Heart Disease Prediction using Regression

Arif Yetik

Abstract

Cardiovascular diseases and chronic respiratory diseases are a global threat. Due to such high deaths, there is a need to tackle the reasons behind these diseases. Most Coronary Heart Diseases can be prevented by addressing behavioral risk factors. It is important to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin. Recent advances in the field of artificial intelligence and softwares have led to the emergence of expert systems for medical applications. Moreover, in the last few decades computational tools have been designed to improve the experiences and abilities of expertss for making decisions about their patients.

In this study we used R program with some packages. Research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using regression models. We will use our data to create a model which tries to predict if a patient has this disease or not.

The dataset is publically available on the Kaggle website. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Keywords

Coronary Heart Disease (CHD), Heart Disease Prediction, Logistic Regression, R Program, Artificial Intelligence

1. Introduction

1.1. Motivation

Heart Diseases are the leading cause of death globally. World Health Organization has estimated 17.9 million people died from Heart Diseases in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. About 659,000 people in the United States die from heart disease each year, that's 1 in every 4 deaths. While many people with heart disease have symptoms such as chest pain and fatigue, as many as 50% have no symptoms until a heart attack occurs. According to the American heart association (AHA), CHD is the leading killer of American men and women, responsible for more than one of every five deaths in 2001 (http://www.americanheart.org, 2008). Many statistics show CHD as the leading cause of premature and permanent disability among American workers.

1.2. Objectives

In this data set, my ultimate goal is to find out the factors that will increase the chances of heart failure or heart disease and create a model that can accurately (hopefully) predict whether a person has the risk of having heart failure or a heart disease based on the given variables. Since this is a classification challenge (high risk or low risk), I will be experimenting with Simple Linear Regressions and Generalized Linear Regression. We will predict 10-year risk of Coronary Heart Diseases.

1.3. Significance

The early prognosis of heart diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using regressions. Use of risk prediction model to estimate total heart disease risk is a major advancement on the older research of identifying and treating individual risk factors, such as gender, age, smoking, BPMeds, stroke, sysBP, and glucose.

2. Relevant Literature

Heart disease also called as coronary heart disease (CHD), is a deposition of fats inside the tubes which supplies blood to the heart muscles. Heart disease actually starts as early as 18 years and patients only came to know about heart disease when the blockage exceeds about 70%. These blockages develop over the years and lead to rupture of the membrane covering the blockage due to pressure increases. If the chemicals released by broken membrane mixed with blood and lead to a blood clot, results to heart disease. The reasons which increase blockage are called as risk factors

Researchers expressed their efforts in finding the best model for predicting cardiovascular disease. In the meantime, various studies give only a glimpse into predicting heart disease using machine learning techniques and fuzzy logic systems.

Logistic regression(LR) is a generalized linear regression model. Therefore, it is similar with multiple linear regression in many aspects. Usually, LR is used for binary classification problems where the predictive variable$y \in [0,1]$, 0 is negative class and 1 is positive class. But it can also be used for multi-classification. Logistic regression is mainly used to for prediction and also calculating the probability of success.

Regression analysis explores the relationship between a quantitative response variable and explanatory variables. Regression model has two main objectives. Firstly, identify the statically significant relationship between these two variables. Secondly, forecast the new observations on response variable based on explanatory variables. In short, these variables are of two types i.e. dependent variable and independent variable. The dependent variable is the one whose value is required to be forecasted (i.e. vital signs values in our case) whereas the independent variable is used to explain dependent variable as input.

3. Methodology

3.1.The Data

In late 1940s, U.S. Government set out to better understand cardiovascular disease (CVD). They track large cohort of initially healthy patients over the City of Framingham, MA selected as site for study. Patients aged 30-59 enrolled. Patients were given questionnaire and exams every 2 years. The exams and questions expanded over time. We will build models using the Framingham data to predict and prevent Coronary Heart Diseases.

