**Instructions:**

**There are a total of 6 (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 to CANVAS.**

**Problems #1** (20 points)

**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 C4.5 decision tree (two levels deep only).**

**Problem #2** (15 points)

**Use R/Python and the NYNJ\_zipcode\_population.csv file to cluster the zip codes of New Jersey and New York into 3 clusters. Use “Low\_income” percent and “Total\_Pop” as clustering attributes. (Algorithm=hierarchical, linkage=simple).**

* **What are the members of each cluster?**
* **Is there an unusual zip code? If so, what is the zip code?**

**Problem #3** (10 points)

**Use R/Python and the** **NYNJ\_zipcode\_population.csv file to cluster the zip codes of New Jersey and New York into 5 clusters. Use “Low\_income” percent and “Total\_Pop” as clustering attributes (Algorithm=kmeans).**

* **What are the members of each cluster?**

**Problem #4** (20 points)

**Use an Artificial Neural Network, with 10 nodes in the hidden layer, learning rate of .001, and the standard scaler (StandardScaler) to develop a classification model for the** **heart\_attack .CSV dataset using 30% of the data as training data. (See the data dictionary below).**

* **What is the model accuracy ?**

**Problem #5** (20 points)

**Use Random Forest to develop a classification model for the heart\_attack .CSV dataset using 30% of the data as training data. (See the data dictionary below).**

* **What is the most important feature? Why?**
* **Show your model metrics.**

**Problem #6** (15 points)

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

* **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).**
* **Adjust the weight if the actual output is 0.8500**

|  |  |  |
| --- | --- | --- |
| **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 Dictionary**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Target/Attribute** |
| Heart\_Attack | Type of Heart Attack (Mild, Light, Massive) | Target |
| RestHR | Resting Heart Rate | Attribute |
| MaxHR | Maximum Heart Rate | Attribute |
| RecHR | Recovery Heart Rate | Attribute |
| BP | Blood Pressure | Attribute |

**Data dependencies: NYNJ\_zipcode\_population.csv,** **Heart\_attack.CSV**