LOLIETT VALDES CASTILLO
valdes.loliett@gmail.com
LinkedIn – linkedin.com/in/loliett-valdes-castillo-3a1801254
https://lolavc.github.io

PREDICTOR MODEL PROJECT

Julio 2023

ABSTRACT

Background: The development of coronary heart disease (CHD) is a complex process involving genetic, lifestyle, and environmental factors (1-4). Monitoring CHD's epidemiology and clinical risk factors is crucial to understand its pathogenesis better. Developing effective strategies to prevent and manage CHD is possible, ultimately improving overall health outcomes for at-risk individuals. This report statistically analyses the results of a retrospective study focusing on risk factors and coronary heart disease (CHD) in South African males. It aims to build a predictor model to evaluate the presence of CHD, considering the risk factors studied.

Method: A data set of 452 cases and eight risk factors were analysed using RStudio 2023.03.1 Build 446 (© 2009-2023 Posit Software, PBC). Two third of the original data was allocated into a training dataset and the rest to a test dataset. A logistic regression model was fitted in the training dataset using backward selection. The optimal balance point for the sensitivity and specificity of the model was obtained through a ROC plot. The sensitivity, specificity and correct classification rate were evaluated in the training and test dataset. In addition, a principal component analysis (PCA) was performed, considering all the numerical variables after scaling the data and using the correlation matrix. The cumulative proportion of variance explained by the sequential components, the Kaiser Criterion and the Scree plot, were observed to decide the number of components to keep. Finally, a biplot was used to analyse the first two components.

Results: The means and medians in all the variables are higher in the group with a CHD diagnosis. The variables tobacco and alcohol are highly skewed due to the high percentage of non-consuming participants versus the low number of consumers. Also, the variable age is skewed in the group with CHD diagnosis due to a higher representation of participants over 40 years. Non-transformations were performed considering the logistic regression would be used. The fitted model included tobacco, ldl, typea and age, the latter being the most correlated with CHD. The training and test datasets' sensitivity, specificity and correct classification rate were below the ideal range, so the model lacks the accuracy sought. In the PCA, there were selected the first four components: PC1 primarily measures risk factors associated with the percentage of body fat, age and BMI; PC2 mainly contrast alcohol and tobacco consumption with LDL concentration and BMI; PC3 primarily measures risk factors associated with Type-A behaviour pattern and alcohol consumption; and PC4 is a contrast between alcohol consumption and BMI versus tobacco, the concentration of LDL and Type-A pattern.

Conclusion: The report showed that the risk factors included in the study are proportional to the risk of coronary heart disease, as proved in previous research (1, 2). Other variables known to be risk factors for the incidence of CHD, such as gender, family history, and diabetes mellitus, should be taken into account in the PCA and logistic regression model (5). Recent approaches to researching CHD also include the genetic risk score (3, 5), the occurrence of mental illness (6) and glycemia intake (7). Including the variables not accounted for in this study can contribute to explaining the highest proportion of variance with less than four components. Overall, this could improve the accuracy of the predictions and, therefore, the sensitivity and specificity of the logistic model.

Keywords: Coronary heart disease; logistic regression; PCA; principal component analysis; risk factors.

1. About the size, nature, and structure of the data.

```
load("CHD.RData")

set.seed(202250524)
z.omit <- sample(1:nrow(CHD), 10, replace=FALSE)

chd_zero <- CHD[-z.omit,]
dim(chd_zero)
## [1] 452  10

head(chd_zero[1,])
##   ind sbp tobacco  ldl adiposity typea obesity alcohol age CHD
## 1   1 160      12 5.73     23.11    49    25.3    97.2  52   1

chd.data <- chd_zero[,-1]

dim(chd.data)
## [1] 452   9
str(chd.data)
## 'data.frame':    452 obs. of  9 variables:
##  $ sbp      : int  160 144 118 170 134 132 142 114 114 132 ...
##  $ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
##  $ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
##  $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
##  $ typea    : int  49 55 52 51 60 62 59 62 49 69 ...
##  $ obesity  : num  25.3 28.9 29.1 32 26 ...
##  $ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
##  $ age      : int  52 63 46 58 49 45 38 58 29 53 ...
##  $ CHD      : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 1 2 1 2 ...
```

The size, nature and structure of the data are exposed in the following points:

- The dataset analysed in this report is a retrospective study focusing on the research of risk factors and coronary heart disease (CHD) in males. The data was collected in Western Cape, South Africa, throughout 2019. The patients could not be identified through scrutiny of the material, protecting the anonymity of the information.
- The original dataset had 462 cases, but after cleaning it (eliminating the unnecessary values), it ended up with 452 cases, ten fewer than the original.
- It had ten variables: Unique number of individuals (ind), Systolic blood pressure in mmHg (sbp), Yearly tobacco use in kg (tobacco), Low-density lipoprotein cholesterol in mmol/L (ldl), Percentage of body fat (adiposity), Type-A behaviour pattern score (typea), the Body mass index (obesity), Current alcohol consumption in litres (alcohol), Age in years (age) and diagnosed with coronary heart disease, (CHD).
- CHD is the response variable. It is a two-level factor, with a value of zero for the participants without a diagnosis of CHD, and one for participants diagnosed with CHD. It was decided to relabel it. The value zero was recorded as "No", and the value one as "Yes".
- The remaining variables are numerical: four discrete (ind, sbp, typea and age) and five continuous (tobacco, ldl, adiposity, obesity, and alcohol).
- The variable ind was not considered for the following analysis, and the dataset, without this variable, was renamed chd.data.

2. Summary statistics and exploring relationship of each potential risk factor with CHD status.

```
stxvar_chd.data <- describeBy(chd.data[,c(-9)], group=chd.data$CHD,
                               mat=TRUE, digits=2)
```

<u>NOTE</u>: The table 1 was obtained using the flextable package. Considering the readability, it is not shown all the R code: only the describeBy command for the descriptive statistics.

TABLE 1. CHD FOR EACH POTENTIAL RISK FACTOR

Risk Factor	CHD	n	Mean	SD	Median	Min	Max	Range
sbp	No	296	135.40	18.00	132.00	101.00	214.00	113.00
	Yes	156	143.54	23.44	138.00	102.00	218.00	116.00
tobacco	No	296	2.61	3.61	1.02	0.00	20.00	20.00
	Yes	156	5.61	5.60	4.19	0.00	31.20	31.20
ldl	No	296	4.33	1.77	3.99	0.98	11.61	10.63
	Yes	156	5.43	2.19	5.04	1.55	14.16	12.61
adiposity	No	296	23.95	7.81	24.67	6.74	42.06	35.32
	Yes	156	28.15	7.14	28.59	9.39	42.49	33.10
typea	No	296	52.37	9.54	52.00	13.00	77.00	64.00
	Yes	156	54.44	10.28	55.00	20.00	78.00	58.00
obesity	No	296	25.74	4.11	25.57	17.75	46.58	28.83
	Yes	156	26.62	4.44	26.41	14.70	45.72	31.02
alcohol	No	296	15.68	23.41	5.61	0.00	145.29	145.29
	Yes	156	18.93	26.01	8.33	0.00	147.19	147.19
age	No	296	38.81	14.87	40.00	15.00	64.00	49.00
	Yes	156	50.57	10.43	53.00	17.00	64.00	47.00

The principal findings about the descriptive statistic, presented in Table 1, are:

- There are 296 participants without a diagnosis of CHD and 156 with CHD, for an odds ratio of 0.527. Therefore, more than half of the participants (52.7%) have been diagnosed with this pathology.
- The means and medians in all the variables are higher in the group with a CHD diagnosis. The group without a previous diagnosis of CHD has a median age of 13 years younger than the group with a diagnosis of CHD (median with CHD = 53 vs without CHD = 40 years).
- The maximums are reached in the group with a CHD diagnosis, except in the variable age, which is the same for both groups (max = 64 years), and in obesity, which is slightly bigger in the group without CHD (max with CHD = 45.72 vs without CHD = 46.58 BMI). The differences between the maximum values are very small also in adiposity (max with CHD = 42.49 vs without CHD = 42.06 % of body fat), typea (max with CHD = 78 vs without CHD = 77 points), and alcohol (max with CHD = 147.19 vs without CHD = 145.29 litres).
- In the same way, the highest minimum values are founded in the group with a diagnosis of CHD, except in obesity (min with CHD = 14.70 vs without CHD = 17.75 BMI). In both groups, there are participants without tobacco use for at least one year and others that currently do not consume alcohol (min = 0 in both groups).
- Analysing the spread of the data, both groups have similar standard deviation values (SD) in adiposity (SD with CHD = 7.14 vs without CHD = 7.81 % of body fat), typea (SD with CHD = 10.28 vs without CHD = 9.54 points) and obesity (SD with CHD = 4.44 vs without CHD = 4.11 BMI), but this difference is clear in the variable age, where the spread is bigger for the group without CHD (SD with CHD = 10.43 vs without CHD
- =14.87 years). This is also noticeable for sbp (SD with CHD = 23.44 vs without CHD = 18.00 mmHg), tobacco (SD with CHD = 5.60 vs without CHD = 3.61 kg), ldl (SD with CHD = 2.19 vs without CHD = 1.17 mmol/L) and

alcohol (SD with CHD = 26.01 vs without CHD = 23.41 litres), where the group with CHD has bigger SD values.

