as the correlation coefficient, is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations.

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy}. Given paired data $(x_1, y_1), ..., (x_n, y_n)$, consisting of n pairs, r_{xy} is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- n: the sample size.

- x_i, y_i: the individual sample points at index i

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: The sample means.

The Pearson correlation coefficient is symmetric: $corr(X, Y) = corr(Y, X)$. The values of Pearson correlation coefficient are on or between -1 and 1, $|r| = 1$ implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying exactly on a line, while when r is closer to 0, it implies that the points are far from the line of best fit:

Figure 1: Perfect positive correlation ($r = 1$) and Perfect negative correlation ($r = -1$)

The correlation sign is determined by the regression slope:

- A value between 0 and 1 is called positive correlation and it is interpreted as when one variable changes, the other variable changes in the same direction.

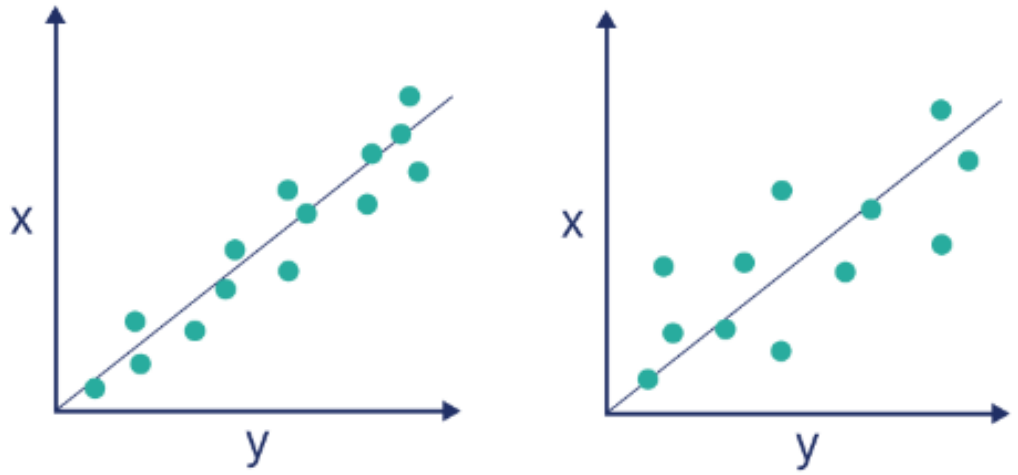


Figure 2: Strong ($r > 0.5$) and Weak positive correlation ($0 < r < 0.3$)

- A value between -1 and 0 is called negative correlation and it is interpreted as when one variable changes, the other variable changes in the opposite direction.

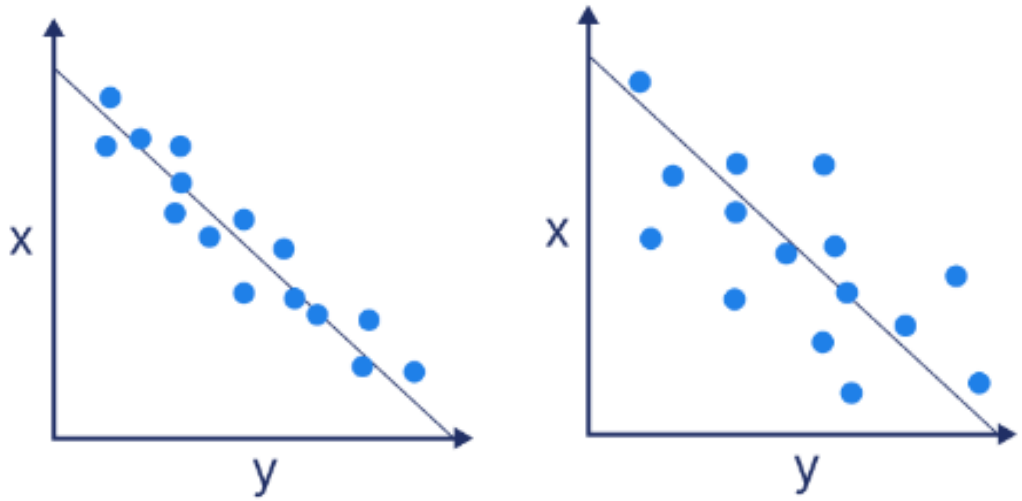


Figure 3: Strong ($r < -0.5$) and Weak negative correlation ($-0.3 < r < 0$)

- A value of 0 implies that there is no linear dependency between the variables.

