PESIT Department of Computer Science and Engineering

Course: Data Mining
Semester: 2016 Spring (January – May)
Instructor: BNR (Dr. B. Narsing Rao)

Assignment: 06
Topic: Classification – Decision Trees
Due by: Midnight on Tuesday, February 23, 2016
Method: Email to bnrao@pes.edu

Use the accompanying file **bank-data2.arff** for this assignment.

Write a Java program that uses the Weka API and performs the following tasks:

1. Take a random sample of 590 of the 600 instances (this will be the training **set**). The remaining 10 will be the **test set**
2. Build two decision trees to predict whether or not a customer will purchase a PEP. The two decision trees should be of the following types:

    a. C4.5 Decision Tree (use the J48 class) (See: http://en.wikipedia.org/wiki/C4.5_algorithm)
    b. Gini Tree (use the BFTree class)

    The output of this task should be the decision trees.
3. For each sample in the training set, use each tree to predict the class. The output of this task should be a table with 10 rows, one for each instance in the test set. The columns should be: Class, Class predicted by C4.5, Class Predicted by Gini

Answer the following questions:

1. What is the difference (if any) between the two decision trees?
2. Why is there a difference?
3. Take one example that you create on your own and explain how each decision tree will be used to predict the class for your example

Submit a zip file (using the standard naming convention) that contains the following files:

- Your program (.java)
- Output from the program (as described above) (.txt)
- Answers to the questions (.pdf)