- Analysing the wide range of the variable alcohol (range with CHD = 147.19 vs without CHD = 145.29 litres), it must be considered that an important proportion of participants do not consume alcohol (min = 0 in both groups). In contrast, others can consume values above 140 litres per year.
- It is also particularly interesting the range of the sbp, since, in both groups, the range is above 100 mmHg. Therefore, the inclusion of participants with very low systolic blood pressure (min around 100 mmHg for both groups) and others with very high values (max around 200 mmHg for both groups) ended up with a wide range for this variable.

```
# BOXPLOTS - INDEPENDENT VARIABLES & CHD
plot_sbp <- ggplot(chd.data, aes(x=CHD, y=sbp, fill=CHD)) +
  geom_boxplot() + labs(title="Variable: Sbp",
                        x = "CHD", y="Systolic blood pressure (mmHg)") +
  theme_bw() + theme(plot.title = element_text(face="bold", colour="#000666", size=12),
                     axis.title = element_text(size=12, face="bold"),
                     legend.title = element_text(size=14, face="bold"),
                     legend.text = element_text(size = 14, face = "bold")) +
  scale_fill_brewer(palette = "Paired")

ggarrange(plot_sbp, plot_tobacco,plot_ldl, plot_adiposity,
          plot_typea, plot_BMI, plot_etoh, plot_age,
          common.legend = TRUE, ncol = 4, nrow = 2)

# BOXVIOLIN PLOTS
plot.ldl <- ggbetweenstats(data = chd.data, x=CHD, y=ldl, results.subtitle=FALSE) +
  labs(x = "CHD", y = "Low density lipoprotein cholesterol (mmol/L)",
       title = "Variable: LDL")
ggarrange(plot.sbp, plot.tobacco,plot.ldl, plot.adiposity,
          plot.typea, plot.obesity, plot.etoh, plot.age,
          ncol = 4, nrow = 2)
```

<u>NOTE</u>: Considering the readability, it is not shown all the R code: only the code for one of the boxplots, one of the violin plots, and the final arrangement that bring all the plots together.

