Heart Disease Prediction

Nazman, Ezgi

2024-10-17

Cleveland Heart Disease Data

Dataset Overview:

1. **Source:** UCI Machine Learning Repository - Heart Disease Dataset
2. **Purpose:** The primary goal is to predict the presence or absence of heart disease in a patient, based on a set of medical attributes.
3. **Original Data:** The dataset was originally collected by the Cleveland Clinic Foundation.

Key Characteristics of Cleveland Data:

1. **Total Instances:** 303 rows
2. **Total Attributes:** 14 columns (13 features and 1 target variable)
3. **Missing Values:** Some instances contain missing data, represented by ?.

Attribute Information:

The dataset contains 13 medical attributes (or features) and 1 target variable that indicates the presence of heart disease.

1. **age:** Age of the patient in years.
2. **sex:** Gender of the patient (1 = Male, 0 = Female).
3. **cp (chest pain type): 1=** Typical angina. **2=** Atypical angina. **3=** Non-anginal pain. **4=** Asymptomatic.
4. ***trestbps:*** Resting blood pressure (in mm Hg) on admission to the hospital.
5. **chol:** Serum cholesterol level (in mg/dL)**.**
6. **fbs** (fasting blood sugar):Whether the fasting blood sugar is greater than 120 mg/dL (1 = True, 0 = False).
7. **restecg** (resting electrocardiographic results): **0=** Normal. **1=** Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV). **2=**Showing probable or definite left ventricular hypertrophy by Estes' criteria.
8. **thalach:** Maximum heart rate achieved during exercise.

9. **exang:** Exercise-induced angina (**1** = Yes, **0** = No).
10. **oldpeak:** ST depression induced by exercise relative to rest (numeric value measured in mm).
11. **slope:**The slope of the peak exercise ST segment: **1=** Upsloping. **2=** Flat. **3=** Downsloping.
12. **ca:** Number of major vessels (0-3) colored by fluoroscopy (higher values indicate more blocked vessels).
13. **thal:** A blood disorder called thalassemia: **3** = Normal. **6** = Fixed defect. **7** = Reversible defect.
14. **num** (target variable): Diagnosis of heart disease (angiographic disease status). Originally a categorical variable ranging from 0 to 4. **0**= No heart disease. **1, 2, 3, 4**= Different levels of heart disease severity. In this study, this is simplified into a binary classification (**0** = No heart disease, **1** = Presence of heart disease).

```r
# Define the URL for the dataset
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

# Define column names
col_names <- c('age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
               'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num')

# Load the dataset directly from the URL
heart_data <- read.table(url, sep = ",", col.names = col_names, na.strings = "?")

# View the first six rows of the dataset

head(heart_data)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal num
## 1  63   1  1      145  233   1       2     150     0     2.3     3  0    6   0
## 2  67   1  4      160  286   0       2     108     1     1.5     2  3    3   2
## 3  67   1  4      120  229   0       2     129     1     2.6     2  2    7   1
## 4  37   1  3      130  250   0       0     187     0     3.5     3  0    3   0
## 5  41   0  2      130  204   0       2     172     0     1.4     1  0    3   0
## 6  56   1  2      120  236   0       0     178     0     0.8     1  0    3   0
```

```r
# Load necessary libraries for plotting
library(ggplot2)
library(gridExtra)
library(reshape2)
```

