Instructions:

You can use Word, Excel, Power Point, R and/or Python to answer the questions in this exam. There are a total of nine (9) multi-part questions, with point values noted for each question.

Please show your calculations, or the details of your program(s) for each problem. You must supply the R/Python programs, and the 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. Measure the distance between (0, 0, 0), (0, 1, 0), (0, 1, 1), and (1, 1, 1)

$$d(x, y) = \Sigma\left(\left|x_{i-Y_i}\right|^{2}\right)$$

Answer: No, the given function is not a proper distance function.

There are 3 conditions to satisfy,

a. **Property 1: Distance is always non-negative ()**
b. **Property 2: Commutative, distance from “A to B” is distance from “B to A”**
c. **Property 3: Triangle inequality holds, distance from “A to C” must be less than or equal to distance from “A to B to C”**

For example: Let us take 3 points as follow - A (0,0), B (0,1), C (1,1)

d (a, b) = $(|0-0|^2 + |0-1|^2) = 1$

d (b, c) = $(|0-1|^2 + |1-1|^2) = 1$

d (a, c) = $(|0-1|^2 + |0-1|^2) = 2$

d (b, a) = $(|0-1|^2 + |0-0|^2) = 1$

d (c, b) = $(|1-1|^2 + |0-1|^2) = 1$

d (c, a) = $(|1-0|^2 + |1-0|^2) = 2$

Now,

Property 1: As there is a mode and square of every function all distances are going to be positive.

Property 2: Here distance from A to B in any case is same to the distance from B to A.

Property 3: Here, triangle inequality does not hold, thus it is not a proper function as 1< 1+2 is not statisfied.

For points (x_1, y_1, z_1) and (x_2, y_2, z_2) in 3-dimensional space, the Euclidean distance between them is $\sqrt{}(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2$. According to that, distance between (0, 0, 0), (0, 1, 0), (0, 1, 1), and (1, 1, 1):

1. d (0, 0, 0) to d (0, 1, 0) = $[(0-0)^2 + (0-1)^2 + (0-0)^2]^{1/2} = (1)^{1/2} = 1$
2. d (0, 0, 0) to d (0, 1, 1) = $[(0-0)^2 + (0-1)^2 + (0-1)^2]^{1/2} = (2)^{1/2} = \sqrt{2}$
3. d (0, 0, 0) to d (1, 1, 1) = $[(0-1)^2 + (0-1)^2 + (0-1)^2]^{1/2} = (3)^{1/2} = \sqrt{3}$
4. d (0, 1, 0) to d (0, 0, 0) = $[(0-0)^2 + (1-0)^2 + (0-0)^2]^{1/2} = (1)^{1/2} = 1$
5. d (0, 1, 0) to d (0, 1, 1) = $[(0-0)^2 + (1-1)^2 + (0-1)^2]^{1/2} = (1)^{1/2} = 1$
6. d (0, 1, 0) to d (1, 1, 1) = $[(0-1)^2 + (1-1)^2 + (0-1)^2]^{1/2} = (2)^{1/2} = \sqrt{2}$
7. d (0, 1, 1) to d (0, 0, 0) = $[(0-0)^2 + (1-0)^2 + (1-0)^2]^{1/2} = (2)^{1/2} = \sqrt{2}$
8. d (0, 1, 1) to d (0, 1, 0) = $[(0-0)^2 + (1-1)^2 + (1-0)^2]^{1/2} = (1)^{1/2} = 1$
9. d (0, 1, 1) to d (1, 1, 1) = $[(0-1)^2 + (1-1)^2 + (1-1)^2]^{1/2} = (1)^{1/2} = 1$
10. d (1, 1, 1) to d (0, 0, 0) = $[(1-0)^2 + (1-0)^2 + (1-0)^2]^{1/2} = (3)^{1/2} = \sqrt{3}$
11. d (1, 1, 1) to d (0, 1, 0) = $[(1-0)^2 + (1-1)^2 + (1-0)^2]^{1/2} = (2)^{1/2} = \sqrt{2}$
12. d (1, 1, 1) to d (0, 1, 1) = $[(1-0)^2 + (1-1)^2 + (1-1)^2]^{1/2} = (1)^{1/2} = 1$

#2 (10 Points)

Load the “COVID19_v3.CSV” dataset, from the raw_data module in CANVAS, into R. This is a fictional COVID19 healthcare workers data set. Perform the EDA analysis by:

(See the data dictionary at the last page of this exam).

