Instructions:

You can use Word, Excel, Power Point and R to answer the questions in this test. There are a total of six (6) multi-part questions, with point values noted for each question.

Please show your calculations, or the details of your program(s) for each problem. The R programs should be commented so that each step is clearly explained.

Combine all your answers/files into a single zipped file and post the zipped file to CANVAS.

#1 (10 Points)

Is the following function a proper distance function? Why? Explain your answer.

$$d(\mathbf{x},\mathbf{y}) = (\sum_i (x_i - y_i)^2)^2$$

Hint: Measure the distance between (0,0), (0,1) and (1,1)

ANSWER:

A function must follow below three conditions in order to be the proper distance function.

(1) d(x,y) >= 0 if d=0 if and only if x=y
(2) d(x,y) = d(y,x)
(3) d(x,y) <= d(x,z) + d(z,y)

As per hint, if we take points (0,0), (0,1) and (1,1), it violates 3rd condition.

d ((0,0), (1,1)) = (1+1)^2 = 4

d ((0,0), (0,1)) = 1

d ((0,1), (1,1)) = 1

d ((0,0), (1,1)) <= d ((0,0), (0,1)) + d ((0,1), (1,1)) is false.

 # 2 (15 Points)

A large department store sells sport shirts in three sizes (Small, Medium and Large), three patterns (plaid/Pl, print/Pr, and stripe/Sr), and two sleeve lengths (long and short). The accompanying tables give the proportions of shirts sold falling in the various category combinations.

Short Sleeves

	Pl	Pr	Sr
S	0.04	0.02	0.05
M	0.08	0.07	0.12
L	0.03	0.07	0.08

Long Sleeves

	Pl	Pr	Sr
S	0.03	0.02	0.03
M	0.1	0.05	0.07
L	0.04	0.02	0.08

- **What is the probability that the next shirt sold is a medium long-sleeved, print shirt? Why?**
  0.05 because we can find it in the table as joint probability of medium and print shirt of long-sleeved
- **What is the probability that the next shirt sold is a medium print shirt? Why?**
  0.12 because medium print shirt is a joint probability and here the size of sleeve is not specified. So, 0.07+0.05. =0.12
- **What is the probability that the next shirt sold is a short sleeved shirt? A long-sleeved shirt? Why?**
  probability of short sleeved shirt: **0.56**
  probability of long-sleeved shirt: **0.44**

  Because, in order to get probability of short sleeved shirt, we need to add all marginal probability of short sleeved table. Same applies for long-sleeved shirt.

- **Given that the shirt just sold was a short sleeved, plaid, what is the probability that its size was medium?**

  0.08/0.15 => **0.53**

  Because, the probability of shirt just sold was a short sleeved, plaid = 0.15 and the probability of its size was medium = 0.0
- **Given that the shirt just sold was medium, plaid, what is the probability that it was short sleeved? Long-sleeved?**
  0.08/0.18 = **0.44**
  0.1/0.18=**0.56**

  Because, the probability of shirt just sold was a medium, plaid = 0.18 and the probability that it was short sleeved= 0.08.
  So, 0.08/0.18= 0.08
  The same applies for long-sleeved. So, 0.1/0.18 = 0.56

#3 (15 Points)

a) Company XYZ is targeting professionals between the ages of 25 to 45 years old with an asset size of 50 to 100K. To estimate the missing income fields, the company is using k-nearest neighbors.

- **What would be the value of income for customer x in the table below if:**

  K = 2 and method = "unweighted vote" is used

  K =3 and method = "distance weighted vote" is used?

ID	Age	Asset Size	Income
X	30	60	?
1	25	50	100K
2	33	60	90K
3	35	80	150K

ANSWER:

ID	Age	Asset Size	Age (N)	Asset size(N)	Distance from X	Income
X	30	60	0.25	0.2	0	?
1	25	50	0	0	$\sqrt{(0.25-0)^2+(0.2-0)^2}=0.32$	100K
2	33	60	0.4	0.2	$\sqrt{(0.25-.4)^2+(0.2-0.2)^2}=0.15$	90K
3	35	80	0.5	0.6	$\sqrt{(.25-.5)^2+(0.2-.6)^2}=0.47$	150K

- **K = 2 and method = "unweighted vote" is used**

  Here, k=2 and ID=2 is the nearest to the missing income record by considering Age and Asset Size and ID=1 is the second nearest.
  As method = "unweighted vote" is given, we will choose ID=2, the nearest to X.
  So, **Income= 90 K**

- **K =3 and method = "distance weighted vote" is used**
  First, we need to normalize the data.
  Here, method = "distance weighted vote". So, we need to calculate Euclidean distance for all the records, which are as shown in the above table and will be inversely proportional to points:

$$Y_{new}=\frac{\sum_i w_i y_i}{\sum_i w_i}=\frac{(100*9.765)+(90*44.44)+(150*4.52)}{9.765+44.44+4.52}$$

$$=\frac{976.5+3999.6+678}{9.765+44.44+4.52}$$

$$=\frac{5654.1}{58.725}$$

$$=96.28$$

The estimated income for ID=X is **100K**

b) The company has decided to classify income by category instead of estimating a number. Furthermore, it has obtained additional customer information with the exact profile of customer X.

