CP322 Machine Learning - Project
Report Due Date: Dec 3, 2025 at 11:59 PM

The objective of this course project is to give you hands-on experience applying machine learning techniques to a meaningful real-world problem. Rather than simply following textbook examples, you'll have the chance to explore a challenge from start to finish—making your own design choices, justifying your approach, and learning how to deal with the kinds of messy, open-ended questions that come up in practical ML work.

Start by forming a collaborative team of 4–5 classmates, and together, identify a real-world problem that genuinely interests your group. This could come from a wide range of domains—whether it's analyzing climate trends, predicting healthcare outcomes, enhancing educational tools, or exploring trends in sports or entertainment. The key is to choose a topic that feels meaningful and engaging to your team. Once your problem is selected, your goal is to work through the complete machine-learning pipeline from start to finish:

1. Understand the challenge and gather the most relevant data.
2. Prepare and explore the data so it's ready for modelling.
3. Design and test creative ML approaches—not just plug-and-play code, but thoughtful tweaks or hybrid ideas you invent.
4. Evaluate performance and draw insights that explain what worked, what didn't, and why.

Everyone can lean into a favourite role—data wrangling, model design, experiment orchestration, or results storytelling—while making sure the final solution shows both breadth (multiple data angles or techniques) and depth (rigorous experimentation and analysis). At the end you will turn in

a) **Report** (IEEE style, $\leq$7 pages),

b) **15 mins** presentation + demo + QA live,

c) slides,

d) **Reproducible code** + README.

Candidate Topics

1. **Transfer Learning on Image Data** Use *two* small image datasets (e.g. CIFAR-10 + EuroSAT). Compare *ResNet-50* vs. *ViT-tiny* fine-tunes; run an ablation on data-augmentation. Include Grad-CAM or LIME visualisations for interpretability.
2. **Sentiment Analysis with BERT Variants** English IMDb + a non-English dataset (e.g. Spanish TASS). Baseline TF-IDF + LogReg, fine-tuned DistilBERT, *and* a multilingual MiniLM. Extra error analysis on minority classes.
3. **Regression on Tabular Data (House Prices & Energy Usage)** *W*ork with *two* regression datasets. Use baselines including XGBoost, neural tabular model (TabNet or FT-Transformer), and hyper-parameter search (Optuna). Provide SHAP plots and discuss feature importance overlap across datasets.
4. **Intro RL – DQN for CartPole & Lunar Lander** Train on *two* Gym tasks. Explore baseline vanilla DQN, and compare to Double-DQN or PPO. Study sensitivity to reward-shaping and include learning-curve variance over 5 seeds.

Timeline (4 Weeks)

Start early! Once teams are formed you should begin work *immediately.* Plan on devoting a **minimum of four continuous weeks** to this project so each stage— from data exploration through model evaluation and demo polish—receives adequate attention. The table below is a *suggested* schedule to keep your group on track; feel free to adjust deadlines internally as long as you meet the final submission date.

Day 1	Kick-off, team role assignment, dataset selection
Day 7	*Mini-proposal*: datasets, baselines, planned "extra depth" element
Day 14	Mid-sprint checkpoint: preliminary results table + demo preview
Day 21	Code and demo due; live presentation and Q&A in class
Day 28	Final report due.

Assessment

It's important to remember that this is a team project, and your success depends on how well you collaborate. Everyone on the team will receive the same final grade, so it's in your best interest to support one another, divide tasks fairly, and communicate openly. Be respectful, check in regularly, and make sure each member has a meaningful role.

That said, active participation is essential. If a student does not contribute meaningfully to the project and cannot provide evidence of their involvement, they may receive a grade of zero for the project, regardless of the team's performance. If you run into any challenges working together, don't wait—reach out to the instructor early so we can help. Here's how your project will be evaluated:

Technical quality & correctness	30 %
Breadth + depth criteria satisfied	25 %
Experimental analysis (baselines, ablation)	20 %
Demo usability & insight	15 %
Report clarity & organisation	10 %

Hints for Success

- Parallelise: one subgroup handles data/EDA, another builds baselines, another polishes demo.
- Use small models first; switch to bigger ones only if time permits.
- Track experiments in a shared spreadsheet or Weights & Biases board.
- Keep the demo simple.

Suggested Structure of the Report

Similar to a research paper, our project report should contain the following components:

1. Introduction
    - What is the problem?
    - Why can't any of the existing techniques effectively tackle this problem?
    - What is the intuition behind the technique that you have developed?
2. Techniques to tackle the problem
    - Brief review of previous work concerning this problem
    - Elaboration of the technique you developed
    - Description of the existing techniques that you will compare to
3. Evaluation

- Analyze and compare (empirically or theoretically) your new approach to existing approaches

4. Conclusion
    - Can your new technique effectively tackle the problem?
    - What future research do you recommend?

Useful Resources

- Kaggle
- UCI Machine Learning Repository
- Latex IEEE Conference Template; Microsoft Word Template
- Course materials and lectures

Academic Integrity

Your work must be original. Plagiarism is a serious violation of academic integrity, whether intentional or unintentional. While Kaggle and other platforms may have code available for research purposes, it is crucial that you do not refer to or copy these resources while working on your project. Start working on the project as early as possible to ensure you have ample time to develop and refine your own solutions.