CPE 213
Data Models

Midterm Project Proposal

Title:
Heart Disease Analysis

Submitted by

Ms. Jidapa Thongnirun
ID: 63070503462

Submitted to

Dr. Santitham Prom-on

Semester 2/2021

King Mongkut's University of Technology Thonburi (KMUTT)

Tables of content

Abstract

Heart disease is a fatal illness. It is the leading cause of death globally, taking an estimated 17.9 million lives each year. By knowing this information, we can make a better decision on how to continue their life. The dataset that we are using for heart disease analysis is named “Personal Key Indicators of Heart Disease” 1984. This dataset is based on the CDC (Centers for Disease Control and Prevention) which started collecting data on heart attacks in 1984 and the most recent data is written in 2020. There are many ways that we can use this information to analyze vital data and summarize the huge dataset into one single analysis, for example, spotting the trend of a person who has heart disease on some other variable or behavior like sleep time, BMI, or age range. The analytic technique that we use is separated into 2 types, relation and distribution, and distribution also divided into two types frequency distribution and probability distribution.

Introduction

Heart disease is a fatal illness. It is the leading cause of death globally, taking an estimated 17.9 million lives each year. It is unable to be seen with the naked eye, so being aware of the risk of contracting the disease can mean the difference between life and death. By analyzing the variables that may have significance to having a heart attack, and representing the analyzed information visually and numerically. The said information may help someone know the trend of them having heart disease. And by knowing this information, they can make a better decision on how to continue their life.

The objective of this analysis is to understand the factors that heart disease may tend to increase the rate which differ from various types of daily activities and it is also analyzed for the risk of each individual person who would get the heart disease in the older age.

Problem Statement

As the dataset on heart disease has been given. There are many ways that we can use this information to analyze the vital data and summarize the huge dataset into one single analysis. We can spot the trend of a person who has heart disease based on some other variable or behavior like sleep time, BMI, or age range. And we can also determine which behavior has a strong correlation to having a heart disease or which didn't. And we can also calculate the risk of heart disease based on certain behavior by using the probability distribution.

Proposed Analytic Technique

From the dataset we acquired, the analytic technique that we use is separated into 2 types, distribution and relation, the distribution will be used when there is a need for frequency stuff and relation will be used when we need to determine the relation between two variables.

First is to spot the trend of a person who has heart disease on some other variable or behavior like sleep time, BMI, or age range. We'll use the distribution technique by cross-tabulation of the two variables, heart disease, and the other interest variable. Then we'll use the stacked bar chart to create a visualization of the trend for us.

Second is to determine which behavior has a strong correlation to having a heart disease or which didn't. In this part we'll still use the relation technique but we'll use the chi-square table after we have done the cross tabulation to the two variables, heart disease, and the other interest variable. And from the chi-square table we'll be able to determine whether those two variables are related or not.

The last one is to calculate and display the risk of heart disease based on certain behavior. In this part, we will use the distribution technique for both frequency and probability. By creating the probability distribution function. Then we'll be able to be informed of the risk of having heart disease by having said behavior.

Data Source Description

The dataset that we are using for heart disease analysis is “Heart Disease Dataset” which was uploaded by David Lapp on Kaggle. Cleveland, Hungary, Switzerland, and Long Beach V are the four databases in this data set, which dates back to 1988. Although there are 76 attributes in total, including the anticipated attribute, all published experiments only use a subset of 14 of these, which are the header or field of table of this dataset. The "target" field indicates whether or not the patient has heart disease. It is integer-valued, with 0 indicating no disease and 1 indicating disease. There are a total 1025 rows in this dataset.

This data set contains 14 columns, so there are age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target.

