**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 your answers/files into a single zipped file and post the zipped file to “HW\_Midterm” in CANVAS.**

**#1** (5 Points)

**Is the following function a proper distance function? Why? Explain your answer.**

****

**Hint: Measure the distance between (0,0), (0,1) and (1,1)**

**#2** (10 Points)

**Load the “IBM\_Employee\_Attrition\_v3” dataset, from the raw\_data module in CANVAS, into R. This is a variation of a fictional data set created by IBM data scientists for predicting employee attrition. Perform the EDA analysis by:**

1. **Summarizing each column (e.g. min, max, mean )**
2. **Identifying missing values**
3. **Displaying the frequency table of “Attrition” vs. “MaritalStatus”**
4. **Displaying the scatter plot of “Age”, “MaritalStatus” and “YearsAtCompany”, one pair at a time**
5. **Showing box plots for columns: “Age”, “MaritalStatus” and “YearsAtCompany”**
6. **Replacing the missing values of “MonthlyIncome” with the “mean” of “MonthlyIncome”.**

**Use EXCEL and the IBM\_attrition\_A.xlxs (Excel file containing another variation of the IBM attrition dataset) to solve the following three problems.**

**#3** (10 Points)

**Use Knn (k=3) to classify the follow three records (test dataset)**

|  |  |  |  |
| --- | --- | --- | --- |
| **JobSatisfaction** | **MartialStatus** | **YearsAtCompany** | **Attrition** |
| 1 | Married | 1 | Yes |
| 3 | Single | 8 | No |
| 2 | Single | 6 | Yes |

**#4** (20 Points)

**Discretize the “YearsAtCompany” into “less than 6 years” and “6 or more years”. Construct a classification and regression tree to classify attrition based on the other variables (only one split level).**

**#5** (20 Points)

**Discretize the “YearsAtCompany” into “less than 6 years” and “6 or more years”. Construct a C4.5 to classify attrition based on the other variables (only one) split level.**

**#6 (**10 Points)

**Load the CANVAS “IBM\_Employee\_Attrition\_v3” dataset into R/Python. Remove the missing values. Use knn(k=5) to predict “attrition rate” for a random sample (30%) of the data (test dataset).**

**#7 (**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 |

**# 8** (10 Points)

**There are three major manufacturing companies that make a product: manufacturers A, B, and C. Manufacturer A has a 50% market share, and manufacturer B has a 30% market share. 5% of A’s products are defective, 6% of B’s products are defective, and 8% of C’s products are defective.**

1. What is the probability that a randomly selected product is defective? P(Defective)?
2. What is the probability that a randomly selected product is defective and that it came from A? P(A and Defective)?
3. What is the probability that a defective product came from B? P(B/Defective)?
4. Are these events (being defective and coming from B) independent? Why?