This data was downloaded from kaggle. We can describe each variable as follows:

1- Sex: male or female (Nominal)

2- Age: Age of the patient (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

3- Current Smoker: whether or not the patient is a current smoker (Nominal)

4- Cigs Per Day: the number of cigarettes that the person smoked on average in one day
5- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
6- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
7- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
8- Diabetes: whether or not the patient had diabetes (Nominal)
9- Tot Chol: total cholesterol level (Continuous)
10- Sys BP: systolic blood pressure (Continuous)
11- Dia BP: diastolic blood pressure (Continuous)
12- BMI: Body Mass Index (Continuous)
13- Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
14- Glucose: glucose level (Continuous)

Predict variable (desired target) 15- 10-year risk of coronary heart disease CHD (binary: "1", means "Yes", "0" means "No")

In this study we used R program and some packages.

```
library(tidyverse)
library(dplyr)
library(cowplot)
library(pROC)
library(caTools)
library(vtree)
library(corrgram)
library(caret)
```

3.1.1 Data Cleaning and Exploratory Data Analysis

```
heart_disease <- read.csv("framingham.csv")
View(heart_disease)
str(heart_disease)
```

```
## 'data.frame':    4238 obs. of  16 variables:
##  $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
##  $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
##  $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
##  $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

```
summary(heart_disease)
```

```
##       male              age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.003   Mean   :0.02963   Mean   :0.005899   Mean   :0.3105
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.00
##  Mean   :0.02572   Mean   :236.7   Mean   :132.4   Mean   : 82.89
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 89.88
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.50
##                    NA's   :50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.97   Mean   :0.152
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.000
##  NA's   :19      NA's   :1        NA's   :388
```

First, we can remove Duplicate Observations and Clean Null Observations.

```
heart_disease <- heart_disease %>% distinct()
colSums(is.na(heart_disease))

heart_disease <- na.omit(heart_disease)
head(heart_disease)

##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 1            0        0     195 106.0    70 26.97        80      77          0
## 2            0        0     250 121.0    81 28.73        95      76          0
## 3            0        0     245 127.5    80 25.34        75      70          0
## 4            1        0     225 150.0    95 28.58        65     103          1
## 5            0        0     285 130.0    84 23.10        85      85          0
## 6            1        0     228 180.0   110 30.30        77      99          0
```

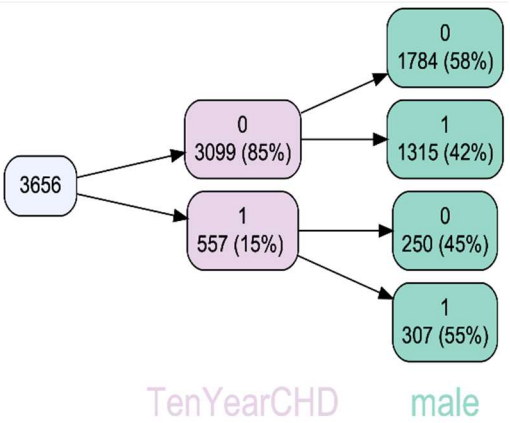
Convert binary variables tonumeric for better visualization

```
heart_disease$currentSmoker <- as.numeric(as.character(heart_disease$currentSmoker))
heart_disease$prevalentHyp <- as.numeric(as.character(heart_disease$prevalentHyp))
heart_disease$diabetes <- as.numeric(as.character(heart_disease$diabetes))
heart_disease$TenYearCHD <- as.numeric(as.character(heart_disease$TenYearCHD ))
```

Data structure: demographic risk factors: male, age, and education behavioral risk factors : currentSmoker and cigsPerDay medical history factors : BPmeds, prevalentStroke, prevalentHyp, diabetes physical exam risk : totChol, sysBP, diaBP, BMI, heartRate, glucose

dependent/outcome variable: CHD in 10 years

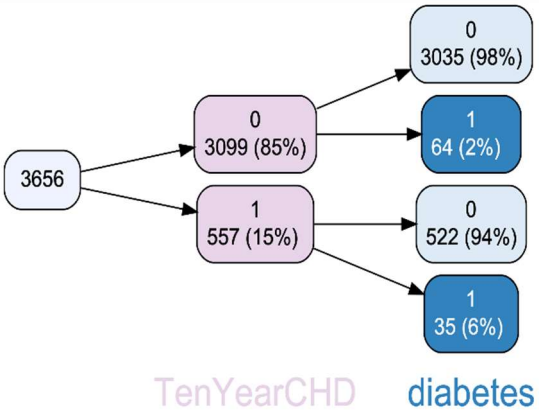
3.1.2. Data Visualization




