Data Science in Health Project: CHD Logistic Prediction Model Comparison

Overview

This data science project aims to analyze health data collected from individuals to predict the likelihood of developing coronary heart disease (CHD) within ten years. Insights produced in this project could save countless lives and provide tremendous benefit in the medical space. The project utilizes logistic regression to build a predictive model based on various health indicators.

This was not the initial project I originally intended to do; it was a mobile health dataset from body sensors containing over ten thousand data points. Because of my computer, it took forever to run and it became evident that I had to switch plans. During that time, I thought about a disease that runs in my family and decided to use it as inspiration for this project.

Literature Review of Previous Works

Nishat, M., Ahmed, S., Hasan, M. M., Ali, M. H., Saha, R., & Mahmud, M. (2021). Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease.

The previous study utilized various machine learning methods to predict CHD. Utilizing data from the University of California, Irvine repository, twelve algorithms were assessed using default hyperparameters, grid search cross-validation, and random search cross-validation methods. Both accuracy and computational time were measured, with hard and soft voting ensemble classifiers achieving 92% accuracy. Adaboost algorithm demonstrated superior precision and specificity compared to ensemble classifiers. The analysis extensively compares algorithm performance across multiple metrics including accuracy, precision, sensitivity, specificity, F1 score, and ROC-AUC.

Even though there were many models that intrigued me, I decided to personally use logistic regression because of its simplicity, efficiency, and clinical acceptance. Attempts of data analysis on the dataset posted on kaggle showed the following:

- Men are more likely to get heart disease than women. As people get older, smoke more cigarettes, or have higher blood pressure, their chances of getting heart disease also go up.
- Having higher total cholesterol doesn’t seem to make much difference in the chance of getting heart disease. This might be because the cholesterol test includes both good and bad cholesterol. Glucose levels also don’t have a big impact on the chance of getting heart disease, only a tiny bit.
- The model we used predicted heart disease correctly 88% of the time. It’s better at saying who doesn’t have heart disease than who does.

Methodology

I will be experimenting with min-max normalization and using the top absolute valued correlated variables associated with the variable ‘TenYearCHD’ which describes whether a subject has cardiovascular disease or not. The top correlated values were selected based on their absolute values.The objective of the project is to play around with the data and double check the claims mentioned in previous analysis.

Preliminaries

Loading Dataset and Packages

The project utilizes several R packages for data manipulation, visualization, and model training. Key packages include caret, pROC, ggplot2, and dplyr. The dataset is loaded using the read.csv() function from the foreign package. R studio was used as the main code editor to compile the project and its resources.

Loading Dataset and Packages

```
#required packages
list.of.packages <- c("foreign","rjags","dplyr","ggplot2","plotly","reshape2","bnl
earn","nnet","caret","pROC","penalized","caret")

#install if necessary
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Pa
ckage"])]
if(length(new.packages)) install.packages(new.packages)

#load all packages
lapply(list.of.packages, library, character.only = TRUE)
```

```
## Loading required package: coda
```

```
## Linked to JAGS 4.3.2
```

```
## Loaded modules: basemod,bugs
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
## Loading required package: lattice
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Welcome to penalized. For extended examples, see vignette("penalized").
```

```
## [[1]]
## [1] "foreign"   "stats"     "graphics"  "grDevices" "utils"     "datasets"
## [7] "methods"   "base"
##
## [[2]]
##  [1] "rjags"     "coda"      "foreign"   "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
##
## [[3]]
##  [1] "dplyr"     "rjags"     "coda"      "foreign"   "stats"     "graphics"
##  [7] "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[4]]
##  [1] "ggplot2"   "dplyr"     "rjags"     "coda"      "foreign"   "stats"
##  [7] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"
##
```

```
## [[5]]
##  [1] "plotly"    "ggplot2"   "dplyr"     "rjags"     "coda"      "foreign"
##  [7] "stats"     "graphics"  "grDevices" "utils"     "datasets"  "methods"
## [13] "base"
##
## [[6]]
##  [1] "reshape2"  "plotly"    "ggplot2"   "dplyr"     "rjags"     "coda"
##  [7] "foreign"   "stats"     "graphics"  "grDevices" "utils"     "datasets"
## [13] "methods"   "base"
##
## [[7]]
##  [1] "bnlearn"   "reshape2"  "plotly"    "ggplot2"   "dplyr"     "rjags"
##  [7] "coda"      "foreign"   "stats"     "graphics"  "grDevices" "utils"
## [13] "datasets"  "methods"   "base"
##
## [[8]]
##  [1] "nnet"      "bnlearn"   "reshape2"  "plotly"    "ggplot2"   "dplyr"
##  [7] "rjags"     "coda"      "foreign"   "stats"     "graphics"  "grDevices"
## [13] "utils"     "datasets"  "methods"   "base"
##
## [[9]]
##  [1] "caret"     "lattice"   "nnet"      "bnlearn"   "reshape2"  "plotly"
##  [7] "ggplot2"   "dplyr"     "rjags"     "coda"      "foreign"   "stats"
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[10]]
##  [1] "pROC"      "caret"     "lattice"   "nnet"      "bnlearn"   "reshape2"
##  [7] "plotly"    "ggplot2"   "dplyr"     "rjags"     "coda"      "foreign"
## [13] "stats"     "graphics"  "grDevices" "utils"     "datasets"  "methods"
## [19] "base"
##
## [[11]]
##  [1] "penalized" "survival"  "pROC"      "caret"     "lattice"   "nnet"
##  [7] "bnlearn"   "reshape2"  "plotly"    "ggplot2"   "dplyr"     "rjags"
## [13] "coda"      "foreign"   "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
##
## [[12]]
##  [1] "penalized" "survival"  "pROC"      "caret"     "lattice"   "nnet"
##  [7] "bnlearn"   "reshape2"  "plotly"    "ggplot2"   "dplyr"     "rjags"
## [13] "coda"      "foreign"   "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
```

```
# Example script to read data
data <- read.csv('/Users/unclenamo/Desktop/Zhaw/Data Science for Health Project /D
ata Science in Health Final Project Folder/framingham.csv')
```

```
head(data,n = 10)
```

```
##    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
## 7     0  63         1             0          0      0               0
## 8     0  45         2             1         20      0               0
## 9     1  52         1             0          0      0               0
## 10    1  43         1             1         30      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
## 7             0        0     205 138.0    71 33.11        60      85          1
## 8             0        0     313 100.0    71 21.68        79      78          0
## 9             1        0     260 141.5    89 26.36        76      79          0
## 10            1        0     225 162.0   107 23.61        93      88          0
```

Conditional Indexing, Selection, and Initial Visualization

Handling Missing Values

Missing values in the dataset are removed using the na.omit() function to ensure data integrity and consistency. Without removing the null values, the dataset’s dimensions would have been problematic in the analysis and training phase.

```
str(data)
```

```
## '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 ...
```

```
clean_data <- na.omit(data)
head(clean_data)
```

```
##   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
```

```
str(clean_data)
```

```
## 'data.frame':	3656 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 ...
##  - attr(*, "na.action")= 'omit' Named int [1:582] 15 22 27 34 37 43 50 55 71 73
...
##   ..- attr(*, "names")= chr [1:582] "15" "22" "27" "34" ...
```

Descriptive Statistics

Descriptive statistics such as mean, median, minimum, maximum, standard deviation, and quartiles are calculated for numeric variables in the dataset to gain insights into the distribution of health indicators. A function was created to compare male and female subjects as well as the average statistics of subjects with CHD and without CHD. I had problems with this function because I had misplaced the variables 'clean_data' and 'data' causing the difference not to be shown.

```
# Check for NULL values
is.null(clean_data)
```

```
## [1] FALSE
```

```
# Summary of dataframe
summary(clean_data)
```

```
##       male             age          education    currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.00   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.00   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.00   Median :0.0000
##  Mean   :0.4437   Mean   :49.56   Mean   :1.98   Mean   :0.4891
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.00   Max.   :1.0000
##    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   : 9.022   Mean   :0.03036   Mean   :0.005744   Mean   :0.3115
##  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
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :113.0   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   :0.02708   Mean   :236.9   Mean   :132.4   Mean   : 82.91
##  3rd Qu.:0.00000   3rd Qu.:263.2   3rd Qu.:144.0   3rd Qu.: 90.00
##  Max.   :1.00000   Max.   :600.0   Max.   :295.0   Max.   :142.50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.08   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.38   Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.78   Mean   : 75.73   Mean   : 81.86   Mean   :0.1524
##  3rd Qu.:28.04   3rd Qu.: 82.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
```

```
# Structure of dataframe
str(clean_data)
```

```
## 'data.frame':    3656 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 ...
##  - attr(*, "na.action")= 'omit' Named int [1:582] 15 22 27 34 37 43 50 55 71 73
...
##   ..- attr(*, "names")= chr [1:582] "15" "22" "27" "34" ...
```

```
calculate_descriptive_statistics <- function(clean_data) {
    # Select specific numeric columns
    numeric_cols <- c("age", "education", "cigsPerDay", "totChol", "sysBP", "diaBP
", "BMI","heartRate", "glucose")

   # Filter the data based on selected numeric columns
   numeric_data <- clean_data[, numeric_cols]


  # Calculate descriptive statistics
  descriptive_stats <- apply(clean_data[, numeric_cols], 2, function(x) {
    mean_val <- mean(x, na.rm = TRUE)
    median_val <- median(x, na.rm = TRUE)
    min_val <- min(x, na.rm = TRUE)
    max_val <- max(x, na.rm = TRUE)
    sd_val <- sd(x, na.rm = TRUE)
    q1 <- quantile(x, probs = 0.25, na.rm = TRUE)
    q3 <- quantile(x, probs = 0.75, na.rm = TRUE)
    iqr <- q3 - q1

    result <- c(mean = mean_val,
                median = median_val,
                min = min_val,
                max = max_val,
                sd = sd_val,
                q1 = q1,
                q3 = q3,
                IQR = iqr)

    return(result)
  })

  # Create a dataframe from the results
  descriptive_stats_df <- t(as.data.frame(descriptive_stats))
  colnames(descriptive_stats_df) <- c("Mean", "Median", "Min", "Max", "SD", "Q1",
"Q3", "IQR")

  return(descriptive_stats_df)
}
```

```
# Assuming 'data' is the name of your dataset
descriptive_stats <- calculate_descriptive_statistics(clean_data)
print(descriptive_stats)
```

```
##                    Mean Median    Min   Max        SD     Q1     Q3   IQR
## age          49.557440  49.00  32.00  70.0  8.561133  42.00  56.00 14.00
## education     1.979759   2.00   1.00   4.0  1.022657   1.00   3.00  2.00
## cigsPerDay    9.022155   0.00   0.00  70.0 11.918869   0.00  20.00 20.00
## totChol     236.873085 234.00 113.00 600.0 44.096223 206.00 263.25 57.25
## sysBP       132.368025 128.00  83.50 295.0 22.092444 117.00 144.00 27.00
## diaBP        82.912062  82.00  48.00 142.5 11.974825  75.00  90.00 15.00
## BMI          25.784185  25.38  15.54  56.8  4.065913  23.08  28.04  4.96
## heartRate    75.730580  75.00  44.00 143.0 11.982952  68.00  82.00 14.00
## glucose      81.856127  78.00  40.00 394.0 23.910128  71.00  87.00 16.00
```