I. Summarizing each column (e.g. min, max, mean)
II. Identifying missing values
III. Displaying the frequency table of “Infected” vs. “MaritalStatus”
IV. Displaying the scatter plot of “Age”, “MaritalStatus” and “MonthAtHospital”, one pair at a time
V. Show box plots for columns: “Age”, “MaritalStatus” and “MonthAtHospital”
VI. Replacing the missing values of “Cases” with the “mean” of “Cases”.

Answer: Please see Solution2.R

Use EXCEL and the “COVID19_A.CSV.xlxs” (Excel file containing another variation of the fictional COVID19 dataset) to solve the following two problems.

#3 (10 Points)

Use unweighted Knn (k=3) to classify the following three records (test dataset)

Use only excel for this problem.

Exposure	MartialStatus	MonthAtHospital	Infected
1	Married	1	Yes
3	Single	4	No
2	Single	6	Yes

Answer: Please see Solution3-4-9.xslx, sheet- Solution3.

#4 (15 Points)

Discretize the "MonthAtHospital" into "less than 6 months" and "6 or more months". Construct a classification and regression tree (CART) to classify infection ("infected') based on the other variables (only one split level). Use only excel for this problem. Do not use the original MonthAtHospital a predictor.

Answer: Please see Solution3-4-9.xslx, sheet- Solution4.

#5 (10 Points)

Load the CANVAS "COVID19_v3.CSV" dataset into R/Python. Remove the missing values. Discretize the "MonthAtHospital" into "less than 6 months" and "6 or more months". Also discretize the age into "less than 35", "35 to 50" and

“51 or over”. Construct a Naïve Bayes model to classify infection (“infected’) based on the other variables. Measure the accuracy of the model.

Do not use the original MonthAtHospital and age variables as predictors.

Answer: Please see Solution5.R

#6 (10 Points)

Load the CANVAS “COVID19_v3.CSV” dataset into R/Python. Remove the missing values. Discretize the “MonthAtHospital” into “less than 6 months” and “6 or more months”. Also discretize the age into “less than 35”, “35 to 50” and “51 and over”. Construct a CART model to classify infection (“infected’) based on the other variables. Measure the accuracy of the model.

Do not use the original MonthAtHospital variable as a predictor.

Answer: Please see Solution6.R

#7 (10 Points)

Load the CANVAS fictional “COVID19_v3.CSV” dataset into R/Python. Remove the missing values. Use unweighted knn(k=5) to predict infection rate (infected) for a random sample (30%) of the data (test dataset).

Answer: Please see Solution7.R

# 8(10 Points)

The following table shows the population and the actual current prevalence rate of COVID19 in the US, Italy and Spain.

Considering <u>only</u> the three countries (US, Italy and Spain) use the table to answer the following questions:

a) **Estimate the number of cases in the US, Italy and Spain.**
b) **Given that a person is living in the US, what is the probability that the person is infected with COVID19.**
c) **Given that a person is diagnosed with the COVID19, what is the probability that the person lives in the US.**

	Population	**Prevalence**
	rounded to	
	nearest	Cases
	Million	**Per Million**
US	331	381.24
Italy	60	1463.97
Spain	47	1590.24

Answer:

A: Estimation of the number of cases in the US, Italy and Spain.

Estimated number of cases in US = Population * Cases Per Million

= 381.24 * 331

= 126190.44

Estimated number of cases in Italy = Population * Cases Per Million

= 1463.97 * 60

= 87838.22

Estimated number of cases in Spain = Population * Cases Per Million

=1590.24 * 47

= 74741.28

B: The probability of a person being infected with COVID-19 who lives in the US

= P (Infected people / People in USA)

= 126190.44 / 331000000

= 0.000381

C: Probability of a person live in US given that the person is diagnosed with COVID-19

= P (People infected in US / People infected in US, Italy & Spain)

= 126190.44 / (126190.44 + 87838.2 + 74741.28)

= 126190.44 / 288769.92

= 0.43699

#9 (15 Points)

a) Company XYZ is targeting professionals between the ages of 20 and 50 years old with an asset size of 50k to 100K. To estimate the missing income fields, the company is using k-nearest neighbors. (Use Excel for this problem) What would be the value of income for customer x in the table below if:

K = 1 and method = ”unweighted vote” is used

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

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

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	Age	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: Please see Solution3-4-9.xslx, sheet- Solution9.

COVID19: Healthcare Workers data dictionary.

Age: Age of healthcare worker

Exposure: Level of exposure to COVID 19 patients

MaritalStatus: Marital Status

Cases: Number of the cases in the county

MonthAtHospital: Number of months that the healthcare worker has been working at the current facility

Infected: Is healthcare worker infected by the COVID19 virus (yes or no?)