```
vtree(heart_disease, c("TenYearCHD", "male"),
   fillcolor = c(TenYearCHD = "#e7d4e8", male = "#99d8c9"),
   horiz = TRUE)
```

In this plot, there are 3656 patients and 557 of them had heart disease in ten years. Among who had heart disease, 307 (55%) of them were male and 250 (45%) were female. More males than females have CHD.

```
vtree(heart_disease, c("TenYearCHD", "diabetes
"),
(TenYearCHD = "#e7d4e8", male
   horiz = TRUE)
```



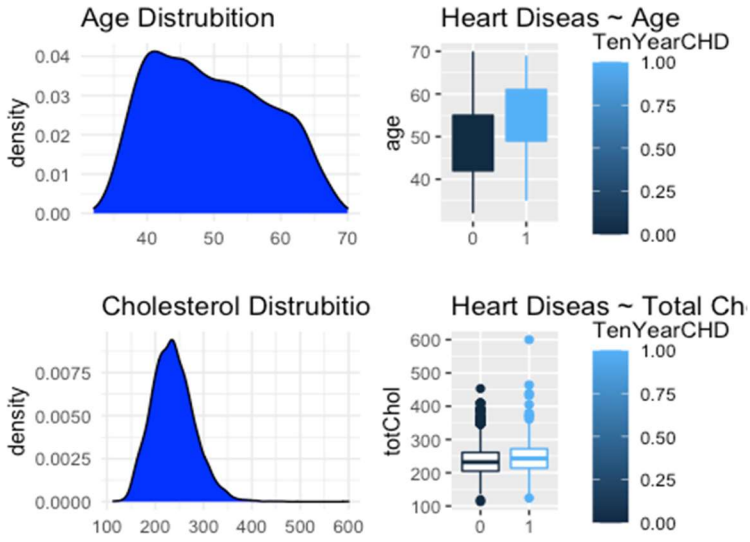

```
rCHD", "diabetes
   fillcolor = c
= "#99d8c9"),
```

In this plot, among 557 patients just 35 (6%) patients have diabetes. had heart disease, More non-diabetics.

diabetics have CHD compared to

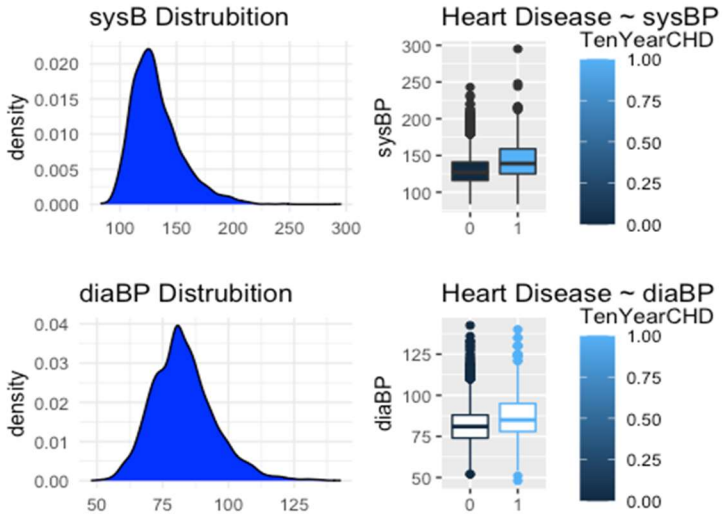
Now, we can look into our variables distributions and effect of heart disease.

