Statistical Learning Project

Andrea Mauro

Introduction

Cardiovascular diseases are among the leading causes of morbidity and mortality worldwide. In particular, **coronary heart disease (`CHD`)** is a critical condition that can lead to severe complications such as myocardial infarction and heart failure. Identifying the risk factors associated with the development of CHD is essential for prevention and management.

This study analyzes a dataset from a cardiovascular study and the following report addresses the following key objectives:

1. **Data Exploration**
2. **Dataset Splitting into Training and Test Sets**
3. **Statistical Modeling and Analysis 1 : logistic regression**
4. **Statistical Modeling and Analysis 2 : k-Nearest Neighbors**
5. **Model Performance Evaluation**
6. **Conclusions and Study Limitations**

Dataset presentation

```
     sex age education smoker cpd stroke HTN diabetes chol DBP   BMI HR CHD
1   Male  39         4      0   0      0   0        0  195  70 26.97 80  No
2 Female  46         2      0   0      0   0        0  250  81 28.73 95  No
3   Male  48         1      1  20      0   0        0  245  80 25.34 75  No
4 Female  61         3      1  30      0   1        0  225  95 28.58 65 Yes
5 Female  46         3      1  23      0   0        0  285  84 23.10 85  No
6 Female  43         2      0   0      0   1        0  228 110 30.30 77  No
```

Preprocessing

Before fitting the logistic regression model, it was necessary to appropriately encode categorical variables. In particular, the variables **`sex`** and `CHD`, originally stored as character data, were transformed into factors. Also was importang to check if the numbers of NAs could compromise the analysis: since the numbers of NAs is negligible the choice was to delete them.

```
Factor w/ 2 levels "Female","Male": 2 1 2 1 1 1 1 1 2 2 ...
```

```
Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 1 ...
```

Variables plotting

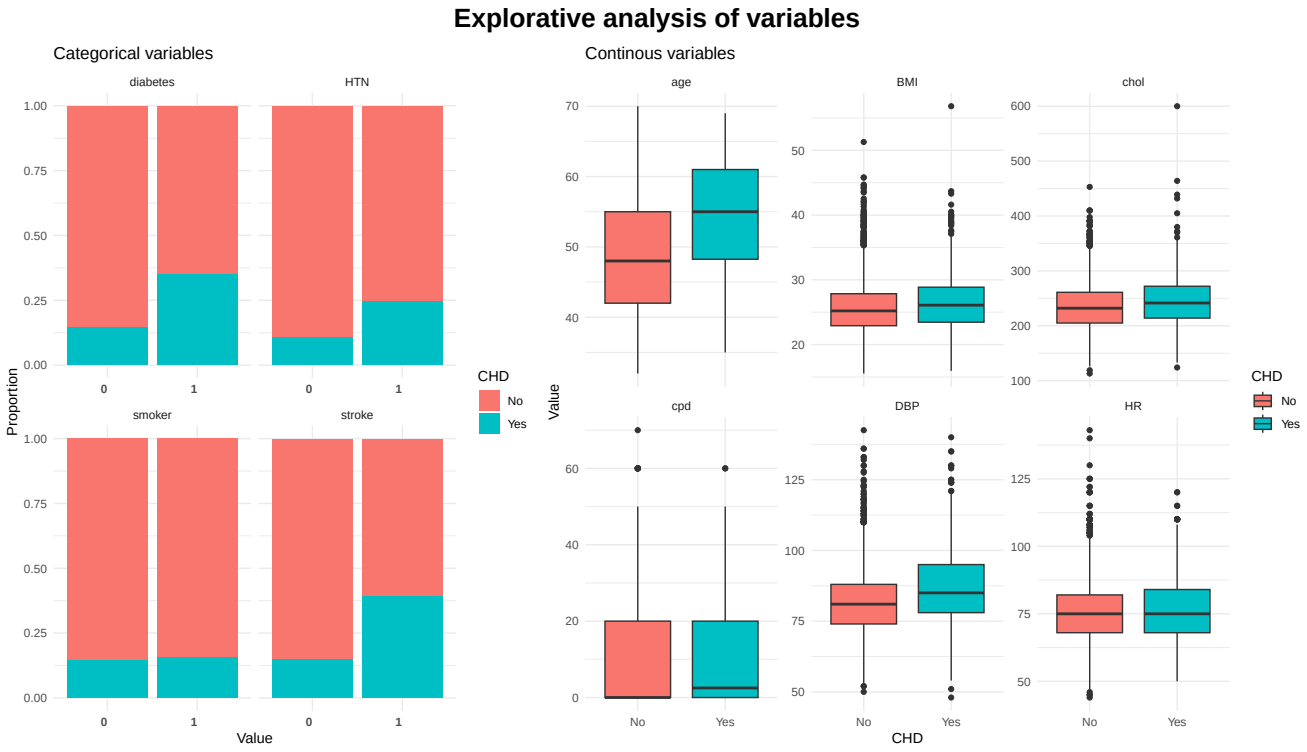
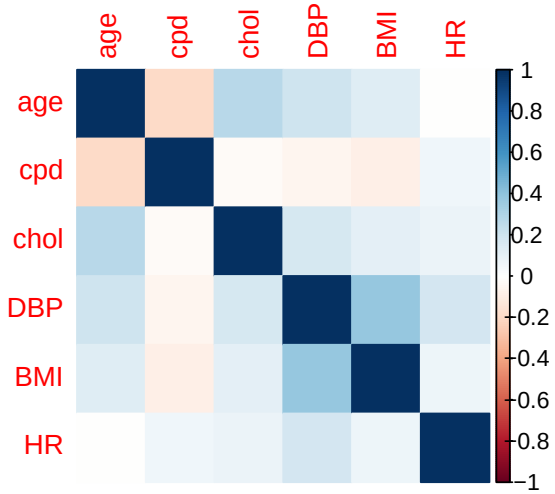