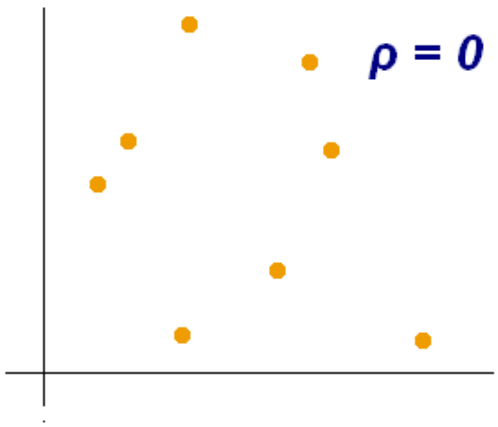


Figure 4: No correlation ($r = 0$)

3 Data analysis

3.1 Data reading

First, we will import necessary library for later use: **ggplot2, dplyr, plotly, cowplot, caret, vcd, ResourceSelection, pROC, corrplot.**

Read data using read.csv and display the data to terminal to check if data is successfully imported.

```
> df=read.csv("C:/Users/dinhq/Desktop/xstk_ass/heart.csv")
> head(df,10)
   age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
1   63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
2   37   1  2    130  250   0       1      187    0     3.5   0   0     2      1
3   41   0  1    130  204   0       0      172    0     1.4   2   0     2      1
4   56   1  1    120  236   0       1      178    0     0.8   2   0     2      1
5   57   0  0    120  354   0       1      163    1     0.6   2   0     2      1
6   57   1  0    140  192   0       1      148    0     0.4   1   0     1      1
7   56   0  1    140  294   0       0      153    0     1.3   1   0     2      1
8   44   1  1    120  263   0       1      173    0     0.0   2   0     3      1
9   52   1  2    172  199   1       1      162    0     0.5   2   0     3      1
10  57   1  2    150  168   0       1      174    0     1.6   2   0     2      1
```

Figure 5: Head of data

3.2 Checking missing values

Using the command `is.na(data)` will return a new data frame which has null value. Therefore, the sum command can be used to calculate the total number of rows having null value.

```
> sum(is.na(df))
[1] 0
```

Figure 6: Result of checking missing values

Our data doesn't have any null value so just skip it and move to next step.

3.3 Data summary

3.3.1 Data statistics

First, we will display the overview of data using **summary(data)**.

```
> summary(df)
      age             sex              cp             trtbps           chol            fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000  
    restecg          thalachh         exng           oldpeak          slp             caa        
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :1.0000   Median :153.0   Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
 Mean   :0.5281   Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
 3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
     thall           output      
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.314   Mean   :0.5446  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
```

Figure 7: Summary of data

For our discrete variables or categorical variables, we will use **as.factor()** to convert a variable or vector to a factor for later use. Then, display to terminal summary of these variables.

```
      sex
output   F   M
     0  24 114
     1  72  93
> xtabs(~output+cp,data=df)
      cp
output   0   1   2   3
     0 104   9  18   7
     1  39  41  69  16
> xtabs(~output+fbs,data=df)
      fbs
output   0   1
     0 116  22
     1 142  23
> xtabs(~output+restecg,data=df)
      restecg
output  0  1  2
     0 79 56  3
     1 68 96  1
```

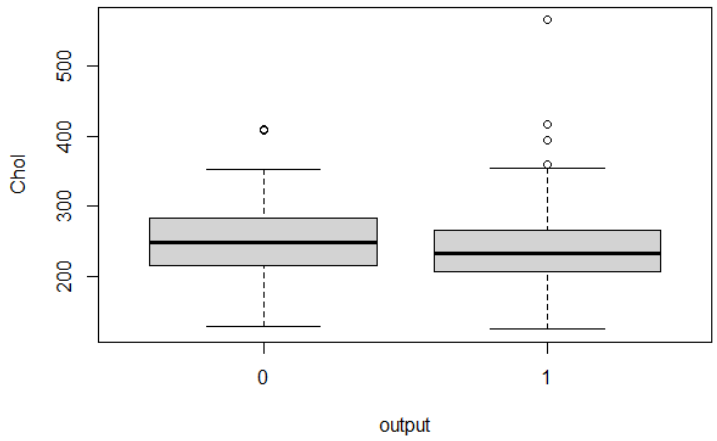
```
> xtabs(~output+exng,data=df)
      exng
output   0   1
     0  62  76
     1 142  23
> xtabs(~output+slp,data=df)
      slp
output   0   1   2
     0  12  91  35
     1   9  49 107
> xtabs(~output+caa,data=df)
      caa
output   0   1   2   3   4
     0  45  44  31  17   1
     1 130  21   7   3   4
> xtabs(~output+thall,data=df)
      thall
output   0   1   2   3
     0   1  12  36  89
     1   1   6 130  28
```

Figure 8: Summary for discrete variables

3.3.2 Data plot

First, we will plot boxplot for our continuous variables: ”trtbps”, ”chol”, ”thalachh”, ”oldpeak”.

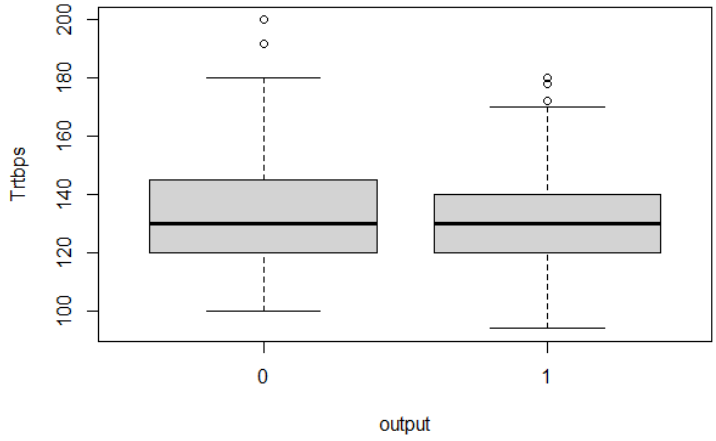
We use boxplot to indicate continuous variables including chol, trtbps, oldpeak and thalachh. These boxplots are used to display the median, identify the interquartile range (IQR), detect skewness, identify outliers, and compare distributions.