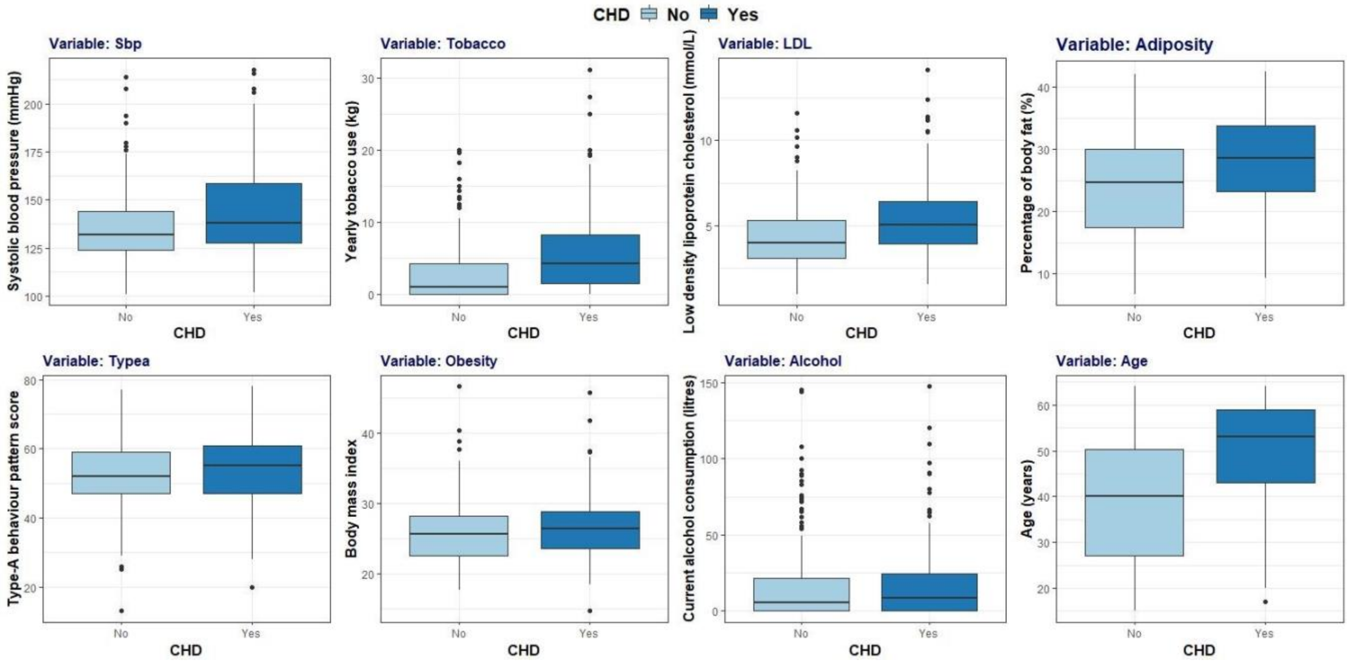

FIGURE 1. BOXPLOTS OF CHD IN FOR EACH INDEPENDENT VARIABLE

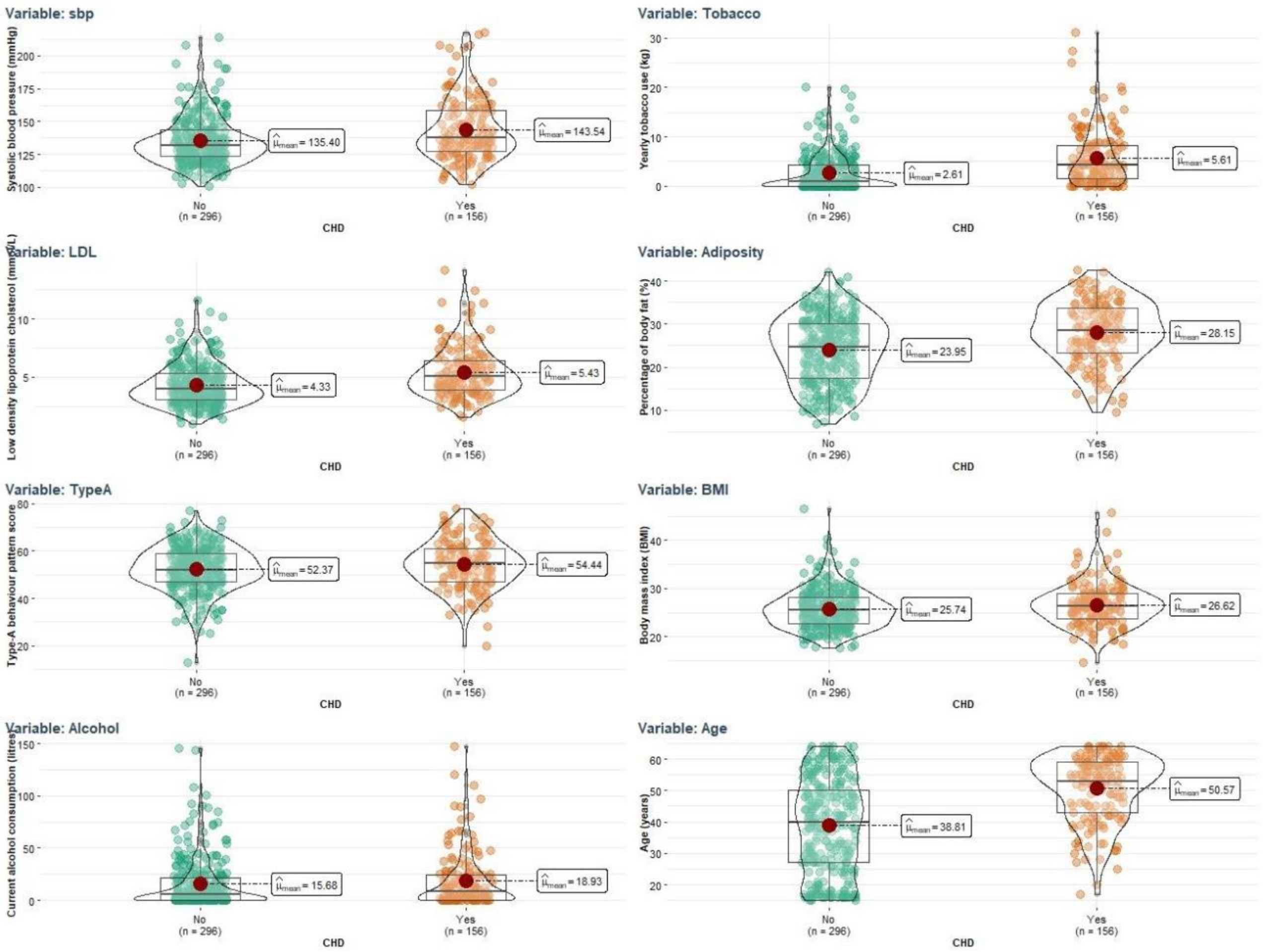

FIGURE 2. DISTRIBUTION OF CHD IN FOR EACH INDEPENDENT VARIABLE

The principal findings about the boxplots and the violin plots, presented in Figure 1 and Figure 2, are:

- The group with CHD diagnosis has higher medians for the nine independent variables (as was mentioned in the descriptive statistic).
- The difference in the medians is particularly clear in the boxplots (Figure 1) for the variables: tobacco, ldl, adiposity and age, and to a lesser extent for sbp, typea, obesity and alcohol.
- The boxplots (Fig.1) show big dispersion of the data, specifically in the variables age and adiposity for both groups and in sbp for the group with a CHD diagnosis.
- Figure 2 shows that the variables tobacco and alcohol are skewed because many participants report not having these habits versus few who do manifest alcohol and tobacco consumption. Also, the variable age, is skewed for the group with CHD diagnosis, where three quartiles of the data correspond to participants older than 40.
- As logistic regression modelling will be performed, it is decided not to transform the data because precisely this is one of the advantages of this kind of modelling versus others, such as linear regression.

3. Examining covariances and correlations of the variables with CHD status.

```
var(chd.data) #  covarian e matrix
sbp                        c       ldl         adiposity  typea       obesity    alcohol     age        CHD
                  tobacco
sbp       415.618326 20.5230581 6.3297050  56.8668208 -10.2696515 20.2900965 59.0741689  119.727582 1.8445735
tobacco   20.523058  21.3495733 1.4506482  10.3833499 -0.7685344  2.3759762  22.6450422  30.502065  0.6785193
ldl       6.329705   1.4506482  3.9566735  7.2478591  0.7213313   2.8931606  -3.0670449  9.607844   0.2499486
adiposity 56.866821  10.3833499 7.2478591  61.3948401 -3.6633205  23.8830666 18.5725123  71.987951  0.9523511
typea     -10.269652 -0.7685344 0.7213313  -3.6633205 96.9001580  3.0967237  10.7034093  -15.208803 0.4683594
obesity   20.290096  2.3759762  2.8931606  23.8830666 3.0967237   17.9937572 4.5649389   18.491958  0.1985168
alcohol   59.074169  22.6450422 -3.0670449 18.5725123 10.7034093  4.5649389  593.3802545 37.623704  0.7363842
age       119.727582 30.5020646 9.6078442  71.9879514 -15.2088034 18.4919584 37.6237036 213.373168 2.6637757
CHD       1.844574   0.6785193  0.2499486  0.9523511  0.4683594   0.1985168  0.7363842   2.663776   0.2265173

eda_chd.data <- glm(CHD=="Yes" ~., data=chd.data, family=binomial) anova(eda_chd.data,
   test = "Chi")
## Analysis of Deviance Table ##
## Model: binomial, link: logit ##
## Response: CHD == "Yes" ##
## Terms added sequentially (first to last) ##
##




##           Df  Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  451       582.52
## sbp       1   16.036   450       566.48     6.215e-05 ***
## tobacco   1   34.958   449       531.53     3.369e-09 ***
## ldl       1   19.980   448       511.55     7.827e-06 ***
## adiposity 1   2.195    447       509.35     0.138461
## typea     1   6.336    446       503.02     0.011833  *
## obesity   1   7.943    445       495.07     0.004828  **
## alcohol   1   0.009    444       495.06     0.923243
## age       1   21.362   443       473.70     3.803e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The covariance matrix shows high values of variance for sbp (415.62), alcohol (593.38) and age (213.37).
- Thinking only about the nature of the relationship and not its magnitude, most variables show a positive correlation. The exceptions are typea, which is negatively correlated with sbp, tobacco, adiposity and age; and alcohol, which negatively correlates with ldl.
- In addition to the covariance matrix, as an exploratory analysis to determine the correlation between the variables CHD and the independent variables, a logistic model with all the data is conducted, and the Anova result of the chi-square test is printed out.
- The variable that correlates the most with CHD is ldl (p-value = 7.827e-06), followed by sbp (p-value = 6.215e-05), age (p-value = 3.803e-06), tobacco (p-value = 3.369e-09), obesity (0.004828) and typea (p-value = 0.011833). The variables adiposity and alcohol have values higher than α= 0.1, the significance level defined for this exploratory analysis level.
- The findings of the exploratory analysis suggest that the variables ldl, sbp, age and tobacco, and to a lesser extent obesity and typea, would be relevant for the fitted model. Adiposity and alcohol could not have significance in the final model.

4. Data splitting and logistic regression modelling of CHD status from the potential risk factors.

```
# SPLITTING THE DATA
set.seed(100)
split_chd.data <- rbinom(nrow(chd.data), 1, prob=0.3333)

chd.training <- subset(chd.data, split_chd.data==0)
chd.test <- subset(chd.data, split_chd.data==1)

dim(chd.training)
## [1] 303   9
dim(chd.test)
## [1] 149   9

glm_chd.training <- glm(CHD=="Yes"~sbp+tobacco+ldl+adiposity+typea+obesity+alcohol+age,
                        data=chd.training,family=binomial)

# Fitting the model - baseline
drop1(glm_chd.training, test="Chi")
## ## adiposity 1  302.69   318.69 0.1958  0.6581069
## ## typea     1  311.94   327.94 9.4447  0.0021176 **
## ## obesity   1  304.81   320.81 2.3156  0.1280783
## ## alcohol   1  303.14   319.14 0.6378  0.4245255      + obesity +
## ## age       1  320.83   336.83 18.3279 1.86e-05  ***
## ## ---
## <none>          302.50 320.50
##
## #fitting the model -sbp
## glm_chd.training_7v <- glm(CHD=="Yes"~tobacco+ldl+adiposity+typea+obesity+alcohol+age,
   data=chd.training,family=binomial) drop1(glm_chd.training_7v, test="Chi")
   ## Single term deletions ##
   ## Model:
   ## CHD == "Yes" ~ tobacco + ldl + adiposity + typea + obesity +
   ## alcohol + age

##
   ##              Df Deviance AIC     LRT     Pr(>Chi)
   ##    <none>       302.51   318.51
   ##    tobacco   1  310.61   324.61 8.0952  0.0044384 **
   ##    ldl       1  316.18   330.18 13.6703 0.0002179 ***
   ##    adiposity 1  302.70   316.70 0.1901  0.6628746
   ##    typea     1  311.95   325.95 9.4334  0.0021307 **
   ##    obesity   1  304.84   318.84 2.3229  0.1274790
   ##    alcohol   1  303.14   317.14 0.6253  0.4290664
   ##    age       1  321.99   335.99 19.4741 1.02e-05  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fitting the model - spb & - adiposity & - alcohol & - obesity
glm_chd.training_4v <- glm(CHD=="Yes"~tobacco+ldl+typea+age,
                           data=chd.training,family=binomial)
drop1(glm_chd.training_4v, test="Chi")
## Single term deletions
##
## Model:
## CHD == "Yes" ~ tobacco + ldl + typea + age
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       306.60 316.60
## tobacco  1   315.97 323.97  9.3714 0.0022039 **
## ldl      1   319.05 327.05 12.4447 0.0004192 ***
## typea    1   315.35 323.35  8.7536 0.0030899 **
## age      1   333.11 341.11 26.5098 2.622e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NOTE: Regarding the readability, some of the backward selection steps were not included. The entire code can be reviewed in the script.