For the details of each columns that we used to analyze are

1. age: patient’s age in years
2. sex: patient’s gender which (1 =male, 0 =female)
3. cp: patient’s chest pain type
    a. Value 0: asymptomatic or no symptoms.
    b. Value 1: atypical angina
    c. Value 2: non-anginal pain
    d. Value 3: typical angina
4. trestbps: patient’s resting blood pressure (measurement in mmHg)
5. chol: patient’s cholesterol measurement in mg/dl
6. fbs: patient’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. restecg: patient’s resting electrocardiographic results
    a. Value 0: showing probable or definite left ventricular hypertrophy by Estes’ criteria
    b. Value 1: normal
    c. Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
8. thalach: patient’s maximum heart rate achieved
9. exang: Exercise-induced angina (1 = yes; 0 = no)

10. oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot.)
11. slope: the slope of the peak exercise ST segment 0: downsloping; 1: flat; 2: upsloping
12. ca: The number of major vessels (0–4)
13. thal: A blood disorder called thalassemia; 3 = normal; 6 = fixed defect; 7 = reversible defect
14. target: diagnosis of heart disease (0= no disease, 1= disease)

Data Preparation

First, we import the dataset and start to clean the dataset. Fortunately, this dataset does not contain any empty cells, so we do not have to remove NA out.

```
colnames(data)[which(names(data) == "target")] <- "hd"
```

Next, we changed the name of the “target” column to “hd”.

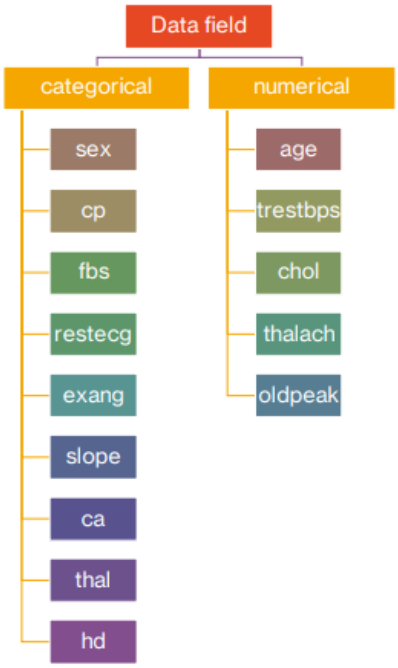
```
data$sex <- ifelse(test = data$sex == 1, yes = "M", no = "F")
```

To make the data easier on the eyes, in Sex column, we convert 0 to “F” for female and convert 1 to "M" for male.

```
data$hd <- ifelse(test = data$hd == 1, yes = "Y", no = "N")
```

In the hd column, we convert 0 to “N” for those who do not have heart disease and convert 1 to "Y" for those who have heart disease.

Data Exploration



This is the diagram of variables in this dataset. The types of data are separated into 2 types which are categorical or numerical. A categorical variable contains data that fits into multiple categories, otherwise a numerical variable is a number. In this dataset, sex, cp, fbs, restecg, exang, slope, ca, thal, and hd are categorical variables whereas the remaining variables are numerical variables. This dataset contains 14 column and 1025 rows.

Data Visualization

```
> summary(data)
      age        sex     cp          trestbps          chol          fbs     restecg      thalach         exang       oldpeak        slope
 Min.   :29.00   F:312   0:497   Min.   : 94.0   Min.   :126   0:872   0:497   Min.   : 71.0   0:680   Min.   :0.000   0: 74
 1st Qu.:48.00   M:713   1:167   1st Qu.:120.0   1st Qu.:211   1:153   1:513   1st Qu.:132.0   1:345   1st Qu.:0.000   1:482
 Median :56.00           2:284   Median :130.0   Median :240           2: 15   Median :152.0           Median :0.800   2:469
 Mean   :54.43           3: 77   Mean   :131.6   Mean   :246                   Mean   :149.1           Mean   :1.072
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:275                   3rd Qu.:166.0           3rd Qu.:1.800
 Max.   :77.00                   Max.   :200.0   Max.   :564                   Max.   :202.0           Max.   :6.200
 ca      thal    hd
 0:578   0:  7   N:499
 1:226   1: 64   Y:526
 2:134   2:544
 3: 69   3:410
 4: 18
```

