PESIT Department of Computer Science and Engineering

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

Assignment:	02
Topic:	Data Characteristics and Discretization
Due by:	Midnight on **Tuesday, January 19, 2016**
Method:	Send zip archive (.zip, .rar, etc.) by email to bnrao@pes.edu The name of the zip archive should be: DM-A02-your USN-your name (USN must be upper case and your name should be in mixed case) The zip archive should contain the following (see below for details): 1. Java program (named DM02YourName) 2. Output from Java program (in a text file) 3. Answer to questions (in a pdf file)

This assignment will use the file **bank-data.csv** that has already been sent to you.

Write a Java program that uses the Weka API to perform the following tasks (use the template on the next page as a guide):

1. Read the data file bank-data.csv
2. Delete the **id** attribute
3. Print, in tabular form the following statistics: minimum, maximum, mean, and standard deviation of **age** and **income**
4. Compute and print the covariance and correlation between age and income using equations 3.4 and 3.5 on page 97 of the textbook (note: do not use an API for this purpose but write your own code)
5. Discretize the income into four bins of equal width (using the class `weka.filters.unsupervised.attribute.Discretize`) and print out the cut-points and the frequencies in each bin
6. (Optional) Include a scatter plot between age and income (write your own code or use any tool)

Answer the following question:

1. What conclusion, if any, can be drawn from the values obtained in task 4 above?
2. Suppose the income were measured in thousands instead of the actual value (for example, and income value of 17,456 now becomes 17.456), how would the results of task 4 above change (if at all)? Explain.

Program Listing

```
import weka.core.Attribute;
import weka.core.Instances;
import weka.core.converters.ConverterUtils.DataSource;
import weka.core.AttributeStats;
import weka.experiment.Stats;
import weka.filters.Filter;
import weka.filters.unsupervised.attribute.Discretize;

// Data Mining - Data Characteristics and Discretization Example
// Written by BNR (PESIT Dept of CSE)

public class DataPreprocessing {

  public static void main (String args[]) {

     String filename = "bank-data.csv";
     DataSource source;
     try {
           // Create new data source
           source = new DataSource(filename);

           // Read instances from the CSV file
           Instances instances = source.getDataSet();

           // Delete the ID attribute
           instances.deleteAttributeAt(0);

           // Get the statistics for attribute at index 0 (now age)
           int index = 0;
           Attribute attr = instances.attribute(index);
           AttributeStats astats = instances.attributeStats(index);
           Stats stats = astats.numericStats;
               …
           // Using unsupervised Discretize (see import)
           Discretize filter = new Discretize();
           filter.setAttributeIndices("1");
           // Set number of bins
           filter.setBins(3);
           filter.setInputFormat(instances);
           Instances output = Filter.useFilter(instances, filter);
           double[] cutPoints = filter.getCutPoints(0);

               …
     } catch (Exception e) {  e.printStackTrace();
     }
  }
}
```