Male vs Female Selection

```
# Select rows where 'male' is equal to 0
female_data <- clean_data[clean_data$male == 0, ]

# Select rows where 'male' is equal to 1
male_data <- clean_data[clean_data$male == 1, ]
```

```
head(female_data)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 2    0  46         2             0          0      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
## 7    0  63         1             0          0      0               0
## 8    0  45         2             1         20      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 2            0        0     250   121    81 28.73        95      76          0
## 4            1        0     225   150    95 28.58        65     103          1
## 5            0        0     285   130    84 23.10        85      85          0
## 6            1        0     228   180   110 30.30        77      99          0
## 7            0        0     205   138    71 33.11        60      85          1
## 8            0        0     313   100    71 21.68        79      78          0
```

```
head(male_data)
```

```
##    male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1     1  39         4             0          0      0               0
## 3     1  48         1             1         20      0               0
## 9     1  52         1             0          0      0               0
## 10    1  43         1             1         30      0               0
## 13    1  46         1             1         15      0               0
## 17    1  48         3             1         10      0               0
##    prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 1             0        0     195 106.0    70 26.97        80      77          0
## 3             0        0     245 127.5    80 25.34        75      70          0
## 9             1        0     260 141.5    89 26.36        76      79          0
## 10            1        0     225 162.0   107 23.61        93      88          0
## 13            1        0     294 142.0    94 26.31        98      64          0
## 17            1        0     232 138.0    90 22.37        64      72          0
```

```
# Assuming 'data' is the name of your dataset
descriptive_stats_female <- calculate_descriptive_statistics(female_data)
print(descriptive_stats_female)
```

```
##                  Mean Median    Min   Max         SD     Q1     Q3   IQR
## age         49.743854  49.00  32.00  70.0  8.5732729  42.00  56.00 14.00
## education    1.963618   2.00   1.00   4.0  0.9660513   1.00   3.00  2.00
## cigsPerDay   5.497050   0.00   0.00  43.0  8.7393385   0.00  10.00 10.00
## totChol    239.638151 237.00 135.00 600.0 46.1683210 206.00 268.00 62.00
## sysBP      133.265241 128.00  83.50 295.0 23.9868093 116.00 146.00 30.00
## diaBP       82.360619  81.00  51.00 142.5 12.3087481  74.00  89.00 15.00
## BMI         25.519651  24.72  15.96  56.8  4.5162673  22.54  27.71  5.17
## heartRate   76.960177  75.00  46.00 143.0 12.1220181  69.00  85.00 16.00
## glucose     81.791052  78.00  40.00 394.0 23.5862453  72.00  86.00 14.00
```

```
# Assuming 'data' is the name of your dataset
descriptive_stats_male <- calculate_descriptive_statistics(male_data)
print(descriptive_stats_male)
```

```
##                 Mean Median    Min    Max        SD       Q1    Q3     IQR
## age         49.32367     48  33.00  69.00  8.542779  42.0000  56.0 14.0000
## education    2.00000      2   1.00   4.00  1.089459   1.0000   3.0  2.0000
## cigsPerDay  13.44266     15   0.00  70.00 13.761522   0.0000  20.0 20.0000
## totChol    233.40567    231 113.00 453.00 41.103254 206.0000 259.0 53.0000
## sysBP      131.24291    128  83.50 232.00 19.406784 118.0000 141.0 23.0000
## diaBP       83.60358     82  48.00 136.00 11.508911  76.0000  90.0 14.0000
## BMI         26.11591     26  15.54  40.38  3.390656  23.9425  28.3  4.3575
## heartRate   74.18866     75  44.00 125.00 11.627529  66.0000  80.0 14.0000
## glucose     81.93773     78  40.00 394.00 24.317237  70.0000  87.0 17.0000
```

```r
# Select rows where 'TenYearCHD' is equal to 0
no_chd_data <-clean_data[clean_data$TenYearCHD == 0, ]

# Select rows where 'TenYearCHD' is equal to 1
chd_data <- clean_data[clean_data$TenYearCHD == 1, ]
```

```r
head(no_chd_data)
```

```
##   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
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
## 8    0  45         2             1         20      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
## 5            0        0     285 130.0    84 23.10        85      85          0
## 6            1        0     228 180.0   110 30.30        77      99          0
## 8            0        0     313 100.0    71 21.68        79      78          0
```

```r
head(chd_data)
```

```
##    male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 4     0  61         3             1         30      0               0
## 7     0  63         1             0          0      0               0
## 16    0  38         2             1         20      0               0
## 18    0  46         2             1         20      0               0
## 26    1  47         4             1         20      0               0
## 29    0  61         3             0          0      0               0
##    prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 4             1        0     225   150    95 28.58        65     103          1
## 7             0        0     205   138    71 33.11        60      85          1
## 16            1        0     221   140    90 21.35        95      70          1
## 18            0        0     291   112    78 23.38        80      89          1
## 26            0        0     294   102    68 24.18        62      66          1
## 29            1        0     272   182   121 32.80        85      65          1
```

```r
count(chd_data)
```

```
##     n
## 1 557
```

```r
count(no_chd_data)
```

```
##      n
## 1 3099
```

```
# Assuming 'data' is the name of your dataset
descriptive_stats_nochd <- calculate_descriptive_statistics(no_chd_data)
print(descriptive_stats_nochd)
```

```
##                   Mean Median    Min    Max        SD     Q1     Q3   IQR
## age          48.708938  48.00  32.00  70.00  8.383279  42.00  55.00 13.00
## education     2.007099   2.00   1.00   4.00  1.019159   1.00   3.00  2.00
## cigsPerDay    8.758632   0.00   0.00  70.00 11.715691   0.00  20.00 20.00
## totChol     235.169732 232.00 113.00 453.00 43.078009 205.00 261.00 56.00
## sysBP       130.280736 127.00  83.50 243.00 20.413624 116.00 141.00 25.00
## diaBP        82.148919  81.00  52.00 142.50 11.320205  74.00  88.00 14.00
## BMI          25.642975  25.23  15.54  51.28  3.965283  23.01  27.86  4.85
## heartRate    75.626331  75.00  44.00 143.00 11.953256  68.00  82.00 14.00
## glucose      80.620200  78.00  40.00 386.00 19.128713  71.00  86.00 15.00
```

```
# Assuming 'data' is the name of your dataset
descriptive_stats_chd <- calculate_descriptive_statistics(chd_data)
print(descriptive_stats_chd)
```

```
##                   Mean Median    Min   Max        SD     Q1     Q3   IQR
## age          54.278276  55.00  35.00  69.0  7.992338  49.00  61.00 12.00
## education     1.827648   1.00   1.00   4.0  1.029645   1.00   2.00  1.00
## cigsPerDay   10.488330   1.00   0.00  60.0 12.904685   0.00  20.00 20.00
## totChol     246.350090 243.00 124.00 600.0 48.336365 214.00 272.00 58.00
## sysBP       143.981149 139.00  83.50 295.0 26.966224 125.00 159.00 34.00
## diaBP        87.157989  85.00  48.00 140.0 14.398497  78.00  95.00 17.00
## BMI          26.569838  26.11  15.96  56.8  4.509435  23.63  28.94  5.31
## heartRate    76.310592  75.00  50.00 120.0 12.141349  68.00  84.00 16.00
## glucose      88.732496  79.00  40.00 394.0 40.785655  72.00  90.00 18.00
```

Comparison Function CHD vs No CHD

```
# Assuming 'data' is the name of your dataset

# Calculate descriptive statistics for individuals with and without CHD
chd_stats <- calculate_descriptive_statistics(clean_data[clean_data$TenYearCHD ==
1, ])
no_chd_stats <- calculate_descriptive_statistics(clean_data[clean_data$TenYearCHD
== 0, ])

# Combine the statistics into a single dataframe for comparison
comparison <- data.frame(Feature = rownames(chd_stats),
                         CHD_Mean = chd_stats[, "Mean"],
                         No_CHD_Mean = no_chd_stats[, "Mean"],
                         Difference = chd_stats[, "Mean"] - no_chd_stats[, "Mean"]
)

# Print the comparison
print(comparison)
```

```
##                Feature   CHD_Mean No_CHD_Mean Difference
## age                age  54.278276   48.708938  5.5693381
## education    education   1.827648    2.007099 -0.1794509
## cigsPerDay cigsPerDay  10.488330    8.758632  1.7296985
## totChol       totChol 246.350090  235.169732 11.1803576
## sysBP           sysBP 143.981149  130.280736 13.7004133
## diaBP           diaBP  87.157989   82.148919  5.0090702
## BMI               BMI  26.569838   25.642975  0.9268633
## heartRate   heartRate  76.310592   75.626331  0.6842614
## glucose       glucose  88.732496   80.620200  8.1122954
```

Visualization Pre-Training

The project includes various data visualizations to explore relationships between different health variables and their impact on the likelihood of developing CHD. Visualizations include scatter plots, bar charts, and violin plots.

```
head(data)
```

```
##   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
```

```
# Select only numeric columns
numeric_data <- clean_data[, sapply(data, is.numeric)]

# Calculate the correlation matrix for the numeric data
correlation_matrix <- cor(numeric_data)

# Print the correlation matrix
print(correlation_matrix)
```

```
##                          male          age   education currentSmoker
## male             1.0000000000 -0.024386991  0.01767684    0.20677793
## age             -0.0243869912  1.000000000 -0.15896134   -0.21086237
## education        0.0176768430 -0.158961341  1.00000000    0.02525285
## currentSmoker    0.2067779295 -0.210862368  0.02525285    1.00000000
## cigsPerDay       0.3312428456 -0.189099490  0.01352711    0.77381894
## BPMeds          -0.0521281205  0.134670170 -0.01364679   -0.05193582
## prevalentStroke -0.0023075218  0.050863869 -0.03035280   -0.03815949
## prevalentHyp     0.0008057437  0.306692997 -0.07909966   -0.10756095
## diabetes         0.0138330267  0.109026510 -0.03954683   -0.04185871
## totChol         -0.0702285291  0.267763684 -0.01295563   -0.05111939
## sysBP           -0.0454844109  0.388550599 -0.12451062   -0.13437098
## diaBP            0.0515751876  0.208880362 -0.05850151   -0.11574796
## BMI              0.0728673292  0.137172104 -0.13728006   -0.15957358
## heartRate       -0.1149234002 -0.002685426 -0.06425396    0.05045182
## glucose          0.0030481786  0.118244733 -0.03187419   -0.05334601
## TenYearCHD       0.0917448852  0.233810450 -0.06306773    0.01917620
##                  cigsPerDay      BPMeds prevalentStroke  prevalentHyp
## male             0.33124285 -0.05212812    -0.002307522  0.0008057437
## age             -0.18909949  0.13467017     0.050863869  0.3066929975
## education        0.01352711 -0.01364679    -0.030352798 -0.0790996577
## currentSmoker    0.77381894 -0.05193582    -0.038159492 -0.1075609504
## cigsPerDay       1.00000000 -0.04647920    -0.036283081 -0.0698895718
## BPMeds          -0.04647920  1.00000000     0.113118955  0.2630468560
## prevalentStroke -0.03628308  0.11311895     1.000000000  0.0660979828
## prevalentHyp    -0.06988957  0.26304686     0.066097983  1.0000000000
## diabetes        -0.03693406  0.04905100     0.009618566  0.0806231104
## totChol         -0.03022238  0.09401050     0.012696639  0.1670744320
```

```
## sysBP           -0.09476371  0.27129113     0.061079638  0.6977899529
## diaBP           -0.05665012  0.19975031     0.055877896  0.6176342217
## BMI             -0.08688806  0.10560316     0.036477739  0.3029168279
## heartRate        0.06354908  0.01289362    -0.017020305  0.1473326726
## glucose         -0.05380272  0.05421037     0.016051252  0.0871291882
## TenYearCHD       0.05215873  0.08911570     0.048350573  0.1815564019
##                    diabetes     totChol       sysBP       diaBP         BMI
## male             0.013833027 -0.07022853 -0.04548441  0.05157519  0.07286733
## age              0.109026510  0.26776368  0.38855060  0.20888036  0.13717210
## education       -0.039546826 -0.01295563 -0.12451062 -0.05850151 -0.13728006
## currentSmoker   -0.041858712 -0.05111939 -0.13437098 -0.11574796 -0.15957358
## cigsPerDay      -0.036934057 -0.03022238 -0.09476371 -0.05665012 -0.08688806
## BPMeds           0.049050998  0.09401050  0.27129113  0.19975031  0.10560316
## prevalentStroke  0.009618566  0.01269664  0.06107964  0.05587790  0.03647774
## prevalentHyp     0.080623110  0.16707443  0.69778995  0.61763422  0.30291683
## diabetes         1.000000000  0.04837075  0.10257419  0.05076727  0.08897004
## totChol          0.048370745  1.00000000  0.22012958  0.17498559  0.12079901
## sysBP            0.102574186  0.22012958  1.00000000  0.78672712  0.33100359
## diaBP            0.050767275  0.17498559  0.78672712  1.00000000  0.38561068
## BMI              0.088970038  0.12079901  0.33100359  0.38561068  1.00000000
## heartRate        0.060995532  0.09305743  0.18490117  0.17900822  0.07440124
## glucose          0.614817444  0.04974867  0.13470173  0.06370364  0.08367110
## TenYearCHD       0.093397417  0.09112675  0.22288534  0.15034173  0.08193118
##                   heartRate      glucose  TenYearCHD
## male            -0.114923400  0.003048179  0.09174489
## age             -0.002685426  0.118244733  0.23381045
## education       -0.064253962 -0.031874187 -0.06306773
## currentSmoker    0.050451822 -0.053346008  0.01917620
## cigsPerDay       0.063549083 -0.053802723  0.05215873
## BPMeds           0.012893624  0.054210370  0.08911570
## prevalentStroke -0.017020305  0.016051252  0.04835057
## prevalentHyp     0.147332673  0.087129188  0.18155640
## diabetes         0.060995532  0.614817444  0.09339742
## totChol          0.093057425  0.049748666  0.09112675
## sysBP            0.184901171  0.134701732  0.22288534
## diaBP            0.179008216  0.063703644  0.15034173
## BMI              0.074401235  0.083671103  0.08193118
## heartRate        1.000000000  0.097025854  0.02052342
## glucose          0.097025854  1.000000000  0.12194204
## TenYearCHD       0.020523424  0.121942043  1.00000000
```

```r
display_correlation_pairs <- function(correlation_matrix) {
  # Convert correlation matrix to a long-form data frame
  df <- reshape2::melt(correlation_matrix)

  # Remove NA and duplicate rows
  df <- df[complete.cases(df), ]
  df <- df[!duplicated(df), ]

  # Sort by absolute correlation value in descending order
  df <- df[order(-abs(df$value)), ]

  # Print the sorted pairs
  print(df)
}

display_correlation_pairs(correlation_matrix)
```

```
##                 Var1            Var2        value
## 1               male            male  1.0000000000
## 18               age             age  1.0000000000
## 35         education       education  1.0000000000
## 52     currentSmoker   currentSmoker  1.0000000000
## 69        cigsPerDay      cigsPerDay  1.0000000000
## 86            BPMeds          BPMeds  1.0000000000
## 103 prevalentStroke prevalentStroke  1.0000000000
## 120     prevalentHyp    prevalentHyp  1.0000000000
## 137         diabetes        diabetes  1.0000000000
## 154          totChol         totChol  1.0000000000
## 171            sysBP           sysBP  1.0000000000
## 188            diaBP           diaBP  1.0000000000
## 205              BMI             BMI  1.0000000000
## 222        heartRate       heartRate  1.0000000000
## 239          glucose         glucose  1.0000000000
## 256       TenYearCHD      TenYearCHD  1.0000000000
## 172            diaBP           sysBP  0.7867271219
## 187            sysBP           diaBP  0.7867271219
## 53        cigsPerDay   currentSmoker  0.7738189372
## 68     currentSmoker      cigsPerDay  0.7738189372
## 123            sysBP    prevalentHyp  0.6977899529
## 168     prevalentHyp           sysBP  0.6977899529
## 124            diaBP    prevalentHyp  0.6176342217
## 184     prevalentHyp           diaBP  0.6176342217
## 143          glucose        diabetes  0.6148174441
## 233         diabetes         glucose  0.6148174441
## 27             sysBP             age  0.3885505989
## 162              age           sysBP  0.3885505989
## 189              BMI           diaBP  0.3856106780
## 204            diaBP             BMI  0.3856106780
## 5         cigsPerDay            male  0.3312428456
## 65              male      cigsPerDay  0.3312428456
## 173              BMI           sysBP  0.3310035899
## 203            sysBP             BMI  0.3310035899
```

```
## 24      prevalentHyp             age  0.3066929975
## 114              age    prevalentHyp  0.3066929975
## 125              BMI    prevalentHyp  0.3029168279
## 200     prevalentHyp             BMI  0.3029168279
## 91             sysBP          BPMeds  0.2712911307
## 166           BPMeds           sysBP  0.2712911307
## 26           totChol             age  0.2677636840
## 146              age         totChol  0.2677636840
## 88      prevalentHyp          BPMeds  0.2630468560
## 118           BPMeds    prevalentHyp  0.2630468560
## 32        TenYearCHD             age  0.2338104505
## 242              age      TenYearCHD  0.2338104505
## 176       TenYearCHD           sysBP  0.2228853419
## 251            sysBP      TenYearCHD  0.2228853419
## 155            sysBP         totChol  0.2201295813
## 170          totChol           sysBP  0.2201295813
## 20     currentSmoker             age -0.2108623681
## 50               age   currentSmoker -0.2108623681
## 28             diaBP             age  0.2088803615
## 178              age           diaBP  0.2088803615
## 4      currentSmoker            male  0.2067779295
## 49              male   currentSmoker  0.2067779295
## 92             diaBP          BPMeds  0.1997503070
## 182           BPMeds           diaBP  0.1997503070
## 21        cigsPerDay             age -0.1890994896
## 66               age      cigsPerDay -0.1890994896
## 174        heartRate           sysBP  0.1849011705
## 219            sysBP       heartRate  0.1849011705
## 128       TenYearCHD    prevalentHyp  0.1815564019
## 248     prevalentHyp      TenYearCHD  0.1815564019
## 190        heartRate           diaBP  0.1790082157
## 220            diaBP       heartRate  0.1790082157
## 156            diaBP         totChol  0.1749855921
## 186          totChol           diaBP  0.1749855921
## 122          totChol    prevalentHyp  0.1670744320
## 152     prevalentHyp         totChol  0.1670744320
## 61               BMI   currentSmoker -0.1595735777
## 196    currentSmoker             BMI -0.1595735777
## 19         education             age -0.1589613409
## 34               age       education -0.1589613409
## 192       TenYearCHD           diaBP  0.1503417292
## 252            diaBP      TenYearCHD  0.1503417292
## 126        heartRate    prevalentHyp  0.1473326726
## 216     prevalentHyp       heartRate  0.1473326726
## 45               BMI       education -0.1372800603
## 195        education             BMI -0.1372800603
## 29               BMI             age  0.1371721044
## 194              age             BMI  0.1371721044
## 175          glucose           sysBP  0.1347017320
## 235            sysBP         glucose  0.1347017320
## 22            BPMeds             age  0.1346701704
## 82               age          BPMeds  0.1346701704
## 59             sysBP   currentSmoker -0.1343709794
## 164    currentSmoker           sysBP -0.1343709794
```

