Statistics for Data Science - Homework #1

Alberto Catalano - 257816

02/04/2025

Introduction

This work analyzes a dataset containing 4238 observations from a cardiovascular study conducted in the United States, investigating potential risk factors associated with the 10-year incidence of **coronary heart disease** (CHD). The dataset includes variables such as: sex, age, education level, smoking status, daily smoked cigarette, previous occurences of strokes or hypertension, presence of diabetes, cholesterol levels, diastolic blood pressure, body mass index and heart rate.
The **objective** of this work is to identify statistically significant predictors of CHD risk development. To achieve this,**generalized linear models** (logistic regression) and **k-nearest neighbors** (KNN) were employed

Methods

Before proceeding with the analysis, this section provides a brief introduction the two models used to assess the relationships between risk factors and CHD incidence:

- GLM (logistic regression), this model assumes a linear relationships between the predictors and the log-odds of the outcome.
- KNN, a non-parametric method that classifies each observation based on the majority class of its nearest neighbors in the feature space.

Data Exploration

As an initial step, the dataset is loaded and a copy is created to preserve the original data. The summary of the observations is then examined.

```
dataset <- read.csv("chd.csv")  #safety copy
dataf <- dataset
summary(dataf)
```

```
##      sex                 age          education         smoker
##  Length:4238        Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  Class :character   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Mode  :character   Median :49.00   Median :2.000   Median :0.0000
##                     Mean   :49.58   Mean   :1.979   Mean   :0.4941
##                     3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##                     Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                     NA's   :105
##       cpd              stroke              HTN            diabetes
##  Min.   : 0.000   Min.   :0.000000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.000   Median :0.000000   Median :0.0000   Median :0.00000
##  Mean   : 9.003   Mean   :0.005899   Mean   :0.3105   Mean   :0.02572
##  3rd Qu.:20.000   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :70.000   Max.   :1.000000   Max.   :1.0000   Max.   :1.00000
##  NA's   :29
```

```
##       chol            DBP              BMI              HR
##  Min.   :107.0   Min.   : 48.00   Min.   :15.54   Min.   : 44.00
##  1st Qu.:206.0   1st Qu.: 75.00   1st Qu.:23.07   1st Qu.: 68.00
##  Median :234.0   Median : 82.00   Median :25.40   Median : 75.00
##  Mean   :236.7   Mean   : 82.89   Mean   :25.80   Mean   : 75.88
##  3rd Qu.:263.0   3rd Qu.: 89.88   3rd Qu.:28.04   3rd Qu.: 83.00
##  Max.   :696.0   Max.   :142.50   Max.   :56.80   Max.   :143.00
##  NA's   :50                       NA's   :19      NA's   :1
##      CHD
##  Length:4238
##  Class :character
##  Mode  :character
##
##
##
##
```

From an initial examination of the dataset, two potential issues were identified:

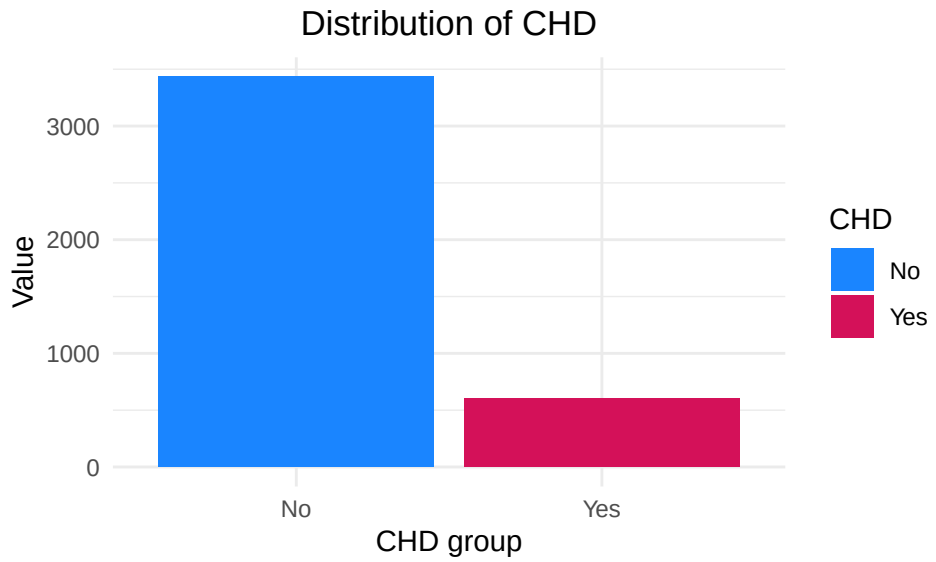
1. Several variables were incorrectly categorized as integers (*e.g. smoker*) or character strings (*e.g. CHD*), despite better fitting categorical data formats. To fix this, all those variables were converted into **factors**.
2. The summary revealed the presence of **missing values** (NAs), particularly in the variables related to education, cholesterol, cigarettes per day and BMI. Since most of these variables are binary, substituting missing values with the mean was not a viable option. Although this decision could be considered controversial, all the observations containing NAs were **removed**. As a result, the dataset was reduced to 4039 observation, corresponding to a 4% data loss. Below is the summary of the dataset after addressing these two issues.

```r
dataf <- na.omit(dataf)
summary(dataf)
```

```
##      sex            age        education smoker        cpd        stroke
##  Female:2297   Min.   :32.00   1:1681    0:2059   Min.   : 0.00   0:4016
##  Male  :1742   1st Qu.:42.00   2:1220    1:1980   1st Qu.: 0.00   1:  23
##                Median :49.00   3: 673             Median : 0.00
##                Mean   :49.53   4: 465             Mean   : 9.01
##                3rd Qu.:56.00                      3rd Qu.:20.00
##                Max.   :70.00                      Max.   :70.00
##  HTN      diabetes      chol            DBP              BMI
##  0:2783   0:3936   Min.   :113.0   Min.   : 48.00   Min.   :15.54
##  1:1256   1: 103   1st Qu.:206.0   1st Qu.: 75.00   1st Qu.:23.05
##                    Median :234.0   Median : 82.00   Median :25.36
##                    Mean   :236.7   Mean   : 82.87   Mean   :25.77
##                    3rd Qu.:263.0   3rd Qu.: 89.50   3rd Qu.:27.99
##                    Max.   :600.0   Max.   :142.50   Max.   :56.80
##        HR          CHD
##  Min.   : 44.00   No :3433
##  1st Qu.: 68.00   Yes: 606
##  Median : 75.00
##  Mean   : 75.87
##  3rd Qu.: 83.00
##  Max.   :143.00
```

We observe that the response variable CHD is a **binary categorical variable**, taking only two possible values: "Yes" or "No". To better understand its distribution, a **bar chart** is provided. It's possible to see that the majority of the individuals did not develop CHD.