- The data was split randomly into a training (chd.training) and a test set of data (chd.test). This was done considering that two third units would be allocated to the training set and the rest to the test set. As a result, a training data set was obtained with 303 cases (67%, i.e., almost two-thirds of de original data) and a test dataset with 149 cases (the 33% remaining).
- A regression modelling of CHD, the binary response variable, was conducted in the training data, considering in the first step the eight independent variables. It must be highlighted that this statistical method is the most commonly used for binary outcomes, as is the case of the study analysed.
- The glm command specifying the family as binomial was used for the regression modelling.
- Using the drop1 command, a backwards selection was performed to select the suitable variables and simplify the model. In each step, the most non-significant variable was eliminated for the analysis, i.e., the variable with the highest p-value. The variables removed, in order, were sbp (p-value=0.91), adiposity (p-value=0.66), alcohol (p-value=0.41) and obesity (p-value=0.08). The backward selection was stopped when all the variables were below the significance level of 5%. At this point, there were four variables: tobacco, ldl, typea and age. The process is summarised in Table 2:

TABLE 2. FITTING THE MODEL: BACKWARD SELECTION seed(100) prob.=0.333

Step	MODEL		AIC	RESULT drop1 Most non-significant	
	Variables considered	Number		Variable	p-value
1	sbp + tobacco + ldl + adiposity + typea + obesity + alcohol + age	8	320.50	sbp	0.9073121
2	tobacco + ldl + adiposity + typea + obesity + alcohol + age	7	318.51	adiposity	0.6628746
3	tobacco + ldl + typea + obesity + alcohol + age	6	316.70	alcohol	0.412687
4	tobacco + ldl + typea + obesity + age	5	315.37	obesity	0.078130
5	tobacco + ldl + typea + age	4	316.60	-	-

- The Akaike information criterion (AIC) measured model goodness-of-fit (Table 2). The baseline model, with the eight independent variables, has an AIC of 320.5. The lowest AIC was obtained with the model comprising five variables (315.37). However, this model included the variable obesity with a p-value still above the significance level considered in the analysis (model 5 variable: obesity p-value = 0.078 > α = 0.05).
- It was decided to eliminate obesity and test the model with four variables: tobacco + ldl + typea + age. In this model, all four remaining variables are significant (p-values > α = 0.05). The AIC is also lowest than in the baseline model (316.60).

```
# Comparing models- ANOVA
anova(glm_chd.training, glm_chd.training_4v, test="Chi")
## Analysis of Deviance Table
##
## Model 1: CHD == "Yes" ~ sbp + tobacco + ldl + adiposity + typea + obesity +
##     alcohol + age
## Model 2: CHD == "Yes" ~ tobacco + ldl + typea + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       294      302.5
## 2       298      306.6 -4  -4.1017   0.3924

anova(glm_chd.training_5v, glm_chd.training_4v, test="Chi")
## Analysis of Deviance Table
##
## Model 1: CHD == "Yes" ~ tobacco + ldl + typea + obesity + age
## Model 2: CHD == "Yes" ~ tobacco + ldl + typea + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)

## 1       297     303.37
## 2       298     306.60 -1   -3.227  0.07243 .
## ---
## Signif. codes:  0 '***' 0.001 '**'  0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The improvement of the model comprising four variables compared to the baseline model (with eight variables) was examined through ANOVA. Additionally, it was examined if the four variables model is effectively suitable compared to the model with five variables, including obesity and has the lowest AIC. The results are shown in Table 3:

TABLE 3. COMPARING MODELS: ANOVA TEST

MODELS TO COMPARE		**p-value**
Model with eight variables (baseline)	Model with four variables	0.3924
Model with five variables	Model with four variables	0.07243

- There is no significant change in the deviance with is compared (through ANOVA) the baseline model (with eight independent variables) versus the final model with four variables (p-value = 0.3924, with α = 0.05). Consequently, the effect more complex model can be substituted for the)¿ final model chosen, the most parsimonious model, with four instead of eight variables.
- There is no significant change in the deviance when is compared (through ANOVA) the model with the lowest AIC, which contain five variables, versus the model with four variables (p-value = 0.072, with α = 0.05). Hence, the effect of the variable obesity (in the model with five variables) is not significantly different from zero.
- **It is concluded that the model with the four variables is the most parsimony model, and it was chosen as the final model. It contains the variables tobacco, ldl, typea and age.**

5. Assessing the fit of the final model and interpreting the coefficients and confidence intervals for the estimates.

```
# FINAL MODEL
library(gtsummary)
library(survival)

glmft_chd <- glm(CHD=="Yes"~tobacco+ldl+typea+age,
                 data=chd.training,family=binomial)
summary(glmft_chd)
##
## Call:
## glm(formula = CHD == "Yes" ~ tobacco + ldl + typea + age, family = binomial,
##     data = chd.training)
##
##
##  Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8626  -0.8555  -0.3886   0.9031   2.3666
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.52389    1.22552  -6.139 8.29e-10 ***
## tobacco      0.09331    0.03240   2.880 0.003979 **
## ldl          0.23903    0.07077   3.377 0.000732 ***
## typea        0.04473    0.01560   2.868 0.004135 **
## age          0.06402    0.01348   4.750 2.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 393.52  on 302  degrees of freedom
## Residual deviance: 306.60  on 298  degrees of freedom
## AIC: 316.6
##
## Number of Fisher Scoring iterations: 5
glmft_chd %>% tbl_regression() # this table is not included

