MATH 4420: Machine Learning and Data Mining 2 **(Due: April 13/14, 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.
- 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 packages).
- 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:

- Project Report (10%)
- Virtual Poster (8%)
- **Extra Credit (3%):** For 3% 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
  - 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.

Project Report

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 **FOUR** sections outlined below. Additional information should easily found in a GitHub repository. The report will be due via Gradescope the evening before the last day of class (**April 13 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)
   - An overview of past work that either (a) addresses the same problem(s) with different methods or (b) uses the same methods to address similar (but not exactly the same) problem(s).

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?

You will need to be concise to keep your report within the page limit. All code and data used should be stored in a GitHub repository shared between both/all project members and Dr. Gerber. You are encouraged to maintain this repository throughout the project and share it with Dr. Gerber early, especially for version control purposes and for keeping track of each project member's commit history.

Use of AI Tools

Use of ChatGPT **is forbidden** on this assignment. However, since 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 ChatGPT. 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 to the existing work to ensure sufficient differences and the logic of your extension(s).
- Give you advice on if 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, **Monday, April 14**, 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 (to allow you a few hours to make final tweaks to it based on feedback in class). **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 a **very** brief text section describing the problem(s) and a **very** brief text section discussing the results.
- 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 attempt to provide some examples of poster styles in the coming weeks.

Project Team Choice

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

- you must both commit to being present for the **ENTIRE** virtual poster session during **Section 3 from 10:30 am - 11:35 am on April 14th**
- any student who is not present for the poster session will not receive full credit for their poster
- only pairs may be made up of students from two sections; groups of 3 (however few there are) may only be made up of students from the same section

All students must fill out this Project Team Choice Form by **Monday, March 3** 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.

Project Schedule (Spring 2025)

Project Assigned (Thursday, February 27)

Team Choice Form Due (Monday, March 3)

Project Check-In (Wednesday, March 19)

- Project topic(s) should be finalized and data source(s) at least identified by the date above.

Final Report Due (11:59 pm Sunday, April 13)

Virtual Poster Session (In-Class Monday, April 14)

Virtual Poster Due (11:59 pm Monday, April 14)