

Machine Learning Mini-Project

Open-Ended Project Instructions

Objective. Apply machine learning to a problem of your choice, from any field. You may use any type of input data and any suitable learning approach. The emphasis is on defining a meaningful problem, building and evaluating a working model, and documenting your work so that someone else can reproduce it.

What to do

1. **Choose a problem and dataset.** Use a publicly available dataset or collect and label your own data. State the problem, the available *inputs*, and the desired *outputs* (or other learning objective, where applicable). Briefly explain why the task matters to your field or application.
2. **Develop and train a model.** Prepare the data, choose an appropriate machine learning model, and train it. You may use existing libraries and pretrained models, provided you clearly describe what you used and what you trained or adapted.
3. **Test, validate, and demonstrate the result.** Evaluate the model on data not used to fit it, using suitable metrics or qualitative evidence. Discuss its performance and limitations. Provide a simple, runnable way to use the trained model on example or new inputs. A web application is *not* required; a script or notebook is sufficient.

What to submit

1. **GitHub repository (share the link).** The repository must contain the code and a clear `README.md` so that the instructor can reproduce the work. At a minimum, include:

- A short description of the problem, dataset, inputs and outputs, and model.
- Instructions to install dependencies and obtain the data (including your labels, if you created them). Do not upload restricted or sensitive data; provide an accessible alternative if necessary.
- Exact commands or notebook steps to **train/retrain** the model and to **run inference or testing** on supplied example inputs.
- The evaluation procedure, main metrics or results, and any required model weights or instructions to produce them.
- Reproducibility details such as the software environment, data split, and random seed where relevant.

Reproducibility check: A reader should be able to follow the README, access the required data, train the model, and run a test or prediction without needing undocumented steps.

2. **Short report (maximum 2 pages, LaTeX).** Use a simple structure:

- **Problem and motivation:** What did you try to solve, and why is it useful?
- **Approach:** What data, model, and evaluation method did you use?
- **Findings:** What worked, what did not, and what evidence supports the conclusions?
- **Reflection:** What did you learn, what are the limitations, and how could this contribute to your field?

Figures and tables may be included within the two-page limit. Cite the dataset and any external code or models used.

Scope and expectations

Keep the project manageable: a clearly defined problem and a reproducible, honestly evaluated baseline are more valuable than an ambitious system that cannot be rerun. You are free to choose the domain, data modality, and model; there is no requirement to develop a novel algorithm or deploy a public website. Clearly acknowledge external resources and any use of generative AI in your repository or report.