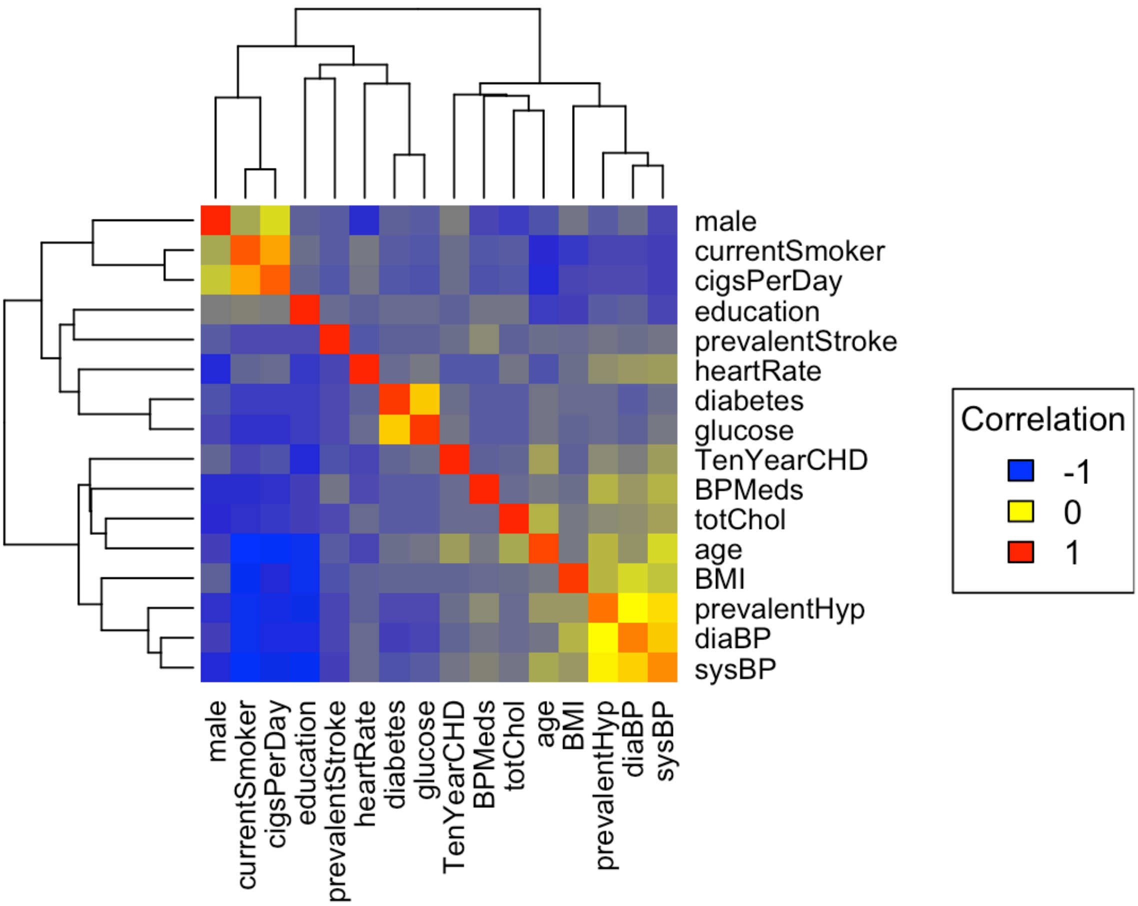
```
## 43            sysBP       education -0.1245106205
## 163       education           sysBP -0.1245106205
## 240      TenYearCHD         glucose  0.1219420426
## 255         glucose      TenYearCHD  0.1219420426
## 157             BMI         totChol  0.1207990064
## 202         totChol             BMI  0.1207990064
## 31          glucose             age  0.1182447325
## 226             age         glucose  0.1182447325
## 60            diaBP   currentSmoker -0.1157479625
## 180   currentSmoker           diaBP -0.1157479625
## 14        heartRate            male -0.1149234002
## 209            male       heartRate -0.1149234002
## 87  prevalentStroke          BPMeds  0.1131189545
## 102          BPMeds prevalentStroke  0.1131189545
## 25         diabetes             age  0.1090265099
## 130             age        diabetes  0.1090265099
## 56     prevalentHyp   currentSmoker -0.1075609504
## 116   currentSmoker    prevalentHyp -0.1075609504
## 93              BMI          BPMeds  0.1056031644
## 198          BPMeds             BMI  0.1056031644
## 139           sysBP        diabetes  0.1025741856
## 169        diabetes           sysBP  0.1025741856
## 223         glucose       heartRate  0.0970258537
## 238       heartRate         glucose  0.0970258537
## 75            sysBP      cigsPerDay -0.0947637083
## 165      cigsPerDay           sysBP -0.0947637083
## 90          totChol          BPMeds  0.0940105008
## 150          BPMeds         totChol  0.0940105008
## 144      TenYearCHD        diabetes  0.0933974173
## 249        diabetes      TenYearCHD  0.0933974173
## 158       heartRate         totChol  0.0930574254
## 218         totChol       heartRate  0.0930574254
## 16       TenYearCHD            male  0.0917448852
## 241            male      TenYearCHD  0.0917448852
## 160      TenYearCHD         totChol  0.0911267540
## 250         totChol      TenYearCHD  0.0911267540
## 96       TenYearCHD          BPMeds  0.0891157036
## 246          BPMeds      TenYearCHD  0.0891157036
## 141             BMI        diabetes  0.0889700379
## 201        diabetes             BMI  0.0889700379
## 127         glucose    prevalentHyp  0.0871291882
## 232    prevalentHyp         glucose  0.0871291882
## 77              BMI      cigsPerDay -0.0868880619
## 197      cigsPerDay             BMI -0.0868880619
## 207         glucose             BMI  0.0836711029
## 237             BMI         glucose  0.0836711029
## 208      TenYearCHD             BMI  0.0819311831
## 253             BMI      TenYearCHD  0.0819311831
## 121        diabetes    prevalentHyp  0.0806231104
## 136    prevalentHyp        diabetes  0.0806231104
## 40     prevalentHyp       education -0.0790996577
## 115       education    prevalentHyp -0.0790996577
## 206       heartRate             BMI  0.0744012355
## 221             BMI       heartRate  0.0744012355
```

```
## 13               BMI            male  0.0728673292
## 193             male             BMI  0.0728673292
## 10           totChol            male -0.0702285291
## 145             male         totChol -0.0702285291
## 72      prevalentHyp      cigsPerDay -0.0698895718
## 117       cigsPerDay    prevalentHyp -0.0698895718
## 104     prevalentHyp prevalentStroke  0.0660979828
## 119  prevalentStroke    prevalentHyp  0.0660979828
## 46         heartRate       education -0.0642539618
## 211        education       heartRate -0.0642539618
## 191          glucose           diaBP  0.0637036444
## 236            diaBP         glucose  0.0637036444
## 78         heartRate      cigsPerDay  0.0635490832
## 213       cigsPerDay       heartRate  0.0635490832
## 48        TenYearCHD       education -0.0630677273
## 243        education      TenYearCHD -0.0630677273
## 107            sysBP prevalentStroke  0.0610796379
## 167  prevalentStroke           sysBP  0.0610796379
## 142        heartRate        diabetes  0.0609955324
## 217         diabetes       heartRate  0.0609955324
## 44             diaBP       education -0.0585015079
## 179        education           diaBP -0.0585015079
## 76             diaBP      cigsPerDay -0.0566501192
## 181       cigsPerDay           diaBP -0.0566501192
## 108            diaBP prevalentStroke  0.0558778962
## 183  prevalentStroke           diaBP  0.0558778962
## 95           glucose          BPMeds  0.0542103700
## 230           BPMeds         glucose  0.0542103700
## 79           glucose      cigsPerDay -0.0538027227
## 229       cigsPerDay         glucose -0.0538027227
## 63           glucose   currentSmoker -0.0533460079
## 228    currentSmoker         glucose -0.0533460079
## 80        TenYearCHD      cigsPerDay  0.0521587275
## 245       cigsPerDay      TenYearCHD  0.0521587275
## 6             BPMeds            male -0.0521281205
## 81              male          BPMeds -0.0521281205
## 54            BPMeds   currentSmoker -0.0519358242
## 84     currentSmoker          BPMeds -0.0519358242
## 12             diaBP            male  0.0515751876
## 177             male           diaBP  0.0515751876
## 58           totChol   currentSmoker -0.0511193922
## 148    currentSmoker         totChol -0.0511193922
## 23   prevalentStroke             age  0.0508638692
## 98               age prevalentStroke  0.0508638692
## 140            diaBP        diabetes  0.0507672746
## 185         diabetes           diaBP  0.0507672746
## 62         heartRate   currentSmoker  0.0504518224
## 212    currentSmoker       heartRate  0.0504518224
## 159          glucose         totChol  0.0497486662
## 234          totChol         glucose  0.0497486662
## 89          diabetes          BPMeds  0.0490509982
## 134           BPMeds        diabetes  0.0490509982
## 138          totChol        diabetes  0.0483707453
## 153         diabetes         totChol  0.0483707453
```

```
## 112      TenYearCHD prevalentStroke  0.0483505730
## 247 prevalentStroke      TenYearCHD  0.0483505730
## 70           BPMeds      cigsPerDay -0.0464791991
## 85       cigsPerDay          BPMeds -0.0464791991
## 11            sysBP            male -0.0454844109
## 161            male           sysBP -0.0454844109
## 57         diabetes   currentSmoker -0.0418587123
## 132   currentSmoker        diabetes -0.0418587123
## 41         diabetes       education -0.0395468261
## 131       education        diabetes -0.0395468261
## 55  prevalentStroke   currentSmoker -0.0381594924
## 100   currentSmoker prevalentStroke -0.0381594924
## 73         diabetes      cigsPerDay -0.0369340566
## 133      cigsPerDay        diabetes -0.0369340566
## 109             BMI prevalentStroke  0.0364777386
## 199 prevalentStroke             BMI  0.0364777386
## 71  prevalentStroke      cigsPerDay -0.0362830812
## 101      cigsPerDay prevalentStroke -0.0362830812
## 47          glucose       education -0.0318741872
## 227       education         glucose -0.0318741872
## 39  prevalentStroke       education -0.0303527976
## 99        education prevalentStroke -0.0303527976
## 74          totChol      cigsPerDay -0.0302223819
## 149      cigsPerDay         totChol -0.0302223819
## 36    currentSmoker       education  0.0252528518
## 51        education   currentSmoker  0.0252528518
## 2               age            male -0.0243869912
## 17             male             age -0.0243869912
## 224      TenYearCHD       heartRate  0.0205234237
## 254       heartRate      TenYearCHD  0.0205234237
## 64       TenYearCHD   currentSmoker  0.0191761963
## 244   currentSmoker      TenYearCHD  0.0191761963
## 3         education            male  0.0176768430
## 33             male       education  0.0176768430
## 110       heartRate prevalentStroke -0.0170203055
## 215 prevalentStroke       heartRate -0.0170203055
## 111         glucose prevalentStroke  0.0160512523
## 231 prevalentStroke         glucose  0.0160512523
## 9          diabetes            male  0.0138330267
## 129            male        diabetes  0.0138330267
## 38           BPMeds       education -0.0136467912
## 83        education          BPMeds -0.0136467912
## 37       cigsPerDay       education  0.0135271093
## 67        education      cigsPerDay  0.0135271093
## 42          totChol       education -0.0129556316
## 147       education         totChol -0.0129556316
## 94        heartRate          BPMeds  0.0128936240
## 214          BPMeds       heartRate  0.0128936240
## 106         totChol prevalentStroke  0.0126966393
## 151 prevalentStroke         totChol  0.0126966393
## 105        diabetes prevalentStroke  0.0096185655
## 135 prevalentStroke        diabetes  0.0096185655
## 15          glucose            male  0.0030481786
## 225            male         glucose  0.0030481786
```

```
## 30        heartRate             age -0.0026854264
## 210             age       heartRate -0.0026854264
## 7    prevalentStroke            male -0.0023075218
## 97             male prevalentStroke -0.0023075218
## 8      prevalentHyp             male  0.0008057437
## 113            male    prevalentHyp  0.0008057437
```

