Advanced Data Mining - Continuous Assessments - Data Mining Project

The continuous assessment component of this course is based on a semester-long project (related to topics discussed in class) that will be selected by students. The projects will be done in groups of students that are randomly assigned by the instructor.

For the project deliverables, you are required to make only one submission per group.

Ideas for Projects:

You might find ideas for your projects by exploring the topics of various **data science competitions**:

- Databank Sri Lanka
- Data science competitions to save the world
- Yelp Dataset Challenge (deadline passed, but you can still get the data): open task, data (including 200K photos) from local businesses in 12 cities across 4 countries
- Various Kaggle competitions
- OSS World Challenge
- Data Science for Social Good
- Recent research papers of interest from top tier journals and conferences within the last two years

Project Deliverables:

1. Survey

(1-2 pages, IEEE Conference Format (also available on overleaf), 15% of the project grade). You will need to pick a research topic for your project and read 6-8 relevant papers. Ideally the survey will help you identify the specific problem you want to address, and will lead to the project proposal naturally. The survey will be part of your final report. It should be a well thought-out synthesis of the papers that you will read, not just a repetition of the paper's abstracts / introductions. Your survey should provide answers to the following questions:

- What is the common theme of the papers you read? Give the problem definition(s).
- What are the challenges of the area?
- How do the papers relate to each other?
- Are they solving a new problem or improving an existing method?
- What are the main techniques that they are using?
- What are the strengths and weaknesses of each paper (try to list at least 3 of each)?
- What are the limitations of each method?

- Think about some future directions. What would you do better? Think about scalability issues, generality (e.g., weighted, directed, time-evolving, attributed networks), applicability to various domains.

>> Include the names of all the group members in the pdf and briefly describe each member's contribution to the deliverable. If you want to submit a longer survey (perhaps to submit to a journal), please ask the instructor first.

2. Project Proposal

(2 pages in PDF format, 15% of the project grade).
Your proposal should include the following sections:

- Problem definition
- Challenges
- Most related prior work and its shortcomings (or research gap or potential to improve)
- Proposed approach
- Data that you will use
- Evaluation plan

>> Include the names of all the group members in the pdf and briefly describe each member's contribution to the deliverable.

3. Mid-term Report

(4-5 pages,IEEE Conference Format (also available on overleaf), 20% of the project grade).
See below for the sections that your final report should have. At this point, for your midterm report, you should start editing the following sections:

- Section 2. Data: Describe the synthetic and real data that you will use, and explain the data collection process (if applicable).
- Section 3. Proposed Method: Introduce the method that you propose, give the necessary definitions, potentially give proof of concept.
- Section 4. Experiments: Give some preliminary experiments (on synthetic or real data).
- Section 5. Progress and Next Steps (temporary section that would change in the final report): Outline your next steps and whether you are on track. Now that you have had time to work on your projects, if anything has changed with respect to your proposal, mention it.
- Section 6. Division of work (your grade will depend on your contribution to the project)

>> Don't forget to include the names of all the group members in the pdf.

4. Final Report, source code, presentation slides, presentation video and viva

50% of the project grade (together with the presentation slides, video and the viva)

Final report should be 8 pages including citations, **IEEE Conference Format (also available on overleaf)**t + CODE

A. Report Structure: Your report should have the form of a paper with (at least) the following sections:

- Section 0. Abstract
- Section 1. Introduction
- Section 2. Data
- Section 3. Proposed Method
- Section 4. Experiments
- Section 5. Related Work
- Section 6. Conclusions (include what you learned)
- Section 7. Division of work (your grade will depend on your contribution to the project)

B. Code: Organize your code in a folder called "CODE". Include a README file
>> Submit a zip file of the CODE/ folder. >> Don't forget to include the names of all the group members in the README file, along with their contribution to the source code.

C. Presentation Slides and Video

- The structure can be similar to the final report
- Time limited - 10 minutes
- Submission video - upload a video (.mp4) to the given moodle link
- Submission of slides - upload a .ppt / .pptx / .pdf
- **There will be a viva (Q & A) session where each team member will be asked some questions about the project**