**Instructions:**

**There are a total of five (5) multi-part questions, with point values noted for each question. You can use R/Python or Excel unless it is specified**

**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 CANVAS.**

**Problem 1 – data prep(20 points)**

The “IBM\_attrition\_v3” CSV dataset on CANVAS, shows whether an employee has left a company (“attrition=yes) or not. Create a dataset (Attrition\_Modified) by engineering the following features:

1. Delete all the rows with missing value.
2. Create four categories (income1, income2, income3, income4) based on monthly income (“MonthlyIncome”) as:
   * 1. Monthly income <=2900
     2. 2900 < Monthly income <=5000
     3. 5000 < Monthly income <=8500
     4. 8500 < Monthly income
3. Create two categories (senior, not-senior) for years at the company (“YearsAtCompany”):
   * 1. Years at the company <=6
     2. 6 < years at the company
4. Create two categories (young, mature) for age:
   * 1. age <=37
     2. 37<age

Drop the original columns: MonthlyIncome, YearsAtCompany and age from the dataframe

**Problem 2 – Random Forest (20 points)**

Use the **Random Forest** methodology to develop a classification model for attrition using the “Attrition\_Modified “ dataset. Create test and training datasets, by selecting every fourth record, starting from the first observation, as the test dataset and the remaining records as the training dataset. Score the test dataset. What is the accuracy of your model?

**Problem 3 – C5.0 (20 points)**

Use the **C5.0**  methodology to develop a classification model for attrition using the “Attrition\_Modified “ dataset. Create test and training datasets, by selecting every fourth record, starting from the first observation, as the test dataset and the remaining records as the training dataset. Score the test dataset. What is the accuracy of your model?

**Use Excel to solve the following two problems.**

**Problem # 4: (20 points)**

Using data in the table below, construct a Neural Network with one Output Layer (z) and one Hidden Layer (two nodes A and B). Calculate the predicted outcome if the inputs to the input nodes are (Node 1=.4, Node 2=.7 Node 3= .7 and Node 4=.2)

Use the actual value of .75 and a learning factor of .1 to adjust the weight for A to z. (Extra credit for using Matrix multiplication)

|  |  |  |
| --- | --- | --- |
| **From** | **To** | **Weight** |
| X | A | 0.5 |
| Node 1 | A | 0.6 |
| Node 2 | A | 0.8 |
| Node 3 | A | 0.6 |
| Node 4 | A | 0.2 |
| x | B | 0.7 |
| Node 1 | B | 0.9 |
| Node 2 | B | 0.8 |
| Node 3 | B | 0.4 |
| Node 4 | B | 0.2 |
| xx | z | 0.5 |
| A | z | 0.9 |
| B | z | 0.9 |

**Problem # 5: 4.5 (20 points)**

Use Excel and the **C4.5** methodology to develop a classification model for the “admitted” outcome using the following training data (one level only):

|  |  |  |  |
| --- | --- | --- | --- |
| **Applicant** | **GRE** | **GPA** | **Admitted** |
| **1** | Medium | High | Yes |
| **2** | Low | Low | No |
| **3** | High | Medium | Yes |
| **4** | Medium | Medium | No |
| **5** | Low | Medium | No |
| **6** | High | High | Yes |
| **7** | Low | Low | No |
| **8** | Medium | Medium | Yes |

Datasets: IBM\_attrition\_v3