MATH 4420: Machine Learning and Data Mining 2 **(Due Dates: Nov 7/Dec 4, 2025)**

Project Guidelines

Professor: Eric Gerber

Project Description

You will work with a partner (or, in rare cases, a group of three) on a final project that involves identifying one or two real data problems which you will attempt to address with **TWO** of the methods learned in this class. You and your partner will:

- Develop some general problem(s) of interest you believe machine learning may be helpful in addressing. If you select more than one problem, there should be some logical link or narrative thread connecting them.
- Perform a high-level literature review of past research done in addressing your problem(s) or similar problem(s).
- Identify existing data which aligns with your problem(s) of interest and collect/curate/prepare it for analysis.
- Choose two machine learning methods under the broad umbrellas of the topics we learn in this course:
  - Neural Networks (MLP and/or CNN)
  - Collaborative Filtering
  - Time Series
  - Bayesian Modeling
- Apply these methods in an attempt at addressing your problem(s). You **must** use Python to apply one method and R to apply the other.
  - You must apply **at least one** of the methods manually (without the use of pre-built modeling packages).
    - "Pre-built modeling" refers to packages such as Scikit-learn, statsmodels, pyMC, rstan, etc. Obviously, packages such as NumPy can (and should) be used, as they allow you to perform the modeling yourself.
- Write a short report and create a virtual poster for disseminating the results of your analysis.

The ultimate deliverables for this project, and their weights in your course grade are:

- Literature Review (due initially as part of Phase I, but will be fully graded as part of the Final Report: 1%)
- Project Report (11%)
  - Includes refined literature review
- Virtual Poster (8%)
- **Extra Credit (2%):** For 2% extra credit on your full course grade, create an application with at least two tabs/pages:
  1. A landing page with a description of your project
  2. A page with an interactive visualization (table or plot) of one of your method's results
  3. The application must be **deployed** to receive full credit; un-deployed applications will receive half credit
  - If you have not built an app before, there are several tutorials on using Streamlit (for python) or Shiny (for R or python) to create basic applications based on .py or .R scripts.

Phase I: Literature Review + GitHub Repo **(Due: Nov 7, 2025)**

As a first phase of the project, you and your partner(s) must provide a short (max 2 page) literature review, as well adding Dr. Gerber to the GitHub repository for your project.

The literature review:

- Should provide an overview of past work (including citations in Chicago or IEEE style) that discusses either:
  - how the previous literature addresses the same problem(s) with different methods, **or**
  - how the previous literature uses the same methods to address similar (but not exactly the same) problem(s)
- Should include a bibliography (again, in either Chicago or IEEE style)
- Should aim for at least 2 references, though ideally between 3-6
- Will be refined and included as part of the introduction of the final report (see next page)

The GitHub repo:

- Should have an informative .README but can be either public or private
- Should host minimally sufficient code to replicate all parts of the final project by the final due date
- Will be used by Dr. Gerber to track teammate contribution to the programming aspects of the project
  - You can add a link to your GitHub repo in the literature review document, but it is your responsibility to invite Dr. Gerber as a collaborator
  - You must add Dr. Gerber (`https://github.com/eaegerber`) as a collaborator by the Phase I due date or your grade will be penalized

Project Report **(Due: Dec 4, 2025)**

You and your partner(s) will write a report in the style of a professional research paper. The typed report should be professional in tone and appearance and consist of the **FIVE** sections outlined below. Any additional information should easily found in a GitHub repository. The report will be due via Gradescope the evening of the last day of class (**December 4 at 11:59 pm**).

The outline and maximum page count for each part of the report:

I. Introduction and Literature Review (2 pages)
   - Recognition and statement of the problem(s)
   - A refined literature review

II. ML Methodology and Data Collection (2 pages)
   - Discuss how the data were chosen, collected, and curated to match the problem(s) of interest
   - Provide a brief overview of the two ML methods used and outline any adjustments made outside of the course content
     * E.g.: if you use a hybrid collaborative filtering approach (for example which combines content-based and user-user), discuss how this works mathematically

III. ML Results and Interpretation (3 pages)
   - Provide the **major** results of each of your methods, reporting things like predictive test error/accuracy, interpretation of coefficients, etc.
   - Include no more than 4 plots/tables (you will want to critically assess which are most important for establishing your results), though these will not count towards your page limit

IV. Conclusions and Future Work (1 page)
   - Summarize the findings of your study
   - Be sure to discuss if all assumptions of your given methods have been met and if the analysis of the data was appropriate
   - What drawbacks are there from analyzing data this way? What would you do differently if you were to continue working on the problem?