- **What would be the income category for X if K=3 and distance weighted vote is used? Why?**

ID	Income	Asset Size	Income
X	30	60	?
1	25	50	Medium
2	33	60	Low
3	35	80	High
4	30	60	Medium
5	30	60	High
6	30	60	High

ANSWER:

Here, K=3 and method distance weighted vote is given.

From table, we can easily conclude that ID 4,5,6 are the three nearest record to X as they have same Age and Asset Size as X. So, distance of all these three records from X is zero.

So, Votes of all are equal. In this case, High gains 2 votes and Medium gets 1 vote. So, **Income Category of X will be High.**

#4 (10 Points)

- **Use R to create a vector of the following 20 numbers**
- **Find maximum, minimum, median, mean and the standard deviation of the follow 20 numbers.**
- **Replace the missing value with the mean of the numbers**
- **Use R to develop a box plot for these numbers**

45	**48**	**6**	**42**	**49**	**63**	**81**	**56**	**21**	**75**
25	**48**	**56**	**24**	**73**	**82**	**NA**	**80**	**86**	**88**

The following two questions refer to the "IBM_attrition_v1.csv" dataset on canvas which is a subset of the "IBM attrition" dataset. The original dataset is used in IBM ML labs to uncover the factors that lead to employee attrition (attrition=yes). The dataset is a fictional data set created by IBM data scientists.

R Program:

```
#####################################################
# Company    : Stevens
# Project    : CS-513
# Purpose    : Midterm Q-4
# First Name : Gopi
# Last Name  : Miyani
# Id     : 10437266
# Date       : 3/31/2019
# Comments   : Following the instruction mentioned in the Midterm Question-4
#####################################################
```

#Use R to create a vector of the following 20 numbers

```
my_vector<-c (45,  48,  6,  42,  49,  63,  81,  56,  21,  75, 25, 48,
        56,  24,  73,  82,  NA,  80,  86,  88)
```

#Find maximum, minimum, median, mean and the standard deviation of the follow 20 numbers.

```
my_vector<-na.omit(my_vector)

min(my_vector,na.rm = TRUE)

max(my_vector,na.rm = TRUE)

median(my_vector,na.rm = TRUE)

mean(my_vector,na.rm = TRUE)

sd(my_vector,na.rm = TRUE)
```

#Replace the missing value with the mean of the numbers

```
my_vector_mean<-as.integer( ifelse(is.na(my_vector), mean(my_vector, na.rm=TRUE), my_vector))
```

#Use R to develop a box plot for these numbers

```
boxplot(my_vector,na.rm=TRUE)
```

#5 (30 Points): **Classification using K Nearest Neighbor:**

Load IBM_attrition_v1.csv into R

a) **Remove any row with missing values**
b) **Select every third record as the test dataset and the remaining records as the training dataset**
c) **Preform K Nearest Neighbor (K=3 unweighted)**
d) **Score the test dataset**
e) **Measure the error rate.**

