HOMEWORK 1

Luisa Porzio (ID: 255069)

2025-03-22

1 Introduction

The data here analyzed originate from a cardiovascular study conducted in the US, investigating the possible risk factors associated to the development of coronary heart disease (CHD) in a 10 years range.

2 Exploratory Analysis

The first step of the analysis is conducting a data pre-processing to check the structure of the data-set and the possible presence of missing values.

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

In addition, in order to proceed it is necessary to transform the nature of some variables from `characters` or `numeric` into `factors`.

```
data <- data %>%
    mutate(CHD = as.factor(CHD), sex = as.factor(sex),
        education = factor(education, levels = c(1,
            2, 3, 4), labels = c("no-high-school",
            "high-school-grad", "college-grad", "post-college")),
        smoker = factor(smoker, levels = c(0, 1), labels = c("smoker",
            "not-smoker")), stroke = factor(stroke,
            levels = c(0, 1), labels = c("No", "Yes")),
        HTN = factor(HTN, levels = c(0, 1), labels = c("No",
            "Yes")), diabetes = factor(diabetes, levels = c(0,
            1), labels = c("No", "Yes")))
```

2.1 Continuous Variables Analysis

The discriminative power of each predictor was then analyzed with the following density plots, showing that the most impacting predictors are Age, which has a lower mean in the group with No CHD compared then in the group with CHD, and DBP, which is slightly lower in the group without CHD compared to the group with CHD.

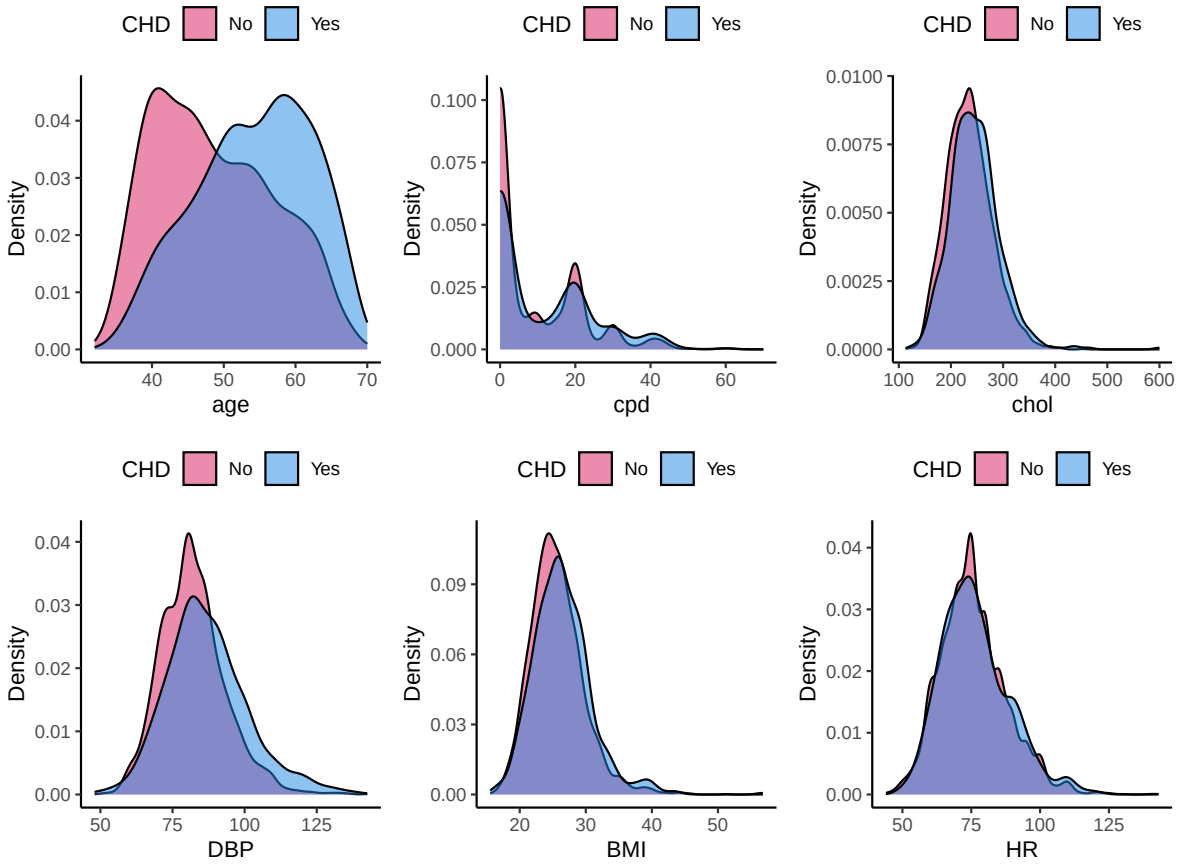

Figure 1: Distribution of Continuos Variables split by CHD

2.2 Categorical Variables Analysis

The following mosaic-plots show the distributions of each variable on the base of CHD presence. For example, we can see that sex present a slight imbalance, showing more females than males in the No CHD group. However, when CHD is present we have the opposite scenario, showing more males than females. Therefore, it is possible to unserstand how impacting is sex for developing CHD or not. The same can be said about Hypertension which behaves similarly to