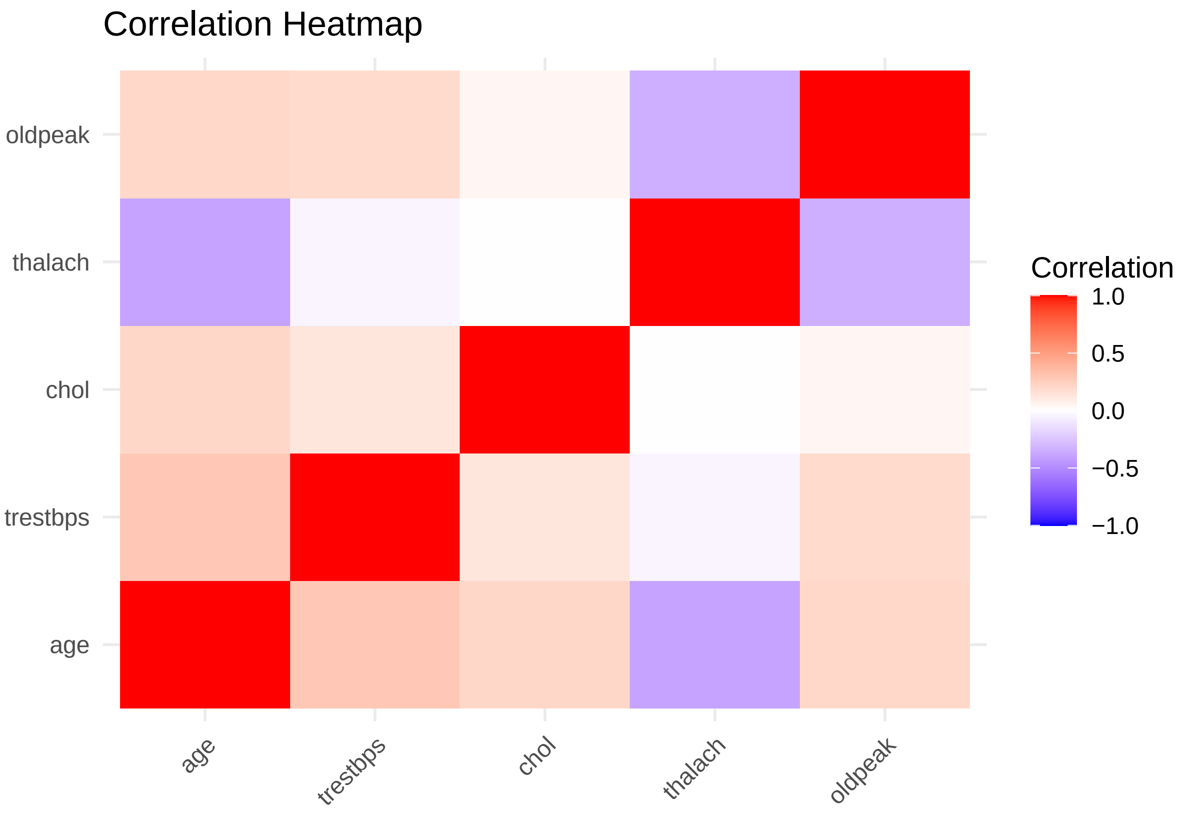
```r
# Select the relevant columns for correlation analysis
cor_data <- heart_data[, c('age', 'trestbps', 'chol', 'thalach', 'oldpeak')]

# Calculate the correlation matrix
cor_matrix <- cor(cor_data, use = "complete.obs")

# Melt the correlation matrix for ggplot
cor_melted <- melt(cor_matrix)

# Create the heatmap
cor_heatmap <- ggplot(data = cor_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), name = "Correlation") +
  theme_minimal() +
  labs(title = "Correlation Heatmap", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the heatmap
print(cor_heatmap)
```



```
#Histogram plot showing the distribution of patient ages
age_plot <- ggplot(heart_data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Frequency") +
  theme_minimal()
```