	0	1
Min	131	126
First Quartile	217	208
Median	249	234
Third Quartile	283	267
Maximum	353	354

Figure 9: Boxplot of cholesterol (mg/dl) and result

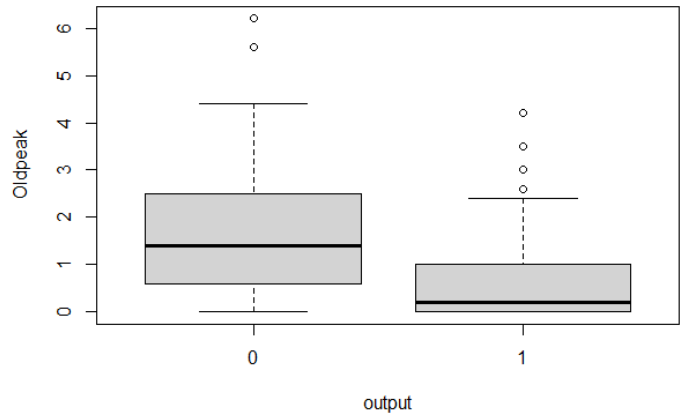
Boxplot of heart disease patients (output = 1) has more outliers. The important point here is that both have a similar distribution. There does not exist a specific range of cholesterol values where the patient has heart disease.



	0	1
Min	100	94
First Quartile	120	120
Median	130	130
Third Quartile	145	140
Maximum	180	170

Figure 10: Boxplot of resting blood pressure (mmHg) and result

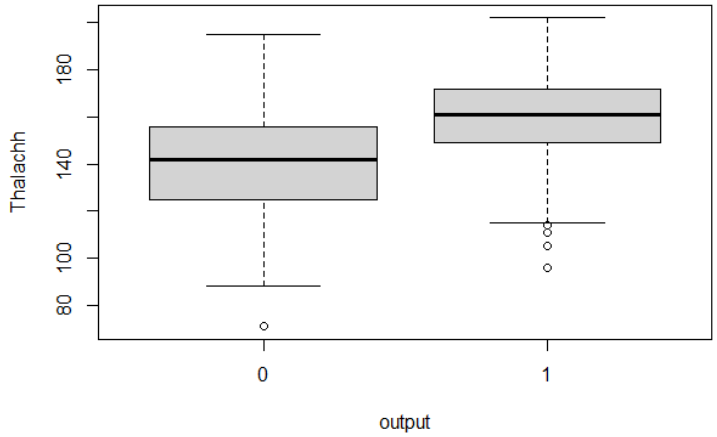
Similar to cholesterol, we also cannot distinguish the range of values that patients have heart disease. They have the same first quartile and median while their third quartile, minimum, and maximum have a little difference.



	0	1
Min	0.0	0.0
First Quartile	0.6	0.0
Median	1.4	0.2
Third Quartile	2.5	1.0
Maximum	4.4	2.4

Figure 11: Boxplot of ST depression and result

Both boxplots have outliers and skewness distribution. For this value, we have a big difference between each factor of boxplot.



	0	1
Min	88	115
First Quartile	125	149
Median	142	161
Third Quartile	156	172
Maximum	195	202

Figure 12: Boxplot of maximum heart rate achieved and result

Boxplot 0 just has one outlier and it is skewness distribution while the orther has more outliers and the boxplot 1 is symmetric distribution. All factors of boxplot 1 are larger than factors of boxplot 0.

We can see that, for each output, the input data is distributed randomly and it is quite similar. We can't determine in which range, patient will have a heart attack if we only consider one variable.

We use the histogram to indicate discrete variables including **cp, restecg, fbs, exng**. These histogram figures are used to see the number of patients of each type, visualize our dataset.

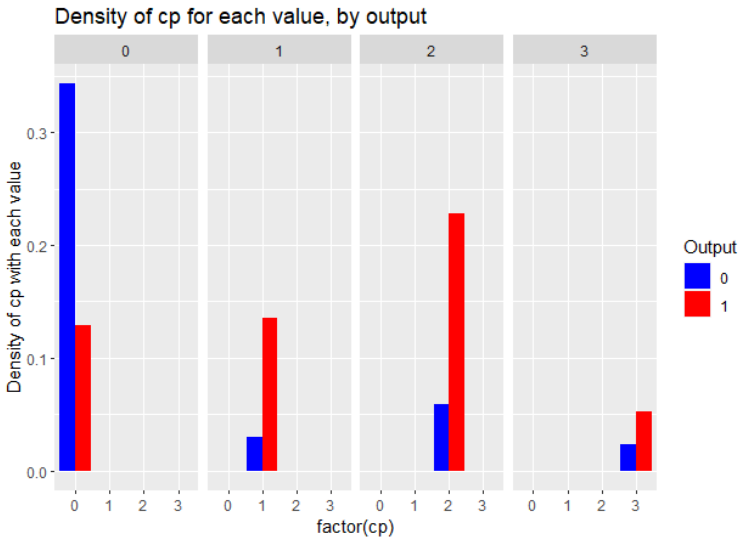


Figure 13: Density of each chest pain type based on 2 outputs

This figure shows that the number of disease heart patients is higher than one who does not have heart disease in all values of cp except the first column($cp = 0$).