the sex variable with respect to CHD presence. In addition, it important to underline that also the target variable is unbalanced. All the other variables do not present great imbalances.

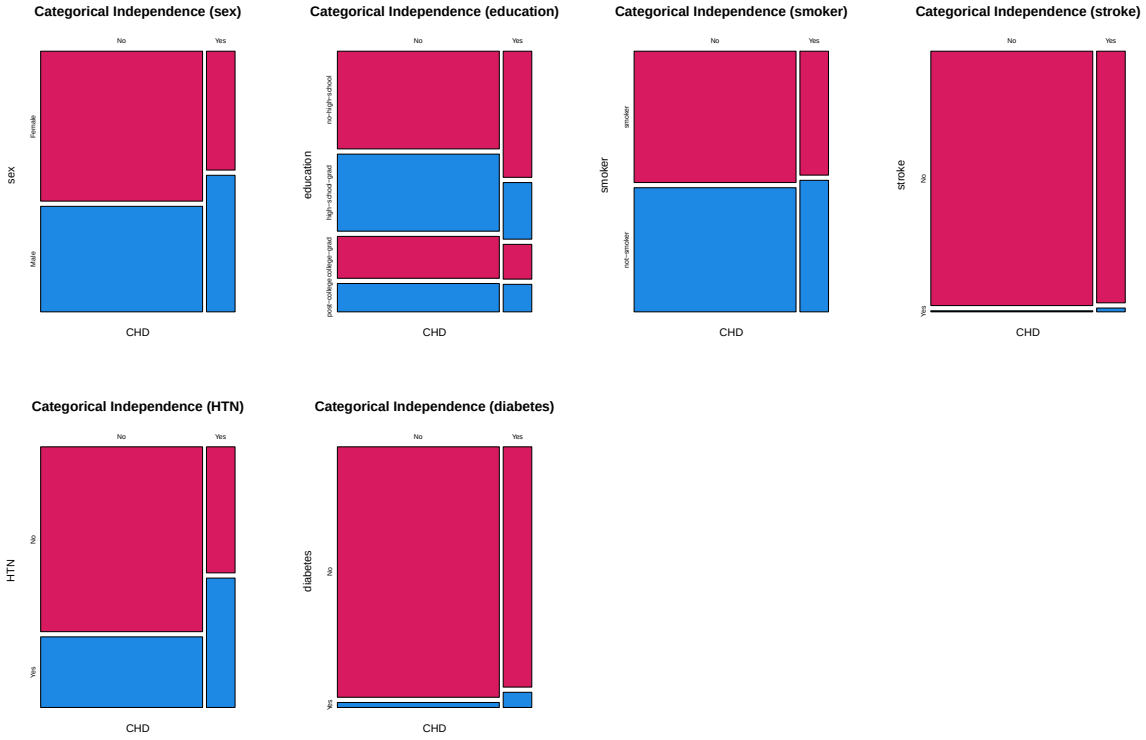

Figure 2: Visualization of Class Composition for Categorical Variables by CHD

3 Main Analysis

Due to the imbalance of the classes, a train and test data split that preserves the proportion of CHD in each split was deemed more appropriate. Therefore, the function below preserves the same proportion of target labels in both subsets.

```
set.seed(6)
train_idx <- createDataPartition(data$CHD, p = 0.75)

train_data <- data[train_idx$Resample1, ]
test_data <- data[-train_idx$Resample1, ]
```

3.1 GLM model - Logistic Regression

In the main analysis we are asked to fit a logistic regression model. Below is the code to replicate the model fitting.

```
mod <- glm(CHD ~ ., data = train_data, family = binomial)
```

