**1) Is the following function a proper distance function? Why? Explain your answer. Measure the distance between (0, 0, 0) and (0, 1, 0)**

**Ans:**

The function given above is **NOT** a valid distance function. A valid distance function should satisfy the following properties

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

Here, the given equation does not satisfy property 1. For instance let’s take 2 points A (0,0,0) and B(0,1,0).

* D(x,y) = D(A,B) = (0-0)^3 + (0-1)^3 +(0-0)^3 = -1

Here, acc. to the distance function, distance from A to B is -1, but we know that distance cannot be negative, so the function is proved to be invalid.

Acc. to the above distance function, the distance between (0, 0, 0) and (0, 1, 0) is -1. The actual distance between (0, 0, 0) and (0, 1, 0) as per euclidean rule:

((0-0)^2 + (0-0)^2 + (0-0)^2)^(1/2) = 1

**2) 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?**

**Ans:** Given,

Prob. of going to England, P(Eng)= 50% =0.5

Prob. of going to Italy, P(Ita)= 20% =0.2

Prob. of going to Spain, P(Spn)= 30% =0.3

Prob. of contacting covid in England, P(covid\_Eng) = 1200/1000000

Prob. of contacting covid in Italy, P(covid\_Ita) = 1500/1000000

Prob. of contacting covid in Spain, P(covid\_Spn) = 1600/1000000

Acc. To the law of total probability:

**Prob. of contacting covid** = P(Eng).P(covid\_Eng)+ P(Spn).P(covid\_Spn)+ P(Ita).P(covid\_Ita)

= 0.5\*1200/1000000 + 0.2\*1500/1000000 + 0.3\*1600/1000000

= 0.0006+0.0003+0.00048 = **0.00138 or 0.138%**

P( person goes to Europe and contact covid) = 0.00138

P( person goes to England and contacts covid) = P(Eng).P(covid\_Eng)= 0.0006

**Prob. that given person contacts covid in europe, he goes to England**=P(Prob. person goes to Eng./ Prob. person contact covid in Europe)

= 0.0006/0.00138 = **0.434 or 43.4%**

**3) 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).**

1. **Summarizing each column (e.g., min, max, mean)**
2. **Identifying missing values**
3. **Replacing the numerical missing values with the “mode” of the corresponding columns**
4. **Displaying the scatter plot of “Age”, “Exposure” and “MonthAtHospital”, one pair at a time**
5. **Showing box plots for columns: “Age”, and “MonthAtHospital”**

**Ans:**

Please find answer in Midterm\_Q3.R file attached.

**4)** **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 |

**Ans:**

Please find the answer attached in Midterm\_Q4.xlsx file.

**5) 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)**

**Ans:**

Please find the answer code in Midterm\_Q5.ipynb file and the output in Midterm\_Q5.png file.

**7) 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).**

**Ans:**

Please find the answer attached in Midterm\_Q7.R file.

**8) 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 |

**Ans:** Please find the answer attached in Midterm\_Q8.xlsx file.