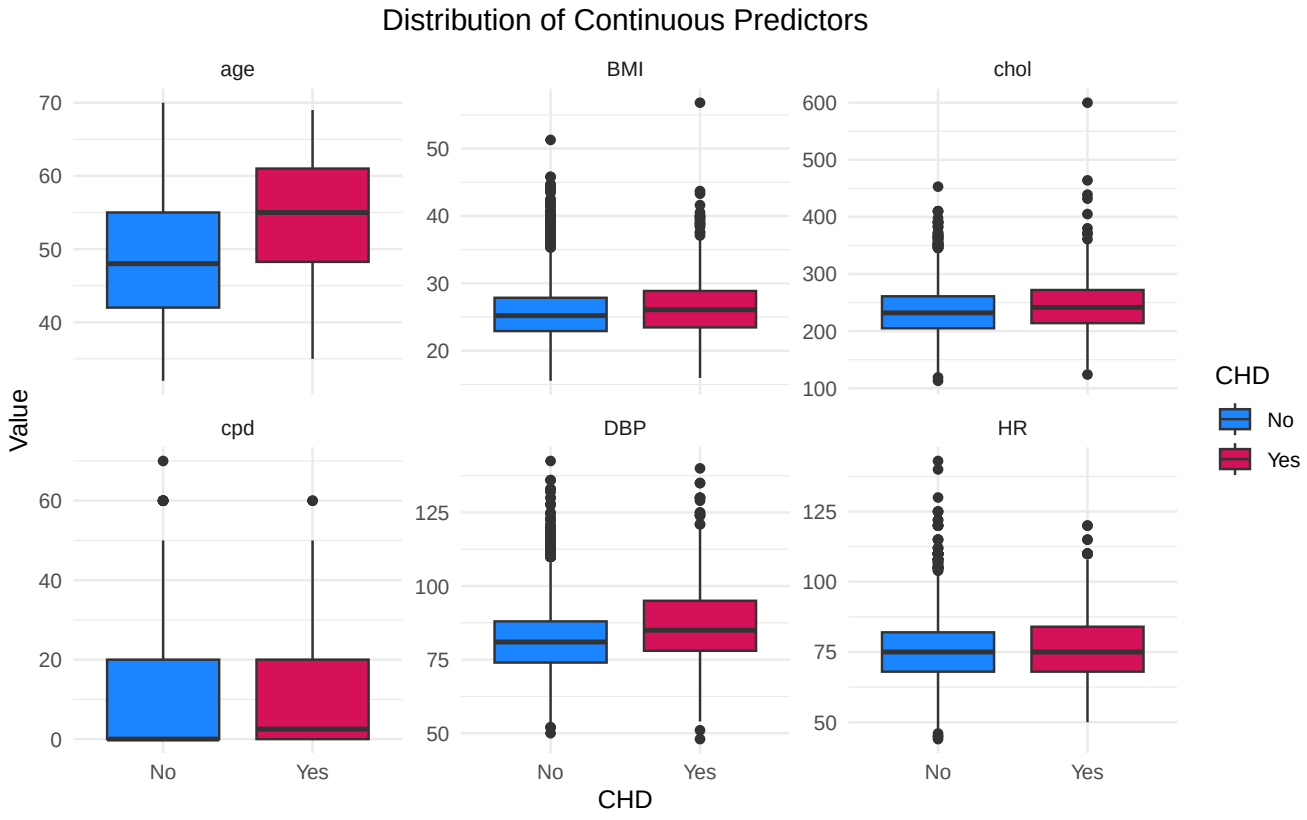
```r
ggplot(dataf, aes(x = CHD, fill = CHD)) + geom_bar() + labs(title = "Distribution of CHD",
    x = "CHD group", y = "Value") + scale_fill_manual(values = c("#1A85FF", "#D41159")) +
    theme_minimal() + theme(plot.title = element_text(hjust = 0.5))
```



