Analysis of Cardiovascular Diseases with R

Contents

1 Introduction

This project aims to help patients prevent the heart disease in advance and increase the accuracy of doctors' diagnosis by analyzing the relationship between characters of patients and whether a man have a heart disease or not.

According to the WHO, cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year, especially heart disease: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1).

The high death rate is since the symptoms of cardiac disease are not always visible and detectable, and it is difficult to draw adequate attention to them:(https://www.nhlbi.nih.gov/health-topics/espanol/enfermedad-coronaria-0).

Therefore, if we can figure out what features the heart disease is based on or find out which feature is more influential than others, it is possible we can prevent it before get any suffering. Our team tried to enable everyone to understand these relationships between different features of real patients and heart disease in the statistical method.

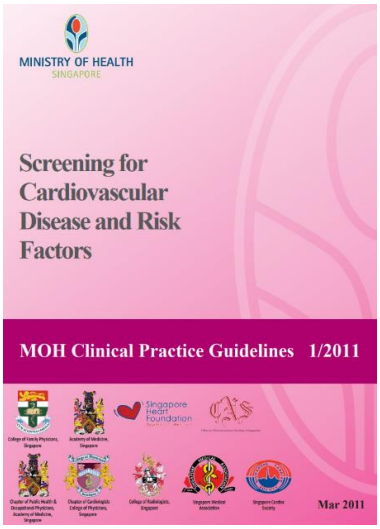
2 Overall Concept

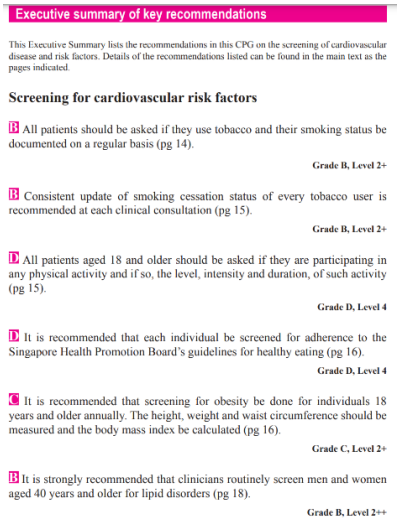
2.1 Research Objectives

This project focus on providing concise visual data analysis tools to potential heart disease patients to assess possible causal relationships between their symptoms and heart disease. Characteristics associated with heart disease in potential patients include age, sex, chest pain, resting blood pressure, Cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate, maximum heart rate, angina during exercise, the ST segment, the blood flow observed.

2.2 Research Innovation

At present, most information about heart disease on the Internet is presented in article form. For example, MOH (Ministry of Health) provides *Clinical Practice Guidelines: Screening for Cardiovascular Disease and Risk Factors* as the guidance on the use of various heart disease risk factors and analysis of test results.





Executive summary of key recommendations

This Executive Summary lists the recommendations in this CPG on the screening of cardiovascular disease and risk factors. Details of the recommendations listed can be found in the main text as the pages indicated.

Screening for cardiovascular risk factors

B All patients should be asked if they use tobacco and their smoking status be documented on a regular basis (pg 14).

Grade B, Level 2+

B Consistent update of smoking cessation status of every tobacco user is recommended at each clinical consultation (pg 15).

Grade B, Level 2+

D All patients aged 18 and older should be asked if they are participating in any physical activity and if so, the level, intensity and duration, of such activity (pg 15).

Grade D, Level 4

D It is recommended that each individual be screened for adherence to the Singapore Health Promotion Board's guidelines for healthy eating (pg 16).

Grade D, Level 4

C It is recommended that screening for obesity be done for individuals 18 years and older annually. The height, weight and waist circumference should be measured and the body mass index be calculated (pg 16).

Grade C, Level 2+

B It is strongly recommended that clinicians routinely screen men and women aged 40 years and older for lipid disorders (pg 18).

Grade B, Level 2++

Source:(https://www.moh.gov.sg/docs/librariesprovider4/guidelines/cpg_screening-for-cardiovascular-disease-mar-2011.pdf)

In comparison, we mainly use descriptive and Inferential Statistics to showcase our results. Descriptive Statistics includes Indicator statistics such as Range, mean, median, mode, standard deviation to make a comparison between indicators for patients and normal indicators. We also conduct shiny to display data including scatter, linear, pie, bar, and box line charts for each indicator. Inferential Statistics is also divided into variable correlation analysis and machine learning models. On the one hand we use ANOVA, Linear Regression Analysis, and Chi-square test to analysis relationship between different variables; On the other hand, we build three models to make future prediction and examine accuracy of these models.