glmft_chd %>% tbl_regression(exponentiate=TRUE)
```

TABLE 4. Final model: OR estimated and their 95% CI

		Estimate	OR[1]	95% CI[1]	p-value
Intercept:	β_0	-7.52389	e^{β_0} = 0.00054		
Variables:					
tobacco	β_1	0.09331	e^{β_1} = 1.098 ≈ 1.10	1.03, 1.17	0.004
ldl	β_2	0.23903	e^{β_2} = 1.270 ≈ 1.27	1.11, 1.47	<0.001
typea	β_3	0.04473	e^{β_3} = 1.046 ≈ 1.05	1.01, 1.08	0.004
age	β_4	0.06402	e^{β_4} = 1.066 ≈ 1.07	1.04, 1.10	<0.001

[1] OR = Odds Ratio, CI = Confidence Interval

Considering the estimates summarised in Table 4, the final model is:

$$\log\left(\frac{CHD^*}{1-CHD^*}\right) = -7.524 + 0.093 \text{ tobacco} + 0.24 \text{ ldl} + 0.045 \text{ typea} + 0.064 \text{ age}$$

OR

- ODDS RATIO OF CHD* = 0.00054 + 1.098 tobacco + 1.270 ldl + 1.046 typea + 1.066 age

** When CHD take the value CHD = "Yes"*

The coefficients and confidence intervals for the estimates (Table 4) are interpreted in the following points:

- The intercept is close to zero. Therefore, the odds of having a CHD are extremely low when all the risk factors included in this study are equal to zero. Biologically, this situation is impossible due to the nature of the variables analysed.
- All the independent variables included in the model have a significant effect on the probability of CHD taking a value of "Yes" (Ho: β1=0, Ha: β1≠0, all p-value < 0.05, with α=0.05). This can be corroborated by the confidence interval values (95% CI).
- Therefore, at the 5% significance level, it is possible to reject H0 and conclude that the odds for the two groups of CHD for each independent variable in the model are different.
- The yearly tobacco consumption (tobacco), the Low-density lipoprotein cholesterol (ldl), the Type-A behaviour pattern score (typea) and the age (age) are associated with an increased risk of having a CHD (all these estimates are positive) relative to not having a CHD.
- Each kilogram of yearly tobacco (tobacco) increase produces an increment of 1.10 in the odds ratio of having a CHD relative to not having a CHD, i.e., a 10% increase in the odds of having a CHD relative to not having a CHD.
- Each mmol/L of increment in the Low-density lipoprotein cholesterol (ldl) produces an increment of 1.27 in the odds ratio of having a CHD relative to not having a CHD, i.e., a 27% increase in the odd of having a CHD relative to not having a CHD.
- Each point of increase in the Type-A behaviour pattern score produces an increment of 1.05 in the odds ratio of having a CHD relative to not having a CHD, i.e., a 5% increase in the odds of having a CHD.
- Each year a person gets older, an increment of 1.07 in the odds ratio of having a CHD relative to not having a CHD occurred, i.e., a 7% increase in the odds of having a CHD relative to not having a CHD.
- It is highly likely, in 95% of the cases, that the increase in the odd ratio of having a CHD relative to not having a CHD lies within 1.03 and 1.17 (3% to 17%) upon a repeated sampling of the population when the odds ratio of CHD* is associated with the yearly increase of one kilogram of tobacco consumption, between
- 1.11 and 1.47 (11% to 47%) when it is associated with an increase of one mmol/L in the concentration of LDL, between 1.01 and 1.08 (1% - 8%) when it is associated with an increase of one point in the Type-A behaviour pattern score, and between 1.04 and 1.10 (4% and 10%) for every year getting older.
- The 95% CI is particularly wide for the ldl variable (11% to 47%) and slightly less for tobacco (3% to 17%). For typea and age, the 95% CI is narrower (1% to 8% and 4% to 10%, respectively), and therefore they are more precise.
- The biggest increment in the odds ratio of having a CHD relative to not having a CHD is associated with the Low-density lipoprotein cholesterol concentration increase (27%), followed by the yearly tobacco consumption (10%), the age (7%), and in the last term, the Type-A behaviour pattern score (5%).
- Considering the findings, preventive interventions focused on controlling LDL levels and campaigns to reduce tobacco consumption and decrease the Type-A score could decrease the cases of CHD in males in Western Cape, South Africa.

6. Finding the optimal balance point for the sensitivity and specificity of model for the fitted data.

```
# Predictions
dim(chd.training)
## [1] 303   9
dim(chd.test)
## [1] 149   9
pre_chd.training <- predict(glmft_chd, newdata=chd.training) #get predictions for trainin
g data
pre_chd.test <- predict(glmft_chd, newdata=chd.test) #get predictions for test data

library(Epi)
ROC(pre_chd.training, chd.training$CHD,plot="ROC")
tab.chd.tr <- table(pre_chd.training > -0.424, chd.training$CHD)
##
##          No Yes
##   FALSE 148  28
##   TRUE   48  79

# Sensitivity & Specificity & Correct Classification - training dataset
# Sensitivity
sens.chd.tr <- (tab.chd.tr[2,2]/(tab.chd.tr[1,2]+tab.chd.tr[2,2]))*100
sens.chd.tr # 73.83%
## [1] 73.83178
#specificity
spec.chd.tr <- (tab.chd.tr[1,1]/(tab.chd.tr[1,1]+tab.chd.tr[2,1]))*100
spec.chd.tr # 75.51%
## [1] 75.5102
#correct classification rate
correct.chd.tr <- sum(diag(tab.chd.tr)/sum(tab.chd.tr))*100
correct.chd.tr # 74.92%
## [1] 74.91749
```

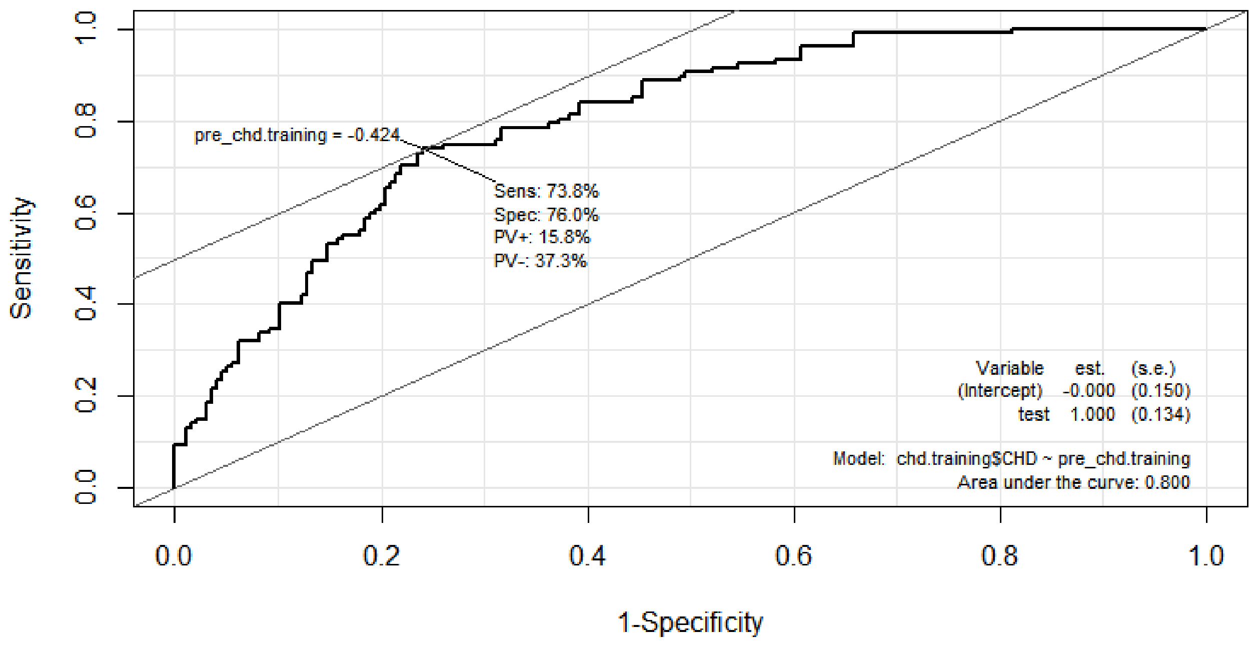

FIGURE 3. ROC PLOT

- The optimal balance point for the sensitivity and specificity of the model for the fitted data was obtained through a ROC plot (Figure 3).
- The area under the curve in the ROC plot is 80%, within the optimal limit (above 80%).
- The best balance between the sensitivity and the specificity is obtained with a cut-off = -0.424. With this cut-off, the sensitivity equals 73.8%, and the specificity equals 76.0%.
- Even when an output of the ROC plot is obtained, the sensitivity and specificity values manually calculated are shown in Table 6, considering the values predicted in Table 5:

TABLE 5. PREDICTED VALUES FOR THE TEST DATASET (CUT-OFF = -0.424)

	No	**Yes**
FALSE	**148**	**28**
TRUE	**48**	**79**

TABLE 6. SENSITIVITY & SPECIFICITY FOR TRAINING SET

DATASET	**Calculation**	**Value**
Sensitivity	$\frac{79}{28+79}$	**0.7383 (73.83%)**
Specificity	$\frac{148}{148+48}$	**0.7551 (75.51%)**

- Considering the value obtained for the sensitivity, the model predicts positive 73.83% of the effectively positive cases, i.e., the model fitted correctly predicts 73.83% of the positive CHD cases in the training dataset.
- Considering the value obtained for the specificity, the model fitted predicts 75.51% of the negative observations that are effectively negative, i.e., the model fitted predicts correctly the 75.51% of the negative CHD cases in the training dataset.
- The sensitivity and specificity values obtained are not particularly high. The ideal values should be above 90%, although above 80% are already acceptable. Both values are below 80%.
- According to the results, a significant proportion of cases are being predicted inaccurately: about a quarter of cases are erroneously predicted to be without CHD when they actually have it, and around the same proportion is predicted to have CHD when they do not.

7. Calculating the overall correct classification rate for both the fitted and test data, and the sensitivity, and specificity for the test data.