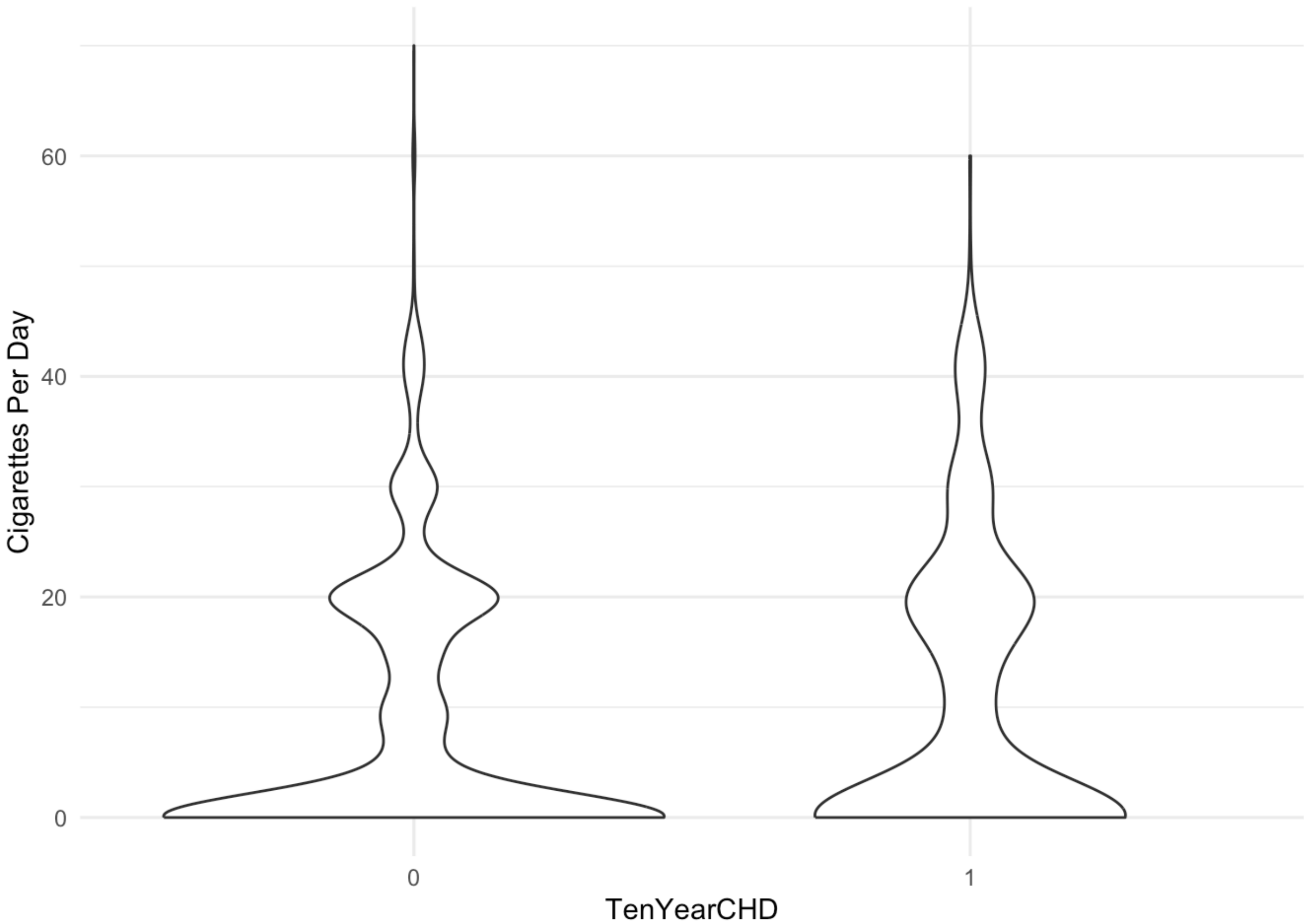
```
# Set the size of the plot
options(repr.plot.width = 30, repr.plot.height = 15) # Adjust width and height as needed

# Create a heatmap of the correlation matrix with color scale
heatmap(correlation_matrix,
        col = colorRampPalette(c("blue", "yellow", "red"))(100),
        scale = "row",    # Add scale for rows
        symm = TRUE,      # To make the heatmap symmetric
        margins = c(10, 10))  # To provide extra space for row and column names

# Add color scale legend
legend("right",     # Position the legend to the right
       legend = c(-1, 0, 1),  # Values for the color scale (simplified)
       fill = colorRampPalette(c("blue", "yellow", "red"))(3),  # Color gradient for the legend
       title = "Correlation")  # Title for the legend
```



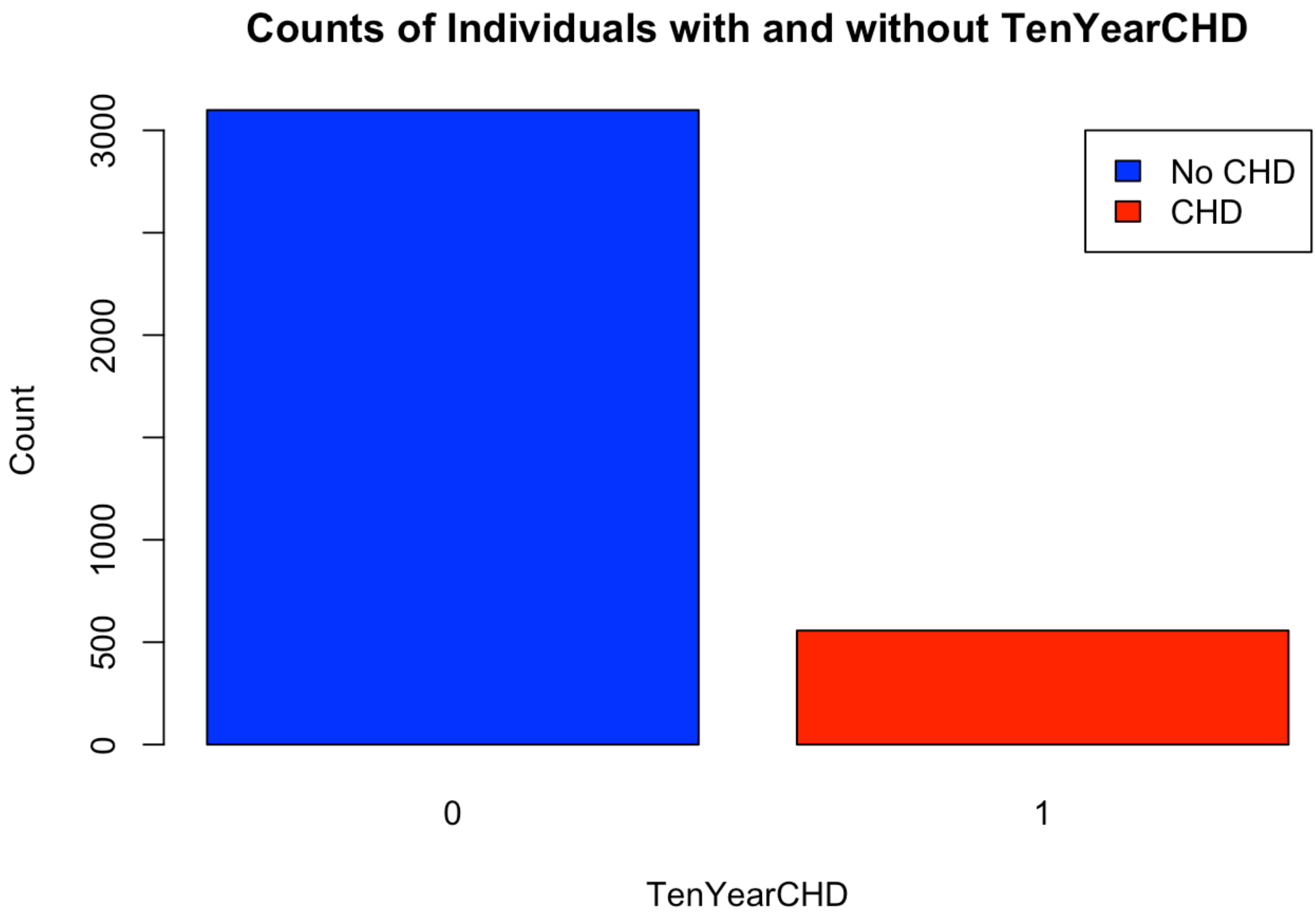
```r
head(clean_data)
```

```
##   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
```

```r
# Create a basic violin plot
violin_plot <- ggplot(clean_data, aes(x = factor(TenYearCHD), y = cigsPerDay)) +
  geom_violin() +
  labs(x = "TenYearCHD", y = "Cigarettes Per Day") +
  theme_minimal()

# Display the violin plot
print(violin_plot)
```