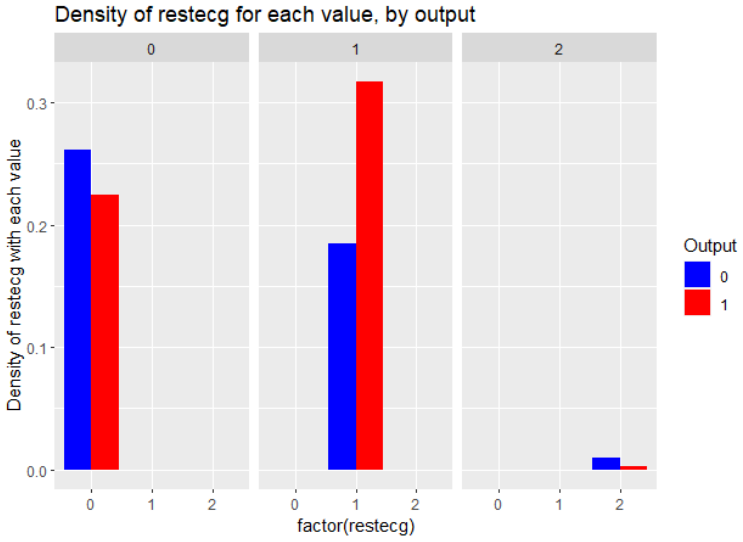


Figure 14: Density of each result of resting electrocardiographic type based on 2 outputs

This figure illustrates that the number of disease heart patients is higher than one who does not have heart disease in the second column($restecg = 1$) while the number of who do not have heart disease is higher than in 2 remained columns. However, the difference is not very significant so it's still quite similar.

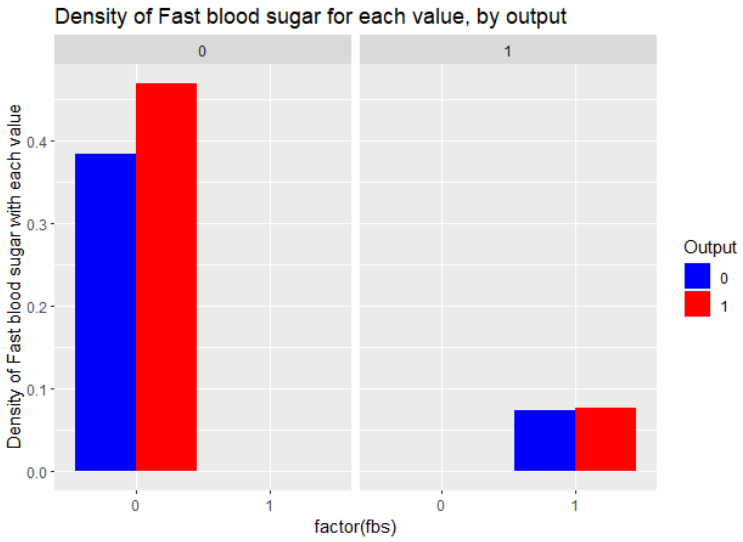


Figure 15: Density of each fasting blood sugar type based on 2 outputs

Same situation happens for this features when the density of each output respect to each type of input is quite similar.

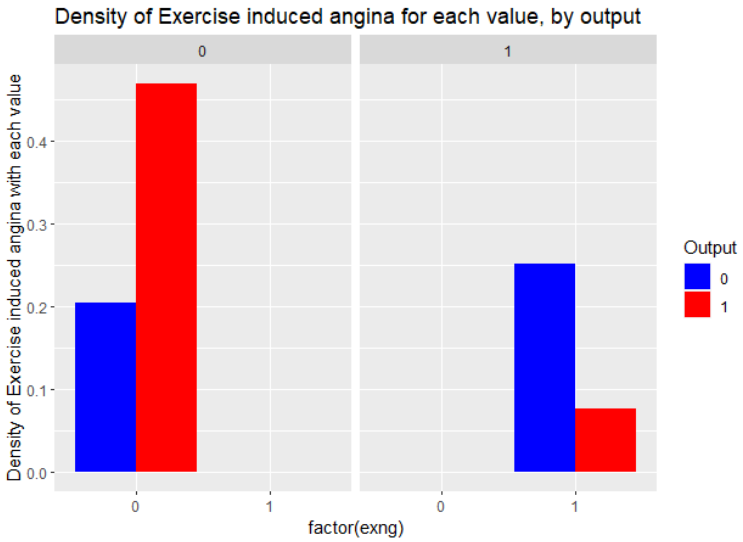


Figure 16: Density of each exercise induced angina type based on 2 outputs

This figure shows that the number of disease heart patients is higher than one who does not have heart disease in the first column(fbs = 0) and vice versa for second column.

