CS 470 Final Project

Due by Monday, December 8, 2025 at 11:59 PM

Submission instructions

- For this project, you will work in a team composed of 3-4 students. You can freely choose your teammates within this semester's CS 470 class, even if you worked with some of those classmates in a previous homework assignment.
- Submit your assignment through the Canvas system. Please use the Group function in Canvas when you submit.

  Upload a single ZIP archive file named **project.zip**, containing all the files of your project:

  1. The program's source files. Include your name(s) in a comment at the top of all source files. If you are working using an online notebook system (such as Jupyter), provide a link to access the code.
  2. A `README.txt` file explaining how to compile and run your program, a web link to download the input dataset, and a web link to watch your presentation video. Do not include the dataset, as it is likely quite large.
  3. The presentation slides in PDF format.
  4. The report in PDF format.
  5. The LaTeX source files used to typeset the report.
- At the top of your report, include the names and IDs of all the members of your team, and a section named "Collaboration statement" in which you acknowledge any collaboration, help, or resource you used or consulted to complete this project.
- All the members of the team are expected to contribute significantly to the work. Later you will be asked to submit a confidential survey in which you rate each team member's contribution to the work. If a student is detected as a "free rider", they will receive a significant reduction in grade.
- No email submissions are accepted. No late submissions are accepted.

1 Project selection

The goal of this project is to give you an opportunity to apply your data mining skills on a problem of your choice. The websites: `https://www.kaggle.com` and `http://snap.stanford.edu/data` contain large collections of data sets. You can also look for other websites containing data sets. Define a problem that you want to solve, then choose a data set, and download it. The data set should be downloadable, and its processing should be feasible using a regular laptop or personal computer. Do not choose data sets that are too large, or tasks that are so computationally intensive that require a supercomputer or cloud computing service to be run.

2 Teams

Teams should be composed of 3 to 4 students. All members of the team are expected to give a significant contribution to the project. At the end of the semester, each member of the team will fill out a confidential "Contribution to the Team" form to rate all team members' contributions.

3 Progress checkpoint (20 points)

Each team should meet with their TA at least once to discuss their progress, sometime halfway through the project (no later than Friday, December 5, but sooner is better). Please plan for a 10-15 minutes meeting, either in person or on Zoom. All team members should participate. Of course feel free to meet with your TA more often if necessary.

4 Presentation in class (10 points)

You will present your project in class on Monday, December 8. The time available will depend on the number of teams, but plan for approximately 5-7 minutes presentation. All members of the team should actively participate in the presentation.

5 Video presentation (10 points)

- **(5 points)** Record a short 5-minute video (minimum: 4:30, maximum: 5:30) in which you present your project, and upload it to a video-sharing service of your choice.
- **(5 points)** Prepare a set of slides to support your presentation.

6 Final report (60 points)

The final report should be a PDF file typeset with LaTeX. It should contain the following sections:

1. **Collaboration statement.** Acknowledge any collaboration, help, or resource you used or consulted to complete this project.
2. **Problem description (5 points).** Describe the problem that you are trying to solve in this project, and the goals you want to achieve.
3. **Data description (5 points).** Describe the dataset that you are using in this project. Provide a link to download the data.

4. **Data pre-processing (15 points).** Describe your work on visualizing and pre-processing the dataset.

5. **Data mining methods (15 points).** Describe the data mining process and algorithms that you used to solve your problem.

6. **Results (15 points).** Describe and discuss your results. Also explain the limitations of your approach.

7. **Future work (5 points).** Discuss a few ideas that could be tried for future work on this problem.

Grading criteria

- 20 points for the progress checkpoints. The team should demonstrate significant progress towards the completion of the project.
- 10 points for a competent, engaging, and clear presentation in class.
- 5 points for a competent, engaging, and clear video presentation.
- 5 points for informative, clear, and well organized video slides.
- 60 points for a complete, clear, and well organized report. Breakdown by section:
  - Problem description: 5 points.
  - Data description: 5 points.
  - Data pre-processing: 15 points.
  - Data mining methods: 15 points.
  - Results: 15 points.
  - Future work: 5 points.
- -20 points if the report is not typeset using LaTeX, or if the LaTeX source code is not provided.
- -10 points for insufficient comments in the code.
- -10 points for each deviation from the submission instructions.
- -10 points for missing collaboration statement.
- A variable penalty (up to -100 points) may be applied to individual team members for insufficient contribution to their team.