```
tab.chd.test <- table(pre_chd.test > -0.424, chd.test$CHD) # from ROC
tab.chd.test
##
##          No Yes
##    FALSE 77  23
##    TRUE  23  26

# Sensitivity & Specificity & Correct Classification - test dataset
# Sensitivity
sens.chd.test <- (tab.chd.test[2,2]/(tab.chd.test[1,2]+tab.chd.test[2,2]))*100
sens.chd.test # 53.06%
## [1] 53.06122
#Specificity
spec.chd.test <- (tab.chd.test[1,1]/(tab.chd.test[1,1]+tab.chd.test[2,1]))*100
spec.chd.test # 77.00%
## [1] 77
#correct classification rate
correct.chd.test <- sum(diag(tab.chd.test)/sum(tab.chd.test))*100
correct.chd.test # 69.13%
## [1] 69.12752
```

Regarding the sensitivity and the specificity of the test data, it can be pointed out that:

- For the test data, it is used the same cut-off as for the training data (-0.424). A two by two table was built with the predicted values for the test dataset (Table 7). The manual calculation of the correct classification rate considering the values predicted for the test dataset is shown in Table 8.

TABLE 7. PREDICTED VALUES FOR THE TEST DATASET (CUT-OFF = -0.424)

	No	Yes
FALSE	**77**	**23**
TRUE	**23**	**26**

TABLE 8. SENSITIVITY & SPECIFICITY FOR TEST DATASET

DATASET	Calculation	Value
Sensitivity	$\frac{26}{23 + 26}$	**0.4694 (46.94%)**
Specificity	$\frac{77}{77 + 23}$	**0.77 (77.00%)**

- Considering the value obtained for the sensitivity in the test dataset, the model fitted predicts as positive the 46.94% of the cases that are effectively positive, i.e. the model fitted predicts correctly the 46.94% of the positive CHD cases in the test dataset. This value is particularly low: less than half is correctly diagnosed as positive. Consequently, more than half of the cases would have a CHD and will not be diagnosed, which could lead to a fatal outcome.
- Considering the value obtained for the specificity in the test dataset, the model fitted predicts 77.00% of the negative observations that are effectively negative, i.e. the model fitted predicts correctly the 77.00% of the negative CHD cases in the test dataset. This value for the specificity is markedly better than the sensitivity obtained, even when it is below the ideal range (from 80% to 90% or higher).
- Compared with the training data, the sensitivity in the test data is noticeably lower than in the training data (sensitivity training data = 73.83 vs test data = 46.94%). However, the specificities are comparable (specificity training data = 75.51% vs test data = 77.00%). Then, there has been a reduction in the sensitivity, where most of the observations are leading to a lower correct classification rate in the test dataset.

Concerning the correct classification rate:

- For calculating the correct classification rate, the predicted values for the training and test dataset are considered (Table 5 and 7, respectively). The manual calculation is shown in Table 9:

TABLE 9. CORRECT CLASSIFICATION RATE

DATASET	Calculation	Correct classification rate
TRAINING DATA	$\frac{148+79}{148+28+48+79}$	**0.7492 (74.92%)**
TEST DATA	$\frac{77+26}{77+23+23+26}$	**0.6913 (69.13%)**

- For the training data, the correct classification rate is 74.92 %, and for the test dataset is 69.13%; i.e. the model predicts 74.92% of CHD cases (either positive or negative) correctly in the training data, and this percentage decrease to 69.13% for the test dataset. Both values are not ideal because they are below 80%.
- Coming back to the sensitivity and specificity results, the values, both on the training and test dataset, are below the ideal range (between 80% to 90%), and the model lacks the accuracy sought.
- Probably, some variables have not yet been accounted for in the model, and that inclusion could help improve the accuracy of the predictions and, therefore, the sensitivity and specificity. Also, co-founders could be presented, affecting the final result.

8. Carrying out a suitable principal component analysis of the original set of potential risk factors excluding CHD status.

```
# General exploration
ggcorrmat(chd.data)
```

Based on the correlation matrix (Figure 4), it is possible correlation between the eight variables analysed:

- The variable adiposity is strongly positively correlated with obesity (0.72) and age (0.63). Also, it is moderately positively correlated with ldl (0.47) and lowly positively correlated with sbp (0.36)
- The variable age is moderately positively correlated with tobacco (0.45), sbp (0.40), and lowly positively correlated with ldl (0.33) and obesity (0.30).
- Obesity has a low positive correlation with ldl (0.34).
- The variable typea is not correlated with the rest of the variables analysed.
- The remaining variables have a very low correlation (below 0.3).
- There are some negative values, but they are extremely low (e.g. -0.11 between typea and age).

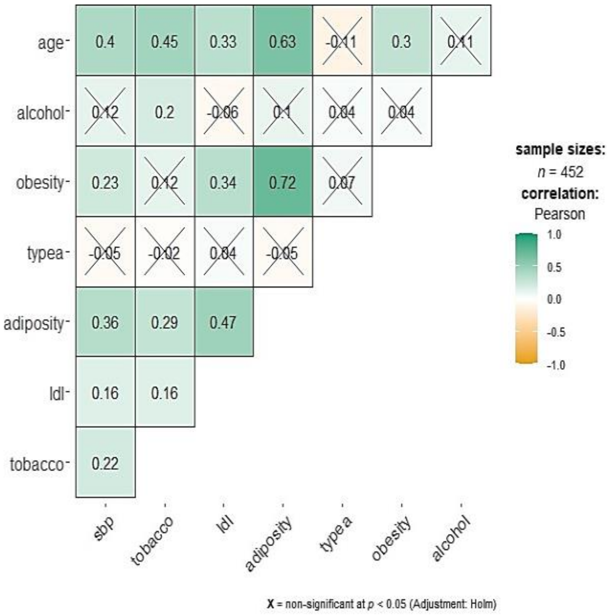

FIGURE 4. CORRELATION MATRIX

```
library(GGally)
ggpairs(chd.data, mapping = aes(color = CHD, alpha = 0.5),
        upper = list(continuous = "points", alpha=0.5))
```

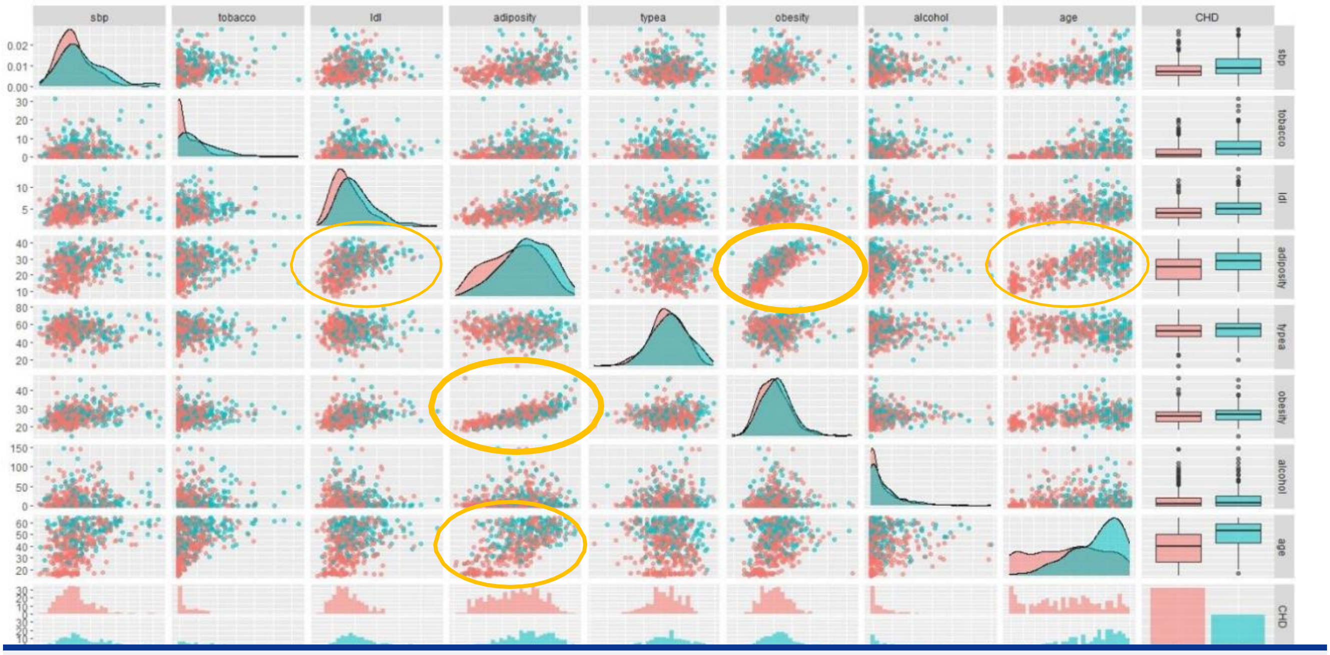

FIGURE 5. LINEAR CORRELATION BETWEEN VARIABLES ON STUDY