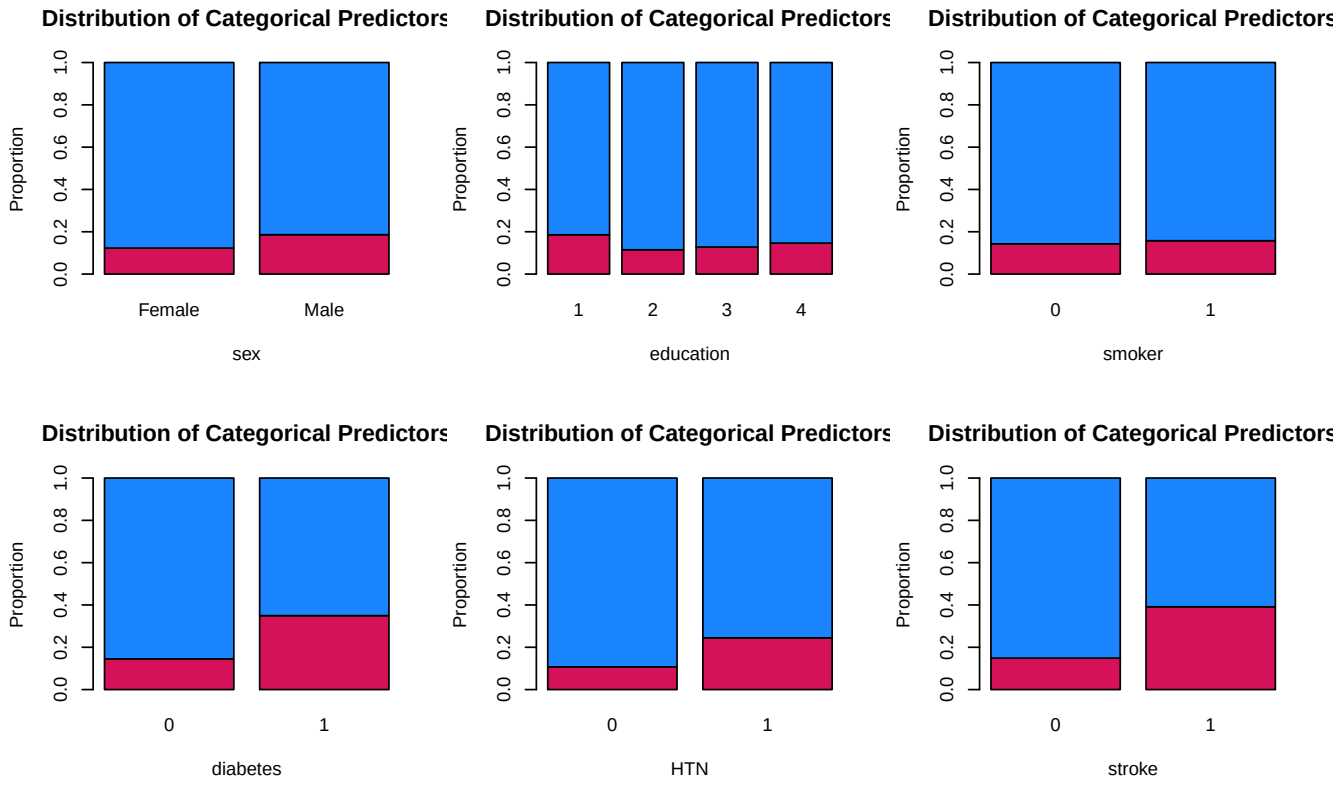
To conclude the preliminary analysis, a visualization of each predictor, to investigate their discriminative power, is made using:

- boxplots for **continuous predictors**
- bar plots for **discrete predictors**

```
cont_pred <- dataf %>%
    pivot_longer(cols = c(age, cpd, chol, DBP, BMI, HR), names_to = "Variable", values_to = "Value") %>%
    mutate(CHD = factor(CHD))
ggplot(cont_pred, aes(x = CHD, y = Value, fill = CHD)) + geom_boxplot() + facet_wrap(~Variable,
    scales = "free_y", ncol = 3) + theme_minimal() + scale_fill_manual(values = c("#1A85FF",
    "#D41159")) + labs(title = "Distribution of Continuous Predictors", y = "Value",
    x = "CHD") + theme(plot.title = element_text(hjust = 0.5))
```



```r
cat_pred <- c("sex", "education", "smoker", "diabetes", "HTN", "stroke")
response <- "CHD"
colors <- c("#D41159", "#1A85FF")
plotting_barplot <- function(predictor, data, response) {
    freq_table <- table(data[[predictor]], data[[response]])
    freq_table <- freq_table[, rev(colnames(freq_table))]
    prob_table <- prop.table(freq_table, margin = 1)
    barplot(t(prob_table), beside = FALSE, main = paste("Distribution of Categorical Predictors"),
        xlab = predictor, ylab = "Proportion", col = colors)
}
par(mfrow = c(2, 3))
invisible(lapply(cat_pred, plotting_barplot, data = dataf, response = response))
```