```
plot_1<- ggplot(heart_disease, aes(age)) + geom_density(fill = "blue") + labs(x="",title = "Age Distrubition") + theme_minimal()
plot_12 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y = age, color = TenYearCHD, fill = TenYearCHD)) +
  geom_boxplot() + labs(x="",title = "Heart Diseas ~ Age")
plot_2 <- ggplot(heart_disease, aes(totChol)) + geom_density(fill = "blue") + labs(x="",title = "Cholesterol Distrubition") + theme_minimal()
plot_22 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y = totChol, color = TenYearCHD)) +
  geom_boxplot() + labs(x="",title = "Heart Diseas ~ Total Chol")
plot_grid(plot_1, plot_12, plot_2, plot_22)
```



In this plot, many patients are 35-55 years old. Among those who are above 55 years old had more heart failure. Many patients have an average 250 cholesterol. Total Cholesterol has effect on heart disease.

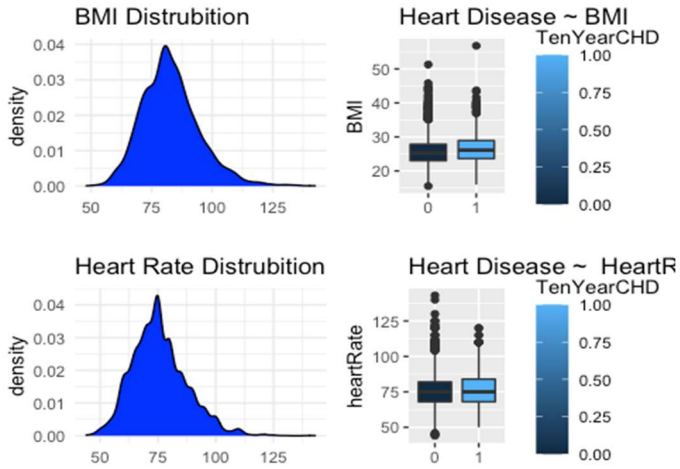
```
plot_3 <- ggplot(heart_disease, aes(sysBP)) + geom_density(fill = "blue") + labs(x="
",title = "sysB Distrubition") + theme_minimal()
plot_33 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y =
sysBP, fill = TenYearCHD)) +
  geom_boxplot() + labs(x="",title = "Heart Disease ~ sysBP")
plot_4 <- ggplot(heart_disease, aes(diaBP)) + geom_density(fill = "blue") + labs(x="
", title = "diaBP Distrubition") + theme_minimal()
plot_44 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y =
diaBP, color = TenYearCHD)) +
  geom_boxplot() + labs(x="", title = "Heart Disease ~ diaBP")
plot_grid(plot_3, plot_33, plot_4, plot_44)
```



In this plot, many patients have average 140 systolic blood pressure. People with CHD have higher mean systolic blood pressure. People with CHD have higher mean diastolic blood pressures.

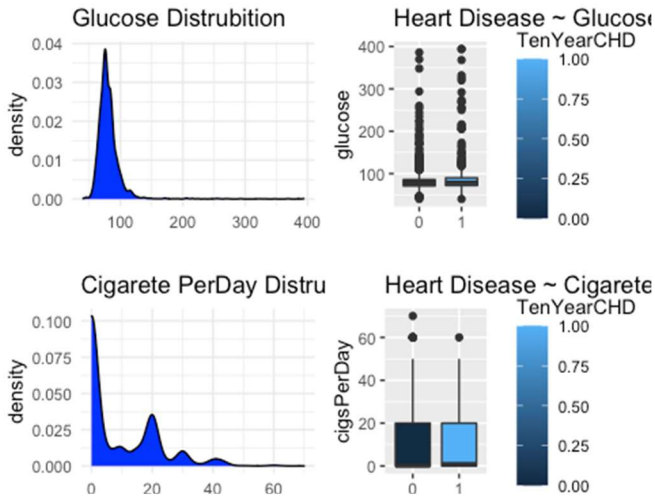
```
plot_5 <- ggplot(heart_disease, aes(diaBP)) + geom_density(fill = "blue") + labs(x="
", title = "BMI Distrubition") + theme_minimal()
```

