**Instructions:**

**There are a total of five (5) 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 CANVAS.**

**Problem 1 - (20 points)**

The “AdmissionB” CSV dataset on CANVAS, shows whether an applicant has been admitted to a college (admit=1), or not (admit=0). There are three predictors. The variables gre and gpa are continuous. The variable rank is categorical and takes on the values 1 through 4.

* Create a dummy variable rankb=1 when an applicant’s rank is 1 or 2 and rankb=0 otherwise.
* Use the kmeans clustering method to create two clusters for the Admission dataset using gre, gpa, and rankb as clustering variables. Tabulate the clustered rows against the “ADMIT” column.
* Use the hierarchical clustering method to create two clusters for the Admission dataset using gre, gpa and rankb as clustering variables. Tabulate the clustered rows against the “ADMIT” column.

**An analyst has categorized the gre and the gpa variables into four categories: low, medium, high, and very high. Use the resulting dataset “Admission\_catB” on CANVAS to develop the following two classification models.**

**Problem 2 - (20 points)**

Use the Random Forest methodology to develop a classification model for the Admission\_cat dataset using gre, gpa and the rank variables as predictors. 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 - (20 points)**

Use the c5.0 methodology to develop a classification model for the Admission\_cat dataset using the gre, gpa and the rank variables as predictors. 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 .78 and a learning factor of .1 to adjust the weight for A to z.

|  |  |  |
| --- | --- | --- |
| **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: (20 points)**

Use the c4.5 methodology to develop a classification model for 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: AdmissionB, Admission\_catB