3 Data Description

3.1 Data Source

This report will use dataset from a heart disease study on the Kaggle website. It comes from the UCI Machine Learning Repository, which recorded 300 patients from Cleveland and some characteristics related to heart disease.

Heart Disease UCI:(https://www.kaggle.com/ronitf/heart-disease-uci).

UCI Machine Learning Repository :(UCI Machine Learning Repository: Heart Disease Data Set).

Some variables can be easily interacted with the user, and they are shown in the table below.

Variable names	Meaning	Type
Target	whether the patient has a heart disease or not	Categorical
Age	Patient age in years	Numerical
Sex	Patient sex	Categorical
Cp	Chest pain type	Categorical
Trestbps	Resting blood pressure in millimeters of mercury (mm Hg) when the patient was admitted to the hospital	Numerical
Chol	Cholesterol level in mg/dl	Numerical
Fbs	Whether the level of sugar in the blood is higher than 120 mg/dl or not	Categorical
Restecg	Results of the electrocardiogram on rest	Categorical
Thalach	Maximum heart rate during the stress test	Numerical
Exang	Whether the patient had angina during exercise	Categorical
Oldpeak	Decrease of the ST segment during exercise according to the same one on rest	Numerical
Slope	Slope of the ST segment during the most demanding part of the exercise	Categorical
Thal	Results of the blood flow observed via the radioactive dye	Categorical
ca	Number of main blood vessels colored by the radioactive dye	Categorical

3.2 Data preparation

The original dataset consisted of 14 variables, including pathological features of heart disease and basic information about the patient. Each row represents information about one patient. All of the variables are of value to us, so we do not exclude any of them.

We import the “heart.csv” file into R and name it “data”. Use distinct function to remove duplicated records.

```
data <-distinct(heart)
```

Then, we drop records with NA values using codes below.

```
missing_ca_indeces <- which(data$ca %in% 4)
missing_thal_indeces <-which(data$thal %in% 0)
missing_values_indeces <- c(missing_ca_indeces, missing_thal_indeces)
data <- data[-missing_values_indeces, ]
```

For categorical variables, we transform data in the original dataset into R factors so that we could analyze them in a suitable way of categorical variables.

```
data$sex <- as.factor(data$sex)
data$cp <- as.factor(data$cp)
data$fbs <- as.factor(data$fbs)
data$restecg <- as.factor(data$restecg)
data$exang <- as.factor(data$exang)
data$slope <- as.factor(data$slope)
data$thal <- as.factor(data$thal)
data$target <- as.factor(data$target)
```

Since these categorical variables are used numerically for statistical convenience, we have changed the values to what they really represent, in order to facilitate our understanding and subsequent visual analysis.

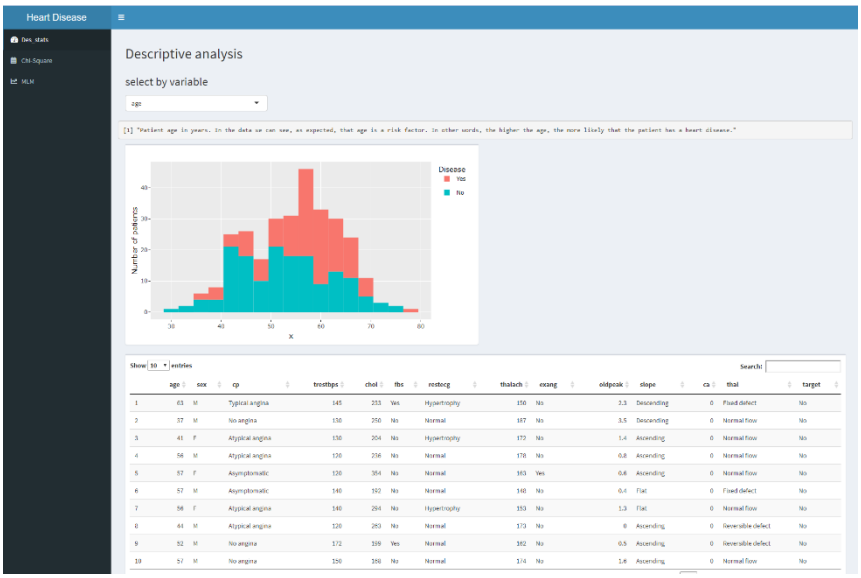
```
levels(data$sex) <- c("Female", "Male")
levels(data$cp) <- c("Asymptomatic", "Atypical angina", "No angina", "Typical angina")
levels(data$fbs) <- c("No", "Yes")
levels(data$restecg) <- c("Hypertrophy", "Normal", "Abnormalities")
levels(data$exang) <- c("No", "Yes")
levels(data$slope) <- c("Descending", "Flat", "Ascending")
levels(data$thal) <- c("Fixed defect", "Normal flow", "Reversible defect")
levels(data$target) <- c("Yes", "No")
```

We would use this cleaned data for data analysis and statistical testing.