Figure 1

Covariance matrix

In evaluating the predictors of the disease, it was useful to configure a covariance matrix to determine whether some variables were collinear and if this could influence predictive models. No variable appears to be particularly correlated, except for `BMI` and `DBP`, which show a correlation coefficient that is relatively low and therefore negligible.



Splitting dataset

Prior to model development, the dataset was divided into training (80%) and testing (20%) portions.

```
# set random seed
set.seed(123)
# trainig and testing division
index <- createDataPartition(data$CHD, p = 0.8, list = FALSE,
    times = 1)
train_df <- data[index, ]
test_df <- data[-index, ]
```

Logistic regression (model 1)

Now, diving into the statistical modeling phase, the first model to be computed was a **logistic regression** using the cross validation method. The model was specified as :

$\text{logit}(E(CHD)) = 0 + 1\text{sex} + 2\text{age} + 3\text{education} + \ldots + 12\text{HR}$

```
# type of training and number of folds(k)
ctrlspecs <- trainControl(method = "cv", number = 5, savePredictions = "all",
    classProbs = TRUE)
# set random seed
set.seed(123)
# logistic regression setting using 'train' from caret
# package
model1 <- train(CHD ~ sex + age + education + smoker + cpd +
    stroke + HTN + diabetes + chol + DBP + BMI + HR, data = train_df,
    method = "glm", family = binomial, trControl = ctrlspecs)

# predict the outcome using model1 applied to test_df
predictions_1 <- predict(model1, newdata = test_df)
# creating a confusion matrix
conf_matrix_1 <- confusionMatrix(data = predictions_1, test_df$CHD,
    positive = "Yes")
conf_matrix_1
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  684 117
       Yes   2   4

               Accuracy : 0.8525
                 95% CI : (0.8262, 0.8763)
    No Information Rate : 0.8501
    P-Value [Acc > NIR] : 0.4457

                  Kappa : 0.0495
```

```
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.033058
            Specificity : 0.997085
         Pos Pred Value : 0.666667
         Neg Pred Value : 0.853933
             Prevalence : 0.149938
         Detection Rate : 0.004957
   Detection Prevalence : 0.007435
      Balanced Accuracy : 0.515071

       'Positive' Class : Yes
```

As we can notice, this first model is quite accurate, but it lacks in predicting the positive cases in the right way.

Resampling (model 2)

In this model, the class imbalance was adjusted through oversampling of the positive class (`CHD`="Yes"). This approach was necessary as the initial model, despite high overall accuracy, struggled to correctly identify true positive cases of coronary heart disease.The `education` variable was excluded from the final model as it demonstrated negligible predictive contribution (importance 0 in preliminary variable analysis)

```
# Oversampling for the class 'Yes' using 'up'
ctrlspecs <- trainControl(method = "cv", number = 5, savePredictions = "all",
    classProbs = TRUE, sampling = "up", summaryFunction = twoClassSummary,
    verboseIter = FALSE)
# Set random seed
set.seed(123)
# Oversampling logistic model
model2 <- train(CHD ~ sex + age + smoker + cpd + stroke + HTN +
    diabetes + chol + DBP + BMI + HR, data = train_df, method = "glm",
    family = binomial, trControl = ctrlspecs, metric = "Sens",
    maximize = TRUE)
# Valuation
predictions <- predict(model2, newdata = test_df, type = "raw")
confusionMatrix(predictions, test_df$CHD, positive = "Yes")
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  466  38
       Yes 220  83

               Accuracy : 0.6803
                 95% CI : (0.6469, 0.7124)
    No Information Rate : 0.8501
    P-Value [Acc > NIR] : 1
```

```
                 Kappa : 0.2255

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6860
            Specificity : 0.6793
         Pos Pred Value : 0.2739
         Neg Pred Value : 0.9246
             Prevalence : 0.1499
         Detection Rate : 0.1029
   Detection Prevalence : 0.3755
      Balanced Accuracy : 0.6826

       'Positive' Class : Yes
```

