Instructions:

You can use Word, Excel, Power Point, R and/or Python to answer the questions in this exam. There are a total of eight (8) 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 of 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) and (0, 1, 0)

$$d(x, y) = \Sigma\left((x_{i-}{}_{Y_i})^{3}\right)$$

 # 2 (10 Points)

An employee of a company is traveling to either England, Italy, or Spain. The employee can travel to only one country. There is a 50% chance the employee will go to England and a 20% chance to Italy.

Assume the chances of contracting COVID to be proportional to the prevalence of the disease in each country, given in the table below. For example, the chances of contracting COVID in England is 1200/1,000,000.

	Prevalence Cases Per Million
England	1200
Italy	1500
Spain	1600

What are the chances that the employee will contract COVID while travelling?

Assume that the employee has traveled to Europe and contracted COVID, what is the probability that he/she traveled to England?

#3 (10 Points)

Load the "COVID19_v4.CSV" dataset, from the raw_data module in CANVAS, into R/Python. 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. **Replacing the numerical missing values with the "mode" of the corresponding columns**
IV. **Displaying the scatter plot of "Age", "Exposure" and "MonthAtHospital", one pair at a time**
V. **Showing box plots for columns: "Age", and "MonthAtHospital"**

#4 (15 Points)

Use Excel and the "COVID19_A.CSV.xlxs" (Excel file containing another variation of the fictional COVID19 dataset) to solve the following problem.

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	12	Yes

#5 (15 Points)

Load the CANVAS "COVID19_v4.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. Predict infection rate (infected) for a random sample (30%) of the data (test dataset). Measure the accuracy of the model.

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

Hint (see 'ifelse' function in R)

#6 (10 Points)

Load the CANVAS "COVID19_v4.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. Predict infection rate (infected) for a random sample (30%) of the data (test dataset).

Measure the accuracy of the model.

Do not use the original MonthAtHospital variable as a predictor.

#7 (15 Points)

Load the CANVAS fictional "COVID19_v4.CSV" dataset into R/Python. Remove the missing values. Develop a knn classifier based on the other variables except "MaritalStatus". Use unweighted knn(k=5) to predict infection rate (infected) for a random sample (30%) of the data (test dataset).

#8 (15 Points)

The table below shows whether an applicant has been rejected, waitlisted, or admitted to a college. There are three predictors. All variables have been categorized to categorical variables.

Use Excel and the CART methodology to develop a classification model for the following training data (one level only):

Applicant	GRE	Gender	Admission	GPA
1	Low	Female	Admitted	High
2	Low	Male	Rejected	Low
3	Low	Male	Waitlisted	Medium
4	Very High	Male	Admitted	Low
5	Very High	Female	Admitted	Medium
6	Very High	Male	Admitted	High
7	Very High	Female	Admitted	High
8	High	Female	Admitted	Medium
9	High	Male	Waitlisted	Low
10	Medium	Female	Waitlisted	High
11	Medium	Male	Rejected	Low

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?)