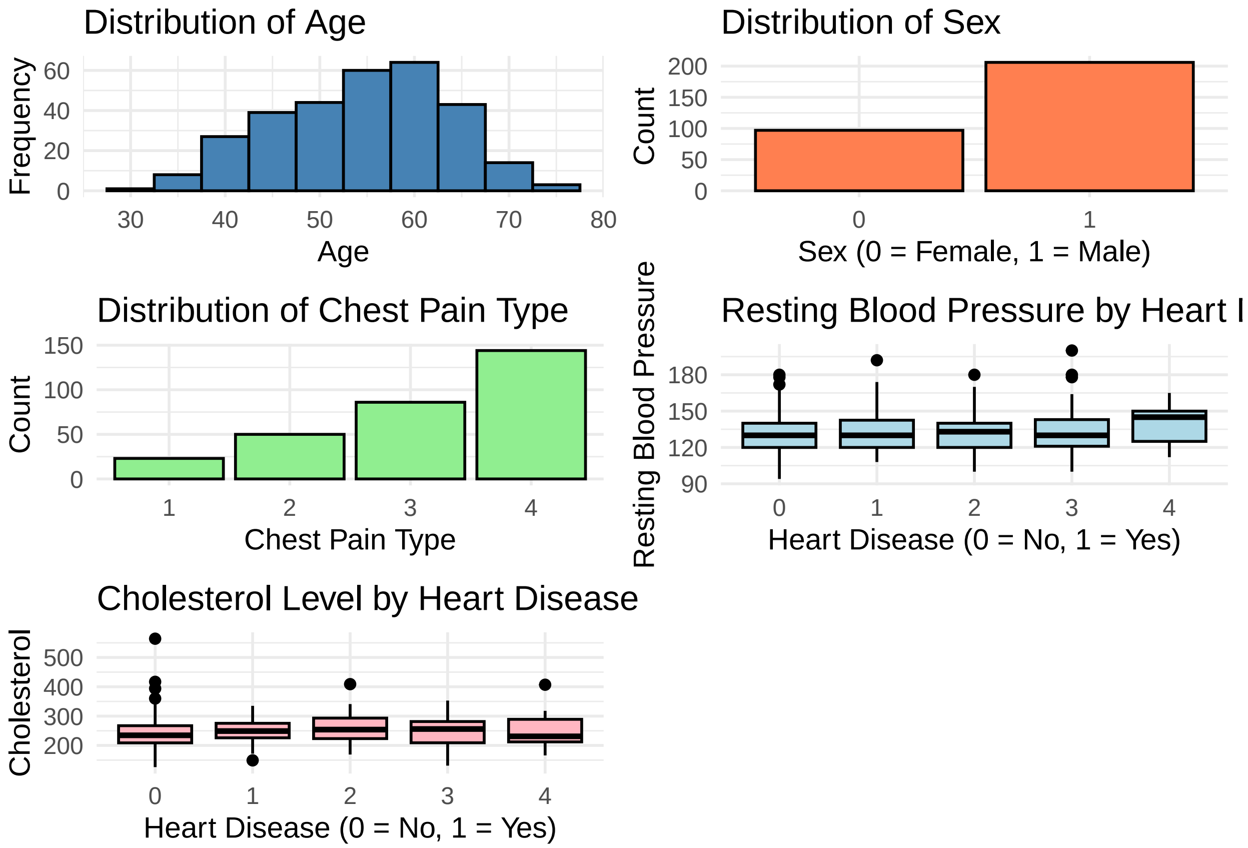
```
#Bar plot showing the count of male and female patients
sex_plot <- ggplot(heart_data, aes(x = factor(sex))) +
  geom_bar(fill = "coral", color = "black") +
  labs(title = "Distribution of Sex", x = "Sex (0 = Female, 1 = Male)", y = "Count") +
  theme_minimal()
```

```
#Bar plot showing different types of chest pain reported by the patients.
cp_plot <- ggplot(heart_data, aes(x = factor(cp))) +
  geom_bar(fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Chest Pain Type", x = "Chest Pain Type", y = "Count") +
  theme_minimal()
```

```
#Box plot comparing the resting blood pressure of patients with and without heart disease.
bp_plot <- ggplot(heart_data, aes(x = factor(num), y = trestbps)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Resting Blood Pressure by Heart Disease", x = "Heart Disease (0 = No, 1 = Yes)", y = "R
  theme_minimal()
```

```
#Box plot comparing the resting blood pressure of patients with and without heart disease.
chol_plot <- ggplot(heart_data, aes(x = factor(num), y = chol)) +
  geom_boxplot(fill = "lightpink", color = "black") +
  labs(title = "Cholesterol Level by Heart Disease", x = "Heart Disease (0 = No, 1 = Yes)", y = "Choles
  theme_minimal()
```

```
# Arrange the above plots in a grid for visualization
grid.arrange(age_plot, sex_plot, cp_plot, bp_plot, chol_plot, ncol = 2)
```



