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

Submission deadline (on or before):
30th September ~~24th September~~ 2023, 10:00:00 PM

Each group should select a dataset from either UCI Machine Learning Repository / Kaggle (Except Iris, Heart Disease, Adult, Titanic, Diabetes) and fill the details in this Google Sheet. You are not allowed to choose a dataset which already one team has chosen (First filled team gets the preference over the dataset). Complete this by 13th September ~~11th September~~ 2023. There will be a penalty on the marks of the groups who delay dataset details-entry and task submission.

Then complete the following Assignment Question over the selected dataset, and must submit your assignment in the moodle (Eduserver) course page, on or before the submission deadline. Only one member among your team should make a submission in Eduserver on behalf of the entire team. Include a README.PDF which contains the name and roll number of the group members. Total of 5 files (4 PDF files related with each of the 4 following tasks and the README file) is expected to submit as part of this assignment.

During evaluation, the genuinity of the submission and contribution of each member will be checked either through viva/quiz. The total marks for the assignment is 8 marks. The marks awarded will be based on the uploaded documents and the viva/quiz.

Assignment Question

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

1. **Dataset Description:**
   Task 1: Briefly describe the chosen dataset (Eg. Size of the dataset, the number of attributes, name and type of attributes). Describe one of the potential data mining applications of the selected dataset briefly (4 to 5 sentences).
   Outcome: Document (T1_<TeamNumber>.PDF).

2. **Data Analysis:**
   Task 2: Select one of the attributes (except the unique ID columns) from the selected dataset and describe the appropriate measures of central tendency and dispersion. Select one the appropriate visualization technique for analyzing the selected attribute. Compute those measures and visualize the data attribute with the help of python code and mention your insights.
   Outcome: Document (T2_<TeamNumber>.PDF) describing the chosen attribute, computed measures and the python codes.

3. **Data Pre-processing:**
   Task 3: Identify if there are any quality issues related with the attributes in the selected dataset. Discuss two data pre-processing techniques required for the dataset (Data cleaning and data reduction techniques), and implement those pre-processing techniques with the help of python code. Select one attribute, and discuss 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 (T3_<TeamNumber>.PDF) describing the mentioned details and the python codes.

4. **Data Mining Tool Usage:**
   Task 4: Make use of Any open source Data Mining Tools like Weka to analyze the dataset. Load the selected dataset on the tool and use the pre-processing and visualization functionalities supported by that tool.
   Outcome: Document (T4_<TeamNumber>.PDF) with screenshots showing the results of two pre-processing operations on the tool, and a visualization technique on the tool for any selected attribute to analyze the nature of the attribute.