Splitting into training and test

To ensure reproducibility, the random seed was first set. The dataset was then divided into training and test sets to build the models. Given the imbalance in CHD, a 70/30 split was selected to maintain sufficient representation of both classes in the training set.
The probability tables demonstrate that the distribution of CHD is preserved in both the training and test datasets.

```r
set.seed(123)  #reproducibility
test <- createDataPartition(dataf$CHD, p = 0.7, list = FALSE)

tr_dataf <- dataf[-test, ]
ts_dataf <- dataf[test, ]

CHD_test <- ts_dataf$CHD
CHD_test <- as.character(CHD_test)

prop.table(table(tr_dataf$CHD))  #Training set distribution
prop.table(table(ts_dataf$CHD))  #Test set distribution
```

```
##
##        No       Yes
## 0.8504132 0.1495868
##
##        No       Yes
## 0.8497702 0.1502298
```

Logistic regression

Now it's possible to fit the logistic regression using all available parameters. The *summary(lreg)* function provides several useful pieces of information. First, it displays the estimated coefficiennts for each predictor in the mode, also with their respective significance levels. Parameters with a p-value < 0.05 are considered statistically significant (marked with *). The other parameters which have p-values greater than 0.05 are not statistically significant and have a weaker impact on predicting CHD risk. The **Null deviance** represents the goodness of fit of a model that includes only the intercept, serving as a baseline. The **residual deviance**, which is slightly lower, suggests that including all predictor variables improves the model's fit and is important for the model's quality. **AIC** (Akaike Information Criterion) is another criterion used for model comparison, with lower AIC values indicating a better-fitting model.

```r
lreg <- glm(CHD ~ sex + age + education + smoker + cpd + HTN + diabetes + chol +
    DBP + BMI + HR, data = tr_dataf, family = binomial)
summary(lreg)
```

```
##
## Call:
## glm(formula = CHD ~ sex + age + education + smoker + cpd + HTN +
##     diabetes + chol + DBP + BMI + HR, family = binomial, data = tr_dataf)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.438726   1.237620  -6.011 1.85e-09 ***
## sexMale      0.290415   0.190164   1.527  0.12671
## age          0.078567   0.011648   6.745 1.53e-11 ***
## education2  -0.522008   0.226791  -2.302  0.02135 *
## education3  -0.288937   0.255189  -1.132  0.25753
## education4  -0.201383   0.286196  -0.704  0.48165
## smoker1     -0.195934   0.291067  -0.673  0.50085
## cpd          0.025947   0.012137   2.138  0.03253 *
## HTN1         0.580919   0.225256   2.579  0.00991 **
## diabetes1    0.532194   0.423712   1.256  0.20911
## chol         0.005067   0.001971   2.571  0.01015 *
## DBP          0.005831   0.008976   0.650  0.51590
## BMI         -0.020887   0.022893  -0.912  0.36159
## HR           0.002518   0.007256   0.347  0.72860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1021.22  on 1209  degrees of freedom
## Residual deviance:  893.84  on 1196  degrees of freedom
## AIC: 921.84
##
## Number of Fisher Scoring iterations: 5
```

Next the predicted probability of CHD for each individual in the test set is computed, as estimated by the logistic regression model. The first 10 predicted probabilities are shown below. These probabilities represent the likelihood that each individual in the test set belongs to the "Yes" category (having CHD).

```
lreg_probs <- predict(lreg, data = ts_dataf, type = "response")
lreg_probs[1:10]
```

```
##          3          7          8         14         16         17         18
## 0.15623253 0.17126731 0.08658591 0.09794070 0.06702465 0.16420831 0.08462268
##         19         27         32
## 0.02069208 0.20162634 0.14249159
```

```
contrasts(CHD)
```

```
##     Yes
## No    0
## Yes   1
```