3.4 Correlation coefficients between variables

To see the linear relationship between each variable, we will plot the correlation coefficient of all variables using **corrplot** function and display these coefficients to terminal.

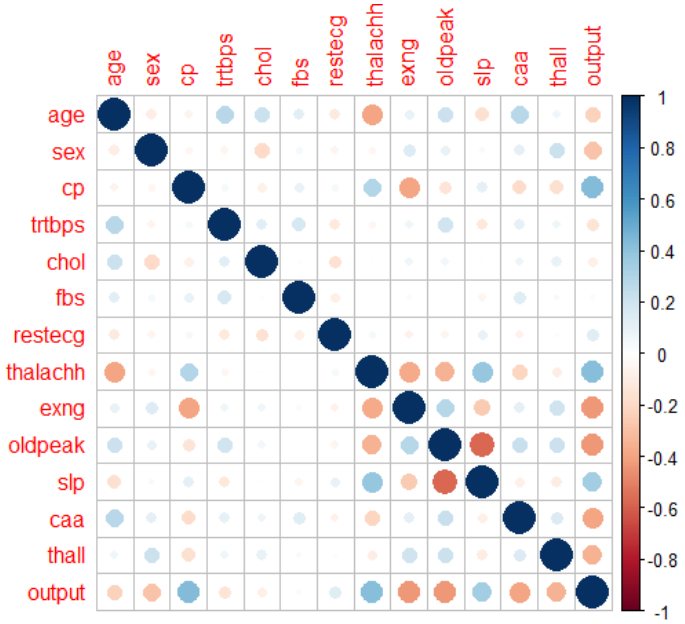


Figure 17: Correlogram of the data

```
> newdf <- sapply(df,as.numeric)
> cor_matrix <- cor(newdf)
> cor_matrix
                age         sex          cp       trtbps        chol         fbs     restecg
age       1.00000000 -0.09844660 -0.06865302  0.27935091  0.213677957  0.12130765 -0.11621090
sex      -0.09844660  1.00000000 -0.04935288 -0.05676882 -0.197912174  0.04503179 -0.05819627
cp       -0.06865302 -0.04935288  1.00000000  0.04760776 -0.076904391  0.09444403  0.04442059
trtbps    0.27935091 -0.05676882  0.04760776  1.00000000  0.123174207  0.17753054 -0.11410279
chol      0.21367796 -0.19791217 -0.07690439  0.12317421  1.000000000  0.01329360 -0.15104008
fbs       0.12130765  0.04503179  0.09444403  0.17753054  0.013293602  1.00000000 -0.08418905
restecg  -0.11621090 -0.05819627  0.04442059 -0.11410279 -0.151040078 -0.08418905  1.00000000
thalachh -0.39852194 -0.04401991  0.29576212 -0.04669773 -0.009939839 -0.00856710  0.04412344
exng      0.09680083  0.14166381 -0.39428027  0.06761612  0.067022783  0.02566514 -0.07073286
oldpeak   0.21001257  0.09609288 -0.14923016  0.19321647  0.053951920  0.00574722 -0.05877023
slp      -0.16881424 -0.03071057  0.11971659 -0.12147458 -0.004037770 -0.05989418  0.09304482
caa       0.27632624  0.11826141 -0.18105303  0.10138899  0.070510925  0.13797932 -0.07204243
thall     0.06800138  0.21004110 -0.16173557  0.06220989  0.098802993 -0.03201934 -0.01198140
output   -0.22543872 -0.28093658  0.43379826 -0.14493113 -0.085239105 -0.02804576  0.13722950
             thalachh        exng      oldpeak         slp         caa       thall      output
age      -0.398521938  0.09680083  0.210012567 -0.16881424  0.27632624  0.06800138 -0.22543872
sex      -0.044019908  0.14166381  0.096092877 -0.03071057  0.11826141  0.21004110 -0.28093658
cp        0.295762125 -0.39428027 -0.149230158  0.11971659 -0.18105303 -0.16173557  0.43379826
trtbps   -0.046697728  0.06761612  0.193216472 -0.12147458  0.10138899  0.06220989 -0.14493113
chol     -0.009939839  0.06702278  0.053951920 -0.00403777  0.07051093  0.09880299 -0.08523911
fbs      -0.008567107  0.02566515  0.005747223 -0.05989418  0.13797933 -0.03201934 -0.02804576
restecg   0.044123444 -0.07073286 -0.058770226  0.09304482 -0.07204243 -0.01198140  0.13722950
thalachh  1.000000000 -0.37881209 -0.344186948  0.38678441 -0.21317693 -0.09643913  0.42174093
exng     -0.378812094  1.00000000  0.288222808 -0.25774837  0.11573938  0.20675379 -0.43675708
oldpeak  -0.344186948  0.28822281  1.000000000 -0.57753682  0.22268232  0.21024413 -0.43069600
slp       0.386784410 -0.25774837 -0.577536817  1.00000000 -0.08015521 -0.10476379  0.34587708
caa      -0.213176928  0.11573938  0.222682322 -0.08015521  1.00000000  0.15183213 -0.39172399
thall    -0.096439132  0.20675379  0.210244126 -0.10476379  0.15183213  1.00000000 -0.34402927
output    0.421740934 -0.43675708 -0.430696002  0.34587708 -0.39172399 -0.34402927  1.00000000
```