3.3 Application Features

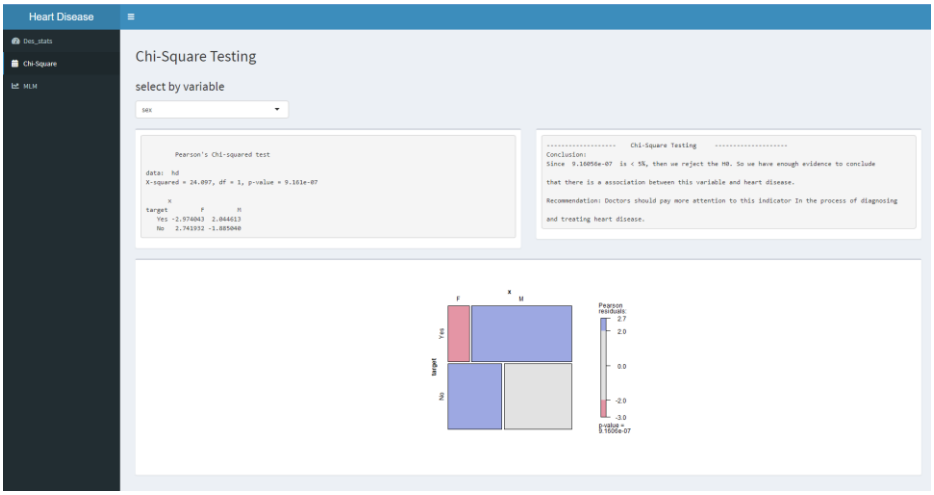
1.Visualize critical heart disease data based on user selected parameters.

By selecting a category of a key variable (e.g., Cp), users could explore the relationship between heart disease and other key variables (e.g., Chol, Age) under that category and this app would automatically generate a histogram. In the table, clear original data is provided for user's reference.



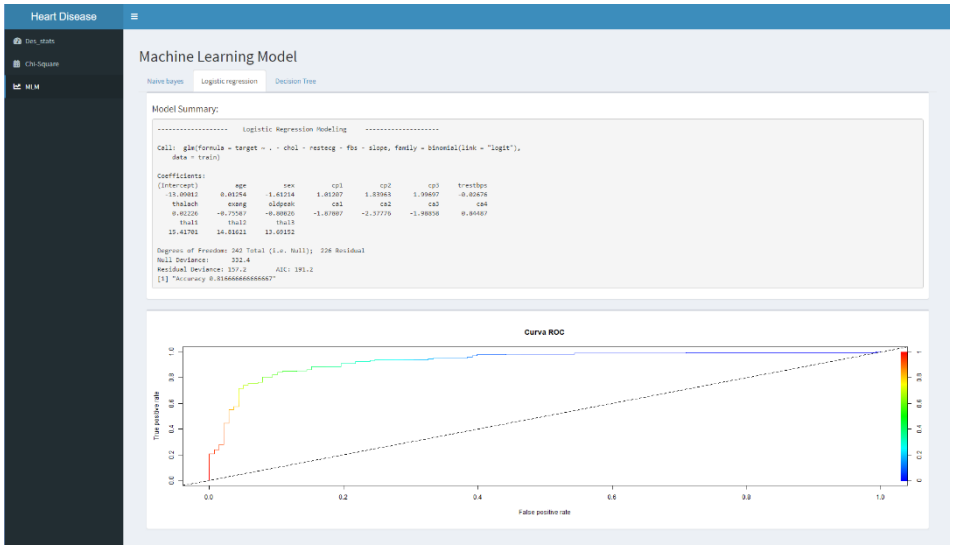
2. Demonstrate the relationship between variables and heart disease.

There are many indicators in cardiovascular diseases testing that are highly correlated with heart disease. By studying the extent to which these indicators contribute to determining the presence or absence of heart disease, we can obtain a more accurate picture of a patient's physical condition.



3. Comparing the training results of different machine learning models

We show here the training results of our different machine learning models. By comparing the accuracy of different models, we can choose the right model to help doctors predict the presence of heart disease in their patients.



4 Methodology

4.1 Methodology Overview

We mainly use descriptive and Inferential Statistics to showcase our results. Descriptive Statistics includes Indicator statistics such as Range, mean, median, mode, standard deviation to make a comparison between indicators for patients and normal indicators. We also conduct shiny to display data including scatter, linear, pie, bar, and box line charts for each indicator. Inferential Statistics is also divided into variable correlation analysis and machine learning models. On the one hand we use ANOVA, Linear Regression Analysis, and Chi-square test to analysis relationship between different variables; On the other hand, we build three models to make future prediction and examine accuracy of these models.

