**Assignment 5 Questions:**

Solve assignment questions below. Be careful to answer exactly what is asked. You can use R, Excel, your calculator or anything else you wish for calculations in answering questions, but include short comments to explain your answers, especially to explain your R commands.

**IMPORTANT:** Name the variables you use in this assignment with an extension including your initials and last three digits of your ID number. To be more specific, if you are importing a file and willing to name the data file as "mydata", then you will name it "mydataMO354", if your initials are MO and last three digits of your ID are 354. Similarly, if you intend to name a variable "pcs", it will be named "pcsMJ907", if your initials are MJ and if last digits of your ID number are 907. You will name all variables in all your answers similarly, so that all of them will have the extension of your initials and last three digits of your ID number. This is crucial and you can lose 50% of your credit if you do not name your variables so. In addition, I want you to add comments to explain what you are doing with your commands in your own words briefly, i.e. short comments like "# I am importing the data file" or "# to compute the principal components". Avoiding to do this will cost 30% of credit.  This assignment will be scanned for plagiarism.

**1.** [30 points, 10 points each] Solve Problems 1.a, 1.b, and 1.c of Chapter 7.

**2.** [20 points, 10 points each] Solve Problems 2.a, and 2.b of Chapter 7.

**3.** [50 points, 10 points each] Solve Problems 1.a, 1.b, 1.c, 1.d, and 1.e of Chapter 8.

**Sketches of Solutions:**

**Problem 7.1.a.**

First create a data frame containing information about two customers and the new prospect with a data frame like:

*custdfMO354 <- data.frame(custid = c(1, 2, 3),*

*Stat = c(1, 0, 0),*

*IT = c(0, 0, 1),*

*Other = c(0, 1, 0),*

*year = c(1, 1.1, 1),*

*course = c(0, 1, NA))*

Here custid #3 is the new prospect.

Exclude custid and response variable (course) from the calculations and get three dummies with a command like:

*custdf3dMO354 <- cust.df[,-c(1, 6)]*

Remove one dummy variable from three dummies (above) to get two dummies with a command like:

*custdf2dMO354 <- cust.df.3dummies[,-3]*

**Problem 7.1.b.**

Use “dist” command with data frames of two and three dummies to compute the Euclidean distances with commands like:

*dist(cust3dMO354)*

**Problem 7.1.c.**

Use the distances calculated above to decide on the classification of the prospect, and answer the question.

**Problem 7.2.** First import the dataset to R and partition the entire dataset into training and validation.

**Problem 7.2.a.**

Input the new customer with a command like:

newcustMO354 <- data.frame(Age = 40,

Experience = 10,

Income = 84,

Family = 2,

CCAvg = 2,

Education = 2,

Mortgage = 0,

Securities.Account = 0,

CD.Account = 0,

Online = 1,

CreditCard = 1)

Identify education as a factor if you use FNN with the “as.factor()”

Normalize the data using “preProcess”, and use “knn” to classify based on the k-nearest neighbors with a command like:

*kNN.predMO354 <- class::knn(train = train.norm.df,*

*test = new.cust.norm,*

*cl = train.df$Personal.Loan, k = 1)*

**Problem 7.2.b.**

Use library “e1071” to select optimal k. Compute accuracy and work out k with the highest accuracy. Then answer the question.

You can define the accuracy first and run a for loop like in the following lines:

*accuracyMO354 <- data.frame(k = seq(1, 15, 1), overallaccuracy = rep(0, 15))*

*for(i in 1:15) {*

*knn.predMO354 <- class::knn(train = train.norm.df,*

*test = valid.norm.df,*

*cl = train.df$Personal.Loan, k = i)*

*accuracyMO354[i, 2] <- confusionMatrix(knn.pred,*

*as.factor(valid.df$Personal.Loan))$overall[1]*

*}*

**Problem 8.1.** Import the data and partition it into training and validation

**Problem 8.1.a.**

Use ftable function to prepare the pivot table of training set with Online as a column variable , CC as a row variable, and Loan as a secondary row variable with a command like:

*ftable(CreditCard, Personal.Loan, Online)*

The table you will return will be like:

#> ftable(CreditCard, Personal.Loan, Online)

# Online 0 1

#CreditCard Personal.Loan

#0 0 793 1140

# 1 79 128

#1 0 312 461

# 1 42 45

**Problem 8.1.b.** Calculate the conditional probability using the table in 8.1.a above.

**Problem 8.1.c.**

Use table() function to prepare the pivot table for Loan (rows) as a function of Online (columns) with a command like:

*table(Personal.Loan, Online)*

Prepare the pivot table for Loan (rows) as a function of CC (columns) similarly with a command like:

*table(Personal.Loan, CreditCard)*

**Problem 8.1.d.** Use the tables prepared above to compute the probabilities

**Problem 8.1.e.** Compute the naive Bayes probability.