

Sta-370 Intro to Machine Learning - Final Project (Fall '26)

Overview: The overarching aim of the final project is for you to engage in a capstone experience that demonstrates your ability to thoughtfully apply modern machine learning methods to novel data of your choosing.

In order to demonstrate this learning aim, you may select one (or more) of the following project pathways:

1. A comparison of deep learning and feature engineering approaches on a complex data structure, such as images/videos, textual, or time-series data.
2. A novel application of deep learning that goes beyond what we've covered in class. For example, image segmentation, image captioning, text generation, etc.
3. An in-depth machine learning analysis of tabular data, including creative/thoughtful pre-processing, model selection, and feature importance steps.

You are free to pursue any topic that allows you to demonstrate your knowledge of machine learning methods within the basic constraints described above. This might include:

1. Attempting a Kaggle competition.
2. Reanalyzing data from a scientific paper or past project.
3. Collecting your own data.
4. Working on any other application you're interested in and can find appropriate data for.

Your project will be evaluated on the following components:

1. **Code Repository** - you are to provide an organized and appropriately documented set of Python files that recreate your main results
2. **Presentation** - a 10-minute presentation providing a high-level overview of your research question, data, and main results
3. **Report** - a 3-5 page report (not including figures or graphics) written in a formal scientific tone providing a detailed overview of your methods and results

Additionally, you will receive a small amount of completion credit for the following:

1. Submitting a topic proposal containing an appropriate level of detail by the stated deadline (Friday 11/6)
2. Sharing a short, informal progress update (1-2 minutes) to the class (Thursday 12/3)

The expectation is that you work collaboratively with one other person (2-person team), or that you work individually.

Timeline:

- Tuesday 11/3 - One member of your intended project team provides the names of both team members. If you do not indicate belonging to a team, it will be assumed that you'll be working alone.
- Friday 11/6 - One member of your team must submit a short project proposal via email. This must include your research question, your data source, and a paragraph explaining your planned methods.
- Thursday 12/3 - You'll share a 1-2 progress update where you share your project topic and current status with the class. The purpose of this step is for your team to solicit ideas from your peers.
- Tuesday 12/8 *and/or* Thursday 12/10 - Formal presentations during class
- Friday 12/18 - Final reports are due by 5:00 PM

Presentation Details:

During class on either Tuesday 12/8 or Thursday 12/10 you will give a roughly 10-minute presentation. Your presentation should summarize the key components of your group's work, including the project's goals, data, methods, and results. You should treat the presentation as if you were at a scientific conference or workshop, meaning you may assume some familiarity with basic machine learning methods and concepts, but you should thoroughly explain any non-standard or more advanced methods you used that were not part of our course content. Your presentation will be scored on the following criteria:

1. *Comprehensiveness* - Did you communicate all of the necessary information that someone would need to understand your work? Were any important details not adequately explained? Could an audience member provide an accurate overview of everything you did?
2. *Coherence* - Was the presentation organized in a logical order? Did it flow smoothly? Was it easy for the audience to follow along?
3. *Professionalism* - Were your presentation materials appropriately prepared? Did you adhere to the stated 10-minute time limit and speak using appropriate volume and pacing? Did each group member contribute to the presentation in a manner that didn't interrupt its flow?

Report Details:

Final reports, code, and data are to be submitted via P-web no later than 5pm on Friday at the end of finals week. Your paper should be 3-5 pages in length (not including figures, tables, references, and any supplemental material). You are encouraged to embed a link to a repository containing your code and data (rather than submitting them directly). Your report will be scored according to the following criteria:

1. *Structure* - Does your report follow a proper scientific structure (Intro, Methods, Results, Discussion, References)? Does the report include appropriate figures and tables that help communicate your main results? Does information appear in the proper section? (ie: results are not given in the Methods section, etc.)
2. *Technical Correctness* - Are your methods and results described correctly? Or there any errors, misleading statements, or important information that is omitted?
3. *Level of Detail* - Are your methods described in sufficient detail? Are your results thorough and comprehensive?
4. *Coherence* - Can a reader of your paper fully understand your work? How easily could a reader summarize the main steps of your project? Are there any gaps, awkward transitions, or omitted sections?
5. *Professionalism* - Are your visualizations and tables professional in appearance? Do you cite scientific sources when appropriate? Is the report written in a scientific tone and free of grammatical/spelling errors?

For additional guidance on scientific writing, I recommend this article, which provides some general advice on how to structure and prepare scientific writing.

Deep Learning Models:

For many tasks it is more productive to fine-tune an existing deep learning model than to build one yourself from scratch. A few places where you could identify a suitable model for your application are:

1. Hugging Face - <https://huggingface.co/models>
2. Kaggle - <https://www.kaggle.com/models>
3. PyTorch - <https://pytorch.org/hub/>
4. TSAI (for time-series data) - <https://timeseriesai.github.io/tsai/>