**Instructions:**

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

**Please show your calculations, or the details of your program(s) for each problem. You must supply the Python/R 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 – C5.0 (15 points)**

**Load the “hepatitis\_A.csv” dataset, from the Raw\_data module in CANVAS, into Python/R (Excel file containing another variation of the hepatitis dataset). This is a dataset used for predicting “patient mortality”. Delete any row with missing values and construct a C5.0 model to classify “patient mortality” based only on the** **“SEX”,”Age\_Quartile”, “STEROID”, “FATIGUE” and “MALAISE” attributes. What is the accuracy of your model? Is this a good model? Why or why not.**

*(30% test, 70% training)*

**Problem 2 – Random Forest (15 points)**

**Load the “hepatitis\_A.csv” dataset, from the Raw\_data module in CANVAS, into Python/R (Excel file containing another variation of the hepatitis dataset). This is a dataset used for predicting “patient mortality”. Delete any row with missing values and construct a Random Forest model to classify “patient mortality” based only on the** **“SEX”,”Age\_Quartile”, “STEROID”, “FATIGUE” and “MALAISE” attributes. Show feature importance**

**What is the accuracy of your model? Is this a good model? Why or why not.**

*(30% test, 70% training)*

**Problem # 3 – Clustering (30 points)**

**Load the “hepatitis\_B2.csv” dataset, from the Raw\_data module in CANVAS,**

**into Python. This is a variation of the hepatitis dataset used for predicting**

**“Patient mortality”. Delete any row with missing value(s) and use Hierarchical and K-means clustering Algorithms to cluster the records into two clusters based on AGE, SEX, ASCITES, BILIRUBIN, ALK\_PHOSPHATE, SGOT, ALBUMIN attributes only**. **Show the “Patient mortality” versus clusters.**

*(Distance = Euclidian , linkage=ward)*

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

**Problem # 4– ANN (20 points)**

**Using the 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=.3, Node 2=.6 Node 3= .6 and Node 4=.3). Use the actual value of .80 and a learning factor of .1 to adjust the weight for A to z and Node 1 to A.**

|  |  |  |
| --- | --- | --- |
| **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 | 0.85 |
| B | z | 0.85 |

**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 |

**Data Dictionary**

1. **Class** – The outcome of the disease (target variable):
   * 1: Die
   * 2: Live
2. **AGE** – The age of the patient (numerical).
3. **SEX** – The gender of the patient:
   * 1: Male
   * 2: Female
4. **STEROID** – Whether the patient was given steroids (Boolean):
   * 1: Yes
   * 2: No
5. **ANTIVIRALS** – Whether the patient was given antiviral medications (Boolean):
   * 1: Yes
   * 2: No
6. **FATIGUE** – Presence of fatigue (Boolean):
   * 1: Yes
   * 2: No
7. **MALAISE** – Presence of general discomfort or unease (Boolean):
   * 1: Yes
   * 2: No
8. **ANOREXIA** – Presence of appetite loss (Boolean):
   * 1: Yes
   * 2: No
9. **LIVER\_BIG** – Whether the liver is enlarged (Boolean):
   * 1: Yes
   * 2: No
10. **LIVER\_FIRM** – Whether the liver is firm (Boolean):

* 1: Yes
* 2: No

1. **SPLEEN\_PALPABLE** – Whether the spleen is palpable (Boolean):

* 1: Yes
* 2: No

1. **SPIDERS** – Presence of spider angiomata (Boolean):

* 1: Yes
* 2: No

1. **ASCITES** – Presence of fluid accumulation in the abdomen (Boolean):

* 1: Yes
* 2: No

1. **VARICES** – Presence of esophageal varices (Boolean):

* 1: Yes
* 2: No

1. **BILIRUBIN** – Bilirubin levels in the blood (continuous, measured in mg/dL).
2. **ALK\_PHOSPHATE** – Alkaline phosphatase levels in the blood (continuous, IU/L).
3. **SGOT** – Serum glutamic-oxaloacetic transaminase (SGOT) levels (continuous, IU/L).
4. **ALBUMIN** – Albumin levels in the blood (continuous, g/dL).
5. **HISTOLOGY** – Result of the histological examination of the liver (Boolean):

* 1: Yes (abnormal)
* 2: No (normal)