#title: "heart\_attack"

#author: "Oumar Kane"

#date: "15/03/2022"

#output: pdf\_document

# Variables selection with heart attack dataset

## We have recuperated a dataset named heart\_attack which is very interesting for a debut

## Let's import # needed libraries fir# r

library(MASS)

library(randomForest)

library(VSURF)

library(corrplot)

library(VGAM)

# # glm function will be use because we have binary categorical target

## We can recuperate # dataset by indicated # path

heart\_data = read.csv("E:/Oumar/Ordinateur Dell/oumar/documents/Cours/IA data forest/master semestre 2/Analyse de données/Projets/Data/Heart\_disease/data/heart.csv")

#Let's attach # dataset to manipulate fastly # columns

attach(heart\_data)

#We have to see # five first lines of # dataset and verify # types

head(heart\_data)

## data is already cleaned. We have some categorical variables but those variables are encoded.

#Let's go deeper to analyzing

## Let's see if # different variables are correlated toge# r with a pairplot

## and a corrplot

pairs(heart\_data)

corrplot(cor(heart\_data))

#We didn't very see well # shapes of # scatter maded beetween variables if we trace # plot

#with # full dataset. But we can do it again with only some variables.

# choose only four first variables

pairs(heart\_data[, 1:4])

#We see that with # categorical variables we did not obtain good distributions. We cannot #interprate # plots with # categorical variables but only with # quantitative variables.

#Let's recuperate only # quantitative variables in a new dataframe.

quanti = heart\_data[, c(1, 4, 5, 8, 10)]

# Let's trace a new pairplot without categorical variables

pairs(quanti)

# We see that some variables like thalach and age or chol and age are correlated

# Let's trace # corrplot to see more clearly # variables interactions

corrplot(cor(quanti))

## Preprocessing

### Let's change # type of # categorical variables to factor

for (i in c(2, 3, 6, 7, 9, 11, 12, 13, 14)){

heart\_data[, i] = as.factor(heart\_data[, i])

}

### Let's see we if obtain a good fit in a logistic regression model

# We use a logistic regression model because # target `output` is a binary categorical variable

summary(glm(output~., data = heart\_data, family = binomial))

## We can begin # selection using # likelihood ratio

# create a function calculating # likelihood ratio

LRT\_dev = function(mod1, mod2){

p\_value = 1 - pchisq(deviance(mod1) - deviance(mod2), df.residual(mod1) - df.residual(mod2))

print(paste("p\_value is : ", p\_value))

print(paste("p\_value < 0.05 : ", as.logical(p\_value < 0.05)))

}

## Forward selection with likelihood ratio

### age

model0 = glm(output~1, data = heart\_data, family = binomial)

model1 = glm(output~age, data = heart\_data, family = binomial)

LRT\_dev(model0, model1)

# age variable is relevant, we can maintain it

### sex

model2 = glm(output~age+sex, data = heart\_data, family = binomial)

LRT\_dev(model1, model2)

# sex variable is relevant, we can maintain it

### cp

model3 = glm(output~age+sex+cp, data = heart\_data, family = binomial)

LRT\_dev(model2, model3)

# cp variable is relevant, we can maintain it

### trtbps

model4 = glm(output~age+sex+cp+trtbps, data = heart\_data, family = binomial)

LRT\_dev(model3, model4)

# trtbps variable is relevant, we can maintain it

### chol

model5 = glm(output~age+sex+cp+trtbps+chol, data = heart\_data, family = binomial)

LRT\_dev(model4, model5)

# chol variable is not relevant, we can exclude it

### fbs

model6 = glm(output~age+sex+cp+trtbps+fbs, data = heart\_data, family = binomial)

LRT\_dev(model4, model6)

# fbs variable is not relevant, we can exclude it

### restecg

model7 = glm(output~age+sex+cp+trtbps+restecg, data = heart\_data, family = binomial)

LRT\_dev(model4, model7)

# restecg variable is not relevant, we can exclude it

### thalachh

model8 = glm(output~age+sex+cp+trtbps+thalachh, data = heart\_data, family = binomial)

LRT\_dev(model4, model8)

