**Instructions:**

**You can use Word, Excel, Power Point and/or R to answer the questions in this exam. There are a total of six (6) 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 program, and the program 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** (10 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** (20 Points)

**Load the “IBM\_Employee\_Attrition\_V3” from canvas into R and 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. **Show histogram box plot for columns: “Age”, “MaritalStatus” and “YearsAtCompany”**
6. **Replacing the missing values of “MonthlyIncome” with the “mean” of “MonthlyIncome”.**

**#3** (20 Points)

**Use the “IBM\_Employee\_Attrition\_v3” dataset in CANVAS into R. This is a subset of a fictional data set created by IBM data scientists. Do not normalize the data. Use knn(k=3) to predict “attrition rate” for a random sample(30%) of the data (test dataset)**

**#4** (30 Points)

**a) Company XYZ is targeting professionals between the ages of 25 and 45 years old with an asset size of 50 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**

**K =3 and method = ”distance weighted 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 |

**# 5** (10 Points)

**There are three major manufacturing companies that make a product: manufactures A , B , and C . Manufacture A has a 50% market share, and manufacture 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?

**# 6** (10 Points)

**True or False:**

1. **In data mining, we usually delete all the data rows that contain a missing value to obtain a clean dataset.**
2. **Supervised data mining methods are those that use expert opinions.**
3. **Usually, low-complexity classifiers/separators need not change very much to accommodate new data points.**
4. **The optimal level of model complexity is obtained at the minimum error rate on the training dataset.**
5. **Data mining processes are autonomous, requiring little or no human oversight.**