**Instructions:**

**You can use Word, Excel, Power Point, Python and R to answer the questions in this test. 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. The R 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.**

**Use the following data descriptions for problems #1 and #2:**

Predict student performance in a statewide exam (e.g. BAR), using the training dataset StateWide.csv (CANVAS). It contains data for 50 students using the following variables. Show your work in an Excel file.

* ID: Student ID
* LSG: Last semester grade (A, B,C)
* CTG: Two preparation tests are conducted and the average of two tests are used to calculate marks. CTG is split into three categories:

Poor: = or < 40%

Average > 40% and < 60%

Good = or >60%.

* GP: General Proficiency seminar.

Yes: Student participated in seminar

No: Student did not participate

* Outcome: Outcome of statewide exam.

**#1 (20 points)**

**Use StateWide.csv (CANVAS), Excel and the CART Methodology to develop a classification model (Two levels).**

**#2 (20 points)**

**Use StateWide.csv (CANVAS), Excel and the C4.5 Methodology to develop a classification model (Two levels)**

**#3 (20 points)**

**Load the “IBM\_Attrition\_v3.csv”. Convert all the variables to numeric variables Perform the following tasks:**

1. **Delete rows with missing values**
2. **Select 30% of the records as the test dataset and the remaining records as the training dataset**
3. **Perform ANN, with 6 hidden nodes, for classification for the “attrition” column**
4. **Score the test dataset**
5. **Measure the error rate.**

**#4 (20 Points):**

**Load the “IBM\_Attrition\_v3.csv”. Read all the columns as numeric variables except the “Attrition”, “JobSatisfaction” and “MaritalStatus” columns. Perform the following tasks:**

1. **Delete rows with missing values**
2. **Select 30% of the records as the test dataset and the remaining records as the training dataset**
3. **Perform Random Forest classification for the “attrition” column**
4. **Score the test dataset**
5. **Measure the error rate.**

**Problem#5: (20 points)**

**Cluster NY zip codes into 4 clusters using the NYNJ\_ZIPCode.csv dataset in CANVAS, R/Python, and features** **Returns\_pct1 to Returns\_pct6. Compare the four clusters for each of the following two methods.**

* **Hierarchical clustering (centroid)**
* **K-means, population (centroid)**

Datasets:NY\_ZIP.csv, IBM\_Attrition\_v3.csv, StateWide.csv