Before proceeding, the performance of the logistic regression model is assessed by comparing its predictions to the actual outcomes in the test set. The confusion matrix below shows the number of instances that were correctly or incorrectly predicted. Additionally, the accuracy and error rate of the model are provided.

```
lreg_pred <- rep("No", nrow(ts_dataf))  # placeholder with all 'No'
lreg_pred[lreg_probs > 0.5] <- "Yes"  # replace with 'Yes' if probability is > 0.5 (threshold)

table(lreg_pred, CHD_test)  #confusion matrix
```

```
##          CHD_test
## lreg_pred   No  Yes
##       No  2377  416
##       Yes   27    9
```

```
mean(lreg_pred == CHD_test)  # accuracy
```

```
## [1] 0.8434076
```

```
mean(lreg_pred != CHD_test)  #error rate
```

```
## [1] 0.1565924
```

KNN

In this step, a k-nearest neighbors classification model is trained and evaluated using only the continuous predictor variables. The parameter k, which represents the number of nearest neighbors considered in the classification, is set to 7. A detailed explanation on how this value was chosen will follow in the next section. As with the logistic regression model, the performance of the knn model is presented using a confusion matrix, accuracy and error rate values.

```
set.seed(123)
train_X <- tr_dataf %>%
    select(age, cpd, chol, DBP, BMI, HR)
test_X <- ts_dataf %>%
    select(age, cpd, chol, DBP, BMI, HR)
train_CHD <- tr_dataf$CHD  #will be the true target for the training set
test_CHD <- ts_dataf$CHD

knn_pred <- knn(train_X, test_X, train_CHD, k = 7, prob = T)
knn_probs <- attr(knn_pred, "prob")
knn_probs <- ifelse(knn_pred == "Yes", knn_probs, 1 - knn_probs)  # Convert to match CHD = 1

table(knn_pred, test_CHD)  # confusion matrix
```

```
##          test_CHD
## knn_pred   No  Yes
##       No  2366  413
##       Yes   38   12
```

```r
mean(knn_pred == test_CHD)  # accuracy
```

```
## [1] 0.8405797
```

```r
mean(knn_pred != CHD_test)  # error rate
```