```
#Variance matrix
covar_chd.data <- data.frame(unclass(round(var(chd.data[,-9]),2)),
                             check.names = TRUE)
covar_chd.data$Variables <- rownames(covar_chd.data)
covar_chd.data <- covar_chd.data %>% relocate(Variables, .before=sbp)
flextable(covar_chd.data) %>% theme_vanilla() %>%
add_header_lines(values = "TABLE 8. Covariance Matrix") %>%
hline(i=1, part="header", border=fp_border_default(color="darkblue", width=2)) %>%
bold(j=1) %>% color(i=1, j=2, "red") %>% color(i=2, j=3, "red") %>%
  color(i=3, j=4, "red") %>% color(i=4, j=5, "red") %>% color(i=5, j=6, "red") %>%
  color(i=6, j=7, "red") %>% color(i=7, j=8, "red") %>% color(i=8, j=9, "red") %>%
  align(align = "left", part="all")
```

- The variables are measured in different units, for example, sbp is in mmHg, tobacco kg and ldl mmol/L (ldl). Also, the magnitudes of the variance differ (Table 8). For example, the variance amongst sbp is 415.62, whereas the variance among the tobacco is 21.35 and among ldl 3.96. Hence, there are noticeable differences between these variance values, with different orders of magnitude. Therefore, it is required to scale and use the correlation matrix for PCA instead of the covariance matrix.

TABLE 8. COVARIANCE MATRIX

Variables	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age
sbp	415.62	20.52	6.33	56.87	-10.27	20.29	59.07	119.73
tobacco	20.52	21.35	1.45	10.38	-0.77	2.38	22.65	30.50
ldl	6.33	1.45	3.96	7.25	0.72	2.89	-3.07	9.61
adiposity	56.87	10.38	7.25	61.39	-3.66	23.88	18.57	71.99
typea	-10.27	-0.77	0.72	-3.66	96.90	3.10	10.70	-15.21
obesity	20.29	2.38	2.89	23.88	3.10	17.99	4.56	18.49
alcohol	59.07	22.65	-3.07	18.57	10.70	4.56	593.38	37.62
age	119.73	30.50	9.61	71.99	-15.21	18.49	37.62	213.37

```
# Scaling
pca_chd <- prcomp(chd.data[,-9], scale=TRUE)
print(pca_chd,digits=3)
#
## Standard deviations (1, .., p=8):
## [1] 1.68 1.10 1.03 0.91 0.88 0.80 0.68 0.42
##
## Rotation (n x k) = (8 x 8):
##              PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
## sbp       -0.331  0.231 -0.089  0.206  0.813  0.240 -0.261 -0.014
## tobacco   -0.306  0.469  0.056 -0.572 -0.175 -0.299 -0.484 -0.041
## ldl       -0.345 -0.377 -0.021 -0.310 -0.229  0.745 -0.168  0.082
## adiposity -0.528 -0.178 -0.005  0.183 -0.121 -0.176  0.161 -0.765
## typea      0.026 -0.223  0.846 -0.314  0.319 -0.067  0.166 -0.047
## obesity   -0.412 -0.372  0.168  0.394 -0.125 -0.375 -0.313  0.504
## alcohol   -0.107  0.572  0.468  0.446 -0.349  0.341  0.061  0.035
## age       -0.466  0.193 -0.161 -0.218  0.053 -0.078  0.717  0.386
summary(pca_chd)
## Importance of components:
   ## PC1                           PC2    PC3    PC4     PC5     PC6    PC7     PC8
   ## Standard deviation     1.6830 1.1020 1.0311 0.9150 0.88038 0.79846 0.6829 0.41677
   ## Proportion of Variance 0.3541 0.1518 0.1329 0.1047 0.09688 0.07969 0.0583 0.02171
   ## Cumulative Proportion  0.3541 0.5059 0.6388 0.7434 0.84030 0.91999 0.9783 1.00000
```

```
# Scree plot
library(factoextra)
```

```
fviz_eig(pca_chd, addlabels=TRUE)
```

To decide how many components to retain in the PCA, there are considered the following points:

- **The proportion of variance explained by the components** (cumulative proportion): By the summary, it is possible to conclude that the first fourth components explain 74.34% of the variability. If we add PC5, this variability increases to 84.03%, but with more complexity, and the goal is to explain about 75% to 80% of the variability in the small number of components possible. With four components, it is still slightly below 75% of the variability.
- **The Kaiser Criterion:** The three first components have eigenvalues (variances) above 1 (PC1 ≈ 1.68, PC2 ≈ 1.10, PC3 ≈ 1.03). With the fourth component, the eigenvalue is below 1 (0.91) and could not be considered. However, with only three components, going back to the previous point, only 63.89 could be explained.
- **The Scree plot:** In Figure 6, it is shown that there is not a clear elbow in the graph. A big decrease at the beginning can be observed, especially between PC1 and PC2. Then the differences are not as acute as in the beginning. Between PC4 and PC5, the difference is very subtle.
- **It is decided to choose the first fourth components: from PC1 until PC4.**

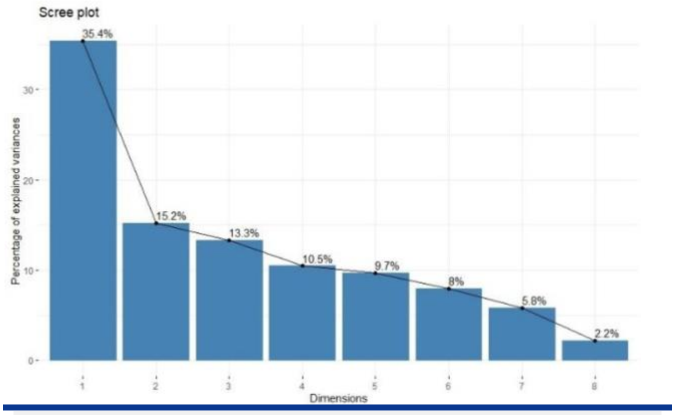

FIGURE 6. SCREE PLOT

TABLE 9. PRINCIPAL COMPONENT ANALYSIS

VARIABLES	PC1	PC2	PC3	PC4
sbp	- 0.331	0.231	~~- 0.089~~	0.206
tobacco	- 0.306	0.469	~~- 0.056~~	- 0.572
ldl	- 0.345	- 0.377	~~- 0.021~~	- 0.310
adiposity	- 0.528	~~- 0.178~~	~~- 0.005~~	~~0.183~~
typea	~~- 0.026~~	- 0.223	0.846	- 0.314
obesity	- 0.412	- 0.372	0.168	0.394
alcohol	~~- 0.107~~	0.572	0.468	0.446
age	- 0.466	~~0.193~~	~~- 0.161~~	- 0.218

Interpreted the component chosen:

- PC1: Two variables have values less than 0.2: typea (0.026) and alcohol (0.107). Hence they make a relatively small contribution and will not be considered. The magnitudes of the rest are between 0.31 and 0.53 and are all negative. Therefore, there is an average component without any contrast. The variable with the highest contribution are adiposity (- 0.528), age (- 0.466), and obesity (- 0.412). Thus, this component primarily measures risk factors associated with BMI (obesity), age and the percentage of fat (adiposity).

- PC2: Adiposity (0.178) and age (0.193) have values below 0.2, and will not be considered. The variables with the highest contribution are alcohol (0.572) and tobacco (0.469). These two variables, plus sbp (0.231), are positive, while ldl (- 0.377), typea (- 0.223) and obesity (- 0.372) are negative. Therefore, there is a contrast between alcohol-tobacco- sbp and ldl-typea-obesity. Then, in one of the extreme values of the second component are located participants with high systolic blood pressure (variable sbp) and high consumption of tobacco and alcohol. On the other side-end are located participants with low blood concentration of LDL (variable ldl), low scores of Type-A behaviour pattern (variable typea), and low values of BMI (variable obesity).
- PC3: Only two variables have values highest than 0.2: typea (0.8456) and alcohol (0.4682). Both are positive. Therefore, it is an average component, as PC1, without any contrast. Thus, this component primarily measures risk factors associated with the Type A behaviour pattern (the variable with the highest contribution to PC3) and the consumption of alcohol.
- PC4: Only adiposity has a value less than 0.2 and will not be considered. The variables sbp (0.206), obesity (0.394) and alcohol (0.446) are positive, differing from tobacco (- 0.572), ldl (- 0.310), typea (- 0.314) and age (- 0.218), which are negative. Then, there is a contrast between them and participants with high values of systolic blood pressure, BMI and alcohol consumption in one extreme, and the youngest individuals with low consumption of tobacco and low values of LDL and Type-A behaviour pattern score in the other extreme. The variable with the highest contribution is tobacco (-0.572), followed by alcohol (0.446).