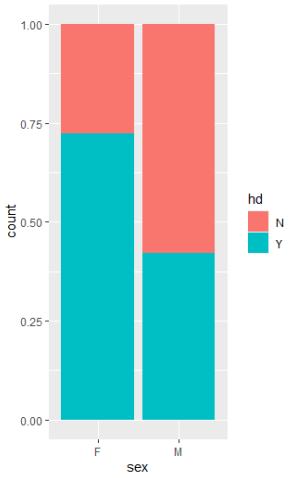
```
> str(data)
'data.frame':   1025 obs. of  14 variables:
 $ age     : int  52 53 70 61 62 58 58 55 46 54 ...
 $ sex     : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 2 2 2 2 ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ trestbps: int  125 140 145 148 138 100 114 160 120 122 ...
 $ chol    : int  212 203 174 203 294 248 318 289 249 286 ...
 $ fbs     : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 1 1 1 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 2 1 2 2 2 1 3 1 1 1 ...
 $ thalach : int  168 155 125 161 106 122 140 145 144 116 ...
 $ exang   : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
 $ oldpeak : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
 $ slope   : Factor w/ 3 levels "0","1","2": 3 1 1 3 2 2 1 2 3 2 ...
 $ ca      : Factor w/ 5 levels "0","1","2","3",..: 3 1 1 2 4 1 4 2 1 3 ...
 $ thal    : Factor w/ 4 levels "0","1","2","3": 4 4 4 4 3 3 2 4 4 3 ...
 $ hd      : Factor w/ 2 levels "N","Y": 1 1 1 1 1 2 1 1 1 1 ...
```

To spot the trend of a person who has heart disease on some other variable. We use the distribution technique by cross-tabulation of two variables, hd, and the other interest variable. Then we use the stacked bar chart to create a visualization of the trend.

For the categorical variables:

1. hd_sex <- xtabs(~ hd + sex, data = data)

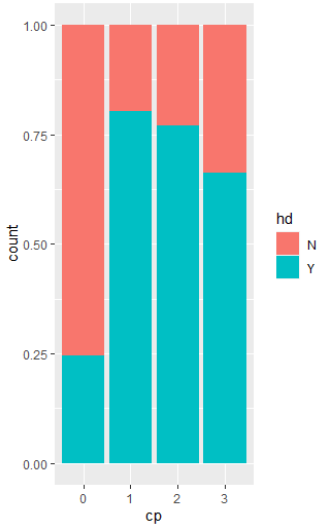
```
    sex
hd    F   M
  N  86 413
  Y 226 300
```



This is a cross-tabulation and stacked bar chart of hd and sex. It shows the number of females and male with and without heart disease. The data is saying that females got higher cases than male.

2. hd_cp <- xtabs(~ hd + cp, data = data)

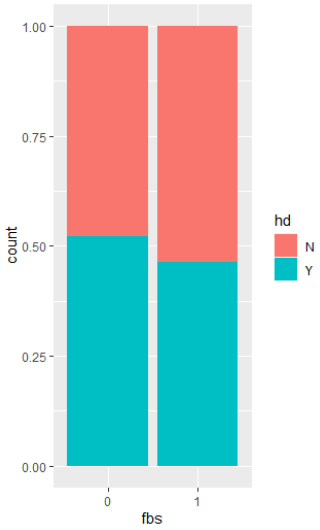
hd	cp 0	1	2	3
N	375	33	65	26
Y	122	134	219	51



This is a cross-tabulation and stacked bar chart of hd and Patient's chest pain type or cp. It shows the number of people with each chest pain type to be diagnosed with heart disease and without disease. The people without heart disease appear to have much fewer cases in variations 1, 2, and 3.

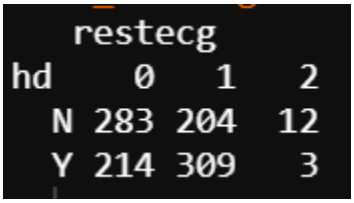
3. hd_fbs <- xtabs(~ hd + fbs, data = data)

hd	fbs 0	1
N	417	82
Y	455	71

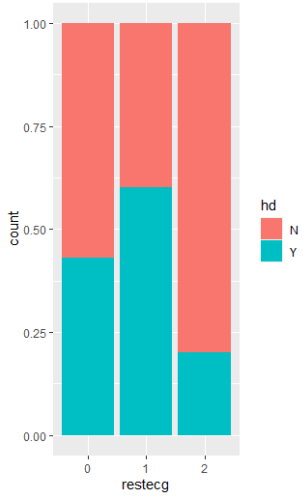


This is a cross-tabulation and stacked bar chart of hd and fbs. It shows the number of people with and without high fasting blood sugar to be diagnosed with heart disease and without disease. The fbs almost appear to have the same proportions for both patients with heart disease and without heart disease.