Figure 18: Correlation coefficients summary

4 Prediction of heart disease based on all features of patients

4.1 Data pre-processing

Take a look at the data frame, we can easily observe that all of the healthy patients are located on the first half of the data set and conversely. Therefore, if we take the raw data and fit directly into the model, the performance of the regression model apparently will be very poor. To handle it, we need a small step: shuffle the data. We also set seed for reproducibility purposes:

```
set.seed(5)
df <- df[sample(nrow(df)), ]
```

Then, we split the data into training set and validation set with ratio 7:3:

```
trainIndex <- createDataPartition(df$output, p = 0.7, list = FALSE)
trainData <- df[trainIndex, ]
testData <- df[-trainIndex, ]
```

Finally, reset index of training set and validation set.

All done, Let's take a look at the training set and validation set:

```
head(trainData)
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
	<int>	<fct>	<fct>	<int>	<int>	<fct>	<fct>	<int>	<fct>	<dbl>	<fct>	<fct>	<fct>	<fct>
1	38	M	2	138	175	0	1	173	0	0.0	2	4	2	Unhealthy
2	67	M	0	120	237	0	1	71	0	1.0	1	0	2	Healthy
3	47	M	2	130	253	0	1	179	0	0.0	2	0	2	Unhealthy
4	60	F	2	120	178	1	1	96	0	0.0	2	0	2	Unhealthy
5	51	M	2	100	222	0	1	143	1	1.2	1	0	2	Unhealthy
6	70	M	2	160	269	0	1	112	1	2.9	1	1	3	Healthy

Figure 19: Training set

```
head(testData)
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
	<int>	<fct>	<fct>	<int>	<int>	<fct>	<fct>	<int>	<fct>	<dbl>	<fct>	<fct>	<fct>	<fct>
1	56	M	2	130	256	1	0	142	1	0.6	1	1	1	Healthy
2	45	M	1	128	308	0	0	170	0	0.0	2	0	2	Unhealthy
3	52	M	0	125	212	0	1	168	0	1.0	2	2	3	Healthy
4	38	M	3	120	231	0	1	182	1	3.8	1	0	3	Healthy
5	41	F	2	112	268	0	0	172	1	0.0	2	0	2	Unhealthy
6	59	M	2	126	218	1	1	134	0	2.2	1	1	1	Healthy

Figure 20: Validation set

4.2 Fitting logistic regression model

Our model can be represented as a function:

$$g(\gamma) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Where:
- $g(\gamma) = log(\frac{\gamma}{1-\gamma})$
- γ : dependent variable
- X_i: independent variable
- β_0: y-intercept (constant term)
- β_i: estimate of each independent variable

We will apply model for our train data and display the result.

```
Call:
glm(formula = output ~ age + cp + trtbps + chol + fbs + restecg +
    thalachh + exng + oldpeak + slp + caa + thall, family = "binomial",
    data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4431  -0.1927   0.0725   0.3678   2.8095
```

Figure 21: The result of model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.671e+01  3.956e+03   0.004 0.996629
age          2.300e-02  3.543e-02   0.649 0.516233
cp1          9.703e-01  7.540e-01   1.287 0.198155
cp2          2.558e+00  7.451e-01   3.433 0.000597 ***
cp3          2.317e+00  9.823e-01   2.359 0.018322 *
trtbps      -3.686e-02  1.607e-02  -2.294 0.021801 *
chol         6.776e-04  4.480e-03   0.151 0.879787
fbs1         1.092e+00  7.957e-01   1.372 0.169991
restecg1     1.336e+00  5.411e-01   2.469 0.013538 *
restecg2     1.345e+00  3.814e+00   0.353 0.724369
thalachh     1.668e-02  1.716e-02   0.972 0.331046
exng1       -4.136e-01  5.994e-01  -0.690 0.490153
oldpeak     -1.029e+00  3.851e-01  -2.672 0.007550 **
slp1        -1.307e+00  1.263e+00  -1.034 0.300983
slp2        -2.707e-01  1.418e+00  -0.191 0.848658
caa1        -2.520e+00  6.868e-01  -3.669 0.000243 ***
caa2        -2.816e+00  1.143e+00  -2.464 0.013733 *
caa3        -2.219e+00  1.390e+00  -1.596 0.110487
caa4         1.563e+01  1.793e+03   0.009 0.993046
thall1      -1.497e+01  3.956e+03  -0.004 0.996980
thall2      -1.359e+01  3.956e+03  -0.003 0.997260
thall3      -1.621e+01  3.956e+03  -0.004 0.996730
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 293.58  on 212  degrees of freedom
Residual deviance: 110.60  on 191  degrees of freedom
AIC: 154.6