Table 1: Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	-7.7087362	0.7594958	-10.1498071	0.0000000
sexMale	0.3987168	0.1182819	3.3709035	0.0007492
age	0.0711157	0.0072013	9.8753370	0.0000000
educationhigh-school-grad	-0.0802532	0.1339935	-0.5989334	0.5492173
educationcollege-grad	-0.1055224	0.1627060	-0.6485466	0.5166315
educationpost-college	0.0417912	0.1795143	0.2328014	0.8159156
smokernot-smoker	-0.0708884	0.1722802	-0.4114712	0.6807270
cpd	0.0288216	0.0067252	4.2855946	0.0000182
strokeYes	0.8193706	0.5099228	1.6068522	0.1080868
HTNYes	0.4194285	0.1382065	3.0347956	0.0024070
diabetesYes	0.7676702	0.2481240	3.0938968	0.0019755
chol	0.0022451	0.0012558	1.7877570	0.0738152
DBP	0.0127283	0.0054235	2.3468760	0.0189316
BMI	0.0022062	0.0139186	0.1585078	0.8740566
HR	0.0002903	0.0045212	0.0642127	0.9488009

In the table above we can see highlighted the coefficients which are statistically significant and that we can interpret.

- **Categorical variables** (e.g., sex, HTN, diabetes): The coefficient reflects the change in log-odds of developing CHD relative to the baseline group, assuming all other variables remain constant.
- **Numerical variables** (e.g., age, cpd, DBP): The coefficient represents the change in log-odds of developing CHD for each one-unit increase in the predictor, while holding all other variables constant.

3.2 K-NN Classifier

This section fits a K-NN classifier by performing a selection of the parameter k ranging from 1 to 50. Before fitting the K-NN, the data was scaled by converting all the categorical variables into numeric, except for the CHD variable. The results of the K-NN fitting are displayed in the confusion matrices in Section 4.1 as well as the process of choice of the optimal k.

```r
set.seed(6)
for (k in 1:50) {
    knn.pred <- knn(x_train, x_train, y_train, k = k)
}
```

4 Performance evaluation

In this section the K-NN fitting is evaluated, and then compared with the GLM method.

4.1 K-NN Performance Evaluation

The K-NN errors plot shown below was designed in order to decide which k minimizes the error. It is possible to see how the train data has a very low error rate when k=1, while their error rate is at maximum peak when k=50. Instead, when analyzing test data, the error rate is at its highest when k=1, which excludes 1 as possible choice of k. Therefore, k=50 was chosen as it minimizes at its best the error rate on the test data.

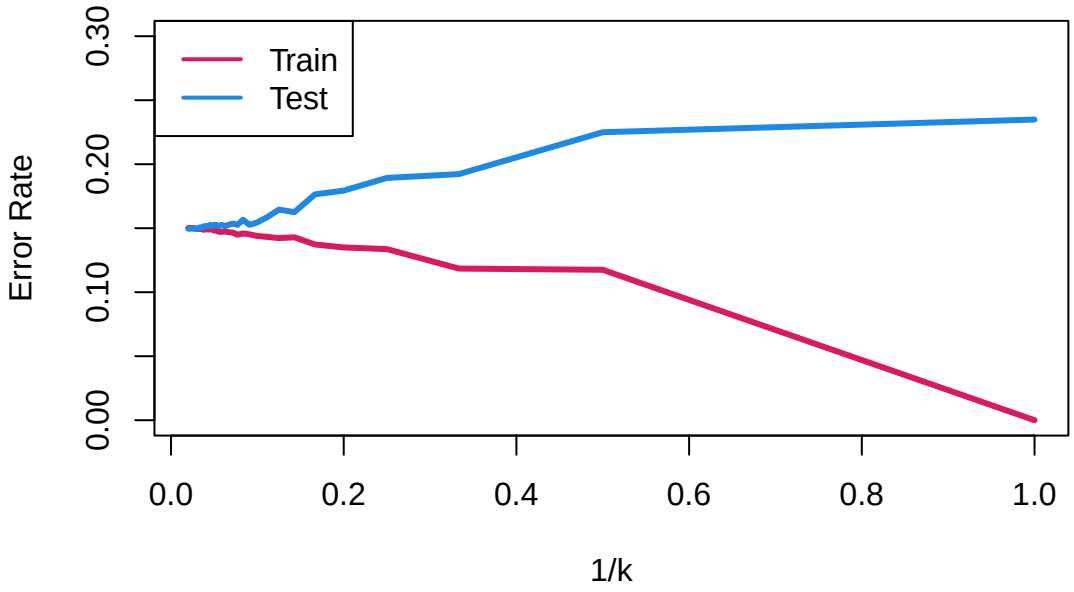

Figure 3: Performance on Train vs Test Data

Table 2: Confusion Matrix for K-NN Model

	No	Yes
No	858	0
Yes	151	0

4.1.1 Confusion Matrix K-NN Test Data

As mentioned in Section 3.2, this confusion matrix shows the overall performance of the K-NN model. It is possible to observe that it never makes a positive classification making a total of 151 false negatives.