# thalachh variable is relevant, we can maintain it

### exng

model9 = glm(output~age+sex+cp+trtbps+thalachh+exng, data = heart\_data, family = binomial)

LRT\_dev(model8, model9)

# exng variable is relevant, we can maintain it

### oldpeak

model10 = glm(output~age+sex+cp+trtbps+thalachh+exng+oldpeak, data = heart\_data, family = binomial)

LRT\_dev(model9, model10)

# oldpeak variable is relevant, we can maintain it

### slp

model11 = glm(output~age+sex+cp+trtbps+thalachh+exng+oldpeak+slp, data = heart\_data, family = binomial)

LRT\_dev(model10, model11)

# slp variable is relevant, we can maintain it

### caa

model12 = glm(output~age+sex+cp+trtbps+thalachh+exng+oldpeak+slp+caa, data = heart\_data, family = binomial)

LRT\_dev(model11, model12)

# caa variable is relevant, we can maintain it

### thall

model13 = glm(output~age+sex+cp+trtbps+thalachh+exng+oldpeak+caa+thall, data = heart\_data, family = binomial)

LRT\_dev(model12, model13)

# thall variable is not relevant, we can exclude it

### Conclusion :

# # final model contains only # variables : age, sex, cp, trtbps, thalachh, exng, oldpeak, slp and caa. # not relevant variables are restecg, fbs, restecg and thall.

final\_model\_forw = model12

# summary of # model

summary(final\_model\_forw)

## Backward selection with likelihood ratio

### age

model0 = glm(output~., data = heart\_data, family = binomial)

model1 = glm(output~.-age, data = heart\_data, family = binomial)

LRT\_dev(model1, model0)

# age variable is not relevant, we can exclude it

### sex

model2 = glm(output~.-age-sex, data = heart\_data, family = binomial)

LRT\_dev(model2, model1)

# sex variable is relevant, we can maintain it

### cp

model3 = glm(output~.-age-cp, data = heart\_data, family = binomial)

LRT\_dev(model3, model1)

# cp variable is relevant, we can maintain it

### trtbps

model4 = glm(output~.-age-trtbps, data = heart\_data, family = binomial)

LRT\_dev(model4, model1)

# trtbps variable is relevant, we can maintain it

### chol

model5 = glm(output~.-age-chol, data = heart\_data, family = binomial)

LRT\_dev(model5, model1)

# chol variable is not relevant, we can exclude it

### fbs

model6 = glm(output~.-age-chol-fbs, data = heart\_data, family = binomial)

LRT\_dev(model6, model5)

# fbs variable is not relevant, we can exclude it

### restecg

model7 = glm(output~.-age-chol-fbs-restecg, data = heart\_data, family = binomial)

LRT\_dev(model7, model6)

# restecg variable is not relevant, we can exclude it

### thalachh

model8 = glm(output~.-age-chol-fbs-restecg-thalachh, data = heart\_data, family = binomial)

LRT\_dev(model8, model7)

# thalachh variable is not relevant, we can exclude it

### exng

model9 = glm(output~.-age-chol-fbs-restecg-thalachh-exng, data = heart\_data, family = binomial)

LRT\_dev(model9, model8)

# exng variable is not relevant, we can exclude it

### oldpeak

model10 = glm(output~.-age-chol-fbs-restecg-thalachh-exng-oldpeak, data = heart\_data, family = binomial)

LRT\_dev(model10, model9)

# oldpeak variable is relevant, we can maintain it

### slp

model11 = glm(output~.-age-chol-fbs-restecg-thalachh-exng-slp, data = heart\_data, family = binomial)

LRT\_dev(model11, model9)

# slp variable is relevant, we can maintain it

### caa

model12 = glm(output~.-age-chol-fbs-restecg-thalachh-exng-caa, data = heart\_data, family = binomial)

LRT\_dev(model12, model9)

# caa variable is relevant, we can maintain it

### thall

model13 = glm(output~.-age-chol-fbs-restecg-thalachh-exng-thall, data = heart\_data, family = binomial)

LRT\_dev(model13, model9)

# thall variable is relevant, we can maintain it

### conclusion :

## final model doesn't contain # following variables : age, chol, fbs, restecg, thalachh and exng.

final\_model\_back = model9

summary(final\_model\_back)