**CSC4008 2019 Term2**

**Assignment #3**

**Deadline:** 2020/3/29, 23:59

(This is a strict deadline. Submissions are not acceptable after the deadline)

**Percentage:** 5%

**Purpose:** to practice how to do data preprocessing, and compare the performance of different classifiers.

**Requirement:** Please give brief descriptions or answers to the following exercises by using Python, R or weka. For Weca, please use the screenshot to prove your answer.

For Python or R, please use the screenshot to prove your answer and attach your code.

1. Load weather.nominal.arff, and answer the following questions.
2. How many instances, attributes in this dataset?
3. The number of the distinct label of the attribute: temperature.
4. Use the filter to remove the value: high in the humidity attribute from the dataset.

Hit: If you use weka, you can use the RemoveWithValuse filter. If you use Python, you can use sklearn.

1. Please build a J48 decision tree (if you use Python or Weka, you can build C4.5 decision tree) to classify the dataset (the whole dataset) and visualize the tree in a figure.
2. Analyze the performance of J48 as we introduced in tutorials (Including but not limited to TPR, FPR, ROC, AUC).
3. Load HR-Employee-Attrition.csv dataset.
4. Implement attribute selection. You should select N attributes from the dataset. You can decide the number N and the attribute selection standard by yourself.
5. Using these N attributes to build a classifier. Please explain why you choose this classifier. You’d better to compare with other classifiers to prove (at least two).
6. You need to compare different test options (Cross-validation and Percentage split), and explain which one is the best.

**Assessment criteria:** Correctness of the answers.

Performance of your classifier.