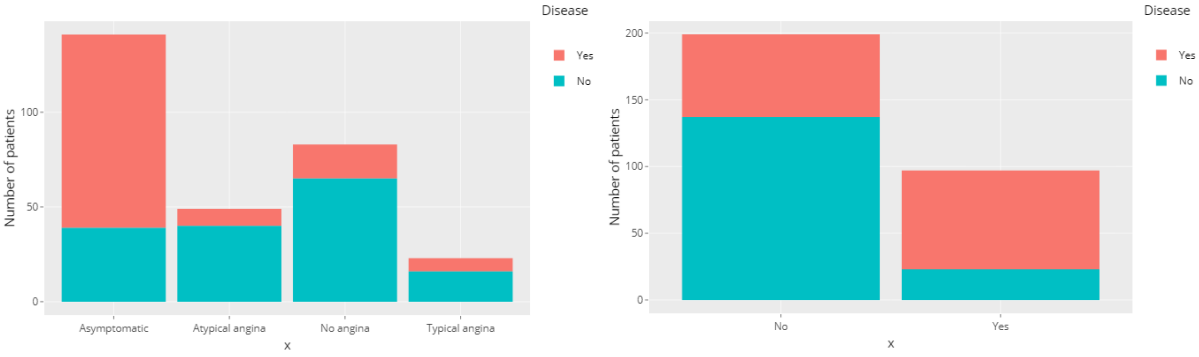
4.2 Descriptive Analysis – Charts

A bar charts is a chart that presents categorical data with rectangular bar with heights proportional to the value they present, which could be generally displayed vertically or horizontally. We could see the distribution of this variable clearly by showing bar charts.

A histogram is an approximate representation of the distribution of numerical data, which divide the entire range into a series of intervals and then count the number of each interval.

we choose some valuable ones to analyse.

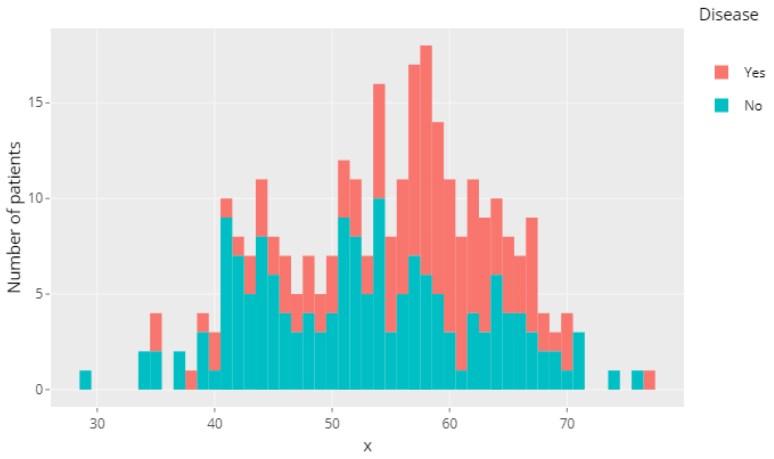
Categorical variables



In terms of cp (chest pain), patient with the type of Asymptomatic are more likely has heart disease than other three types. But it is hard to conclude that people with this type must have heart disease as other disease may also have this symptoms.

Exang is great indicator for the presence of heart disease. When Exang (yes), patients are more likely have heart disease.

Numerical variables



From the chart we can recognize that the age is a risky factor, the older the patient is, the more likely he has a heart disease. This is consistent with our expectation as old people are more likely to suffer from cardiovascular diseases due to the ageing of body tissues.

4.3 Inferential Analysis - Correlation Analysis

```
install.packages("corrplot")
library(corrplot)
cor=cor(data)
cor
```

We calculated multiple regression relationships between numerical variables, finding that the variables of thalach and age shows an important relationship whose Correlation coefficient is -0.3985219. we'd like to know whether the age has a significant impact on thalach, as well as whether there is significant difference in thalach between different age groups.

