SECOND SEMESTER 2023-24
CS F415: DATA MINING - Project Description

Project Logistics: The following guidelines need to be followed

1. Size of the project group: Max 3 students in a group.
2. Submissions: One team member will submit the deliverable on CMS per group on or before the deadlines.
3. **For all submissions, save the file name with all team members IDs separated by "-"**

Overview

Your project evaluation will consist of five elements.

- Project Proposal and Data Collection (2% - 4 Marks): 18th Feb 2024
  - **You are free to choose any dataset /problem of your choice. Table -1 gives you a sample for each algorithm.**
  - **You must select one category from the following: Association Rule Mining, Clustering, Classification, or Outlier. For example, if you choose a classification, each member of the team will implement either Naive Bayes or SVM or Logistic or Decision tree, etc and finally compare the performance of all the implemented algorithms.**
- Phase-1 Presentations (4% - 8 Marks): 2nd March 2024 (schedule will release later)
  - **You are expected to pre-process the dataset identified and make an in-person presentation covering the required Pre-processing methods that apply to your dataset identified.**
- Final Presentations (8% - 16 Marks): 27th April 2024 (schedule will release later)
  - **Each team member will implement one algorithm and give a Demo of the same and finally comparative analysis will also be presented.**
- Research paper submission (6% - 12 Marks): 21st April 2024

Scale of Project

The specifics of the project will be very flexible. I expect each team to perform data mining on some real data set. The goal is to gain more in-depth and hands-on experience with a few algorithms taught in the classroom (Association Rule Mining, Clustering, Outlier Detection, and Classification).
Project options may include:

- apply advanced techniques from the class towards a real data set
- compare several basic techniques from the class to a real data set
- propose and test extensions to techniques from class on a real data set

Project Proposal (2% - 4 Marks) Due 18th Feb 2024:

Prepare a 250-word document outlining your plan. This should contain:

1. who is in your group?
2. what data do you plan to use?
3. what is the problem you are trying to solve?
4. why this problem is interesting?
5. what is new, you will implement/learn? (It is expected that each student in the team will implement one algorithm)

Table 1: **Sample data Sources**

Algorithm	Sample data sources
Association rule Mining	Sample Dataset 1 Sample Dataset 2
Clustering	Sample Dataset 1 Sample Dataset 2
Outlier Analysis	Sample Dataset 1 Sample Dataset 2
Classification	Sample Dataset 1 Sample Dataset 2

Research paper Template:

Title (This must be title cased)

BITS Pilani Hyderabad Campus

CS F415 Data Mining Project

Authors names and Emails

Abstract

(about 250 words)

Abstract Structure:

- A general background statement about the concept worked out in the paper. (1 sentence)
- Why is this problem important and or challenging? (1 sentence)
- What is done in this paper? (2-3 sentences)
- Outcome of your experimental results (1 – sentence)
- Keywords (At most 5)

1. **Introduction** *(1 to 1.5 pages with five paragraphs covering the questions below)*
   - What is the problem and objective?
   - Why is it interesting and important?
   - Why is it hard? (E.g., why do naive approaches fail?)
   - Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions? How does mine differ?)
   - What are the key components of my approach and results?

2. **Related Work** *(summary of research papers, one paragraph for research paper, roughly half a page)*
   - Discuss existing research and methods related to your problem, highlighting their strengths, gaps and limitations. (Cite each source) It can be arranged in chronological order or thematically.

3. **Approach/Methodology** *(you may have subsections if required - one page)*
   - What problem are you tackling?
   - What do you need to solve this problem? How would you get this information?
   - Discuss the dataset(s) and its properties.

- Detail the algorithm or technique you will apply - Why did you select this approach? Why it would be better than what has been done before? Describe the various steps in the algorithm or technique.

4. **Experiments** *(1 - 1.5 pages)*

   4.1 Dataset

   - What pre-processing methods do you need to apply?
   - What would the final processed dataset look like?

   4.2 Evaluation method / Metrics

   - What methods would you use to evaluate the proposed methodology? Discuss these methods in brief.

   4.3 **Experimental setup** *(hyperparameters, etc.)*

   - What are the characteristics and properties of the Data Mining techniques used in the proposed method?

5. **Results and discussion** *(1.5 to 2 pages)*

   - What were the outcomes of your proposed method?
   - What were the outcomes of your evaluation metrics?
   - Comparative analysis of the proposed method with current state of art/ existing techniques.
   - Test your method with another dataset and discuss the efficiency of your proposed method.

6. **Conclusion** *(About 250 words)*

   - Give a brief summary of the work done in your project. (Problem, method used, and results).
   - Discuss future scope of work. (How could your proposed method be further improved?)

7. **References**
   - IEEE referencing style must be used. Please refer to the following guide.