```
plot_55 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y =
BMI, fill = TenYearCHD)) +
  geom_boxplot() +labs(x="", title = "Heart Disease ~ BMI")
plot_6 <- ggplot(heart_disease, aes(heartRate)) + geom_density(fill = "blue") + labs
(x="", title = "Heart Rate Distrubition") + theme_minimal()
plot_66 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y =
heartRate, fill = TenYearCHD)) +
  geom_boxplot() +labs(x="", title = "Heart Disease ~  HeartRate")
plot_grid(plot_5, plot_55, plot_6, plot_66)
```



People with CHD have a higher mean BMI. People with CHD have very similar mean heart rates as people without CHD.

```
plot_7 <- ggplot(heart_disease, aes(glucose)) + geom_density(fill = "blue") + labs(x
="", title = "Glucose Distrubition") + theme_minimal()
plot_77 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y =
glucose, fill = TenYearCHD)) +
  geom_boxplot() +labs(x="", title = "Heart Disease ~ Glucose")
plot_8 <- ggplot(heart_disease, aes(cigsPerDay)) + geom_density(fill = "blue") + lab
s(x="", title = "Cigarete PerDay Distrubition") + theme_minimal()
plot_88 <- ggplot(data = heart_disease, mapping = aes(x = as.factor(TenYearCHD), y =
cigsPerDay, fill = TenYearCHD)) +
  geom_boxplot() +labs(x="", title = "Heart Disease ~ Cigarete PerDay")
plot_grid(plot_7, plot_77, plot_8, plot_88)
```



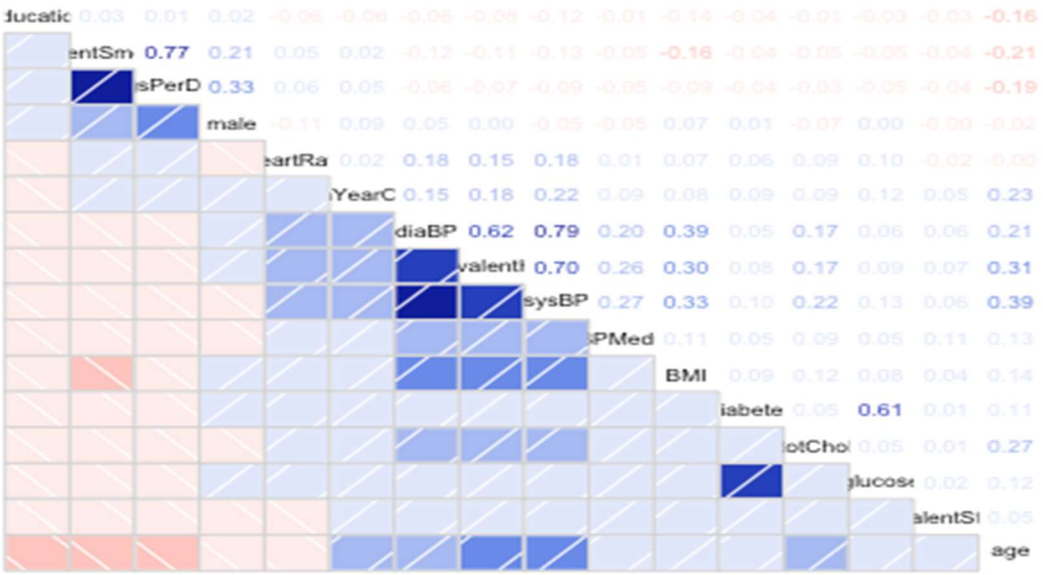
In this plot, people with CHD and without CHD have very similar mean glucose levels

3.2.1 Linear Regressions

We can compare the HeartDisease output with all the numeric variables within our data set and see whether we can find any correlations by using a correlogram.

We can check correlation between variables.

