```
#Exercise 7 in Classification and Discriminant analysis
#Carim El-Cheschin

require(mgcv)
require(ElemStatLearn)
require(mvpart)

# data loading and preparation
data(SAheart)
dat <- SAheart[,-5]     # removing the "famhist" variable
n<-nrow(dat)
set.seed(100)
i_train<-sample(n,300)
train_dat<-dat[i_train,]
test_dat<-dat[-i_train,]

################################################################
# 1a)

m_gam <- gam(chd~ s(sbp) + s(tobacco) + s(ldl) + s(adiposity) + s(typea) + s(obesity) +
s(alcohol) + s(age),data=train_dat,family=binomial)
summary(m_gam)
# typea and age are marked as significant and ldl only on a significance level of 10%
m_res <- predict(m_gam,test_dat,type="response") > 0.5
TAB <- table(test_dat$chd,as.numeric(m_res))
TAB
mkrgam <- 1 - sum(diag(TAB))/sum(TAB)
mkrgam
# missclassification rate of ~32.7 %

# 1b)

plot(m_gam,page=1,shade=TRUE,shade.col="yellow")
# shows the influence of the regressors on the response, it can be seen that at the tails
the confidence intervals are
# wide since there aren't many data points
# the pattern of the influence on the response is visualized

################################################################
# 2a)

filepath <- "http://www.statistik.tuwien.ac.at/public/filz/students/klassdis/bank.zip"
temp <- tempfile()
download.file(filepath,temp)
d <- read.csv2(unz(temp,"bank.csv"))
train_size <- 3000                      # has taken 3000 observations
set.seed(230)
i_train <- sample(dim(d)[1],train_size)
d_train <- d[i_train,]
d_test <- d[-i_train,]

d_baum <- rpart(y ~.,data=d_train, cp=0.005, xval=20)
d_baum
# 2b)

plot(d_baum)
text(d_baum)
# seems to be an overfit since there are to many knots and further investigations are
needed
```

```
# duration seems to be a important seperation variable

# 2c)

b1_pred <- predict(d_baum,d_test,type="class")
b1_tab <- table(d_test$y,b1_pred)
b1_tab
mkr_baum <- 1 - sum(diag(b1_tab)) / sum(b1_tab)
mkr_baum
# missclassification rate of ~11.8 %

# 2d)
printcp(d_baum)
# root node error of 0.10567 which is around 11 like our missclassification rate
# cp        ... the complexity parameter in every step
# nsplit    ... number of splits in the recent tree
# xerror    ... error according to the lost of information by classifiying
# xstd      ... standard error of the xerror


plotcp(d_baum)
# minimas are marked with red bullets and the optimal cp is at the yellow bullet, which
is the first within the area of
# the minima + 1 standard error
# so the optimal is at 0.0283912 according to the plot and printcp

# 2e)
d_baum2 <- prune(d_baum,cp=0.0283912)
plot(d_baum2)
text(d_baum2)
# this tree is much less complex to the initial tree and only duration, poutcome and
martial are the splitting variables

# 2f)
b2_pred <- predict(d_baum2,d_test,type="class")
b2_tab <- table(d_test$y,b2_pred)
b2_tab
mkr_baum2 <- 1 - sum(diag(b2_tab)) / sum(b2_tab)
mkr_baum2
# missclassification rate of ~11.7 % (first ~11.8%), therefore a little improvement of
the missclass. rate could be observed
# and also the intepretation of this tree is much easier than the initial tree
```