4. hd_restecg <- xtabs(~ hd + restecg, data = data)

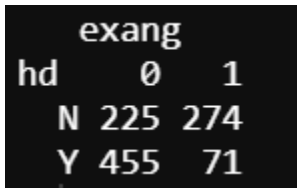


hd	restecg 0	1	2
N	283	204	12
Y	214	309	3

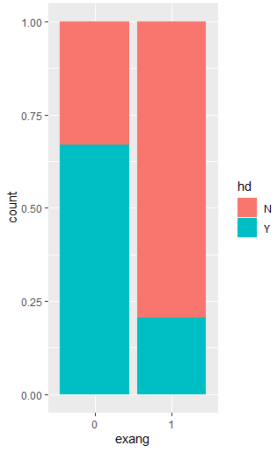


This is a cross-tabulation and stacked bar chart of hd and rest ecg. There are about 10% more people with heart disease having a definite left ventricular hypertrophy. And about 40% more people without heart disease have a restecg normal.

5. hd_exang <- xtabs(~ hd + exang, data = data)



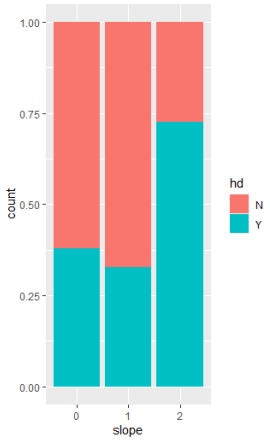
hd	exang 0	1
N	225	274
Y	455	71



This is a cross-tabulation and stacked bar chart of hd and exercise-induced angina. It shows the number of people with and without exercise-induced angina to be diagnosed with heart disease and without disease. 80% of not heart disease patients are not getting it for exercising too much.

6. hd_slope <- xtabs(~ hd + slope, data = data)

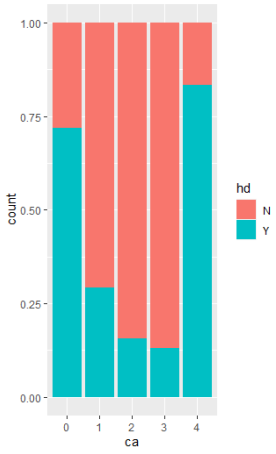
```
   slope
hd    0   1   2
  N  46 324 129
  Y  28 158 340
```



This is a cross-tabulation and stacked bar chart of hd and slope. It shows the number of people with each type of slope of the peak exercise ST segment to be diagnosed with heart disease and without disease. It shows that people who got upsloping in the peak exercise ST segment got more chances of having heart disease.

7. hd_ca <- xtabs(~ hd + ca, data = data)

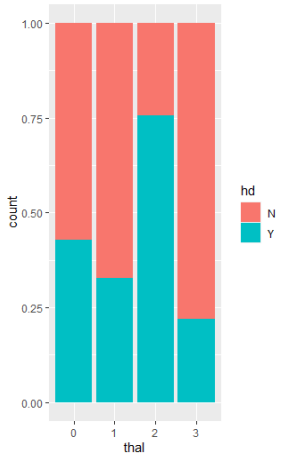
```
   ca
hd    0   1   2   3   4
  N 163 160 113  60   3
  Y 415  66  21   9  15
```



This is a cross-tabulation and stacked bar chart of hd and ca. It shows the number of people with each major blood vessel to be diagnosed with heart disease and without disease. People who have 4 major blood vessels and 0 major blood vessels tend to have more chances of having heart disease.

