**Assignment 3**

**Part A.** The following terms are used in chapter 7. Please do some research and provide a definition and examples for each of them.

* numerical variable
* categorical variable
* nominal variable
* ordinal variable
* binary variable
* dummy variable

**Part B.**

**Please read and understand the case study in chapter 7 before doing this part of assignment.**

In this part, we will take the following steps to predict cancer diagnosis using features related to the shape and size of the cell nuclei.

1. **Download the dataset from Blackboard to your computer. It is a breast cancer dataset that includes 569 examples of cancer biopsies, each with 32 features.**
2. **Load the data from your local drive to R.**

(Hint: use read.cvs() function. Please note the augment should be the path where the file is stored on your computer, and you should use”/” instead of “\”. For example, cancer\_data<-read.csv("C:/INFS6343/cancer\_data.csv”)

1. **Display the first five observations, summary, and structure of the data**

(Hint: head(), summary(), and str() function)

1. **Delete the ID columns since it does not provide useful information for prediction.**

(Hint: see section 7.3.2 on page 272 of the textbook for an example)

1. **Normalize all predictor variables. (We use diagnosis as target/class variable and others as predictor variables)**

(Hint: You may have to create a new function and apply it to predictors. Please see section 7.3.3 on page 273 for an example.)

1. **Display summary again to see if all predictors’ values are in the range of 0 to 1** (Hint: summary() function)
2. **Randomly split the data into training and testing sets, with 80% observations in training set and 20% in testing set.**

(Hints: set.seed() and sample() functions. The textbook has the following codes as comments in section 7.3.4 on page 274. Note that set.seed() can use any integer as its argument (e.g. set.seed(2) or set.seed(345)). Set.seed() function is used to get the same number from sample() function every time.

set.seed(1)

subset\_int <- sample(nrow(boystown\_n), floor(nrow(boystown\_n)\*0.8)) bt\_train<- boystown\_n [subset\_int, ]

bt\_test<-boystown\_n[-subset\_int, ]

)

1. **Create four data frames for (1) training set predictors, (2) testing set predictors, (3) training set class variable, and (4) testing set class variable. The first two only have predictor variables and the last two only have class variable.**

(Hint: see section 7.3.4 on page 274 for an example. The first three data frames will be used as arguments for knn() function in the next step, and the last data frame will be used as the second argument for confusionMatrix() function in the following step.)

**9. Using KNN algorithm with K=22 to predict class variable labels for the training set.**

(Hint: knn()function in “class” package.

Please note that knn() takes four arguments:

train, the predictors for the training set. test, the predictors for the testing set. cl, the true class labels for the train set.

k, the number of neighbors to consider.

The output of knn() is predicted class variable labels for testing data set, which will be the first argument for confusionMatrix() function in the next step. See section 7.3.5 on page 274 for an example.)

**10. Show a confusion matrix to evaluate the performance of the model.**

(Hint: confusionMatrix() function in “caret” package. See section 7.3.9 on page 283-284 for an example)

Note that you can find hints for all question, including function names or textbook page numbers. If you don’t know how to use a function, please try the following (1) use HELP() or ? to get details of the function in R

1. search the function in your textbook.
2. Google it
3. Search it in the online books posted on Blackboard “R Learning Resources” section. (5) ask it on Blackboard General Help forum

**Please submit all your answers in one Word document. Answers to part B should be created by R Markdown.**