Mid Term Exam

Souvik Samanta

12/11/2020

# Clearing environment  
rm(list=ls())  
  
# Setting Working Directory  
setwd("C:/Users/souvi/Desktop/DMAM")  
  
# Importing Dataset  
data <- read.csv("Question1 end term.csv")  
  
#Structure of Data  
str(data)

## 'data.frame': 299 obs. of 13 variables:  
## $ age : num 75 55 65 50 65 90 75 60 65 80 ...  
## $ anaemia : int 0 0 0 1 1 1 1 1 0 1 ...  
## $ creatinine\_phosphokinase: int 582 7861 146 111 160 47 246 315 157 123 ...  
## $ diabetes : int 0 0 0 0 1 0 0 1 0 0 ...  
## $ ejection\_fraction : int 20 38 20 20 20 40 15 60 65 35 ...  
## $ high\_blood\_pressure : int 1 0 0 0 0 1 0 0 0 1 ...  
## $ platelets : num 265000 263358 162000 210000 327000 ...  
## $ serum\_creatinine : num 1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...  
## $ serum\_sodium : int 130 136 129 137 116 132 137 131 138 133 ...  
## $ sex : int 1 1 1 1 0 1 1 1 0 1 ...  
## $ smoking : int 0 0 1 0 0 1 0 1 0 1 ...  
## $ time : int 4 6 7 7 8 8 10 10 10 10 ...  
## $ DEATH\_EVENT : int 1 1 1 1 1 1 1 1 1 1 ...

#Removing TIME column  
data$time <- NULL  
#Converting factor variables  
data$anaemia <- factor(data$anaemia)  
data$diabetes <- factor(data$diabetes)  
data$high\_blood\_pressure <- factor(data$high\_blood\_pressure)  
data$sex <- factor(data$sex)  
data$smoking <- factor(data$smoking)  
data$DEATH\_EVENT <- factor(data$DEATH\_EVENT)  
str(data)

## 'data.frame': 299 obs. of 12 variables:  
## $ age : num 75 55 65 50 65 90 75 60 65 80 ...  
## $ anaemia : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...  
## $ creatinine\_phosphokinase: int 582 7861 146 111 160 47 246 315 157 123 ...  
## $ diabetes : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...  
## $ ejection\_fraction : int 20 38 20 20 20 40 15 60 65 35 ...  
## $ high\_blood\_pressure : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...  
## $ platelets : num 265000 263358 162000 210000 327000 ...  
## $ serum\_creatinine : num 1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...  
## $ serum\_sodium : int 130 136 129 137 116 132 137 131 138 133 ...  
## $ sex : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...  
## $ smoking : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...  
## $ DEATH\_EVENT : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

head(data)

## age anaemia creatinine\_phosphokinase diabetes ejection\_fraction  
## 1 75 0 582 0 20  
## 2 55 0 7861 0 38  
## 3 65 0 146 0 20  
## 4 50 1 111 0 20  
## 5 65 1 160 1 20  
## 6 90 1 47 0 40  
## high\_blood\_pressure platelets serum\_creatinine serum\_sodium sex smoking  
## 1 1 265000 1.9 130 1 0  
## 2 0 263358 1.1 136 1 0  
## 3 0 162000 1.3 129 1 1  
## 4 0 210000 1.9 137 1 0  
## 5 0 327000 2.7 116 0 0  
## 6 1 204000 2.1 132 1 1  
## DEATH\_EVENT  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1

summary(data)