V. Appendix: Implementation (Does not count towards page count)
   - Provide at the end of your report (in a labeled appendix) the python and R code for your implementations of the methods used in the project
   - Only include the training implementation; no need to provide any data collection, preprocessing, testing or results work (those should be available in the GitHub repo)
   - Recall that one of your methods **must** be implemented manually (without the use of pre-built packages except for standard linear algebra libraries, e.g. NumPy is of course okay, but Scikit-learn is not)

You will need to be concise to keep your report within the page limit. If you are having trouble cutting things down, please feel free to discuss with Dr. Gerber and/or the TAs.

There will be a mandatory project check-in the week of **November 17** where each group and all members must quickly meet (approx. 5 minutes) with a TA to confirm they are on track/address any issues.

Use of AI Tools

Use of more general generative AI tools, such as ChatGPT, Claude, Perplexity, etc. **is forbidden** on this assignment. However, since a specific use-case of generative AI may be helpful, especially in conducting the literature review section, you are **encouraged** to make use of Google's NotebookLM, which is trained on a more specific corpus (for research tasks) than the other more general use chatbots. You are encouraged to use NotebookLM to:

- Give notes on important topics from relevant articles to mention in your literature review.
- Compare your own work with existing work to ensure sufficient differences and the logic of your extension(s).
- Give you advice on whether you are missing anything important from your report.

As with any assignment, no generative AI should be used to write any portion of your final report. If there is any evidence that parts of your report are not your own, you will **fail the course** and be reported to **OSCCR**.

Virtual Poster

On the final day of class, **Thursday, December 4**, you and your partner(s) will present a virtual poster of your work on one of your computers. This poster will be due by the end of the day, along with the final report. **All team members should be present on the final day.** For the first 30 minutes of class, one team member will be in charge of presenting their poster to their classmates, while the other moves around encountering other projects; the team members will switch roles for the second half of the class.

The poster should:

- Provide brief text sections introducing the problem(s) and the data (including at least a link to the data set, or a snippet of the data), discussing the results, and conclusions and future work.
- Provide several visual (tables, plots, etc.) representations of the methods and results of the analysis.

It can be either .pptx or .pdf format and will be graded both on content as well as design. Posters should not be too dense nor too sparse. Dr. Gerber will to provide some examples of poster styles in the coming weeks.

Project Team Choice/Evaluation Form

You may work with any student, from any section, on this project. **HOWEVER:**

- Only pairs may be made up of students from across two sections; groups of 3 (however few there are) may only be made up of students from the same section
- If you work across sections, you must both commit to being present for the **ENTIRE** virtual poster session during the section Dr. Gerber assigns to you
- Any student who is not present for the poster session will not receive credit for their poster

All students must fill out this Project Team Choice Form by **Monday, October 27** so that (a) Dr. Gerber knows who you are working with and so that (b) anyone who does not know who they will work with can be assigned a partner.

ANYONE WHO DOES NOT FILL OUT THE FORM WILL FAIL THE PROJECT, SINCE THEY WILL OTHERWISE NOT BE ASSIGNED TO A TEAM.

All students will fill out a teammate evaluation form at the end of the semester, establishing whether there was equal contribution of work. Deviations from equal contribution will result in individual grade adjustments.

Project Schedule (Fall 2025)

Project Assigned (Monday, October 20)

Team Choice Form Due (Monday, October 27)

Phase I due (Friday, November 7)

Project Check-In (Week of November 17)

Virtual Poster Session (In-Class Thursday, December 4)

Virtual Poster Due (11:59 pm Thursday, December 4)

Final Report Due (11:59 pm Thursday, December 4)