**Instructions:**

**You can use Word, Excel, Power Point, R and/or Python to answer the questions in this exam. There are a total of seven (7) multi-part questions, with point values noted for each question. You must use Excel if it is specified in the problem statement.**

**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 of 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 between (0, 0, 0) and (0, 1, 0)**

**#2** (15 Points)

**Load the “IBM\_attrition\_v3” dataset, from the raw\_data module in CANVAS, into R/Python. This is a fictional, IBM supplied, employee data set. Perform the EDA analysis by:**

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

**#3** (15 Points)

**Use Excel and the “****IBM\_Attrition\_missing.csv” (Excel file containing another variation of the fictional employee dataset) to replace the missing “MonthlyIncome” using knn (k=3).**

**#4** (15 Points)

**Load the CANVAS “IBM\_attrition\_v2.CSV” dataset into R/Python. Remove the missing values. Discretize the “MonthlyIncome” into “up to3,000”, “3000 up to 5,000.00”, “5000 up to 8,500” and “8500 or more” per months. Also discretize the age into “less than 31”, “31 up to 38”, “38 up to 48”, and “48 or over”. Construct a Naïve Bayes model to classify attrition (attrition=’yes’) based on the other variables. Predict attrition for a random sample (30%) of the data (test dataset). Measure the accuracy of the model.**

**Hint (see ‘ifelse’ function in R)**

**#5** (15 Points)

**Load the CANVAS “IBM\_attrition\_v2.CSV” dataset into R/Python. Remove the missing values. Discretize the “MonthlyIncome” into “up to3,000”, “3000 up to 5,000”, “5,000 up to 8,500” and “8500 or more” per months. Also discretize the age into “less than 31”, “31 up to 38”, “38 up to 48”, and “48 or over”. Construct a CART model to classify attrition (attrition=’yes’) based on the other variables. Predict attrition for a random sample (30%) of the data (test dataset). Measure the accuracy of the model.**

**Hint (see ‘ifelse’ function in R)**

**#6 (**15 Points)

**Load the CANVAS “IBM\_attrition\_v2.CSV” dataset into R/Python. Remove the missing values. Construct a knn (k=3) model to classify attrition (attrition=’yes’) based on the other variables. Predict attrition for a random sample (30%) of the data (test dataset). Measure the accuracy of the model.**

**Hint (see ‘ifelse’ function in R)**

**#7** (15 Points)

**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 | Male | Waitlisted | Medium |
| **9** | High | Male | Waitlisted | Low |
| **10** | Medium | Female | Waitlisted | High |
| **11** | Medium | Male | Rejected | Low |

**Dataset dependency: IBM\_attrition\_V2, IBM\_attrition\_v3, IBM\_Attrition\_missing.csv**