```
heart_disease %>% corrgram(order=TRUE, upper.panel=panel.cor)
```



Based on the given data, we can make a regression model. We will make predictions on the target variable coronary heart disease (CHD).

```
model <- lm(TenYearCHD ~ ., data = heart_disease)
summary(model)

##
## Call:
## lm(formula = TenYearCHD ~ ., data = heart_disease)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73036 -0.18936 -0.10687 -0.01153  1.08018
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.758e-01  7.895e-02  -7.293 3.70e-13 ***
## male             5.742e-02  1.248e-02   4.601 4.35e-06 ***
## age              7.052e-03  7.695e-04   9.165  < 2e-16 ***
## education       -5.581e-03  5.673e-03  -0.984 0.325252
## currentSmoker    8.008e-03  1.816e-02   0.441 0.659297
## cigsPerDay       2.167e-03  7.833e-04   2.767 0.005688 **
## BPMeds           4.585e-02  3.469e-02   1.321 0.186435
## prevalentStroke  1.368e-01  7.537e-02   1.815 0.069587 .
## prevalentHyp     2.838e-02  1.746e-02   1.626 0.104132
## diabetes         2.585e-02  4.423e-02   0.585 0.558849
```

```
## totChol          1.284e-04  1.353e-04   0.948 0.342976
## sysBP            2.473e-03  4.962e-04   4.984 6.53e-07 ***
## diaBP           -1.168e-03  8.169e-04  -1.430 0.152943
## BMI              6.917e-05  1.547e-03   0.045 0.964328
## heartRate       -3.597e-04  4.928e-04  -0.730 0.465560
## glucose          1.144e-03  3.023e-04   3.784 0.000157 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.3414 on 3640 degrees of freedom
## Multiple R-squared:  0.1012, Adjusted R-squared:  0.09745
## F-statistic: 27.31 on 15 and 3640 DF,  p-value: < 2.2e-16
```

Adjusted R-squared is too low. We can remodel with significant factors of male, age, sysBP, and glucose. cigsPerDay is almost significant factor. We can make new model.

```
re_model2 <- lm(TenYearCHD ~ male + age + sysBP + glucose, data = heart_disease)
summary(re_model2)
```

```
##
## Call:
## lm(formula = TenYearCHD ~ male + age + sysBP + glucose, data = heart_disease)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75535 -0.18858 -0.10809 -0.01844  1.12653
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6558797  0.0431653 -15.195  < 2e-16 ***
## male         0.0741235  0.0114178   6.492 9.62e-11 ***
## age          0.0070478  0.0007203   9.785  < 2e-16 ***
## sysBP        0.0024623  0.0002799   8.796  < 2e-16 ***
## glucose      0.0012235  0.0002398   5.101 3.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3426 on 3651 degrees of freedom
## Multiple R-squared:  0.09231,    Adjusted R-squared:  0.09132
## F-statistic: 92.83 on 4 and 3651 DF,  p-value: < 2.2e-16
```

The new model only achieved 0.09 adjusted R-squared which is same as previous model. Adjusted R-squared indicate a weak fit.

For this challenge, since heart disease is a classification problem, We will be using Generalized Linear Regression (Logistic Regression) to see which model would have the highest accuracy. But before that, we will need to split our data into training and testing data sets.

3.2.2 Generalized Linear Regression

1. step -> Randomly split patients into training and testing sets
2. step -> Logistic regression on training set to predict whether or not a patient experienced CHD within 10 years of first examination
3. step -> Evaluate predictive power on test set

Randomly split the data into training and testing sets. We may put 70% of the data in the training set. When you have more data like we do here, you can afford to put less data in the training set and more in the testing set. This will increase our confidence in the ability of the model to extend to new data since we have a larger test set, and still give us enough data in the training set to create our model.