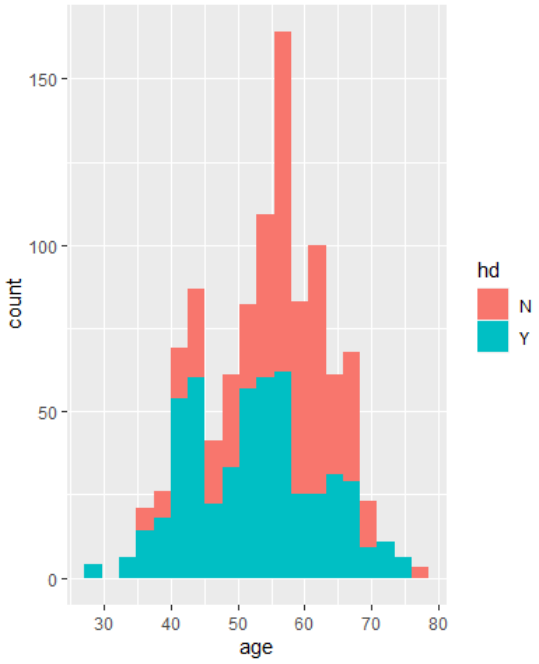
8. hd_thal <- xtabs(~ hd + thal, data = data)

```
    thal
hd     0   1   2   3
  N    4  43 132 320
  Y    3  21 412  90
```

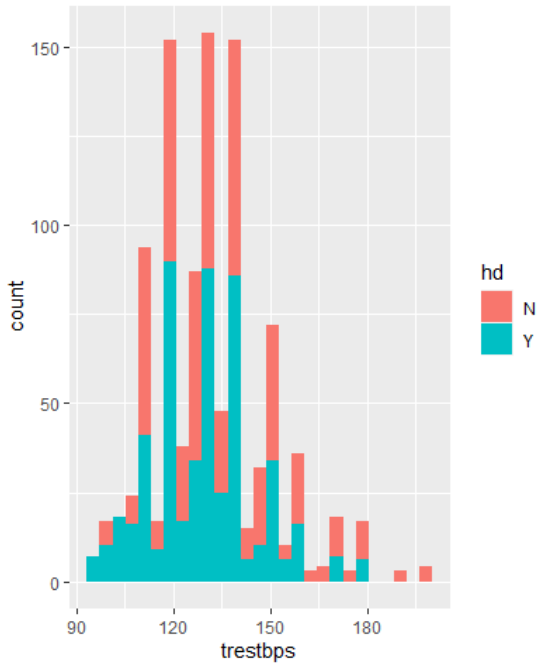


This is a cross-tabulation and stacked bar chart of hd and thal. It shows the number of people with each type of thalassemia to be diagnosed with heart disease and without disease. The data is saying that people with fixed defect thalassemia have higher chance of heart disease.

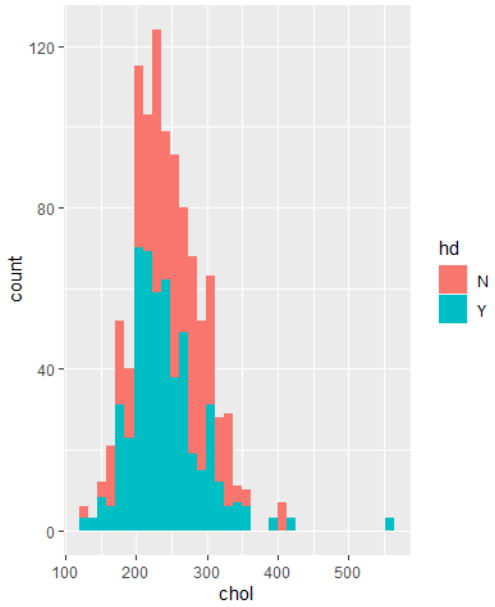
For the numerical variables:



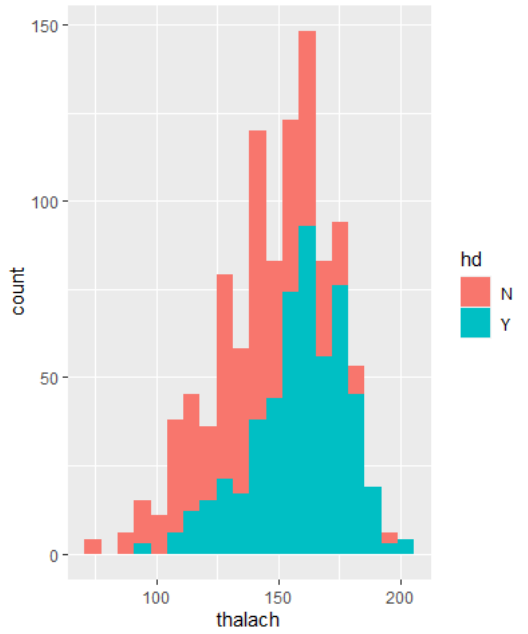
In this graph, it is the relationship between hd and age. As you can see People over 55 have a lower risk of getting heart disease.



