**Instructions:**

**There are a total of six (six) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s), for each problem. Your program(s) 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.**

**Problems #1** (30 points – two parts, 15 points each noted below)

**Using an “Addiction” dataset, a researcher has prepared the following table of patient counts:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ethnicity** | **Age Category** | **Alcohol** | **Cocaine** | **Heroin** | **Row Total** |
| **Black** | **Old** | **30** | **48** | **17** | **95** |
|  | **Young** | **25** | **72** | **13** | **110** |
| **Hispanic** | **Old** | **7** | **0** | **5** | **12** |
|  | **Young** | **8** | **7** | **19** | **34** |
| **White** | **Old** | **60** | **2** | **17** | **79** |
|  | **Young** | **26** | **10** | **34** | **70** |
| **Column Total** |  | **156** | **139** | **105** | **400** |

**Use the table above and Excel to classify patient addiction type (alcohol, cocaine, heroin) using Ethnicity and Age Category:**

* **Construct a classification and regression tree (CART) (two levels only).**  (15 Points)
* **Construct a C4.5 decision tree (two levels only).**  (15 Points)

**Problem 2 -** (15 points)

**The** **“KY\_NJ\_Zip” CSV dataset on CANVAS categorizes the percentage of tax returns of families in the states of Kentucky 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 four clusters for the “KY\_NJ\_ZIP” dataset.**
* **Show the cross tabulation of the clusters versus the State.**

**Problem 3 -** (15 points)

**The “KY\_NJ\_Zip” CSV dataset on CANVAS categorizes the percentage of tax returns of families in the states of Kentucky and New Jersey into six categories (Returns\_pct1 to Returns\_pct6). Use these six categories and average Euclidian distance, to perform the following analysis:**

* **Use the hierarchical clustering method to create four clusters for the “KY\_NJ\_Zip” dataset.**
* **Show the cross tabulation of the clusters versus the State.**

**Problem # 4:** (10 points)

**Use the Random Forest methodology to develop a classification model for the “Income” (target), using “Education\_years”, “Gender”, “Hours\_Worked\_Perweek” in the “Adult\_income\_v2.CSV” dataset (30% test 70% training).**

* **What is the accuracy of your model?**
* **What are “the mean decrease accuracies” of the features?**
* **What are the “mean decrease Gini Indices” of the features?**

**Problem # 5:** (10 points)

**Use the C5.0 methodology to develop a classification model for the “Income” (target), using “Education\_years”, “Gender”, “Hours\_Worked\_Perweek” in the “Adult\_Income\_v2.CSV” dataset (30% test 70% training).**

* **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?**

**6) Using data in the table below, construct a Neural Network with one Output Layer (z) and one Hidden Layer (A and B).** (20 points)

1. **Calculate the predicted outcome if the inputs to the input nodes are (x=1, Node 1=.4, Node 2=.7 Node 3= .7 and Node 4=.2).**
2. **Adjust the weight if the actual output is 0.7500**

|  |  |  |
| --- | --- | --- |
| **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 |

**Data dependencies: KY\_NJ\_Zip, Adult\_income\_v2**