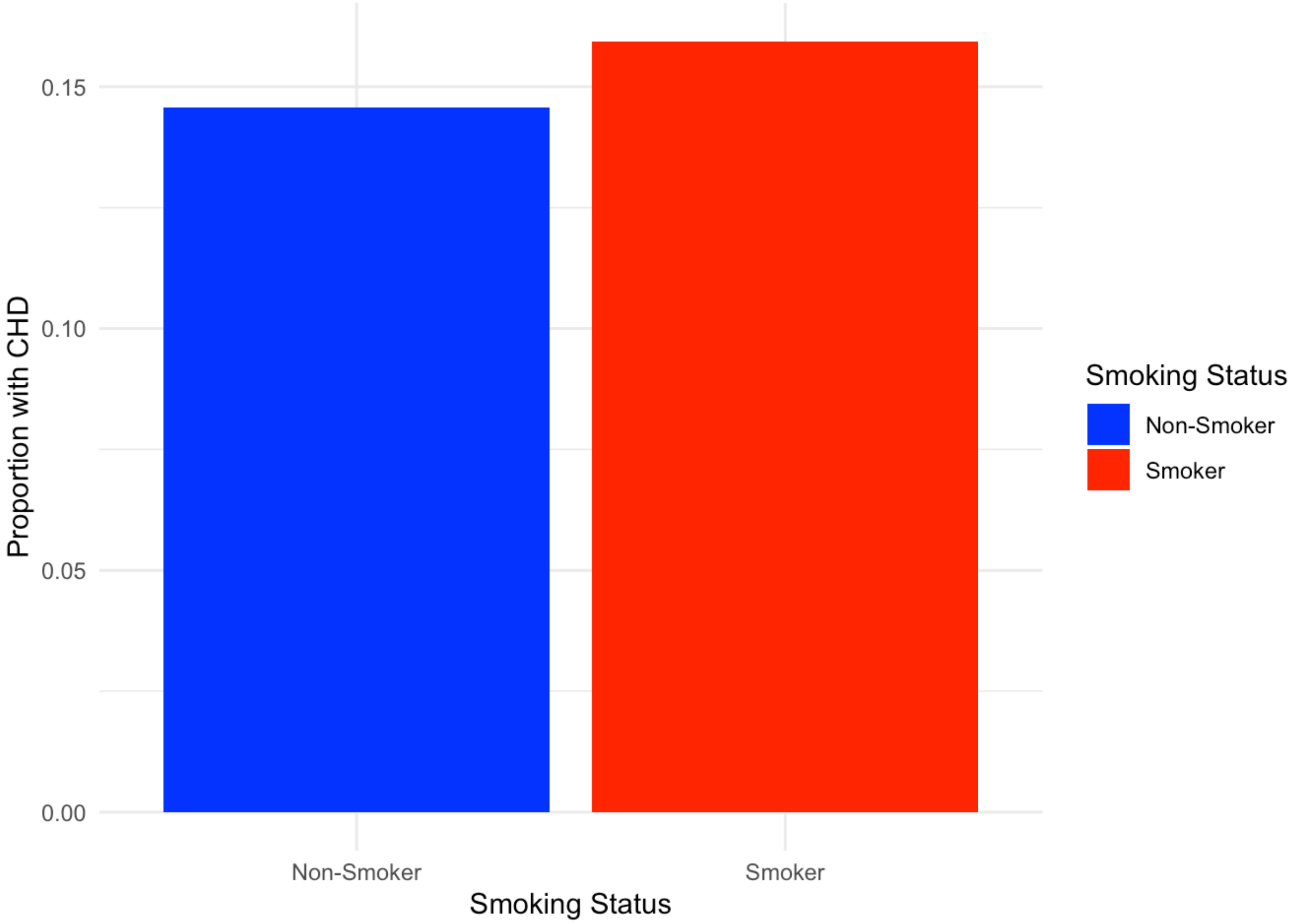



```
# Calculate the counts of individuals with and without TenYearCHD
chd_counts <- table(clean_data$TenYearCHD)

# Create the bar chart
barplot(chd_counts, col = c("blue", "red"), main = "Counts of Individuals with and
without TenYearCHD",
        xlab = "TenYearCHD", ylab = "Count", legend = c("No CHD", "CHD"))
```

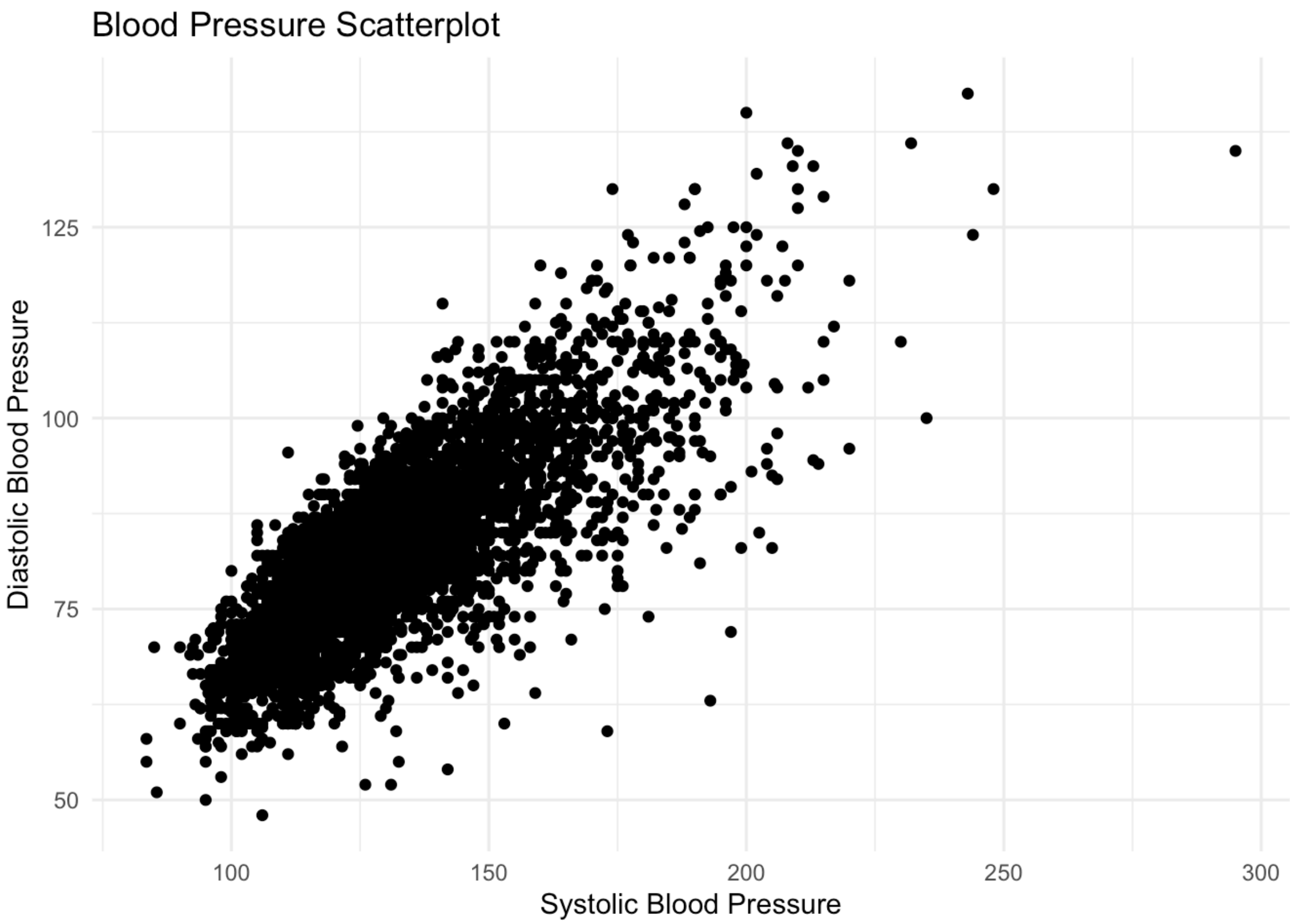



```
# Calculate the proportion of individuals with and without CHD among smokers and n
on-smokers
chd_prop <- aggregate(TenYearCHD ~ currentSmoker, data = clean_data, FUN = functio
n(x) sum(x == 1) / length(x))
names(chd_prop) <- c("currentSmoker", "CHD_Proportion")

# Convert currentSmoker to factor for better visualization
chd_prop$currentSmoker <- factor(chd_prop$currentSmoker, levels = c(0, 1), labels
= c("Non-Smoker", "Smoker"))

# Create the grouped bar chart
grouped_bar_chart <- ggplot(chd_prop, aes(x = currentSmoker, y = CHD_Proportion, f
ill = currentSmoker)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Smoking Status", y = "Proportion with CHD", fill = "Smoking Status") +
  scale_fill_manual(values = c("Non-Smoker" = "blue", "Smoker" = "red")) +  # Cust
omizing fill colors
  theme_minimal()
```

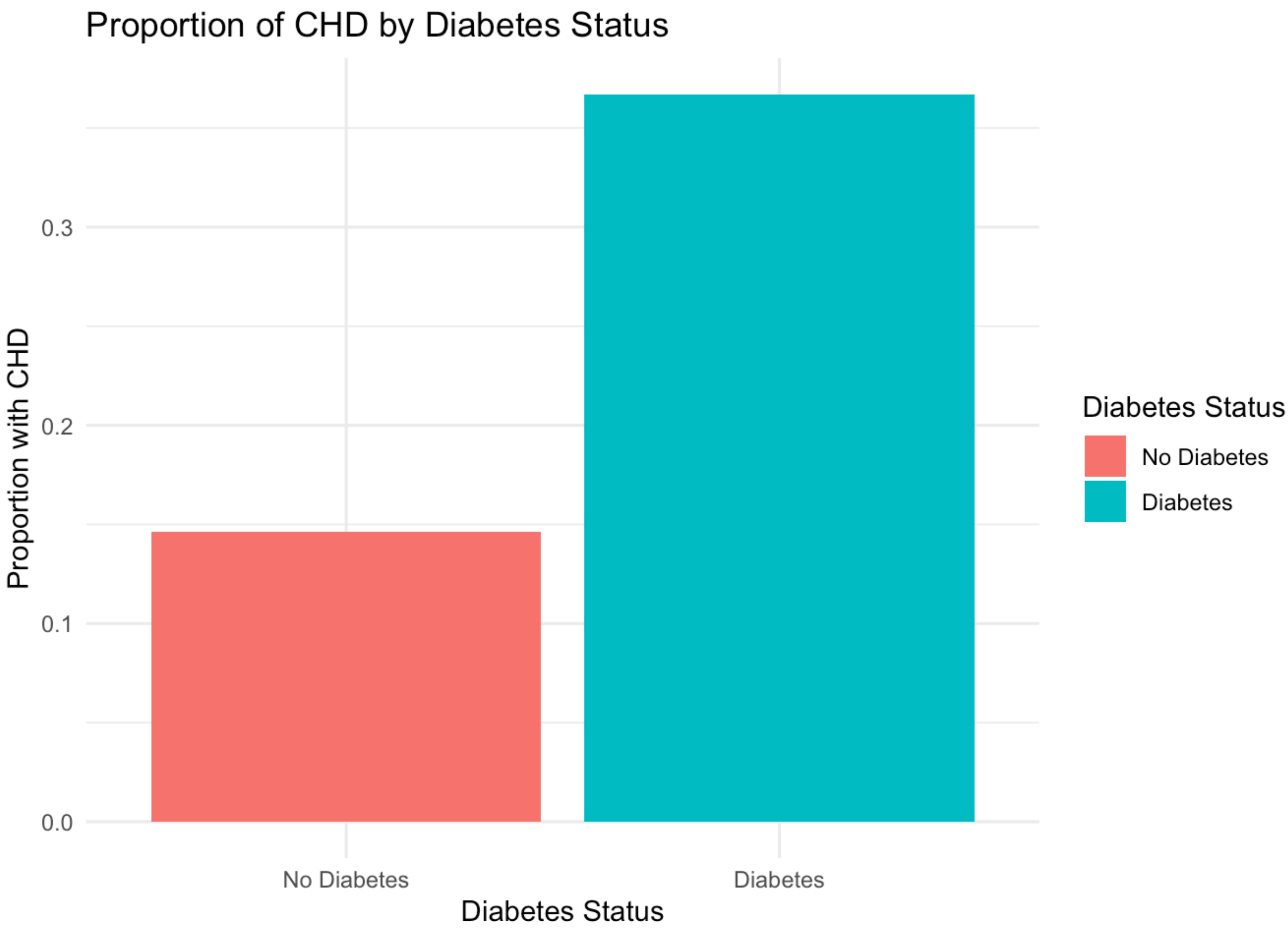
```
# Display the grouped bar chart
print(grouped_bar_chart)
```




```
# Load the ggplot2 library
library(ggplot2)

# Assuming 'data' is the name of your dataset
# Scatterplot for Blood Pressure (sysBP, diaBP)
bp_scatterplot <- ggplot(data, aes(x = sysBP, y = diaBP)) +
  geom_point() +
  labs(x = "Systolic Blood Pressure", y = "Diastolic Blood Pressure", title = "Blo
od Pressure Scatterplot") +
  theme_minimal()


print(bp_scatterplot)
```