Firstly, we show its linear model.

```
> fit <- lm(heart_disease$thalach~heart_disease$age)
> summary(fit)

Call:
lm(formula = heart_disease$thalach ~ heart_disease$age)

Residuals:
    Min      1Q  Median      3Q     Max
-65.949 -11.954   3.975  15.921  44.985

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       204.2892     7.3485  27.800  < 2e-16 ***
heart_disease$age  -1.0051     0.1333  -7.539 5.63e-13 ***
---
```

We explore the difference between age groups by analysis of variance (ANOVA), we use code to group age into three parts, under 50 & 50~60 & above 60.

```
anova_age=data$age
anova_thalach=data$thalach
ano<-cbind(anova_age,anova_thalach)
ano_v<-ano[order(ano[,1]),]
ano_v[,2]

my_data <- data.frame("thalach" = ano_v[,2],
                      "age" = c(rep("under 50",85),rep("50-60",122),rep("above 60",89)))

library(dplyr)

group_by(my_data, age) %>%
  summarise(count = n(), mean = mean(thalach),sd = sd(thalach))
```

And then we show the result of ANOVA.

```
> # Summary of the analysis
> summary(res.aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
age           2  20436   10218   22.14 1.11e-09 ***
Residuals   293 135223     462
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> # Identify which one(s) is different from others
> TukeyHSD(res.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = thalach ~ age, data = my_data)

$age
                       diff        lwr       upr     p adj
above 60-50-60    -8.154817 -15.209460 -1.100174 0.0187457
under 50-50-60    13.317358   6.167383 20.467333 0.0000475
under 50-above 60 21.472174  13.797165 29.147184 0.0000000
```

Because the p-value <0.001, we have enough evidence to support that there is a significant difference between different groups. And TukeyHSD show that the difference between age under 50 and above 60 is the most significant.

So, the essential insight we found is that, for a specific pathological features, older people and young people would have differently numerical range of normal number, so we should take age into consideration seriously in the diagnosis and treatment of heart disease.

4.4 Inferential Analysis - Chi-square Test

In our report, we used the chi-square test to examine the association between different categorical variables and whether the patient had heart disease or not.

We use x-tabs function to transform data frame to a contingency table.

```
> hd<-xtabs(~target+cp,data=data)
> hd
      cp
target Asymptomatic Atypical angina No angina Typical angina
   Yes          102                9        18              7
   No            39               40        65             16
```

Firstly, we choose to exam the association between cp (chest pain) between target.

The X^2 of independence between these two variables could be found by “chisq.test”:

```
> Result <- chisq.test(hd,correct=FALSE)
> print(Result)

        Pearson's Chi-squared test

data:  hd
X-squared = 76.454, df = 3, p-value < 2.2e-16
```

We can see that R shows the X-squared=76.454, the degree of freedom=3, and the p-value<2.2e-16. The p-value is highly significant in statistic (p<0.001), which indicates that there is a significant difference in the type of chest pain between heart disease patients and the general person.

We also want to know the extent to which different types of chest pain contribute to the x-square. We use Results $ residuals to get specific data.

```
> Result$residuals
          cp
target     Asymptomatic Atypical angina No angina Typical angina
   Yes        4.623800       -2.848044 -3.260558       -1.097450
   No        -4.262933        2.625767  3.006086        1.011799
```

As show above. There is an obvious number difference in Asymptomatic, indicating that people with Asymptomatic are more likely to have heart disease. **So, we could draw a conclusion that the Asymptomatic is a key indicator to the heart disease, once a patient has this type of chest pain, the doctor should recommend him to do further screening for heart disease.**

After that, we'd also like to know the correlation between other categorical variable and heart disease. Since the number of variables is not very large, we run a chi-square test on each of them

```
      fbs
target  No Yes
   Yes 116  20
   No  137  23
> Result <- chisq.test(hd,correct=FALSE)
> print(Result)

        Pearson's Chi-squared test

data:  hd
X-squared = 0.006482, df = 1, p-value = 0.9358

> Result$residuals
      fbs
target          No         Yes
   Yes -0.02256093  0.05472465
   No   0.02080015 -0.05045363
```

```
      restecg
target Hypertrophy Normal Abnormalities
   Yes          78     55            3
   No           67     92            1
> Result <- chisq.test(hd,correct=FALSE)
Warning message:
In chisq.test(hd, correct = FALSE) :
  Chi-squared approximation may be incorrect
> print(Result)

        Pearson's Chi-squared test

data:  hd
X-squared = 9.2624, df = 2, p-value = 0.009743

> Result$residuals
      restecg
target Hypertrophy     Normal Abnormalities
   Yes   1.3940321 -1.5259278     0.8572611
   No   -1.2852341  1.4068359    -0.7903557
```

```
      exang
target  No Yes
   Yes  62  74
   No  137  23
> Result <- chisq.test(hd,correct=FALSE)
> print(Result)

        Pearson's Chi-squared test

data:  hd
X-squared = 53.486, df = 1, p-value = 2.604e-13

> Result$residuals
      exang
target         No        Yes
   Yes -3.078052  4.408762
   No   2.837824 -4.064678
```