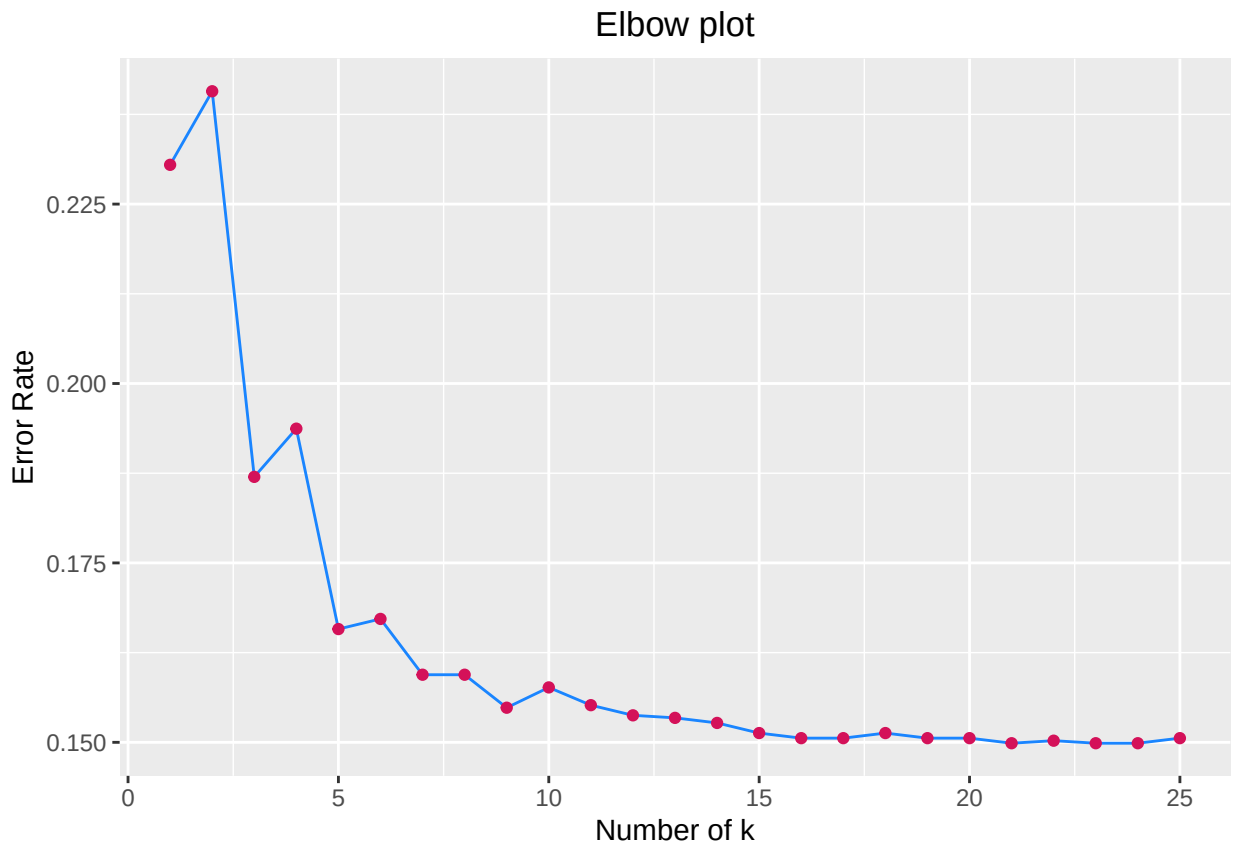
```
## [1] 0.1594203
```

The parameter K was selected using an **elbow plot**. The method involves plotting the model's error rate against different values of k and identifyng the point where the error rate stops decreasing significantly, forming an "elbow". This point represents the optimal value of K, as increasing K further brings minimal improvement. Additionally an odd number was chosen to prevent ties, since the classification problem involves only two classes. Using an odd value helps avoiding ties when the neighbors are evenly split between the two classes.

```r
error_rate <- numeric()
for (i in 1:25) {
    knn_pred <- knn(train_X, test_X, train_CHD, k = i)
    error_rate <- c(error_rate, mean(knn_pred != test_CHD))
}

error_df <- data.frame(k = 1:25, error_rate = error_rate)

ggplot(error_df, aes(x = k, y = error_rate)) + geom_line(color = "#1A85FF") + geom_point(color = "#D41159"
    labs(title = "Elbow plot", x = "Number of k", y = "Error Rate") + theme(plot.title = element_text(hjus
```



Conclusion

To draw some conclusions, first thing to do is comparing the accuracy of the two models.

```
mean(lreg_pred == CHD_test)  #logistic regression
```

```
## [1] 0.8434076
```

```
mean(knn_pred == test_CHD)  #knn
```

```
## [1] 0.8494168
```

In this analysis, the knn model slightly outperforms the logistic regression model. However, the difference in performance is marginal, indicating that the two models are quite comparable. Therefore, it would not be appropriate to assert that one method is definitively more suitable than the other for predicting the risk of developing CHD.

An additional point of concern is the significance of the predictor variables in the logistic regression model. It is surprising that a variable like education holds more significance in predicting CHD risk compared to other more clinically relevant variables, such as smoking status, BMI, heart rate, and diastolic blood pressure (all of which are well-documented risk factors for CHD in the existing literature).

Last important aspect is the distribution of variables such as diabetes, HTN, and stroke. The barplots suggest a stronger correlation with CHD; however, this intuition is not supported by statistical significance in the logistic regression.The same pattern is observed in discrete variables such as age and DBP.

Note: The following libraries were used in this analysis. With these libraries and the dataset file, the analysis can be fully reproduced.

```
library(tidyverse)
library(tidymodels)
library(ISLR2)
library(class)  #knn
library(caret)
library(ggplot2)
```