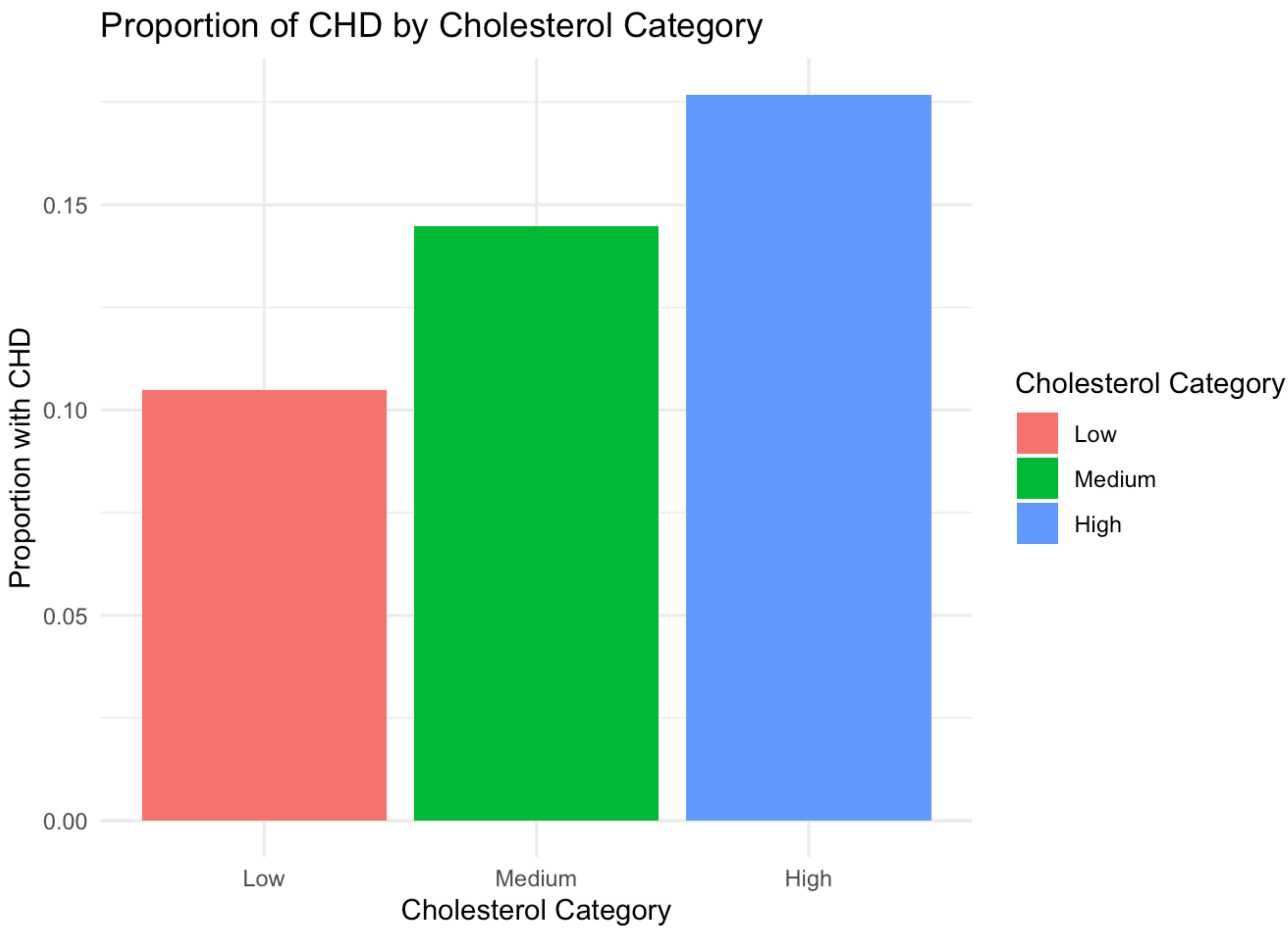



```
# Calculate the proportion of individuals with and without CHD for each level of d
iabetes status
diabetes_chd_prop <- aggregate(TenYearCHD ~ diabetes, data = data, FUN = function(
x) mean(x == 1))
names(diabetes_chd_prop) <- c("Diabetes_Status", "CHD_Proportion")

# Convert diabetes status to factor for correct ordering in the plot
diabetes_chd_prop$Diabetes_Status <- factor(diabetes_chd_prop$Diabetes_Status, lev
els = c(0, 1), labels = c("No Diabetes", "Diabetes"))

# Create a grouped bar plot
grouped_bar_plot <- ggplot(diabetes_chd_prop, aes(x = Diabetes_Status, y = CHD_Pro
portion, fill = Diabetes_Status)) +
  geom_bar(stat = "identity") +
  labs(x = "Diabetes Status", y = "Proportion with CHD", fill = "Diabetes Status",
title = "Proportion of CHD by Diabetes Status") +
  theme_minimal()

# Display the grouped bar plot
print(grouped_bar_plot)
```



- **Low Cholesterol**: Total cholesterol level below 200 mg/dL.
- **Desirable/Medium Cholesterol**: Total cholesterol level between 200 mg/dL and 239 mg/dL.
- **High Cholesterol**: Total cholesterol level 240 mg/dL or higher.

```r
# Create a new variable to categorize cholesterol levels
data$Cholesterol_Category <- cut(data$totChol,
                                 breaks = c(-Inf, 200, 239, Inf),
                                 labels = c("Low", "Medium", "High"),
                                 right = FALSE)

# Calculate the proportion of individuals with and without CHD for each level of c
holesterol category
cholesterol_chd_prop <- aggregate(TenYearCHD ~ Cholesterol_Category, data = data,
FUN = function(x) mean(x == 1))
names(cholesterol_chd_prop) <- c("Cholesterol_Category", "CHD_Proportion")

# Create a grouped bar plot
grouped_bar_plot_cholesterol <- ggplot(cholesterol_chd_prop, aes(x = Cholesterol_C
ategory, y = CHD_Proportion, fill = Cholesterol_Category)) +
  geom_bar(stat = "identity") +
  labs(x = "Cholesterol Category", y = "Proportion with CHD", fill = "Cholesterol
Category", title = "Proportion of CHD by Cholesterol Category") +
  theme_minimal()

# Display the grouped bar plot
print(grouped_bar_plot_cholesterol)
```

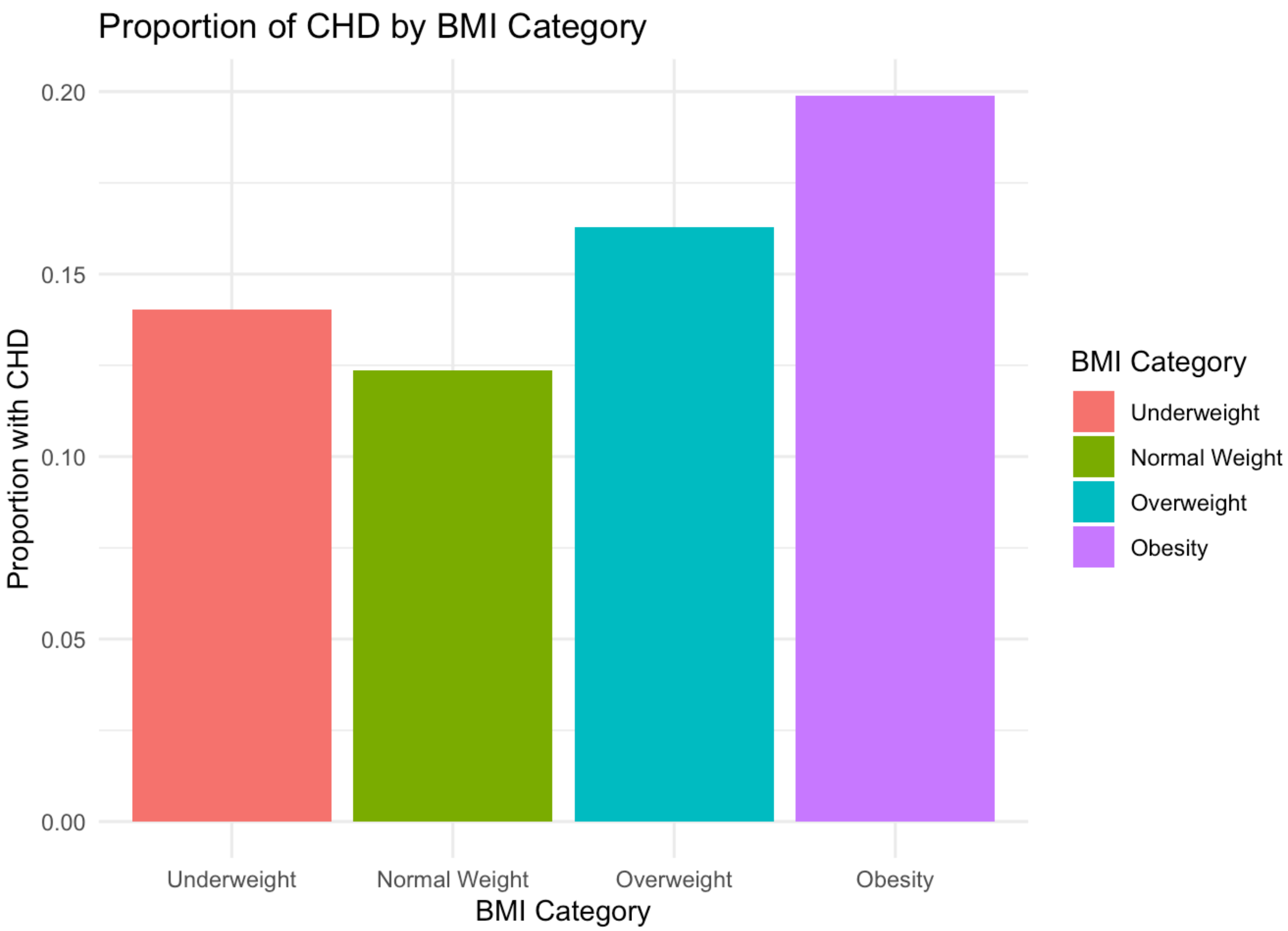


- Age Range
    - Young Adult: Age < 40

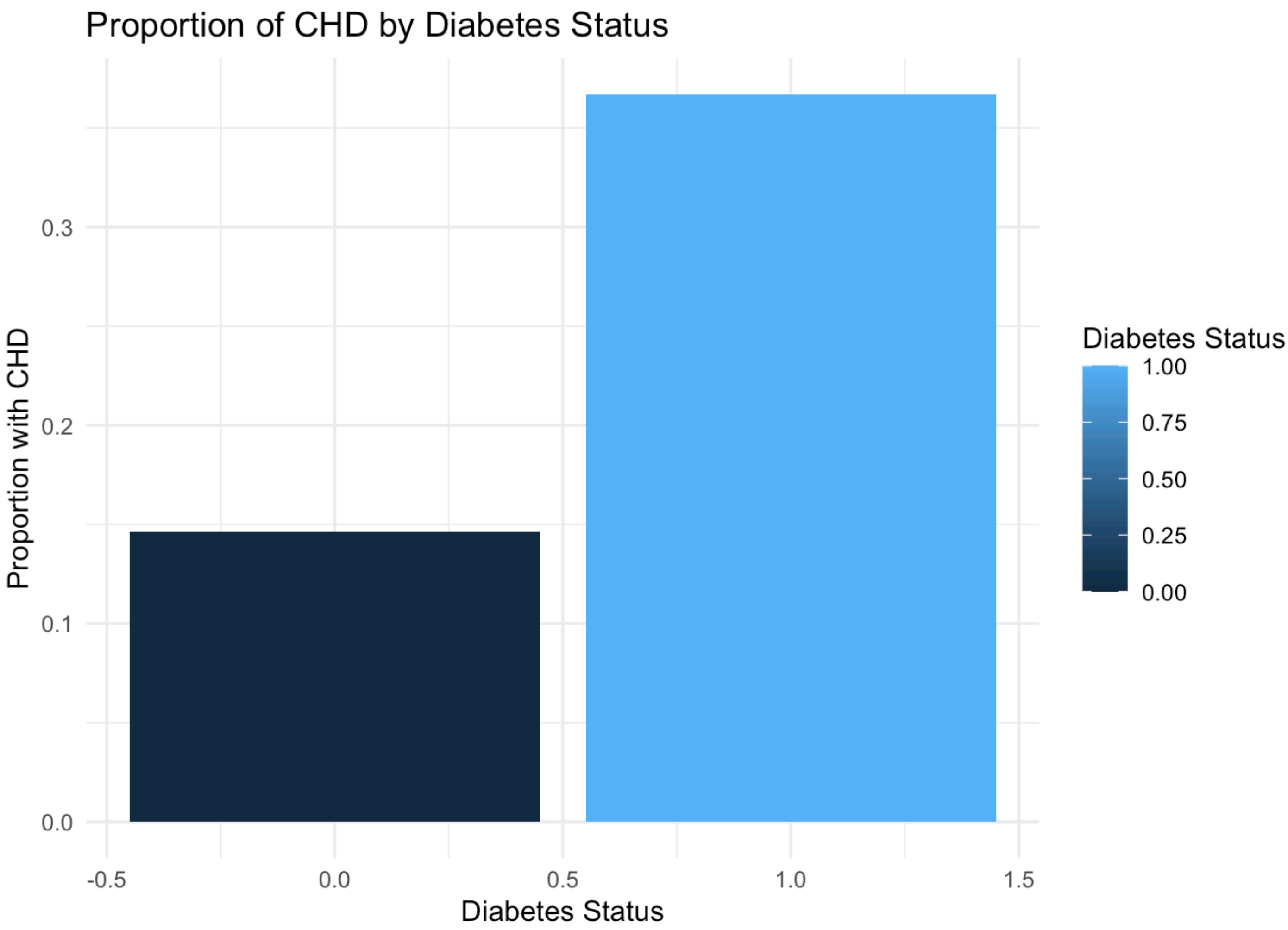
  - Middle-Aged Adult: 40 ≤ Age < 65
  - Elderly: Age ≥ 65
- BMI
  - Young Adult: Age < 40
  - Middle-Aged Adult: 40 ≤ Age < 65
  - Elderly: Age ≥ 65

```
# Categorize BMI
data$BMI_Category <- cut(data$BMI,
                         breaks = c(-Inf, 18.5, 24.9, 29.9, Inf),
                         labels = c("Underweight", "Normal Weight", "Overweight",
"Obesity"),
                         right = FALSE)

# Categorize Age
data$Age_Category <- cut(data$age,
                         breaks = c(-Inf, 40, 65, Inf),
                         labels = c("Young Adult", "Middle-Aged Adult", "Elderly")
,
                         right = FALSE)

# Calculate the proportion of individuals with and without CHD for each level of B
MI category
bmi_chd_prop <- aggregate(TenYearCHD ~ BMI_Category, data = data, FUN = function(x
) mean(x == 1))
names(bmi_chd_prop) <- c("BMI_Category", "CHD_Proportion")

# Create a grouped bar plot for BMI
grouped_bar_plot_bmi <- ggplot(bmi_chd_prop, aes(x = BMI_Category, y = CHD_Proport
ion, fill = BMI_Category)) +
  geom_bar(stat = "identity") +
  labs(x = "BMI Category", y = "Proportion with CHD", fill = "BMI Category", title
= "Proportion of CHD by BMI Category") +
  theme_minimal()

# Calculate the proportion of individuals with and without CHD for each level of d
iabetes status
diabetes_chd_prop <- aggregate(TenYearCHD ~ diabetes, data = data, FUN = function(
x) mean(x == 1))
names(diabetes_chd_prop) <- c("Diabetes_Status", "CHD_Proportion")

# Create a grouped bar plot for Diabetes Status
grouped_bar_plot_diabetes <- ggplot(diabetes_chd_prop, aes(x = Diabetes_Status, y
= CHD_Proportion, fill = Diabetes_Status)) +
  geom_bar(stat = "identity") +
  labs(x = "Diabetes Status", y = "Proportion with CHD", fill = "Diabetes Status",
title = "Proportion of CHD by Diabetes Status") +
  theme_minimal()

# Calculate the proportion of individuals with and without CHD for each age catego
```