## age anaemia creatinine\_phosphokinase diabetes ejection\_fraction  
## Min. :40.00 0:170 Min. : 23.0 0:174 Min. :14.00   
## 1st Qu.:51.00 1:129 1st Qu.: 116.5 1:125 1st Qu.:30.00   
## Median :60.00 Median : 250.0 Median :38.00   
## Mean :60.83 Mean : 581.8 Mean :38.08   
## 3rd Qu.:70.00 3rd Qu.: 582.0 3rd Qu.:45.00   
## Max. :95.00 Max. :7861.0 Max. :80.00   
## high\_blood\_pressure platelets serum\_creatinine serum\_sodium sex   
## 0:194 Min. : 25100 Min. :0.500 Min. :113.0 0:105   
## 1:105 1st Qu.:212500 1st Qu.:0.900 1st Qu.:134.0 1:194   
## Median :262000 Median :1.100 Median :137.0   
## Mean :263358 Mean :1.394 Mean :136.6   
## 3rd Qu.:303500 3rd Qu.:1.400 3rd Qu.:140.0   
## Max. :850000 Max. :9.400 Max. :148.0   
## smoking DEATH\_EVENT  
## 0:203 0:203   
## 1: 96 1: 96   
##   
##   
##   
##

# Creating Training and validation dataset  
set.seed(1)  
train.index <- sample(c(1:dim(data)[1]), dim(data)[1]\*0.8)  
# Dataset is not that huge and for getting maximum training data, taking train-test split of 80:20  
valid.index <- setdiff(c(1:dim(data)[1]), train.index)  
train.df <- data[train.index, ]  
valid.df <- data[valid.index, ]  
  
dim(train.df)

## [1] 239 12

t<- table(train.df$DEATH\_EVENT); t

##   
## 0 1   
## 164 75

dim(valid.df)

## [1] 60 12

v<-table(valid.df$DEATH\_EVENT); v

##   
## 0 1   
## 39 21

############################################  
# ADABOOST  
############################################  
library(ada)

## Warning: package 'ada' was built under R version 4.0.3

## Loading required package: rpart

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

cvcontrol <- trainControl(method="repeatedcv", number = 10,repeats = 10)  
train.ada <- train(DEATH\_EVENT ~ ., data=train.df, method="ada",trControl=cvcontrol)  
  
#computing confusion matrix for training data  
pred.train2 <- predict(train.ada,train.df,type = "raw")  
confusionMatrix(pred.train2, as.factor(train.df$DEATH\_EVENT),positive="1")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 152 24  
## 1 12 51  
##   
## Accuracy : 0.8494   
## 95% CI : (0.7976, 0.8922)  
## No Information Rate : 0.6862   
## P-Value [Acc > NIR] : 5.378e-09   
##   
## Kappa : 0.6344   
##   
## Mcnemar's Test P-Value : 0.06675   
##   
## Sensitivity : 0.6800   
## Specificity : 0.9268   
## Pos Pred Value : 0.8095   
## Neg Pred Value : 0.8636   
## Prevalence : 0.3138   
## Detection Rate : 0.2134   
## Detection Prevalence : 0.2636   
## Balanced Accuracy : 0.8034   
##   
## 'Positive' Class : 1   
##

# We got an accuracy of 84.9%  
# we got sensitivity of 68%  
  
#computing confusion matrix for validation data  
pred.valid2 <- predict(train.ada,valid.df,type = "raw")  
confusionMatrix(pred.valid2, as.factor(valid.df$DEATH\_EVENT),positive="1")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 36 12  
## 1 3 9  
##   
## Accuracy : 0.75   
## 95% CI : (0.6214, 0.8528)  
## No Information Rate : 0.65   
## P-Value [Acc > NIR] : 0.06561   
##   
## Kappa : 0.3902   
##   
## Mcnemar's Test P-Value : 0.03887   
##   
## Sensitivity : 0.4286   
## Specificity : 0.9231   
## Pos Pred Value : 0.7500   
## Neg Pred Value : 0.7500   
## Prevalence : 0.3500   
## Detection Rate : 0.1500   
## Detection Prevalence : 0.2000   
## Balanced Accuracy : 0.6758   
##   
## 'Positive' Class : 1   
##

# We got an accuracy of 75%  
# we got sensitivity of 42.9%