4.2 GLM Performance Evaluation

This section evaluates the Logistic Regression Model.

4.2.1 Confusion Matrix Logistic Regression

The results displayed in the following confusion matrix provide a detailed breakdown of the model showing that due to class imbalance false negatives appear to be more frequent, meaning the model struggles to correctly predict CHD cases.

Table 3: Confusion Matrix for Logistic Regression

	No	Yes
No	854	4
Yes	146	5

The model provides an accuracy of 85,13% which is very high and can be misleading; indeed, due to the highly imbalanced dataset the model only learns to predict more "No CHD" cases as they are more present.

Table 4: Model's Accuracy, Precision, Recall and F1-score

Metric	Value
Accuracy	0.85
Precision	0.56
Recall	0.03
F1-Score	0.06

However, from the confusion matrix showed above it is possible to extract other metrics such as precision, recall, and F1-score. From this analysis it is possible to conclude that this model is not good for classification as it fails to correctly identify the majority of CHD positive cases. Possible solution to such problem will be covered in Section 5.

4.2.2 ROC Curve and AUC

The ROC curve illustrates the trade-off between the *true positive rate (sensitivity)* and *false positive rate (1-specificity)* at various threshold levels. Specifically, it shows how the Logistic Regression model performs on unseen (test) data.

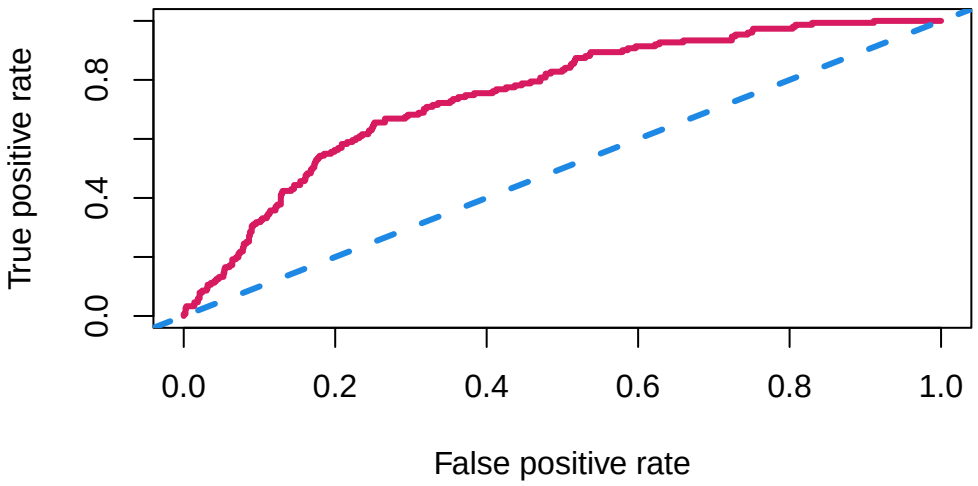

Figure 4: ROC Curve for the Logistic Regression Model

The overall AUC for the model described below has a value around 0.75, which is neither good nor bad, as it is neither closer to 1 nor closer to 0.5. Therefore, it is possible to consider this model not very strong and rather randomly guessing a response. However, as mentioned in Section 4.2.1, the model does not perform well at all as the few times that it correctly classifies CHD cases, they mainly fall in the "No-CHD" group.

```
Area under the curve: 0.7505
```

5 Discussion and Conclusion

The dataset here analyzed presents several issues stemming from class imbalance, affecting both the target variable predictions and other predictors influence. Logistic Regression and K-NN were employed to classify CHD cases, with results indicating that neither model performed optimally under the given dataset conditions. Indeed, the dataset contains significantly fewer CHD cases than non-CHD cases, leading to a bias where both models (K-NN and Logistic Regression) may favor predicting the majority class (No CHD). This imbalance affects both model training and coefficient interpretation.

5.0.1 Possible Improvements for the K-NN Model

As mentioned in Section 4.2.1, there are different possible solutions to this bias, such as:

1. Re-sampling techniques such as *Synthetic Minority Over-sampling* to generate more positive CHD cases
2. Threshold adjustment
3. Try different models such Gradient Boosting that could allow for more precision in capturing trends in the data. Also a Random Forest model could be implemented, even though it performs better on balanced dataset.

5.0.2 Possible Improvements for the Logistic Regression Model

The Logistic Regression model is equally affected by the bias generated by the data. One potential improvement that was not explored for the sake of the length of this report is the use of regularization techniques such as L1 (Lasso) or L2 (Ridge) penalties. These methods could help mitigate issues related to class imbalance by adjusting the decision boundary and preventing the model from being overly influenced by the majority class.