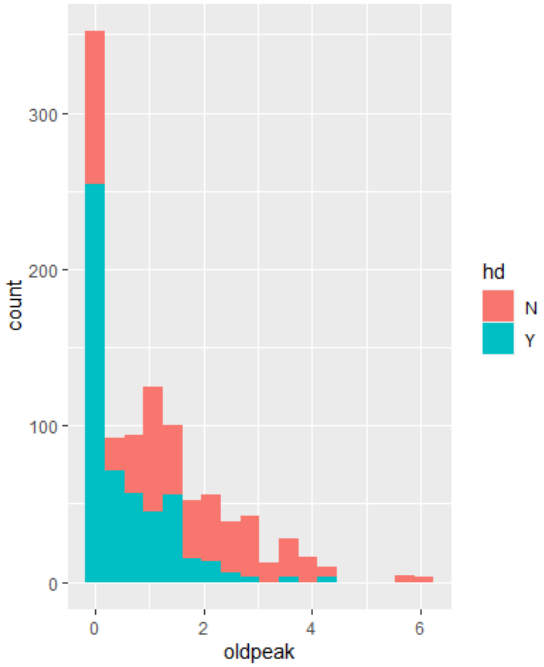
In this graph, it is the relationship between hd and trestbps (Patient's resting blood). The graph of patient's resting blood pressure of people with heart disease and without disease appear to have the same shape. However, patients with heart disease have higher resting blood pressure levels.



Next is the relationship between hd and cholesterol. People with high cholesterol levels of 200 to 260 are more likely to be diagnosed with heart disease.



Next is the relationship between hd and thalach (patient's maximum heart rate). People with a high maximum heart rate are more likely to be diagnosed with heart disease.



Next is the relationship between hd and oldpeak. As the value of oldpeak increases, the number of people with heart disease decreases.

Model Explanation

About model explanation, this method will show about what data model that we selected or used for create the predictive model so, since the heart disease data that we used in this project are related numerical and categorical data and each of them are related in each other so, we use logistic regression to create predictive model.

Why we use logistic regression and change some part in data set to catecorical instead of using the raw data that are all numerical and use linear regression to create a predictive model because when we use yes and no as 1 and 0, it is not appropriate way to fix the range of value and also want to decrease a negative range of value

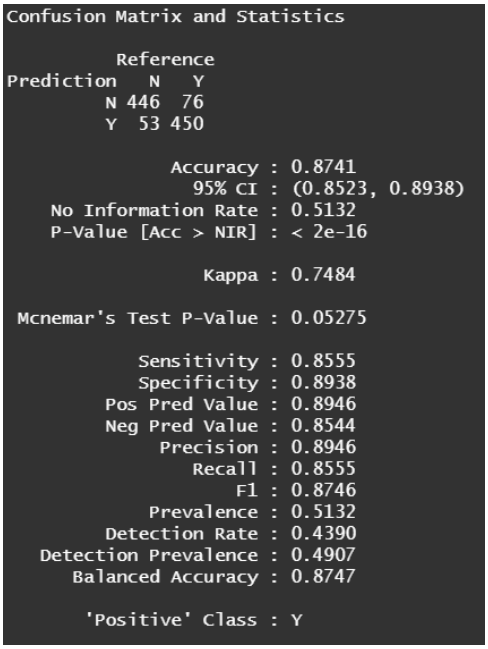
After we have explored and visualized our data. We start by selecting the appropriate model that will be used. In this case, our problem is about finding the probability of having heart disease or not. Which means the outcome of having the disease or not is a discrete outcome. So, logistic regression, which is the process of modeling the probability of a discrete outcome given an input variable, is the most appropriate model for our problem.