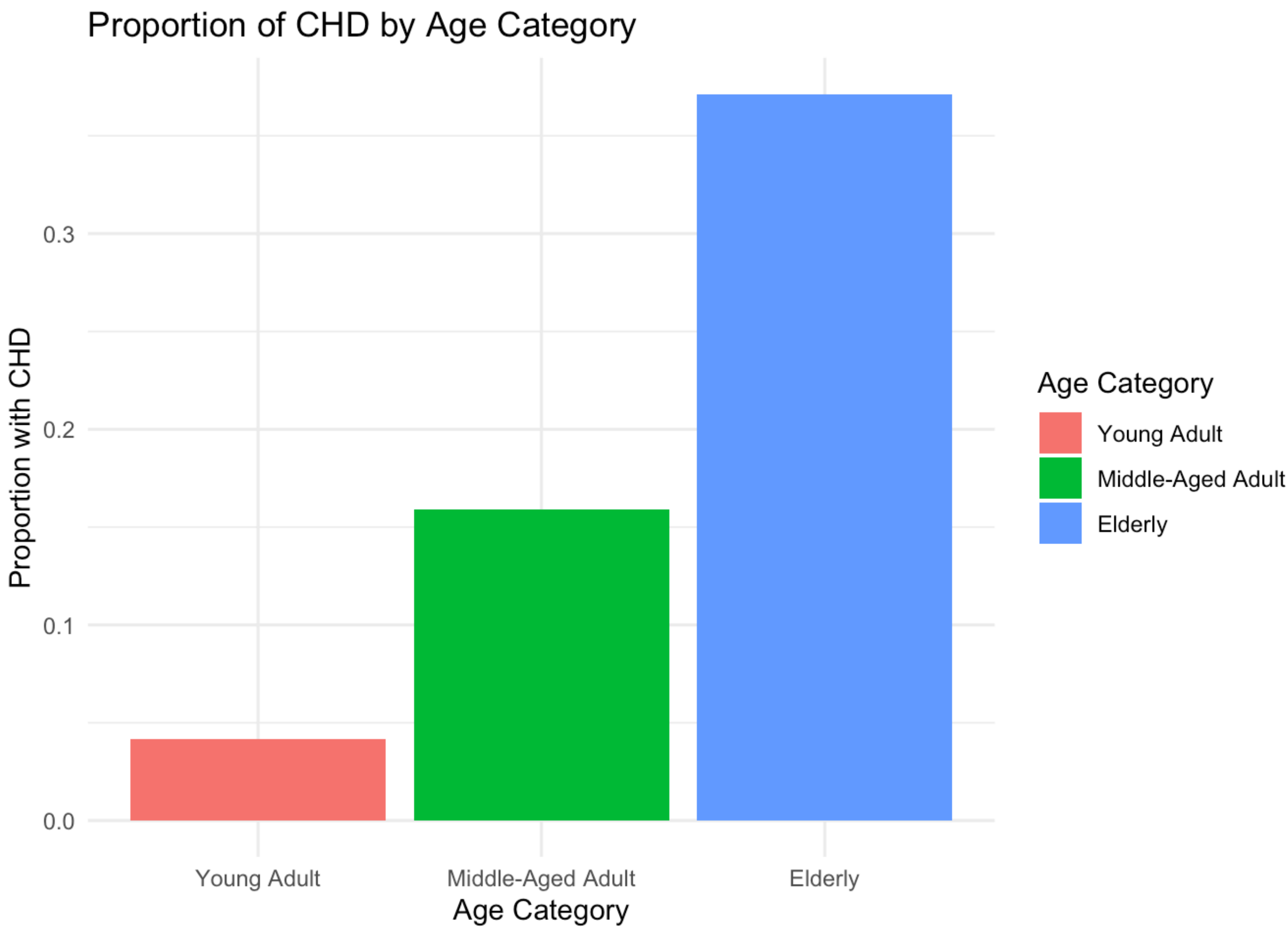
```
ry
age_chd_prop <- aggregate(TenYearCHD ~ Age_Category, data = data, FUN = function(x
) mean(x == 1))
names(age_chd_prop) <- c("Age_Category", "CHD_Proportion")

# Create a grouped bar plot for Age
grouped_bar_plot_age <- ggplot(age_chd_prop, aes(x = Age_Category, y = CHD_Proport
ion, fill = Age_Category)) +
  geom_bar(stat = "identity") +
  labs(x = "Age Category", y = "Proportion with CHD", fill = "Age Category", title
= "Proportion of CHD by Age Category") +
  theme_minimal()

# Display the grouped bar plots
print(grouped_bar_plot_bmi)
```




```
print(grouped_bar_plot_diabetes)
```




```
print(grouped_bar_plot_age)
```




```
head(clean_data)
```

```
##   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
```

Data Normalization

Numeric variables in the dataset are normalized using min-max scaling to ensure uniformity and prevent any single variable from dominating the model due to differences in scale. Min-max normalization was used to ensure ranges of zero to one since ROC utilizes probability.

```r
# Select only numeric columns except for the "activity" column from the dataset
numeric_data <- clean_data[, sapply(clean_data, is.numeric)]

# Min-max scaling to normalize between 0 and 1
min_max_scaled <- apply(numeric_data, 2, function(x) (x - min(x)) / (max(x) - min(
x)))

# Convert the scaled data back to a data frame
min_max_normalized_data <- as.data.frame(min_max_scaled)

head(min_max_normalized_data)
```

```
##   male       age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1 0.1842105 1.0000000             0  0.0000000      0               0
## 2    0 0.3684211 0.3333333             0  0.0000000      0               0
## 3    1 0.4210526 0.0000000             1  0.2857143      0               0
## 4    0 0.7631579 0.6666667             1  0.4285714      0               0
## 5    0 0.3684211 0.6666667             1  0.3285714      0               0
## 6    0 0.2894737 0.3333333             0  0.0000000      0               0
##   prevalentHyp diabetes   totChol     sysBP     diaBP       BMI heartRate
## 1            0        0 0.1683778 0.1063830 0.2328042 0.2770238 0.3636364
## 2            0        0 0.2813142 0.1773050 0.3492063 0.3196801 0.5151515
## 3            0        0 0.2710472 0.2080378 0.3386243 0.2375182 0.3131313
## 4            1        0 0.2299795 0.3144208 0.4973545 0.3160446 0.2121212
## 5            0        0 0.3531828 0.2198582 0.3809524 0.1832283 0.4141414
## 6            1        0 0.2361396 0.4562648 0.6560847 0.3577315 0.3333333
##      glucose TenYearCHD
## 1 0.10451977          0
## 2 0.10169492          0
## 3 0.08474576          0
## 4 0.17796610          1
## 5 0.12711864          0
## 6 0.16666667          0
```

Training Logistic Regression Models

Model Training

Logistic Regression

Logistic regression is employed as the primary machine learning model for predicting the probability of developing CHD. The glm() function is used to train the logistic regression model, and evaluation metrics such as accuracy, precision, recall, specificity, F1 score, and Matthews correlation coefficient (MCC) are computed to assess model performance.

I had a tremendously difficult time with a different library during the training process with cross validation, so I decided to take a step back and use the basic glm() function instead. It was an error due to multiplications of incorrect object dimensions.

- Error in dimnames(out) <- *vtmp* : length of ‘dimnames’ [2] not equal to array extent

The formula that was used for the logistic regression were the highly correlated values with the variable TenYearCHD.

TenYearCHD ~ male + age + sysBP + prevalentHyp + diaBP + glucose + diabetes

```
# Check unique values in TenYearCHD
unique_values <- unique(min_max_normalized_data$TenYearCHD)

# Check if there are any unexpected values
print(unique_values)
```

```
## [1] 0 1
```

```
head(min_max_normalized_data)
```

```
##   male       age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1 0.1842105 1.0000000             0  0.0000000      0               0
## 2    0 0.3684211 0.3333333             0  0.0000000      0               0
## 3    1 0.4210526 0.0000000             1  0.2857143      0               0
## 4    0 0.7631579 0.6666667             1  0.4285714      0               0
## 5    0 0.3684211 0.6666667             1  0.3285714      0               0
## 6    0 0.2894737 0.3333333             0  0.0000000      0               0
##   prevalentHyp diabetes   totChol     sysBP     diaBP       BMI heartRate
## 1            0        0 0.1683778 0.1063830 0.2328042 0.2770238 0.3636364
## 2            0        0 0.2813142 0.1773050 0.3492063 0.3196801 0.5151515
## 3            0        0 0.2710472 0.2080378 0.3386243 0.2375182 0.3131313
## 4            1        0 0.2299795 0.3144208 0.4973545 0.3160446 0.2121212
## 5            0        0 0.3531828 0.2198582 0.3809524 0.1832283 0.4141414
## 6            1        0 0.2361396 0.4562648 0.6560847 0.3577315 0.3333333
##      glucose TenYearCHD
## 1 0.10451977          0
## 2 0.10169492          0
## 3 0.08474576          0
## 4 0.17796610          1
## 5 0.12711864          0
## 6 0.16666667          0
```

```
# Check the number of rows for the column TenYearCHD
num_rows <- nrow(min_max_normalized_data$TenYearCHD)
print(num_rows)
```

```
## NULL
```

