**Instructions:**

**You can use Word, Excel, Power Point, or Python/R to answer the questions in this exam. There are a total of seven (7) 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 Python/R programs, and the programs should be commented on 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 from (0, 0, 0) to (0, 1, 0) and from (0, 1, 0) to (0, 0, 0)**

**The function is not a proper distance function because the function can return negative numbers and that breaks one of the laws of distance functions.**

**= 1**

**# 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 1,200/1,000,000.**

|  |  |
| --- | --- |
|  | **Prevalence** |
|  | Cases |
|  | **Per Million** |
| **England** | 1,200 |
| **Italy** | 1,500 |
| **Spain** | 1,600 |

**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?**

The overall probability of contracting COVID during the trip:

Probability of going to England: 50% = 0.50

Probability of contracting COVID in England: 1200/1000000 = 0.0012

Probability of going to Italy: 20% = 0.20

Probability of contracting COVID in Italy: 1500/1000000 = 0.0015

Probability of going to Spain: 30% = 0.30

Probability of contracting COVID in Spain: 1600/1000000 = 0.0016

The total probability is the sum of the probabilities for each country:

P(COVID) = (0.5 x 0.0012) + (0.20 x 0.0015) + (0.30 x 0.0016)

P(COVID) = 0.0006 + 0.0003 + 0.00048

P(COVID) = 0.00138

Therefore, the chance that the employee will contract COVID while traveling is **0.00138 or 0.138%.**

Probability of Having Traveled to England Given COVID Contraction

P(E) = 0.50 (probability of going to England)

P(C|E) = 0.0012 (probability of contracting COVID in England)

P(C) = 0.00138 (total probability of contracting COVID, calculated earlier)

Therefore, assuming the employee traveled to Europe and contracted COVID, the probability that they traveled to England is approximately **0.4348 or 43.48%.**

| **#3** (10 Points)  **Load the “hepatitis\_A.csv” dataset, from the Raw\_data module in CANVAS, into Python (see the data dictionary at the bottom of this document). This is a dataset used for predicting “patient mortality”. Perform the EDA analysis by:**   | 1. **Summarizing each numerical column (e.g., min, max, mean)** 2. **Displaying scatter plots of “BILIRUBIN”,”** **SGOT” and “ALBUMIN” one pair at a time** 3. **Showing box plots for columns “BILIRUBIN”,”** **SGOT” and “ALBUMIN”** |  | | --- | --- | |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |

**#4** (15 Points)

**Load the “hepatitis\_A.csv” dataset, from the Raw\_data module in CANVAS, into Python/R (Excel file containing another variation of the hepatitis dataset). This is a dataset used for predicting “patient mortality”. Construct a CART model to classify “patient mortality” based only on the** **“SEX”,”Age\_Quartile”, “STEROID”, “FATIGUE” and “MALAISE” attributes**

**#5 (15 Points)**

**Load the “hepatitis\_B2.csv” dataset, from the Raw\_data module in CANVAS,**

**into Python. This is a variation of the hepatitis dataset used for predicting**

**“Patient mortality”. Construct a knn model to classify “patient death” based on only the AGE, SEX, ASCITES, BILIRUBIN, ALK\_PHOSPHATE, SGOT, ALBUMIN**

**attributes (K=1,3,5,7)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |

**#6** (20 Points)

**Use Excel and the k-Nearest Neighbors (k-NN) algorithm with k=1,2,3 to classify the first five rows of hepatitis\_C2.csv as test data. Use the remaining records as training data. Evaluate the model's performance by measuring, accuracy**, **precision, recall, and F1-score.**

**(Excel file hepatitis\_C2.csv is another variation of the hepatitis dataset).**

**#7** (20 Points)

**Use Excel and the training data** **in *hepatitis\_D2.csv* to construct the first-level split of a CART (Classification and Regression Tree) classification model.**

**(Excel file *hepatitis\_D2.csv* is another variation of the hepatitis dataset).**

**Data Dictionary**

1. **Class** – The outcome of the disease (target variable):
   * 1: Die
   * 2: Live
2. **AGE** – The age of the patient (numerical).
3. **SEX** – The gender of the patient:
   * 1: Male
   * 2: Female
4. **STEROID** – Whether the patient was given steroids (Boolean):
   * 1: Yes
   * 2: No
5. **ANTIVIRALS** – Whether the patient was given antiviral medications (Boolean):
   * 1: Yes
   * 2: No
6. **FATIGUE** – Presence of fatigue (Boolean):
   * 1: Yes
   * 2: No
7. **MALAISE** – Presence of general discomfort or unease (Boolean):
   * 1: Yes
   * 2: No
8. **ANOREXIA** – Presence of appetite loss (Boolean):
   * 1: Yes
   * 2: No
9. **LIVER\_BIG** – Whether the liver is enlarged (Boolean):
   * 1: Yes
   * 2: No
10. **LIVER\_FIRM** – Whether the liver is firm (Boolean):

* 1: Yes
* 2: No

1. **SPLEEN\_PALPABLE** – Whether the spleen is palpable (Boolean):

* 1: Yes
* 2: No

1. **SPIDERS** – Presence of spider angiomata (Boolean):

* 1: Yes
* 2: No

1. **ASCITES** – Presence of fluid accumulation in the abdomen (Boolean):

* 1: Yes
* 2: No

1. **VARICES** – Presence of esophageal varices (Boolean):

* 1: Yes
* 2: No

1. **BILIRUBIN** – Bilirubin levels in the blood (continuous, measured in mg/dL).
2. **ALK\_PHOSPHATE** – Alkaline phosphatase levels in the blood (continuous, IU/L).
3. **SGOT** – Serum glutamic-oxaloacetic transaminase (SGOT) levels (continuous, IU/L).
4. **ALBUMIN** – Albumin levels in the blood (continuous, g/dL).
5. **HISTOLOGY** – Result of the histological examination of the liver (Boolean):

* 1: Yes (abnormal)
* 2: No (normal)