Number of Fisher Scoring iterations: 16
```

Figure 22: The result of model

So now, we have the relationship between all inputs with output, we will replace theses coefficients to model and predict for test dataset.

4.3 Prediction for test dataset

By using model above, we will assume that for probability which is greater than 0.5, patient will have a heart disease. The result will be:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 28  9
         1 13 40

               Accuracy : 0.7556
                 95% CI : (0.6536, 0.84)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 2.879e-05

                  Kappa : 0.5033

 Mcnemar's Test P-Value : 0.5224

            Sensitivity : 0.6829
            Specificity : 0.8163
         Pos Pred Value : 0.7568
         Neg Pred Value : 0.7547
             Prevalence : 0.4556
         Detection Rate : 0.3111
   Detection Prevalence : 0.4111
      Balanced Accuracy : 0.7496

       'Positive' Class : 0
```

Figure 23: Result for prediction using logistic regression

The accuracy of model is quite low, so to improve the performance of this model, we will draw the ROC curve to choose a better threshold.

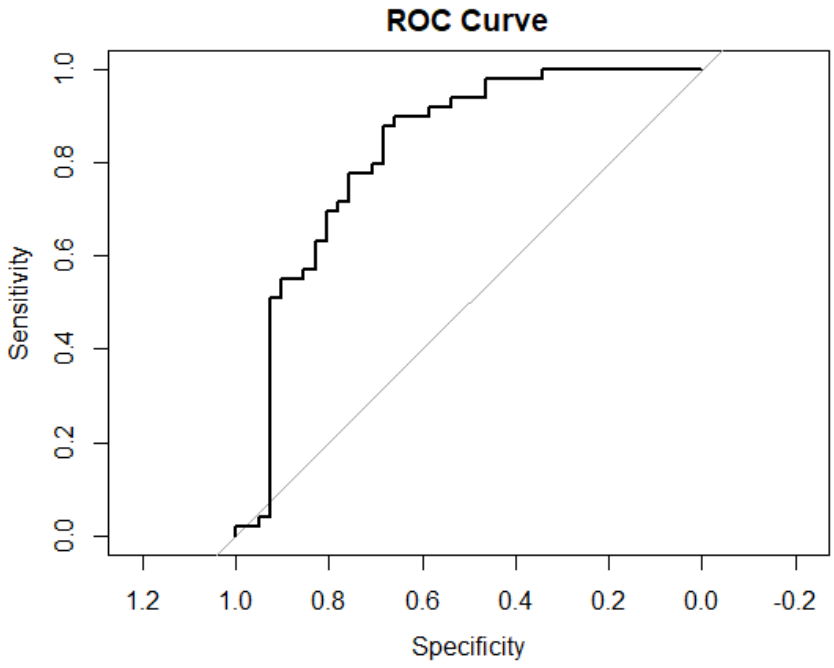


Figure 24: ROC curve of model

Take threshold from ROC curve (top left point), apply to model and predict again. The result will be:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 28  6
         1 13 43

               Accuracy : 0.7889
                 95% CI : (0.6901, 0.8679)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 1.206e-06

                  Kappa : 0.5684

 Mcnemar's Test P-Value : 0.1687

            Sensitivity : 0.6829
            Specificity : 0.8776
         Pos Pred Value : 0.8235
         Neg Pred Value : 0.7679
             Prevalence : 0.4556
         Detection Rate : 0.3111
   Detection Prevalence : 0.3778
      Balanced Accuracy : 0.7802

       'Positive' Class : 0
```

Figure 25: Improved model for prediction using logistic regression

4.4 Test the assumption

We can see that although we choose threshold from ROC curve, the situation has not improved much yet. Let's do Hosmer and Lemeshow test to see the goodness of fit:

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  output_int, fitted_numeric
X-squared = 17.185, df = 8, p-value = 0.02824
```

Figure 26: Hosmer and Lemeshow test for model

As we know, a small p-value, which is less than 0.05, indicates that we have enough evidence to conclude the lack of fit. So we reject the null hypothesis that the model fits the data well. So that the reason why when we apply this model to predict, we have an unexpected result.

4.5 Summary

In summary, logistic regression was utilized to analyze the occurrence of heart attacks. During the data visualization process, it became evident that the data exhibited a high level of variability. Consequently, it proved challenging to determine which factors, or their respective ranges, had the most significant impact on the individuals surveyed. Armed with this knowledge, we proceeded to employ a generalized logistic regression model that incorporated all the data features. Remarkably, this model achieved an accuracy of 78.89% on the test set and 80% on training set, showcasing a commendable performance without succumbing to overfitting. Subsequently, we conducted the Hosmer-Lemeshow test to assess the model's goodness-of-fit. Notably, the obtained p-value was exceptionally small, standing at only 0.02, our model need improving more, especially on more advanced architecture model and more data.

5 Discussion and Extension

5.1 Discussion

Forecasting the probability of a person getting sickness has never been easy, especially in heart related issues. In this assignment, machine learning algorithms approach has been used in heart attack analysis and prediction. Machine learning algorithms can analyze large amounts of data from various sources, including cholesterol, resting electrocardiographic, thalach and so on, to predict the likelihood of a heart attack. However, one of the major challenges in our analysis and prediction is the high variability of symptoms and risk factors among individuals. Some people may have no apparent symptoms or risk factors, while others may have multiple risk factors or symptoms.