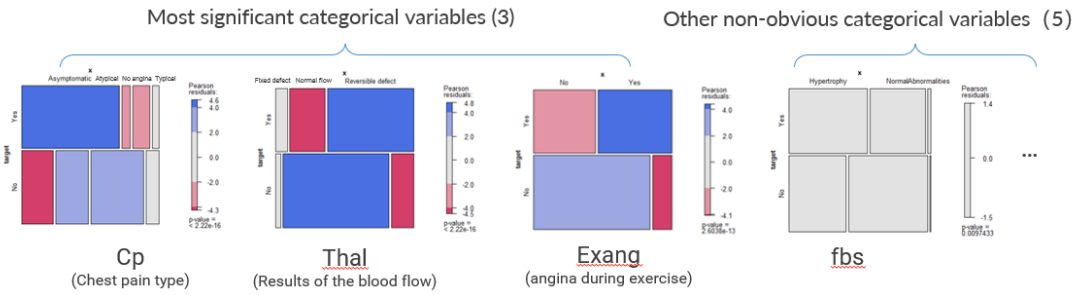
As we can see above, we are able to draw the following conclusion.

the p-value of fbs (Whether the level of sugar in the blood is higher than 120 mg/dl or not) =0.9358, and the residuals of this variable between patients and general person is not highly significant in statistical. So, it seems to be a useless indicator in diagnosis of heart disease.

For testing about restecg (Results of the electrocardiogram on rest), we can see R send a warning message, which may be due to the fact that more than 20% of the variables in the contingency table have an expected frequency of less than 5, so we may need

more samples to test this variable.

Finally, we get three significant indicators for heart disease by Chi-Square Testing.



4.5 Inferential Analysis - Logistic Regression

By logistic regression, we can predict whether the patient has heart disease or not. Firstly, we check for missing values by using sapply() and apply(). There is no missing value in the dataset.

```
> sapply(data,function(x) sum(is.na(x)))
  锘緁ge       sex        cp trestbps      chol       fbs   restecg
       0         0         0         0         0         0         0
 thalach     exang   oldpeak     slope        ca      thal    target
       0         0         0         0         0         0         0
> 
```

```
> apply(data,2,anyNA)
  锘緁ge       sex        cp trestbps      chol       fbs   restecg
   FALSE     FALSE     FALSE     FALSE     FALSE     FALSE     FALSE
 thalach     exang   oldpeak     slope        ca      thal    target
   FALSE     FALSE     FALSE     FALSE     FALSE     FALSE     FALSE
```

We used as.factor() to encode the variables as factors in data preparation.

We used 80% of data as the training set and 20% of data as the testing set.

```
train <- data[60:302,]
test <- data[1:60,]
```

Then we run the logistic regression model.

```
model1=glm(formula=target~.,  family="binomial"(link='logit'),data=train)
```

We used anova(). By looking at the P-values, we can infer that chol , *fbs, restecg and slope* are not statistically significant. We should remove them.

```
> anova(model1, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: target

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       236     323.36
锘緁ge     1   12.565       235     310.80 0.0003931 ***
sex        1   28.441       234     282.36 9.662e-08 ***
cp         3   59.677       231     222.68 6.889e-13 ***
trestbps   1    6.100       230     216.58 0.0135210 *
chol       1    1.771       229     214.81 0.1832912
fbs        1    0.200       228     214.61 0.6543679
restecg    2    1.766       226     212.84 0.4134576
thalach    1    9.476       225     203.37 0.0020819 **
exang      1    2.723       224     200.65 0.0989392 .
oldpeak    1   18.131       223     182.51 2.062e-05 ***
slope      2    2.089       221     180.43 0.3519462
ca         3   27.832       218     152.59 3.938e-06 ***
thal       2    6.773       216     145.82 0.0338321 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We exclude chol, fbs, restecg and slope to create a new model.

From the result below we can figure out that *sex*, *cp*, *trestpbs*, *oldpeak* and ca are statistically significant. *cp* and *ca* have the lowest p-value. Given that *target*=1 means the patient doesn't have heart disease, the positive coefficient for *cp*2 predictor suggests that all other variables being equal, the patient who has pain without relation to angina is less likely to have heart disease. the negative coefficient for *ca*1 predictor suggests that all other variables being equal, the patient who has one main blood vessel colored by the radioactive dye is more likely to have heart disease.

