**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 “AL\_NJ\_Income\_pct” CSV dataset on CANVAS categorizes the tax returns of families in the states of Alabama and New Jersey into six categories (Returns\_pct1 to Returns\_pct6). Use these six categories and Euclidian distance, to perform the following analysis

* Use the kmeans clustering method to create two clusters for the “AL\_NJ\_Income\_pct” dataset.
* Show the cross tabulation of the clusters versus the State feature.
* Use the hierarchical clustering method and single linkage to create 4 clusters for the the “AL\_NJ\_Income\_pct” dataset.
* Identify the outliers (if any).

**Problem 2 - (20 points)**

Use the Random Forest methodology to develop a classification model for the “State” (target), using the Returns\_pct1 to Returns\_pct6 features in the “AL\_NJ\_Income\_pct dataset.

* Show the cross tabulation of the classification.
* What is the accuracy of your model?
* What is the precision of the model?
* What is the recall of the model?
* What is the F1 of the model?

**Problem 3 - (20 points)**

Use the C5.0 Forest methodology to develop a classification model for the “State” (target), using the Returns\_pct1 to Returns\_pct6 features in the “AL\_NJ\_Income\_pct dataset.

* Show the cross tabulation of the classification.
* What is the accuracy of your model?
* What is the precision of the model?
* What is the recall of the model?
* What is the F1 of the model?

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

Use theCART methodology to develop a classification model for the “State” (target), using the Returns\_pct1 to Returns\_pct6 features in the “AL\_NJ\_Income\_pct dataset.

* Show the cross tabulation of the classification.
* What is the accuracy of your model?
* What is the precision of the model?
* What is the recall of the model?
* What is the F1 of the model?

**Problem # 5: (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 xx 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 |

Datasets: AL\_NJ\_Income\_pct.csv