```
set.seed(1000)
split = sample.split(heart_disease$TenYearCHD, SplitRatio = 0.70)
train = subset(heart_disease, split==TRUE)
test = subset(heart_disease, split==FALSE)
```

Convert binary variables to numeric for better visualization train and test datas.

```
train$currentSmoker <- as.numeric(as.character(train$currentSmoker))
train$prevalentHyp <- as.numeric(as.character(train$prevalentHyp))
train$diabetes <- as.numeric(as.character(train$diabetes))
train$TenYearCHD <- as.numeric(as.character(train$TenYearCHD))
test$currentSmoker <- as.numeric(as.character(test$currentSmoker))
test$prevalentHyp <- as.numeric(as.character(test$prevalentHyp))
test$diabetes <- as.numeric(as.character(test$diabetes))
test$TenYearCHD <- as.numeric(as.character(test$TenYearCHD))
```

Now, we can make new generalized model as follows:

```
glm_model <- glm(TenYearCHD ~ ., data=train, family=binomial, na.action=na.omit)
round(summary(glm_model)$coefficients, 3)
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.886      0.846  -9.325    0.000
## male               0.551      0.131   4.202    0.000
## age                0.059      0.008   7.436    0.000
## education         -0.088      0.059  -1.483    0.138
## currentSmoker      0.002      0.188   0.011    0.991
## cigsPerDay         0.022      0.007   3.025    0.002
## BPMeds            -0.094      0.299  -0.316    0.752
## prevalentStroke    0.704      0.526   1.337    0.181
## prevalentHyp       0.301      0.166   1.811    0.070
## diabetes          -0.068      0.393  -0.174    0.862
## totChol            0.003      0.001   2.356    0.018
## sysBP              0.016      0.005   3.506    0.000
## diaBP             -0.005      0.008  -0.667    0.505
## BMI                0.001      0.015   0.098    0.922
## heartRate         -0.006      0.005  -1.104    0.270
## glucose            0.007      0.003   2.433    0.015
```

It looks like male, age, cigsPerDay, total cholesterol, systolic blood pressure, and glucose are all significant in our model. The diaBP is almost significant .

All of the significant variables have positive coefficients, meaning that higher values in these variables contribute to a higher probability of 10-year coronary heart disease.

Remove the insignificant variables and retrain the model.

```
new_glm_model <- glm(TenYearCHD ~ male + age + totChol + cigsPerDay + sysBP + glucos
e, data=train, family=binomial)
```

3.3. Statistical Inference

```
summary(new_glm_model)

##
## Call:
## glm(formula = TenYearCHD ~ male + age + totChol + cigsPerDay +
##     sysBP + glucose, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9005  -0.6002  -0.4335  -0.2862   2.8855
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.141103   0.570099 -16.034  < 2e-16 ***
## male         0.564310   0.128640   4.387 1.15e-05 ***
## age          0.062417   0.007591   8.223  < 2e-16 ***
## totChol      0.002986   0.001329   2.248  0.02460 *
## cigsPerDay   0.021256   0.004953   4.291 1.78e-05 ***
## sysBP        0.017979   0.002581   6.965 3.27e-12 ***
## glucose      0.006654   0.002258   2.946  0.00322 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2184.6  on 2558  degrees of freedom
## Residual deviance: 1939.7  on 2552  degrees of freedom
## AIC: 1953.7
##
## Number of Fisher Scoring iterations: 5
```

Our model formula can be generated as follow:

Heart Disease = -8.703 + 0.531*male1 + 0.067*age + 0.0029*totChol + 0.019*cigsPerDay + 0.017*sysBP + 0.007*glucose

Confidence Intervals and P-values

We can check our model's confidence interval.