```
> model2=glm(formula=target~.-chol-restecg-fbs-slope,  family="binomial"(link='logit'),data=train)
> summary(model2)

Call:
glm(formula = target ~ . - chol - restecg - fbs - slope, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.49202  -0.40994  -0.09251   0.46534   2.84832

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.427170   2.788659   0.870 0.384097
锘縜ge       0.009561   0.026456   0.361 0.717803
sex         -1.540858   0.550240  -2.800 0.005105 **
cp1          0.995754   0.596586   1.669 0.095100 .
cp2          1.873598   0.564585   3.319 0.000905 ***
cp3          2.010854   0.802731   2.505 0.012245 *
trestbps    -0.026026   0.012364  -2.105 0.035292 *
thalach      0.021193   0.011999   1.766 0.077375 .
exang       -0.655866   0.502375  -1.306 0.191713
oldpeak     -0.812911   0.249660  -3.256 0.001130 **
ca1         -1.932776   0.538836  -3.587 0.000335 ***
ca2         -2.319929   0.746466  -3.108 0.001884 **
ca3         -1.997220   0.987575  -2.022 0.043140 *
thal2       -0.560506   0.912916  -0.614 0.539233
thal3       -1.679856   0.894321  -1.878 0.060332 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 323.36  on 236  degrees of freedom
Residual deviance: 154.82  on 222  degrees of freedom
  (因为不存在，6个观察量被删除了)
AIC: 184.82


Number of Fisher Scoring iterations: 6
```

We assessed the predictive ability of the model. Decision boundary will be 0.5, means if the probability of the patient having a heart disease is greater than 0.5, the model well predicts target=1, this patient doesn't have heart disease, the accuracy is 81.67%

```
> fitted.values <- predict(model2,newdata=test,type='response')
> fitted.values <- ifelse(fitted.values > 0.5,1,0)
> ClassificError <- mean(fitted.values != test$target)
> print(paste('Accuracy',1-ClassificError))
[1] "Accuracy 0.816666666666667"
```

Finally, by using library "ROCR" we plot the ROC curve and calculate the AUC. This model has the good predictive ability because the AUC is 0.93.

```
> library(ROCR)
> prediction=predict(model2,data,type="response")
> pred = prediction(prediction, data$target)
> perf = performance(pred,"tpr", "fpr")
> plot(perf,colorize = TRUE,main="Curva ROC")
> abline(0, 1, lty = 2)
> (performance(pred,"auc")@y.values)[[1]]
[1] 0.9281548
```

4.6 Naive Bayes

As part of the Inferential Statistics of the project we applied Naive Bayes to help predict whether a person has heart disease, relying on several factors, such as age, chest pain type, fasting blood sugar, maximum heart rate achieved, etc. Here the dataset has 303 observations and 14 variables shown as follows:

We start by setting random numbers which is used to ensure that the results remain unchanged during each iteration. Secondly, we divided the dataset into two parts: training and testing. We choose 80% of dataset for training and the other 20% for testing dataset.

We observed the details of our model, and we find categorical data which is different from numeric data. The former is still the traditional probability, the first field of the latter is the average value, and the second field is the standard deviation.

```
nb_default <- naiveBayes(target ~ ., data=train)
```

```
> nb_default

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.4504132 0.5495868
```

```
   ca
Y        [,1]      [,2]
  0 1.1376147 1.0581946
  1 0.3984962 0.9123716

   thal
Y       [,1]      [,2]
  0 2.541284 0.7009532
  1 2.090226 0.4680496
```

Then we used R code as follows to make Bayes' prediction for validation and reserved three decimal and we could see from the following examples that the probability has been normalized.

```
> test.y_hat <- predict(nb_default, test, type="class")
> test.y_hat_prob <- round(predict(nb_default, test, type="raw"),3)
>
> cbind(Prediction=as.character(test.y_hat), test.y_hat_prob)
```

```
      Prediction 0       1
 [1,] "0"        "0.884" "0.116"
 [2,] "1"        "0.001" "0.999"
 [3,] "1"        "0.014" "0.986"
 [4,] "1"        "0.061" "0.939"
 [5,] "1"        "0.001" "0.999"
 [6,] "0"        "0.736" "0.264"
 [7,] "0"        "0.51"  "0.49"
 [8,] "1"        "0.102" "0.898"
 [9,] "1"        "0.302" "0.698"
[10,] "1"        "0.264" "0.736"
[11,] "0"        "0.528" "0.472"
[12,] "1"        "0.395" "0.605"
[13,] "1"        "0.024" "0.976"
[14,] "1"        "0.001" "0.999"
[15,] "1"        "0.341" "0.659"
[16,] "1"        "0.001" "0.999"
```

We then evaluated the accuracy of the Bayes model, and the value could as high as 80%.