```
# Biplot
fviz_pca_biplot(pca_chd,        label="var",        habillage        =        chd.data$CHD)
```

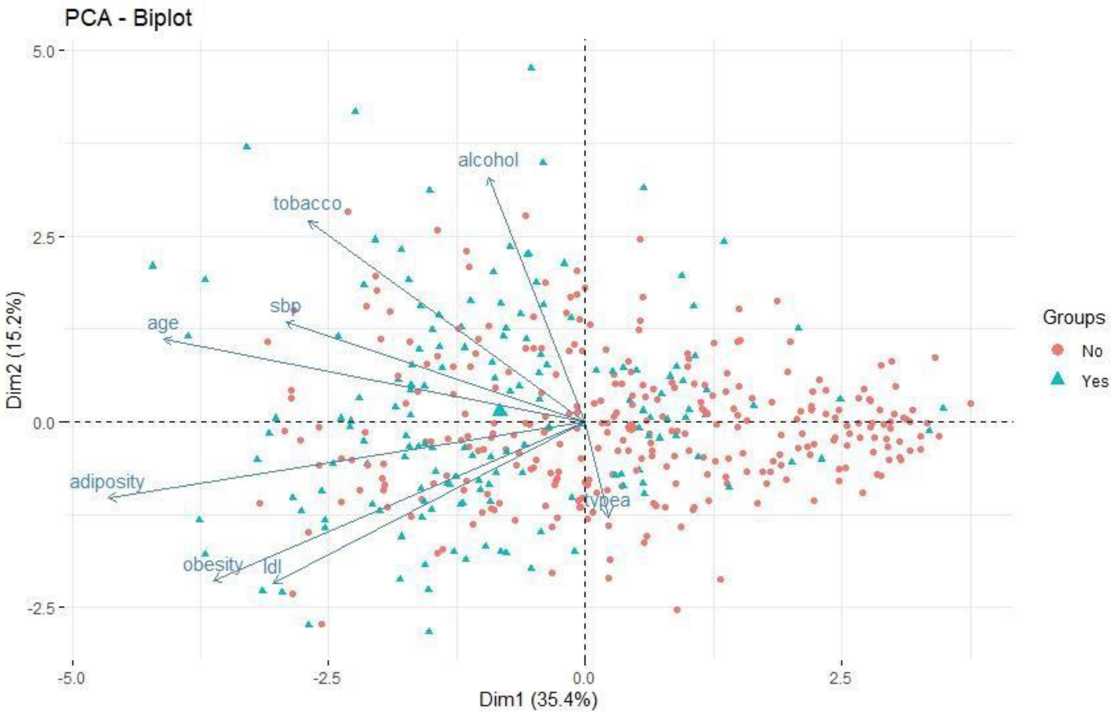

FIGURE 7. BIPLOT

In Figure 7, it is shown the first two PCs. It was decided to use figures to differentiate between individuals with and without CHD, not identification numbers, because it makes the interpretation difficult. The principal observations related to the biplot are:

- Age and adiposity are strongly positively correlated to PC1 (arrows with the smallest non-obtuse angles to the x-axis). The variables typea and alcohol have almost a right angle with PC1, hence do not contribute much to PC1. These findings are consistent with the interpretation of the components chosen, where typea and alcohol were not analysed because they had values below 0.2, and the variables with the highest contribution were precisely adiposity (- 0.528) and age (- 0.466). The individuals with a low percentage of body fat (adiposity) and the youngest ones will be located on the right side of the graph, where are the biggest values for PC1.
- All arrows, except typea, point to the left. However, what is mentioned in the previous point about typea must be considered: it does not contribute much to PC1. Therefore, participants with high values of alcohol, tobacco, sbp, adiposity, obesity, ldl and the older ones will be located on the left side of the graph.
- The variable alcohol is strongly correlated to PC2 (arrows with small angles, close to 45°, with the y-axis), followed by tobacco. The variables ldl, obesity and typea form obtuse angles with the PC2 axis; therefore, they negatively correlate with the second component. Conversely, age and adiposity are close to forming a right angle with the PC2 axis and, consequently, do not have a crucial contribution to the second component. All these results also are congruent with the values shown in Table 9, where the highest contributions were alcohol (0.572) and tobacco (0.469); ldl (- 0.377), obesity (- 0.372) and typea (- 0.223) had negative values and adiposity and age had values bellow 0.2. The individuals with high consumption of alcohol and tobacco will be located at the top side of the graph, where are the biggest values for PC2.
- In Figure 7, the red point corresponded to the participants without a CHD and the blue one with a CHD. As pointed out before, the individuals with low values of the risk factors analysed (without considered typea) will be located on the right side of the graph, where the biggest density of red points are, i.e., those without a CHD. On the other hand, the individuals with high values of the risk factors will be located on the left hand of the graph, and just where the high density of blue points can be observed, i.e., those with CHD.
- Also, it can be observed that at the top of the graph, there are several blue points, corresponding with individuals with CHD and also with high consumption of alcohol and tobacco.

9. Comparing results of the PCA to those of the logistic regression.

- The variable age included in the final model of the logistic regression is the second one with the highest contribution to PC1 (- 0.466), and tobacco to PC2 (0.469). The variable typea, also included in the model, has an important contribution to the PC3 (0.846). The exception is ldl, which contributes to PC1, PC2 and PC4 at around 0.3, and it is not a protagonist in any of the three components.
- In the PCA, the variable adiposity contributes the most to the PC1 (- 0.528), and obesity is the third (- 0.412). Neither of these two variables is included in the logistic model. This could be for the presence of cofounded between the variables, negatively affecting the result of the logistic model but not the PCA.
- Future analysis could include the variable adiposity and exclude ldl to evaluate if there are improvements in the predictive model's sensitivity, specificity, and correct classification rate.
- Future analysis could link the PCA with a cluster analysis.

References

1. Galimudi RK, Mudigonda S, Gantala SR, Hanumanth SR. Impact of Epidemiological and Clinical Risk factors in the Patho-genesis of Coronary Heart Disease. 2023.
2. Hasbani NR, Ligthart S, Brown MR, Heath AS, Bebo A, Ashley KE, et al. American Heart Association's Life's Simple 7: Lifestyle Recommendations, Polygenic Risk, and Lifetime Risk of Coronary Heart Disease. Circulation. 2022;145(11):808-18.
3. Wang Z, Zhu C, Nambi V, Morrison AC, Folsom AR, Ballantyne CM, et al. Metabolomic Pattern Predicts Incident Coronary Heart Disease. Arteriosclerosis, Thrombosis, and Vascular Biology. 2019;39(7):1475-82.
4. Voutilainen A, Brester C, Kolehmainen M, Tuomainen TP. Effects of data preprocessing on results of the epidemiological analysis of coronary heart disease and behaviour-related risk factors. Ann Med. 2021;53(1):890-9.
5. Glovaci D, Fan W, Wong ND. Epidemiology of Diabetes Mellitus and Cardiovascular Disease. Current Cardiology Reports. 2019;21(4):21.
6. De Hert M, Detraux J, Vancampfort D. The intriguing relationship between coronary heart disease and mental disorders. Dialogues in Clinical Neuroscience. 2018;20(1):31-40.
7. Sieri S, Agnoli C, Grioni S, Weiderpass E, Mattiello A, Sluijs I, et al. Glycemic index, glycemic load, and risk of coronary heart disease: a pan-European cohort study. The American Journal of Clinical Nutrition. 2020;112(3):631-43.