# PESIT Department of Computer Science and Engineering

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

| | |
|---|---|
| Assignment: | 03 |
| Topic: | Data Cubes |
| Due by: | Midnight on **Tuesday, January 26, 2016** |
| Method: | See below for details; email to bnrao@pes.edu |
| | The name of the zip file should be: DM-A03-your USN-your name |

For this assignment, use the file **bank.data.csv**. Write a Java program that uses the Weka API and preforms the following tasks:

1. Preprocess the bank-data in the following manner:
   a) Delete all attributes **except** age, sex, region, and income
   b) Discretize age using 3 equal width bins and name the attribute values as
      - YOUNG (representing young people)
      - MIDDLE (representing middle aged people)
      - OLD (representing old people)

2. Compute all data cuboids for the facts **count** and **avg_income**, where **avg_income** represents the average income for a cell. The program should take a command line parameter **n** representing the dimension of the cuboid and print out the corresponding cuboids.

For example

DataCube 0 will print out the apex cuboid
DataCube n will print out all level n cuboids (n <= 3 in this case)

Either print the cuboids for both **count** and **avg_income** or use that as another command line parameter.

3. Answer the following questions based on your output:

Determine the average income for:

a) Inner City Males
b) Middle Aged Rural Females
c) Young Suburban People

In the 2-D cuboids, which cells have a support of less than 5% (i.e. count less than 5% of the total count)?

Please submit **three** files in text format:

- Code listing with proper comments, including your name
- Sample output (cut and paste from console or write to file)
- Answers to question 3 above

The above files must be in a zip archive.