Selected model: Logistic Model 2 (oversampled)

While the initial model demonstrated high overall **accuracy** (85.3%), its critical limitation was an extreme imbalance in class-specific performance: near-perfect **specificity** (99.7% for 'No') but negligible **sensitivity** (3.3% for 'Yes'). This renders it clinically inadequate for CHD prediction, as it fails to identify 96.7% of actual cases. The model's apparent discriminative capacity (**AUC** = 0.727) is misleading, as it stems entirely from correct classification of negative cases while failing to rank positive cases correctly . In contrast, the selected model (with up-sampling) achieves balanced **sensitivity** (68.6%) and **specificity** (67.9%) with comparable **AUC** (0.721), but crucially demonstrates actual ability to distinguish positive cases. The trade-off in nominal **accuracy** (68.0% vs 85.3%) is justified by the clinical priority of identifying at-risk patients.

K-nn

The KNN model was trained on the same variables as the logistic regression, following standardization (mean=0, SD=1). The k parameter (number of neighbors) was optimized through 5-fold cross-validation by maximizing sensitivity. To address class imbalance, oversampling of the minority class ('Yes') was applied during training.

```r
# 3.KNN training
set.seed(123)
model_knn <- train(CHD ~ sex + age + smoker + cpd + stroke +
    HTN + diabetes + chol + DBP + BMI + HR, data = train_df,
    method = "knn", trControl = ctrl_knn, metric = "Sens", tuneLength = 10,
    preProcess = c("center", "scale"))
# 4. Valuation
knn_pred <- predict(model_knn, test_df)
confusionMatrix(knn_pred, test_df$CHD, positive = "Yes")
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
```

```
       No  427  44
       Yes 259  77

               Accuracy : 0.6245
                 95% CI : (0.5901, 0.6581)
    No Information Rate : 0.8501
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1495

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.63636
            Specificity : 0.62245
         Pos Pred Value : 0.22917
         Neg Pred Value : 0.90658
             Prevalence : 0.14994
         Detection Rate : 0.09542
   Detection Prevalence : 0.41636
      Balanced Accuracy : 0.62941

       'Positive' Class : Yes
```

The KNN model selected k=23 through systematic evaluation, as this neighborhood size optimally balanced detection capability (61.1% sensitivity) with specificity (63.3%) for reliable CHD screening. The 0.5 threshold maintained this equilibrium - sufficiently sensitive to identify true cases while avoiding excessive false positives that could overwhelm clinical workflows.

Conclusions and study limitations

1. Optimal Model Selection
The logistic regression demonstrates superior performance to KNN for CHD prediction, with clinically meaningful advantages:

- **Higher discriminative power**: AUC 0.721 vs KNN's AUC 0.664 (at k=23)
- **Better sensitivity-balanced accuracy tradeoff**:
  - Sensitivity: 68.6% (Logistic) vs 63.6% (KNN)
  - Balanced Accuracy: 68.3% (Logistic) vs 62.9% (KNN)
- **Stronger negative predictive value**: 92.5% (Logistic) vs 90.7% (KNN), critical for ruling out CHD

Clinical implication: The logistic model identifies ~5% more true CHD cases while maintaining better specificity - a decisive advantage for preventive cardiology where false negatives carry high risks.

2. Key Limitations
a) **Class imbalance**: Both models struggle with the low CHD prevalence (14.99%), despite upsampling. External validation in balanced cohorts is needed.

b) **Modest positive predictive values**:

- Logistic: 27.4%
- KNN: 22.9%

Statistical Summary Table

Metric	Logistic Regression	KNN	Clinical Preference
AUC	0.721	0.664 (at k=23)	Logistic
Sensitivity	68.6%	63.6%	Logistic
Specificity	67.9%	62.2%	Comparable
NPV	92.5%	90.7%	Logistic
PPV	27.4%	22.9%	(Both inadequate)