For example, consider a following false negative data. As we can see, most of indexes of this patient indicate that this person is healthy as well as machine learning algorithm, but in reality, she has heart attack issue:

```
age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
43   1  0    150  247   0       1       171    0     1.5   2   0     2
```

In contrast, a patient with many symptoms has no heart-related issues but the regression indicates ”yes”, so the following case is considered as false positive:

```
age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
64   1  2    140  335   0       1       158    0     0.0   2   0     2
```

Therefore, it is crucial to develop personalized approaches that can take into account individual differences in symptoms and risk factors.

5.2 Extension

In the realm of binary classification problems, the use of multiple algorithms for prediction and comparison goes beyond the traditional approach of relying solely on logistic regression. In this section, we will explore the performance and suitability of three alternative algorithms: XGBoost, Deep Neural Networks (NN), and K-Nearest Neighbors (KNN). By comparing these algorithms, we aim to uncover their respective strengths and weaknesses, enabling informed decision-making in choosing the most appropriate model for a given binary classification task.

Name	**Packages**	**Hyperparameter**	**Accuracy**
Logistic Regression	glm	Threshold: 0.2961293	78.89%
Deep Neural Network	tensorflow keras	4 Hidden layers such that: - Layer 1: 64 units, activation = "ReLu" - Layer 2: 64 units, activation = "ReLu" - Layer 3: 32 units, activation = "ReLu" - Layer 4: 32 units, activation = "ReLu" In the output layer, we use sigmoid activation function. Then compile with Adam optimizer.	75%
KNN	class	23 neighbors	65%
XGBoost	xgboost	23 rounds	65%

As discussion above, after carefully analyzing confusion matrix, we observe that all of the models face the challenge of high variability of symptoms and risk factors among individuals. Therefore, not only mathematical models are considered, we also need the huge amount of data of patients as well as data specialized for individuals to deliver better performance and prevent overfitting.

One promising area of research is the use of wearable devices and mobile health technologies for continuous monitoring of cardiovascular health. Wearable devices such as smartwatches and fitness trackers can monitor physiological signals such as heart rate, blood pressure, and activity levels. Mobile health technologies such as smartphone apps can collect and analyze data on lifestyle factors such as diet and exercise.

Another area of research is the use of genomics and genetic testing for heart attack analysis and prediction. Genetic factors play a significant role in cardiovascular disease, and several genes have been identified as risk factors for heart attacks. By analyzing genetic data, researchers can develop personalized risk assessments and identify individuals at high risk for heart attacks.

6 Code and data availability

The source code can be accessed here: heartV2.r

The source data can be accessed here: heart.csv

7 Conclusion

With the topic of predicting heart disease with logistic regression using the R programming language, our team had a more intuitive view of how to extract data, process and analyze raw data, turn them into valuable data sources long-term, or better yet, being able to generalize the general situation and make predictions about the data set.

Besides, after learning R programming language and using IDE RStudio to apply analytic calculations and graphing, we significantly gain more skills in the programming process, know how to arrange the correct sequence of implementation and what to do when encountering a problem, as well as having more tools to support calculations and solve complex problems with the help of a computer. The cooperation in the implementation of the project has improved the ability, and the responsibility at work of each of the members.

Certainly, the process of implementing the project cannot definitely avoid any minor errors. Therefore, we are really looking forward to receiving comments and suggestions from the lecturer to carry out more accurate and professional topics in the foreseeable future.

Finally, we, as the authors of this project, hope that our solutions will satisfy the given problems and we wish you all the best.

Sincerely,

Members of Group 5 ./.

References

[1] Vijay K. Rohatgi, A. K. Md. Ehsanes Saleh, *An Introduction to Probability and Statistics*, 2015.

[2] David W. Hosmer, Stanley Lemeshow, *Applied Logistic Regression*, 2000.

[3] Josh Starmer, *The StatQuest Illustrated Guide to Machine Learning*, 2022.

[4] David Diez, Christopher Barr, Mine Cetinkaya-Rundel, *OpenIntro Statistics*, 2019.

[5] IBM, *What is logistic regression?*, Available at: https://www.ibm.com/topics/logistic-regression.

[6] Jonathan Bartlett, *The Hosmer-Lemeshow goodness of fit test for logistic regression*, Available at: https://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/

[7] Morten W. Fagerland, David W. Hosmer, *A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistoc regression models*, The Stata Journal, 2012.

[8] Nguyen Ngoc Binh, *Danh gia kiem dinh Hosmer-Lemeshow*, Available at: https://rpubs.com/nguyenngocbinhneu/hosmer_lemeshow_test

[9] Shaun Turney, *Pearson Correlation Coefficient (r) — Guide & Examples*, Available at: https://www.scribbr.com/statistics/pearson-correlation-coefficient

[10] Laerd Statistics, *Pearson Product-Moment Correlation*, Available at: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[11] John Uebersax, *McNemar Tests for Marginal Homogeneity*, Available at: https://john-uebersax.com/stat/mcnemar.htm

[12] Mayo Clinic, *Angina, Symptoms & Causes*, Available at: https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373