**MI6426 Homework Assignment 3 – due no later than 11:59 pm (ET) on Tuesday, 16 June 2020**

The purpose of this problem is to give you some experience in clustering of data sets. There are a large number of clustering methods available, but a very common one is k means clustering (which is considered a machine learning method).

This assignment gives you a data set of 768 patients. Each of the patients is a female member of the Pima Indian tribe in the southwest of the U.S. This tribe is known for its very high prevalence of diabetes.

You will use the patients in the data set to develop k means clustering models (using the R programming language) of whether the patient has diabetes or not for k =2 and k = 4 (note that in general, you need to examine a number of different k values to determine the optimal one, but for this assignment just use the two k values provided). In general, with machine learning methods, you need to divide the data set into training and test sets, but for simplicity you will not be asked to do so for this problem and only need to train the k means clustering model using all the data set as your training set.

The data set, named MI6426\_hw\_3\_diabetes\_data\_set, is in the Course Content area of the Modules. This is a text file with comma delimiters between variables for a given patient (with each patient on a separate row) There are nine entries for each patients: eight for various measured characteristics of the patient and one for the diagnosis (not diabetes/diabetes). A number of the characteristics were not recorded and you need to account for that in your analysis. Below is a description of the text file entries and their possible values:

Column Description

1 Number of times pregnant

2 Plasma glucose concentration after 2 hours in an oral glucose

tolerance test (mg/dl)

3 Diastolic blood pressure (mm Hg)

4 Triceps skin fold thickness (mm)

5 2-Hour serum insulin (mu U/ml)

6 Body mass index (weight in kg/(height in m)^2)

7 Diabetes pedigree function

8 Age (years)

9 Class variable (0 for not diabetes or 1 for diabetes)

The R source code to solve this problem is as follows (note that you will need to change the file reference for the data file since it refers to my directory path and file name):

library(cluster)

install.packages("factoextra")

library("factoextra")

dm\_data <- read.table("C:/Users/gcravens/Documents/NSU/MI6426\_hw\_3\_diabetes\_data\_set.txt", sep = ",")

col\_nmes <- c('preg', 'glc', 'dbp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'class')

colnames(dm\_data) <- col\_names

for(i in c(2, 4)) {

kmeans\_results <- kmeans(dm\_data, i, iter.max = 10)

clusplot(dm\_data, kmeans\_results$cluster, color = TRUE, shade = TRUE, labels = 4, main = paste("Original data, K = ", i),

col.clus = c(1, 3, 5, 7), col.p = c(1, 3), plotchar = TRUE, lines = 0)

}

pca.tot <- prcomp(dm\_data[, 1:8])

fviz\_contrib(pca.tot, choice = "var")

fviz\_pca\_var(pca.tot, col.var = "contrib", gradient.cols = c("yellow", "blue", "red"), repel= TRUE)

dm\_data <- dm\_data[-which(dm\_data$glc == 0), ]

dm\_data <- dm\_data[-which(dm\_data$dbp == 0), ]

dm\_data <- dm\_data[-which(dm\_data$skin == 0), ]

dm\_data <- dm\_data[-which(dm\_data$insulin == 0), ]

dm\_data <- dm\_data[-which(dm\_data$bmi == 0), ]

summary(dm\_data)

for(i in c(2, 4)) {

kmeans\_results <- kmeans(dm\_data, i, iter.max = 10)

clusplot(dm\_data, kmeans\_results$cluster, diss = FALSE,

color = TRUE, shade = TRUE, labels = 4, main = paste("Missing data deleted, K = ", i), col.clus = c(2, 4, 6, 8),

col.p = c(2, 4), plotchar = TRUE, lines = 0, xlab = paste("Component = 1"), ylab = paste("Component = 2"))

}

pca.tot <- prcomp(dm\_data[1:8,])

fviz\_contrib(pca.tot, choice = "var")

fviz\_pca\_var(pca.tot, col.var = "contrib", gradient.cols = c("yellow", "blue", "red"), repel= TRUE)

Note that I have placed one or two bugs in the above code to give you experience debugging code. So, you need to debug the code and provide an exact description of the process you used to debug the code. You also must provide comments for each line of code as to what it is doing and why. Please execute the code and submit the resulting plots.