R Program:

```
#####################################################
#  Company    : Stevens
#  Project    : CS-513
#  Purpose    : Midterm Q-5
#  First Name : Gopi
#  Last Name  : Miyani
#  Id         : 10437266
#  Date       : 3/31/2019
#  Comments   : Following the instruction mentioned in the Midterm Question-5
#####################################################

rm(list=ls())

#Load IBM_attrition_v1.csv into R
file_name<-file.choose()
IBM_data<-read.csv(file_name)
View(IBM_data)
#a) Remove any row with missing values
IBM_data<-na.omit(IBM_data)
str(IBM_data)
View(IBM_data)
#Minmax Normalization function
mnnorm <-function(x,minx,maxx) {z<-((x-minx)/(maxx-minx))
return(z)
}

# Normalize the data using Minmax Normalization function
IBM_data_normalized <- as.data.frame(
  cbind(Age= mnnorm(IBM_data[,1],min(IBM_data[,1]),max(IBM_data[,1])),
```

```
      JobSatisfaction= mnnorm(IBM_data[,2],min(IBM_data[,2]),max(IBM_data[,2])),
      MonthlyIncome= mnnorm(IBM_data[,3],min(IBM_data[,3]),max(IBM_data[,3])),
      YearsAtCompany= mnnorm(IBM_data[,4],min(IBM_data[,4]),max(IBM_data[,4])),
      Single = IBM_data[,5],
      Gender = IBM_data[,7],
      Attrition = IBM_data[,6]
  )
)

#b) Select every third record as the test dataset and the remaining records as the training dataset
index<-seq(from=1,to=nrow(IBM_data_normalized),by=3)
test<-IBM_data_normalized[index,]
train<-IBM_data_normalized[-index,]

View(test)
View(train)

summary(test)
summary(train)

#c) Preform K Nearest Neighbor ( K=3 unweighted)

library(kknn)
predict <- kknn(formula=factor(Attrition)~. ,  train,test  , kernel="rectangular", k=3)

#d) Score the test dataset
fit <- fitted(predict)
table(kknn=fit,test$Attrition)

#e) Measure the error rate.
knn_error_rate=sum(fit!=test$Attrition)/length(test$Attrition)
print('Error Rate: ')
print(knn_error_rate)
knn_accuracy=1-knn_error_rate
print('Accuracy: ')
print(knn_accuracy)
'
```

```
for(i in 1:20) {
  predict <- kknn(formula=factor(Attrition)~. ,  train,test  , kernel="rectangular", k=i)

  #Extract fitted values from the object " "
  fit <- fitted(predict)
  table(kknn=fit,test$Attrition)
  knn_error_rate=sum(fit!=test$Attrition)/length(test$Attrition)
  print(i)
  print(knn_error_rate)
}
'
```

#6 (20 Points): **Naïve Bayes:**

Load IBM_attrition_v1.csv into R

a) **Remove any row with missing values**
b) **Select every third record as the test dataset and the remaining records as the training dataset**
c) **Preform Naïve Bayes using only the following columns: "Job Satisfaction", "Single" and "Gender"**
d) **Score the test dataset**
e) **Measure the error rate.**

R Program:

```
#####################################################

#  Company    : Stevens

#  Project    : CS-513

#  Purpose    : Midterm Q-6

#  First Name : Gopi

#  Last Name  : Miyani

#  Id     : 10437266
```

```
#  Date      : 3/31/2019
#  Comments   : Following the instruction mentioned in the Midterm Question-6
#####################################################


rm(list=ls())
#Load IBM_attrition_v1.csv into R
file_name<-file.choose()
IBM_data<-read.csv(file_name)
View(IBM_data)


#a)       Remove any row with missing values
IBM_data<-na.omit(IBM_data)
View(IBM_data)
##Define max-min normalization function
mnnorm <-function(x,minx,maxx) {z<-((x-minx)/(maxx-minx))
return(z)
}
IBM_data_normalized <- as.data.frame(
  cbind(JobSatisfaction= mnnorm(IBM_data[,2],min(IBM_data[,2]),max(IBM_data[,2])),
        Single = IBM_data[,5],
        Gender = IBM_data[,7],
        Attrition = IBM_data[,6]
  )
)
```

```
is.factor(IBM_data_normalized$Attrition)

IBM_data_normalized$JobSatisfaction=as.factor(IBM_data_normalized$JobSatisfaction)

IBM_data_normalized$Single=as.factor(IBM_data_normalized$Single)

IBM_data_normalized$Gender=as.factor(IBM_data_normalized$Gender)

IBM_data_normalized$Attrition=as.factor(IBM_data_normalized$Attrition)


#b)	Select every third record as the test dataset and the remaining records as the training dataset

index<-seq(from=1,to=nrow(IBM_data_normalized),by=3)

test<-IBM_data_normalized[index,]

train<-IBM_data_normalized[-index,]

View(test)

View(train)

summary(test)

summary(train)

#c)	Preform Naïve Bayes using only  the following columns: “JobSatisfaction”,  “Single” and “Gender”

library(class)

library(e1071)


nBayes <- naiveBayes(factor(Attrition)~JobSatisfaction+Single+Gender, data =train)

#d)	Score the test dataset

category_all<-predict(nBayes,test )

table(NBayes=category_all,Attrition=test$Attrition)
```

```
#e)	Measure the error rate.

NB_wrong<-sum(category_all!=test$Attrition )

NB_error_rate<-NB_wrong/length(category_all)

print('Error Rate:')

print(NB_error_rate)

NB_accuracy=1-NB_error_rate

print('Accuracy:')

print(NB_accuracy)
```