```
> #Bayes model evaluation
> accuracy.nb_default <- sum(test.y_hat==test$target) / length(test$target)
>
> accuracy.nb_default
[1] 0.8032787
```

We finally built a cross table to compare the actual situation and the Bayesian model prediction. From the result, we found the accuracy of the model is very high at 80% percent; and for testing part, it can also be found that the established Naive Bayes model has high predictive performance. The target is actual to be 0, 24 predicted to be 0, and only 5 are predicted to be 1; and the target are actual to be 1, there are 25 predicted to be 1, and 7 predicted to be 0.

```
> table(test.y_hat, test$target, dnn=c("Prediction","Actual"))
          Actual
Prediction  0  1
         0 24  7
         1  5 25
```

Here, we choose Naive Bayes to make prediction due to the small size of the sample and the characters are independent to each other. And the Pros Naive Bayes model are:

Firstly, it is simple and convince to forecast the test data set and it also good at multi-class prediction which was used in this project. secondly, a Naive Bayes model outperforms alternative models such as logistic regression and decision tree as well as the less training data was needed when independence hypothesis holds. thirdly, in comparison to numerical input variables, it performs well with categorical input variables (s). A normal distribution is assumed for numerical variables (bell curve, which is a strong assumption).

However, the important limitation of Naive Bayes is the assumption of independent predictors. In actual life, getting a collection of predictors that are totally independent is sometimes impossible.

4.7 Decision Tree

Decision tree models use classification to make predictions about the model, and the basis and results of the classification are very straightforward for data users.

```
output$lb3_2 <- renderPlot({
  heart <- data
  ind <- createDataPartition(heart$target,times = 1,p=0.75,list = F)
  heart_train <- heart[ind,]
  heart_test <- heart[-ind,]
  fitControl <- trainControl(method="cv", number=10)
  set.seed(8)
  model.tree <- train(target ~ .,
                      data = heart_train,
                      method = "rpart",
                      trControl = fitControl)
  plot(model.tree)
```

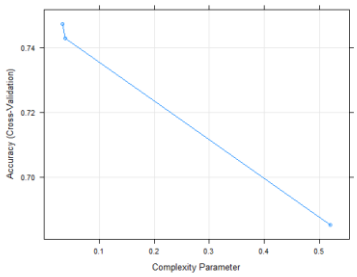
```
------------------    Desicion Tree Modeling    ------------------
CART

222 samples
 13 predictor
  2 classes: 'Yes', 'No'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 199, 200, 200, 199, 200, 200, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.03921569  0.7525692  0.5029648
  0.06372549  0.7211462  0.4354312
  0.50000000  0.6256917  0.2091210

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.03921569.
```



5 Conclusion

5.1 Recommendation

After all those analyses, we got some conclusions and made the following three recommendations

1. Normal indicators and abnormal indicators at different ages are not the same. Doctors need to pay attention to indicators-related variables, such as age when judging whether they are abnormal.
2. In the examination and screening of heart disease, doctors and patients can focus on these indicators: Chest pain type, Results of the blood flow, and angina during exercise.
3. After comparing different machine learning models, we believe that for this data set, logistic regression has the highest accuracy, and doctors can use this model to predict their patient's heart disease.

5.2 Limitations

However, our research still has some limitations. We did not have multiple datasets, so it does not represent the overall level very accurately. Our dataset only recorded a

sample size of 296, which is inadequate in terms of quantity. And we only used 13 variables in determining whether a patient had heart disease. What's more, the impact of related diseases and regional differences had not been considered. Apart from that, most of the data in the dataset are categorical and lack sufficient quantitative analysis, making the model less accurate.

5.3 Future Work

Due to those limitations, there are also some improvements for future work that can be done in several aspects.

1. For the sample size, we plan to increase the number of hospitals selected and conduct multiple tests on the same sample to obtain more accurate data and get more medical data indicators.
2. For the other related diseases, more datasets of other diseases can be brought in to find the relationship and influence on heart disease. Such as hyperlipidemia, hypertension, and so on.
3. For the different regional research, take heart disease datasets for different regions, such as developing countries, into consideration and explore the impact of different regions on heart disease.

6 REFERENCES

Michy Alice. (13 September, 2015), R bloggers: How to perform a Logistic Regression in R

Retrieved from: https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/

Jake VanderPlas. (November, 2016), Python Data Science Handbook: In Depth: Naive Bayes Classification

Retrieved from: https://jakevdp.github.io/PythonDataScienceHandbook/05.05-naive-bayes.html

Clemens R., Peter F., Robert G., and Rudolf D. (4 April, 2008), Statistical Data Analysis Explained: Applied Environmental Statistics with R

Retrieved from:chttps://onlinelibrary.wiley.com/doi/book/10.1002/9780470987605

Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R, Packt Publishing 2017