```
confint(new_glm_model)

## Waiting for profiling to be done...

##                      2.5 %      97.5 %
## (Intercept) -1.027783e+01 -8.04187156
## male         3.128590e-01  0.81744906
## age          4.762674e-02  0.07739852
## totChol      3.671333e-04  0.00558251
## cigsPerDay   1.151938e-02  0.03095406
## sysBP        1.293808e-02  0.02306330
## glucose      2.238286e-03  0.01114030
```

We can do ANOVA as well:

```
anova(new_glm_model, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: TenYearCHD
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         2558     2184.7
## male         1   25.100      2557     2159.5 5.442e-07 ***
## age          1  128.570      2556     2031.0 < 2.2e-16 ***
## totChol      1   11.323      2555     2019.7 0.0007657 ***
## cigsPerDay   1   17.457      2554     2002.2 2.938e-05 ***
## sysBP        1   53.847      2553     1948.3 2.168e-13 ***
## glucose      1    8.661      2552     1939.7 0.0032510 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the ANOVA, age, male, restecg, totChol, cigsPerDay, sysBP, and glucose are significant factors for predicting heart disease. P- values are extremely close to zero.

Prediction

We'll call our predictions predictTest and use the predict function, which takes as arguments the name of our model, new_glm_model, then type = "response", which gives us probabilities, and lastly newdata = test, the name of our testing set.

We'll use the table function and give as the first argument, the actual values, test$TenYearCHD, and then as the second argument our predictions, predictTest > 0.5.

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such a system is commonly evaluated using the data in the matrix. Predictions on the test set and Confusion matrix with threshold of 0.5

```
predictTest = predict(new_glm_model, type="response", newdata=test)
table(test$TenYearCHD, predictTest > 0.5)
```

```
##
##     FALSE TRUE
##   0   924    6
##   1   150   17
```

```
##   83.22%
```

Our model has 83.22% accuracies. It shows 924 NO heart diseases accurately and 150 NO death events were wrongly predicted as the death events. Also, 6 heart diseases were wrongly predicted as NO heart diseases. 17 death events are accurately predicted.

4. Results and Discussion

The study cohort accumulated 3,656 patients-years of observation with 10 years. The statistically independent predictive risk factors in our model are age, male, restecg, totChol, cigsPerDay, sysBP, and glucose. With every extra cigarette one smokes there is a 2% increase in the odds of CDH. Smoking and aging are factors which can greatly cause heart disease. For Total cholesterol level and glucose level there are no significant change. There is a 1.7% increase in odds for every unit increase in systolic Blood Pressure.

5. Conclusions

We have an accuracy of about 83.22% on our test set, which means that the model can differentiate between low risk patients and high risk patients pretty well. Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease. Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol(HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%).

6. References:

1. http://www.who.int/mediacentre/factsheets/fs317/en/
2. Wajid Shah, Cardiovascular and Chronic Respiratory Diseases Prediction System
3. Sumit Sharma, Heart Diseases Prediction Using Hybrid Ensemble Learning, Dublin Business School
4. Qi Zhenya & Zuoru Zhang, A hybrid cost-sensitive ensemble for heart disease prediction, Open Access Published: 25 February 2021
5. Fatma Zahra Abdeldjouad, A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques, Open Access Published: 23 June 2020
6. José A. Piniés, Fernando González-Carril, Development of a prediction model for fatal and non-fatal coronary heart disease and cardiovascular disease in patients with newly diagnosed type 2 diabetes mellitus, Published: 12 September 2014
7. Saaol times, Monthly magazine, Modifiable risk factors of heart disease, pp. 6–10, July (2015), Google Scholar
8. M. A. Jabbar, Prediction of Heart Disease Using Random Forest and Feature Subset Selection, 15 December 2015
9. https://www.heart.org/?identifier=4726
10. https://www.cdc.gov/heartdisease/facts.htm