National Institute of Technology Calicut
Department of Computer Science and Engineering
CS4038/CS4038D DATA MINING
Assignment 1

Submission deadline (on or before):
06th October 2019, 10:00:00 PM
Policies for Submission and Evaluation
You can form a group of one or two students among yourself and should inform the course instructor regarding the group split up through the Google sheets uploaded along with the assignment. You are allowed to **make modifications on the Google sheet till 15 September 2019**. You must submit your assignment in the moodle (Eduserver) course page, on or before the submission deadline. During evaluation, your uploaded documents will be checked and will be asked for explaining the tasks conducted for checking the genuinity of the submission. The total marks for the assignment is 10 marks. The marks awarded will be based on the uploaded documents and the viva examination

Assignment Question
Perform the following tasks and submit the outcomes described for each task

1. **Dataset Selection:**
   Task 1: Identify one of the existing datasets and briefly describe the dataset (Eg. Size of the dataset, the number of attributes). Describe one of the potential data mining applications of the selected dataset briefly.
   Outcome: Document describing the details of the dataset.
2. **Data Analysis:**
   Task 2: Select one of the attributes from the selected dataset and describe the appropriate measures of central tendency and dispersion. Select one the appropriate visualization technique for analysing the selected attribute. Compute those measures and visualize the data attribute with the help of python code and mention your insights.
   Outcome: Document describing the computed measures and the python codes.
3. **Data Pre-processing:**
   Task 3: Identify if there is any quality issues related with the attributes in the selected dataset. Discuss two data pre-processing techniques required for the dataset (Preferably data cleaning and data reduction techniques), and implement those pre-processing techniques with the help of python code. Select one attribute, and discuss about the appropriate normalization technique required by that attribute. Implement data normalization on that attribute with the help of python code and provide insights.
   Outcome: Document describing the mentioned details and the python codes.
4. **Weka/OpenRefine:**
   Task 4: Make use of Weka or OpenRefine to analyse the dataset. Load the selected dataset on the tool and use the pre-processing and visualization functionalities.
   Outcome: Document with screenshots of the tool usage on the selected dataset.