Model Implementation

From the model explanation that we chose and selected what type of data model is most suitable and appropriate for applying to create a predictive model, next step we will make a modeling implementation for fitting our model. To train the predictive model we use each variable in the dataset and then apply cross validation to prevent overfitting of the predictive model and use the model of logistic regression to train the model. From this step, when we train the model we will get the correct predictive value than before. After that we use a confusion matrix to evaluate the model to see the relationship between each variable and heart disease.

As for how we implement our logistic regression model into our work, we use K-Fold cross validation to ensure that our model is not overfitting. In R, we can implement this part by using the train control functions that come with the caret packages. By providing the method as "cv" for cross validation and providing how many folds we want to operate when training.
We've successfully created the control for training. It's time to train our data. In this part, we will also be using the train function, which comes from the caret packages. By providing the input for the model, which is every column except the "hd" column, with the train control that we've created, the method that we'll be using for training, which is a "glm", and the family that is a binomial.

Evaluation



```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 446  76
         Y  53 450

               Accuracy : 0.8741
                 95% CI : (0.8523, 0.8938)
    No Information Rate : 0.5132
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7484

 Mcnemar's Test P-Value : 0.05275

            Sensitivity : 0.8555
            Specificity : 0.8938
         Pos Pred Value : 0.8946
         Neg Pred Value : 0.8544
              Precision : 0.8946
                 Recall : 0.8555
                     F1 : 0.8746
             Prevalence : 0.5132
         Detection Rate : 0.4390
   Detection Prevalence : 0.4907
      Balanced Accuracy : 0.8747

       'Positive' Class : Y
```

To evaluate the model that we have, we use a confusion matrix to evaluate the model's prediction so we will get the prediction accuracy, prediction precision and prediction recall so we can evaluate the train model at this step. Thus from any value of prediction that we get from the train model, we don't use only accuracy for the prediction of the train model because the accuracy is more suitable for balance classification or balance categorical so, we will include precision value and recall value to determine the prediction of the train model. When we get the result from confusion matrix we get that only accuracy is not enough for define and indicated the value of prediction so, to test and improve the train model we can indicate other factor that we get from confusion matrix by using 95% CI and No Information Rate (from our project we get that 95% confidence interval is 0.7846 to 0.8713 while No Information Rate are 0.5292 thus, we get that both value are not overlap each other which mean the train model is better by significance) and for P-value will test about the accuracy when prediction are better that No Information rate if less than 0.05 which mean we can say that build predictive model are better than no model. After we have created a model, it is essential that we evaluate our model so that we understand its performance, as well as its strengths and weaknesses.

To evaluate the model that we have created, we are using the confusion matrix to evaluate the model's predictions. From these values here, we can see that the accuracy is pretty high. But looking only at the accuracy alone isn't a practical way to evaluate the model. So, we try to look at other values. In this case, it will be precision and recall, which are pretty high. We can assume that this model is functioning somewhat properly. Let's look at the accuracy. Together with 95% CI(confidence interval) and NIR (no information rate), we can see that accuracy is higher and the 95% CI (confidence interval) range does not fall within the NIR (no information rate) value. which is a good sign, showing that our model has some significance.

We can conclude that the rate of predicting when they are and not mispredicting when they are not is very high. because both of the sensitivity and specificity values are very high. From the PPV and NPV values, we see that both values are high enough so that we can trust the model when the results indicate that we are positive or negative.

Discussion and Conclusion

For discussion and improvement, there are many ways to improve our data analysis. For the data set we inherited, it is based on 1988 information, which nowadays, the information might be inaccurate, so the first improvement we believed to get a better result is to take a data set from the current year or if we have time, we could make our own data gathering process by doing a interview form. And to be specific, the data set we are using is from Cleveland, Hungary, Switzerland, and Long Beach V but we could scope the data set into asia country for better accuracy or even using Thailand as a test subject.

We established an analysis that is able to predict the chances of a person having a heart disease given the input information. So, we can use this to approximately have a patient know their risk. But the result that we got from this project isn't a perfect example of a modern AI model that is performing efficiently enough. It still needs a lot more tweaking and improvement. Like, increasing the sample size and more.

In conclusion, our project is able to indicate the risk of a patient having a heart disease.