Fall 2023 CS 484 Data Mining
Final Project

Project Proposal Due Date: September 24, 2023, 11:59 pm
Project Progress Report Due Date: October 22, 2023, 11:59 pm
Final Project Due Date: November 26, 2023, 11:59 pm

For the project you need to choose a topic which is related to the topics covered in the class, explore it in great detail. It involves additional reading and implementation of the chosen methods and demonstrate it in action.

The project can be done in groups of **three to five people**. Project format is

- **Kaggle competition**: Find a competition on Kaggle and implement a data mining system to enter in it. Here are some options of data mining competitions [1].

Project Deliverables (submissions on GMU Blackboard Portal)

- **Proposal**: One page project proposal should be uploaded to GMU Blackboard Portal in PDF format by one group member. Here is a template [2]. You will receive feedback on your proposal but not a formal grade. However, failure to turn in the proposal on time will result in a penalty on the overall project grade. The due date is September 24.

- **Project Progress Report:** A summary of your current efforts, with notes on any modifications to your original project goals. At the very least, you should show evidence of successfully running some basic methods (e.g., training an off-the-shelf model) on your target data. **The target length is a three-pages PDF document**. As with the proposal, you will receive feedback, but not a formal grade, and failure to turn in the update on time will result in a penalty on the overall project grade. The due date is October 22.

- **Final Project Report**: The final report with detailed implementation (data processing, model training) and analysis. The due date is November 26.

[1] `https://www.kaggle.com/competitions?hostSegmentIdFilter=5`
[2] `https://docs.google.com/document/d/1VfLnP8xxGjO4w3UXzUZK1Dj2wZOW5aAiqxPLDyseR-8/edit?usp=sharing`

Format for the Report

The final report should be submitted in PDF format by one designated group member on GMU Blackboard Portal. It should be (the equivalent of) **at least six pages** (single-spaced, 11 point font, 1 inch margins, excluding Cover page, Statement of individual contribution and References) and mimic the style of a research paper. Here is the outline to follow for the report:

- **Cover page**: executive summary: List title and authors. Briefly summarize your problem, line of attack, and most interesting/surprising findings. Be sure to include at least one diagram or example result figure. This is not counted in the six-page minimum.
- **Introduction**: Define and motivate the problem, discuss background material or related work, and briefly summarize your approach.
- **Details of the approach**: Include any formulas, pseudocode, diagrams – anything that is necessary to clearly explain your system and what you have done. If possible, illustrate the intermediate stages of your approach with results images.
- **Results**: Clearly describe your experimental protocols and identify any external code and datasets used. Present your quantitative evaluation (if any) and show some example outputs.
- **Discussion and conclusions**: Summarize the main insights drawn from your analysis and experiments. You can get a good project grade with mostly negative results, as long as you show evidence of extensive exploration, thoughtfully analyze the causes of your negative results, and discuss potential solutions.
- **Statement of individual contribution**: Required if there is more than one group member. This is not counted in the six-page minimum.
- **References**: including URLs for any external code or data used. This is not counted in the six-page minimum.

Grading

You have to submit your source code (with a README file to describe the code), but grading will be be based primarily on the quality of the report (strength of idea, clarity, thoroughness, extent of analysis, etc.). **More will be expected of larger groups**. You can still get a good grade if your ideas do not work out, as long as your report shows evidence of extensive analysis and exploration, and provides thoughtful explanations of the observed outcomes.