STA5104 2022/23 2nd term Assignment 1

The dataset “bank-market.csv” is related to direct marketing campaigns of a foreign bank:

Column	**Name**	**Description**
1	age	continuous: age of the customer
2	marital	categorical: "married","divorced","single" (divorced include widowed)
3	education	categorical: "unknown","secondary","primary","tertiary"
4	balance	continuous: average yearly balance, in euros
5	housing	has housing loan? (binary: "yes","no")
6	loan	has personal loan? (binary: "yes","no")
7	duration	continuous: last contact duration, in seconds
8	campaign	no of contacts performed during this campaign (numeric)
9	pdays	no of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
10	previous	no of contacts performed before this campaign and for this client (numeric)
11	poutcome	outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")
12	deposit	has the client subscribed a term deposit? (binary: "yes","no")

Note that the last column deposit is the binary target variable.

Q1 (Outlier detection)

(a) Read in the data and save them in d. Select all the records in d such that deposit="no" and save them in d0. Similarly select all records in d such that deposit="yes" and save them in d1.

(b) Modify the function mdist() in my notes to delout(d,id,prob=0.99) so that d[,id] will be used to compute the Mahalnobis distance and the output is the cleaned dataset . Some R codes in this function is as follow but you need to fill in the missing part.

```
  # delout will detect and delete outiler using mahalanobis dist
  # input matrix d, output cleaned dataset dc, prob is level for chisq (default=0.99)
  # input vector id = the column index used to compute mahalanobis dist.
  delout<-function(d,id,prob=0.99) {

                    # *** fill in the missing part ***
  cat('size of input=',n,' size of cleaned data=',nc, 'no of outliers deleted=',n-nc,'\n')
  dc                              # output cleaned dataset
}
```

(c) Using **only** columns 1,4,7 from d0 and d1 and the delout() function in part (b) with default level 0.99 to clean d0 and d1 and save them in x0 and x1 respectively.

(d) Combine x0 and x1 into x and save it in a file “bank-market1.csv” with option row.names=F.

Q2 (CTREE, continue from Q1)

(a) Use the last 5 digits of your student id as random seed, select 85% of records from x (part (d) of Q1) as training dataset and save them in d0. The rest of the records are saved in d1 as testing dataset.

(b) Use rpart() to build a classification tree with deposit as target and other 11 variables as the input variables. Save your result in ctree.

(c) Plot and print ctree. Write down and compute the support, confidence and capture of each rule.

(d) Produce the classification table for d0 and d1 and compute their training and testing error rate.

Submit your assignment on or before **February 19, 2023.**

You have to save and submit all the R commands in Q1 and Q2 in asg1.r. Your R commands should be commented as clearly as possible. Save your answers, outputs and plots in asg1.doc or asg1.pdf format.