Data Preparation for binary Logistic Regression

Here the num column indicates the presence of heart disease (0 = No, 1 = Yes).

```
# Convert 'num' column to binary: 0 for no heart disease, 1 for any level of heart disease
heart_data$num <- ifelse(heart_data$num > 0, 1, 0)

# Remove rows with missing values
heart_data <- na.omit(heart_data)
```

```r
# Check the structure of the dataset
str(heart_data)
```

```
## 'data.frame':    297 obs. of  14 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ restecg : num  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : num  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ ca      : num  0 3 2 0 0 0 2 0 1 0 ...
##  $ thal    : num  6 3 7 3 3 3 3 3 7 7 ...
##  $ num     : num  0 1 1 0 0 0 1 0 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:6] 88 167 193 267 288 303
##   ..- attr(*, "names")= chr [1:6] "88" "167" "193" "267" ...
```

Model Fitting

1. Significant predictors (marked with * or *)**,** such as sex, chest pain type, blood pressure, maximum heart rate, exercise-induced angina, and thalassemia.

2. Coefficients indicate how each variable influences the likelihood of heart disease.

3. The p-values show the statistical significance of each predictor in the model.

```r
# Fit a binary logistic regression model to predict heart disease
model <- glm(num ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach +
             exang + oldpeak + slope + ca + thal,
             data = heart_data, family = binomial)

# View the summary of the model
summary(model)
```

```
## 
## Call:
## glm(formula = num ~ age + sex + cp + trestbps + chol + fbs + 
##     restecg + thalach + exang + oldpeak + slope + ca + thal, 
##     family = binomial, data = heart_data)
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.372042   2.879476  -2.560  0.01046 * 
## age         -0.014164   0.023970  -0.591  0.55459   
## sex          1.312073   0.488474   2.686  0.00723 **
## cp           0.575898   0.191197   3.012  0.00259 **
## trestbps     0.024044   0.010730   2.241  0.02504 * 
## chol         0.004995   0.003774   1.324  0.18561   
## fbs         -1.021918   0.555330  -1.840  0.06574 .
## restecg      0.245153   0.185005   1.325  0.18513
## thalach     -0.020665   0.010225  -2.021  0.04327 *
## exang        0.926104   0.413343   2.241  0.02506 *
## oldpeak      0.247386   0.211832   1.168  0.24287
## slope        0.570009   0.363085   1.570  0.11644
## ca           1.267719   0.265384   4.777 1.78e-06 ***
## thal         0.343936   0.100361   3.427  0.00061 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 204.69  on 283  degrees of freedom
## AIC: 232.69
##
## Number of Fisher Scoring iterations: 6
```

Key interpretation of the results:

1. The intercept is significant, suggesting a meaningful baseline log-odds for heart disease when all predictors are zero.
2. Sex (male) is a significant predictor, with males being more likely to have heart disease (p-value = 0.00723).
3. Chest pain type (cp) is highly significant (p-value = 0.00259), with more severe types of chest pain strongly associated with heart disease.
4. Resting blood pressure (trestbps) is also significant, showing that higher blood pressure slightly increases the risk (p-value = 0.02504).
5. Maximum heart rate achieved (thalach) is significant (p-value = 0.04327), with lower maximum heart rates increasing heart disease risk.
6. Exercise-induced angina (exang) is significant (p-value = 0.02506), increasing the likelihood of heart disease.
7. The number of major vessels colored by fluoroscopy (ca) is one of the most significant predictors (p-value = 1.78e-06), showing a strong relationship with heart disease.
8. Thalassemia (thal) is also highly significant (p-value = 0.00061), with certain types of thalassemia being associated with higher risk.

Non-significant predictors: Age, cholesterol, fasting blood sugar, resting electrocardiographic results, ST depression (oldpeak), and slope of the ST segment were not significant. Note that this might be due to characteristics specific to this dataset.

Model fit:

1. The model provides a good fit, with a substantial reduction in deviance (from Null deviance of 409.95 to Residual deviance of 204.69).
2. AIC (Akaike Information Criterion) = 232.69, suggesting that this model strikes a balance between goodness of fit and complexity.