**Instructions:**

**You can use Word, Excel, Power Point, 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 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** (5 Points)

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

**# 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 “****UCI\_heart.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 “Heart\_problems” in patients. Perform the EDA analysis by:**

| 1. **Summarizing each column (e.g., min, max, mean)** |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |

1. **Displaying the scatter plot of “Age”, “trestbps”, “chol”, and “thalach” one pair at a time**
2. **Showing box plots for columns: “Age”, “trestbps” , “chol”, and “thalach”**

**#4** (15 Points)

**Load the “UCI\_heart.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 “Heart\_problems” in patients. Construct a Naïve Bayes model to classify Heart\_problems based only on** **the “sex”, “fbs”, “restecg”, “exang” and “Age\_quantile” variables and 25%- 75% test-training splits**

**#5** (15 Points)

**Load the “UCI\_heart.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 “Heart\_problems” in patients. Construct a CART model to classify Heart\_problems based only on the “sex”, “fbs”, “restecg”, “exang” and “Age\_quantile” variables and 25%-75% test-training splits!**

**#6** (15 Points)

**Load the “UCI\_heart.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 “Heart\_problems” in patients. Construct a knn model to classify Heart\_problems based only on the “age”, “sex”, “trestbps”, “chol”, “thalach” variables and 25%-75% test-training split**

**#7** (15 Points)

**Use knn (k=3), Excel and the training data in** **“ UCI\_heart\_v2.csv” (Excel file containing another variation of the dataset UCI heart dataset) to score the test data in “UCI\_heart\_v2”. Measure the accuracy, precision, recall and f1 of the results.**

**#8** (15 Points)

**Use Excel and the training data in “ UCI\_heart\_v3.csv” (Excel file containing another variation of the dataset UCI heart dataset) to construct the first level split for a CART classification model.**

**Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| age | Feature | Integer | Age |
| sex | Feature | Categorical | Gender |
| trestbps | Feature | Integer | resting blood pressure (on admission to the hospital) |
| chol | Feature | Integer | serum cholestoral |
| fbs | Feature | Categorical | fasting blood sugar > 120 mg/dl |
| restecg | Feature | Categorical | Resting electrocardiogram |
| thalach | Feature | Integer | maximum heart rate achieved |
| exang | Feature | Categorical | exercise induced angina |
| Heart\_problem | Target | Integer | diagnosis of heart disease |
| Age\_quantile | Feature | Categorical | Age quantile |