Training Test Data Split

```
# Split data into training and test sets
set.seed(123)  # for reproducibility
train_index <- createDataPartition(min_max_normalized_data$TenYearCHD, p = 0.8, li
st = FALSE)
train_data <- min_max_normalized_data[train_index, ]
test_data <- min_max_normalized_data[-train_index, ]
```

```
head(min_max_normalized_data)
```

```
##   male       age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1 0.1842105 1.0000000             0  0.0000000      0               0
## 2    0 0.3684211 0.3333333             0  0.0000000      0               0
## 3    1 0.4210526 0.0000000             1  0.2857143      0               0
## 4    0 0.7631579 0.6666667             1  0.4285714      0               0
## 5    0 0.3684211 0.6666667             1  0.3285714      0               0
## 6    0 0.2894737 0.3333333             0  0.0000000      0               0
##   prevalentHyp diabetes   totChol     sysBP     diaBP       BMI heartRate
## 1            0        0 0.1683778 0.1063830 0.2328042 0.2770238 0.3636364
## 2            0        0 0.2813142 0.1773050 0.3492063 0.3196801 0.5151515
## 3            0        0 0.2710472 0.2080378 0.3386243 0.2375182 0.3131313
## 4            1        0 0.2299795 0.3144208 0.4973545 0.3160446 0.2121212
## 5            0        0 0.3531828 0.2198582 0.3809524 0.1832283 0.4141414
## 6            1        0 0.2361396 0.4562648 0.6560847 0.3577315 0.3333333
##      glucose TenYearCHD
## 1 0.10451977          0
## 2 0.10169492          0
## 3 0.08474576          0
## 4 0.17796610          1
## 5 0.12711864          0
## 6 0.16666667          0
```

```
head(test_data)
```

```
##    male       age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 3     1 0.4210526 0.0000000             1 0.28571429      0               0
## 14    0 0.2368421 0.6666667             0 0.00000000      1               0
## 24    0 0.5263158 0.6666667             1 0.28571429      0               0
## 49    0 0.8157895 0.3333333             1 0.57142857      0               0
## 54    0 0.7894737 0.0000000             0 0.00000000      0               0
## 58    1 0.4473684 0.0000000             1 0.02857143      0               0
##    prevalentHyp diabetes   totChol     sysBP     diaBP       BMI heartRate
## 3             0        0 0.2710472 0.2080378 0.3386243 0.2375182 0.3131313
## 14            1        0 0.4496920 0.1914894 0.4232804 0.3822104 0.2121212
## 24            0        0 0.2094456 0.2293144 0.3597884 0.2319438 0.2727273
## 49            0        0 0.1355236 0.1536643 0.2222222 0.1602036 0.5151515
## 54            0        0 0.2607803 0.2907801 0.3650794 0.3085313 0.1919192
## 58            1        0 0.2915811 0.2836879 0.3492063 0.2450315 0.3131313
##       glucose TenYearCHD
## 3  0.08474576          0
## 14 0.12429379          0
## 24 0.09887006          0
## 49 0.09887006          1
## 54 0.09887006          0
## 58 0.11299435          0
```

```
head(train_data)
```

```
##   male       age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1 0.1842105 1.0000000             0  0.0000000      0               0
## 2    0 0.3684211 0.3333333             0  0.0000000      0               0
## 4    0 0.7631579 0.6666667             1  0.4285714      0               0
## 5    0 0.3684211 0.6666667             1  0.3285714      0               0
## 6    0 0.2894737 0.3333333             0  0.0000000      0               0
## 7    0 0.8157895 0.0000000             0  0.0000000      0               0
##   prevalentHyp diabetes   totChol     sysBP     diaBP       BMI heartRate
## 1            0        0 0.1683778 0.1063830 0.2328042 0.2770238 0.3636364
## 2            0        0 0.2813142 0.1773050 0.3492063 0.3196801 0.5151515
## 4            1        0 0.2299795 0.3144208 0.4973545 0.3160446 0.2121212
## 5            0        0 0.3531828 0.2198582 0.3809524 0.1832283 0.4141414
## 6            1        0 0.2361396 0.4562648 0.6560847 0.3577315 0.3333333
## 7            0        0 0.1889117 0.2576832 0.2433862 0.4258362 0.1616162
##     glucose TenYearCHD
## 1 0.1045198          0
## 2 0.1016949          0
## 4 0.1779661          1
## 5 0.1271186          0
## 6 0.1666667          0
## 7 0.1271186          1
```

```
# Ensure dimensions match
length(test_data$TenYearCHD)
```

```
## [1] 731
```

```
length(train_data$TenYearCHD)
```

```
## [1] 2925
```

```
str(test_data)
```

```
## 'data.frame':    731 obs. of  16 variables:
##  $ male           : num  1 0 0 0 0 1 1 1 0 0 ...
##  $ age            : num  0.421 0.237 0.526 0.816 0.789 ...
##  $ education      : num  0 0.667 0.667 0.333 0 ...
##  $ currentSmoker  : num  1 0 1 1 0 1 1 0 0 0 ...
##  $ cigsPerDay     : num  0.286 0 0.286 0.571 0 ...
##  $ BPMeds         : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : num  0 1 0 0 0 1 0 0 1 0 ...
##  $ diabetes       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : num  0.271 0.45 0.209 0.136 0.261 ...
##  $ sysBP          : num  0.208 0.191 0.229 0.154 0.291 ...
##  $ diaBP          : num  0.339 0.423 0.36 0.222 0.365 ...
##  $ BMI            : num  0.238 0.382 0.232 0.16 0.309 ...
##  $ heartRate      : num  0.313 0.212 0.273 0.515 0.192 ...
##  $ glucose        : num  0.0847 0.1243 0.0989 0.0989 0.0989 ...
##  $ TenYearCHD     : num  0 0 0 1 0 0 0 1 0 0 ...
```

```
str(train_data)
```

```
## 'data.frame':    2925 obs. of  16 variables:
##  $ male           : num  1 0 0 0 0 0 0 1 1 0 ...
##  $ age            : num  0.184 0.368 0.763 0.368 0.289 ...
##  $ education      : num  1 0.333 0.667 0.667 0.333 ...
##  $ currentSmoker  : num  0 0 1 1 0 0 1 0 1 0 ...
##  $ cigsPerDay     : num  0 0 0.429 0.329 0 ...
##  $ BPMeds         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : num  0 0 1 0 1 0 0 1 1 0 ...
##  $ diabetes       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : num  0.168 0.281 0.23 0.353 0.236 ...
##  $ sysBP          : num  0.106 0.177 0.314 0.22 0.456 ...
##  $ diaBP          : num  0.233 0.349 0.497 0.381 0.656 ...
##  $ BMI            : num  0.277 0.32 0.316 0.183 0.358 ...
##  $ heartRate      : num  0.364 0.515 0.212 0.414 0.333 ...
##  $ glucose        : num  0.105 0.102 0.178 0.127 0.167 ...
##  $ TenYearCHD     : num  0 0 1 0 0 1 0 0 0 0 ...
```

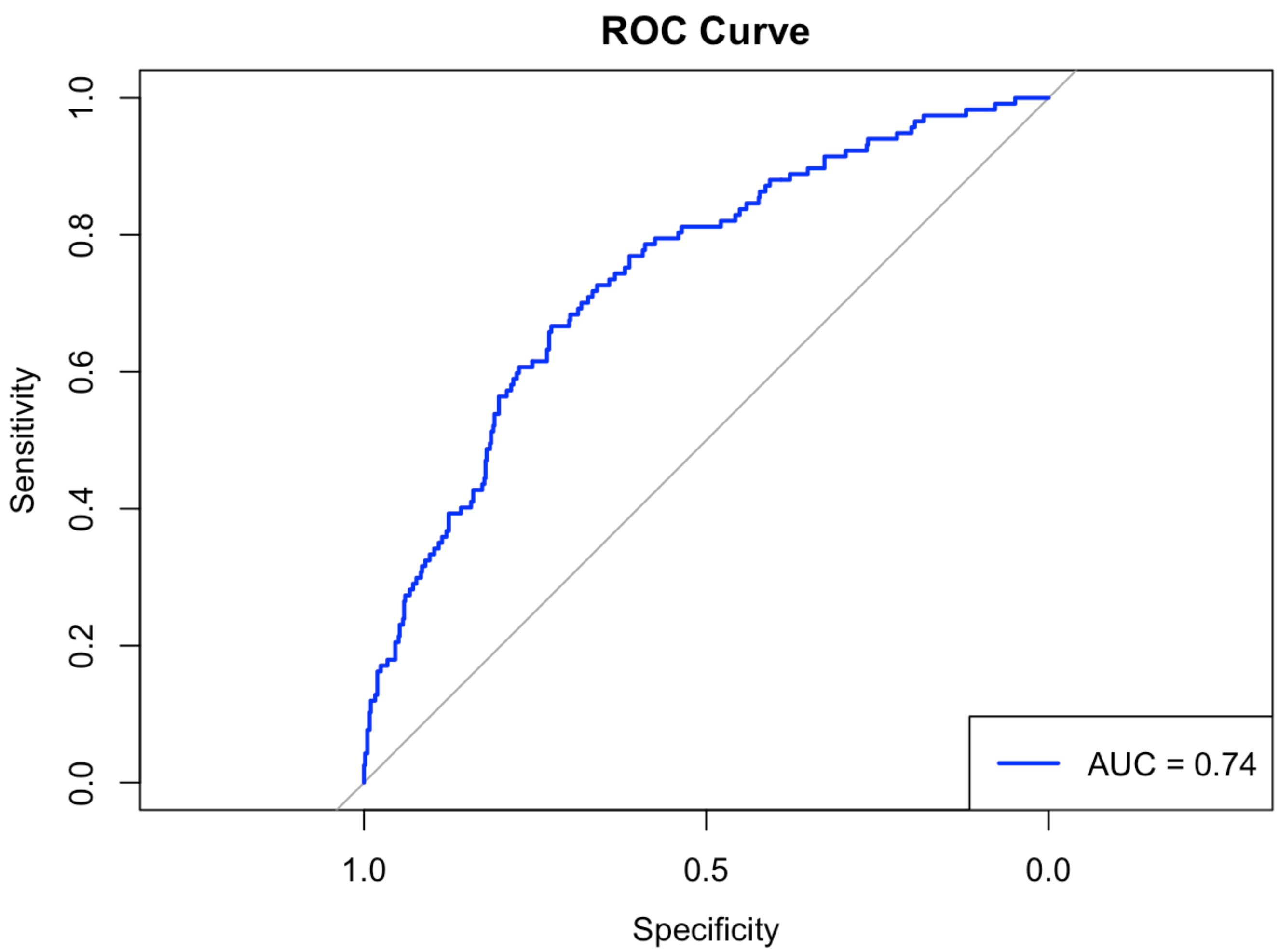
Model Training

```
# Define the formula with specific variables
formula <- as.formula("TenYearCHD ~ male + age + sysBP + prevalentHyp + diaBP + gl
ucose + diabetes")

# Train the logistic regression model
model <- glm(formula, data = train_data, family = binomial)

# Predict probabilities of positive class (1)
predictions <- predict(model, newdata = test_data, type = "response")

# Compute ROC curve
roc_curve <- roc(test_data$TenYearCHD, predictions)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
# Plot ROC curve
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)

# Add AUC to the plot
legend("bottomright", legend = paste("AUC =", round(auc(roc_curve), 2)), col = "bl
ue", lwd = 2)
```



```
summary(model)
```

```
##
## Call:
## glm(formula = formula, family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.1131     0.2542 -16.180  < 2e-16 ***
## male           0.6544     0.1125   5.815 6.07e-09 ***
## age            2.2340     0.2679   8.338  < 2e-16 ***
## sysBP          3.3312     0.8811   3.781 0.000156 ***
## prevalentHyp   0.3302     0.1542   2.142 0.032229 *
## diaBP         -0.6968     0.6577  -1.059 0.289396
## glucose        1.7789     0.8747   2.034 0.041986 *
## diabetes       0.2024     0.3608   0.561 0.574903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2477.2  on 2924  degrees of freedom
## Residual deviance: 2220.8  on 2917  degrees of freedom
## AIC: 2236.8
##
## Number of Fisher Scoring iterations: 5
```

```
levels(factor(round(predictions)))
```

```
## [1] "0" "1"
```

```
levels(test_data$TenYearCHD)
```

```
## NULL
```

```
# Get unique levels from both factors
all_levels <- union(levels(factor(round(predictions))), levels(test_data$TenYearCH
D))

# Set the same levels for both factors
predictions_factor <- factor(round(predictions), levels = all_levels)
actual_values_factor <- factor(test_data$TenYearCHD, levels = all_levels)
```

```
conf_mat <- confusionMatrix(predictions_factor, actual_values_factor)
# Extracting specific metrics
accuracy <- conf_mat$overall['Accuracy']
precision <- conf_mat$byClass['Pos Pred Value']
recall <- conf_mat$byClass['Sensitivity']
specificity <- conf_mat$byClass['Specificity']
f1_score <- (2 * precision * recall) / (precision + recall)
mcc <- cor(test_data$TenYearCHD, round(predictions))


# Print the evaluation metrics
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.8440492
```

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.8462604
```

```
cat("Recall:", recall, "\n")
```

```
## Recall: 0.995114
```

```
cat("Specificity:", specificity, "\n")
```

```
## Specificity: 0.05128205
```

```
cat("F1 Score:", f1_score, "\n")
```

```
## F1 Score: 0.9146707
```

```
cat("Matthews Correlation Coefficient:", mcc, "\n")
```

```
## Matthews Correlation Coefficient: 0.1542653
```

```r
# Set the threshold for predicting positive class
threshold <- 0.5

# Predict classes based on the threshold
predicted_classes <- ifelse(predictions > threshold, 1, 0)

# Create the confusion matrix
conf_matrix <- table(Actual = test_data$TenYearCHD, Predicted = predicted_classes)

# Print the confusion matrix
print(conf_matrix)
```

```
##       Predicted
## Actual   0   1
##      0 611   3
##      1 111   6
```