CMSE 492 Homework: Project Setup and Proposal

Due Date: Nov 2, 2025

Total points: 49 points

Overview

This assignment serves multiple critical purposes for your final project:

1. **Technical Infrastructure:** Set up your GitHub repository and development environment
2. **Project Planning:** Develop a comprehensive, professional project proposal
3. **LaTeX Proficiency:** Practice using LaTeX/Overleaf for technical writing
4. **Peer Review Preparation:** Create a proposal that will be reviewed by your peers

Your proposal will follow a professional format and must be written using the provided LaTeX template. This document will be distributed to your peers for review in next week's class, so clarity, completeness, and technical feasibility are essential.

Important: This assignment requires using the provided LaTeX template. Non-compliant submissions will receive a zero.

Use of AI Tools

You are **encouraged to use AI tools** (such as ChatGPT, Claude, Copilot, etc.) to help you complete this homework and write your report. AI can be valuable for:

- Brainstorming ideas and refining your problem statement
- Debugging code and improving your implementations
- Generating LaTeX syntax and formatting
- Improving clarity and grammar in your writing
- Creating initial drafts of sections

However, your final report must meet professional standards:

- **No bullet points** in the narrative sections (except where explicitly required, like in lists of requirements)
- **No emojis** or informal language
- **Professional, technical tone** appropriate for a scientific report
- **Coherent narrative flow:** Your report should read as a unified document, not as disconnected sections. Each section should connect logically to the next, building a complete story of your project from problem statement through proposed solution.
- **Integrated transitions:** Use transitional sentences to connect sections. For example, after describing your data, naturally lead into how this informs your preprocessing choices.

Think of your proposal as a paper that you will present to stakeholders or at conference, not a homework assignment. Each section contributes to a cohesive argument for why your project is important, feasible, and well-planned. The narrative should guide the reader through your thinking process.

Remember: AI is a tool to enhance your work, not replace your critical thinking. You are responsible for the accuracy, coherence, and quality of your final submission. Read the Learn with AI document, I put in D2L.

Learning Objectives

By completing this assignment, you will:

- Establish version control and collaboration practices for your project
- Articulate your project's goals, methodology, and evaluation strategy
- Develop proficiency in LaTeX for technical and scientific writing
- Create a professional-quality proposal suitable for peer review
- Conduct preliminary data analysis and baseline experiments

Part A: GitHub Repository Setup (9 points)

A1. Repository Creation and Structure (6 points)

Create a new GitHub repository named `cmse492_project` with the following structure:

```
cmse492_project/
├── README.md
├── .gitignore
├── data/
│   ├── raw/
│   └── processed/
├── notebooks/
│   └── exploratory/
├── src/
│   ├── preprocessing/
│   ├── models/
│   └── evaluation/
├── figures/
├── docs/
└── requirements.txt
```

Requirements:

- Initialize with a README.md file (2 points)
- Create an appropriate .gitignore file for Python/ML projects (2 points)
  - Should ignore: `__pycache__/`, `*.pyc`, `.ipynb_checkpoints/`, `data/raw/` (if large), `.env`, etc.
- Set up the directory structure shown above (2 points)

A2. Repository Configuration

If your repository is private, you must add collaborators so your instructor and TA can access your work:

- Add your instructor's GitHub username: `lucianogsilvestri`

- Add your TA's GitHub username: `maryambrj`

Go to: Repository → Settings → Collaborators → Add people

Note: We recommend making your repository public to build your professional portfolio, but private repositories are acceptable if you have concerns about sharing your work publicly. If your repository is public, you do not need to add collaborators.

A3. README Documentation (3 points)

Your README.md must include:

- Project title and one-paragraph description (1 point)
- Directory structure explanation. Note that the directory structure can vary from that provided above, however, it must be explained. (1 point)
- Setup instructions (dependencies, environment setup) (1 point)

Note: You'll expand this README as your project progresses.

Part B: Preliminary Analysis and Baseline (8 points)

Complete the following in a Jupyter notebook in the `notebooks/exploratory/` directory:

B1. Data Acquisition and Loading (2 points)

- Document where your data comes from
- Load and display basic dataset information (shape, dtypes, etc.)
- Save a sample of your data to `data/processed/` (if applicable)

B2. Initial Exploratory Data Analysis (4 points)

Create at least **4 meaningful visualizations** that reveal:

- Distribution of key features
- Relationships between variables
- Class balance (for classification) or target distribution (for regression)
- Missing values or data quality issues

All figures should be saved to the `figures/` directory with descriptive names.

B3. Baseline Model (2 points)

Implement a simple baseline model:

- For **classification:** Majority class classifier or simple logistic regression
- For **regression:** Mean predictor or simple linear regression
- For **RL:** Random policy or simple heuristic
- Report baseline performance using appropriate metrics

Deliverable: Commit your notebook to GitHub before starting the proposal.

Part C: Project Proposal in LaTeX (22 points)

Write a comprehensive project proposal using the **CMSE_492_Project_Template.tex** provided. Your proposal should be at least **2 pages** (excluding references).

Required Sections and Point Distribution:

1. Title and Author Information (1 point)

- Descriptive project title
- Your name, email, and date
- GitHub repository link

2. Abstract (2 points)

Write a 150-250 word abstract summarizing:

- The problem you're addressing
- Your proposed ML approach
- Expected contributions
- Preliminary findings (if any)

3. Background and Motivation (4 points)

- Clear problem statement
- Importance and real-world impact
- Existing approaches and their limitations
- How ML can address this problem
- Your project's potential contribution

4. Data Description (4 points)

- Dataset source and provenance (not just "Kaggle"). Who/What collected this data, Why was the data collected, and How was the data collected.
- Size, features, and data types
- Data quality assessment (missing values, outliers, balance)
- **Include 2 figures** from your EDA showing key dataset properties
- Planned preprocessing steps

5. Proposed Methodology (5 points)

- Detailed description of your ML approach
- Justification for model selection
- If using RL: specify states, actions, rewards, environment
- If using supervised/unsupervised: justify algorithm choices
- **Include:** Architecture diagram or methodological flowchart
- Comparison of at least 3 models (increasing complexity)
- Define the complexity of each model

6. Evaluation Framework (4 points)

- Selected metrics with justification
- Train/validation/test split strategy
- Baseline comparisons
- Success criteria

7. Timeline and Milestones (2 points)

Create a **Gantt chart** showing your project timeline from now through the final project deadline (Monday, December 8, 2025).

Your Gantt chart should include:

- Weekly breakdown of major tasks and milestones
- Parallel activities (what can be done simultaneously)
- Dependencies between tasks
- Key deliverables and checkpoints
- Buffer time for unexpected challenges
- Final presentation preparation (Week 15: Dec 2-4)
- Final report and code submission (Dec 8)

Insert the Gantt chart as a figure in this section and provide a brief narrative (1-2 paragraphs) explaining:

- Your critical path and key milestones
- How you've allocated time for data collection, model development, experimentation, and writing
- Contingency plans for potential obstacles
- Alignment with course deadlines and commitments

Tip: Ask AI to create a Gantt chart for you. Even better ask AI to write code to produce the Gantt chart. Make sure that the text in the image is readable, color coordinated.

Organization Tip: Save the LaTeX file as CMSE492_ProjectProposal.tex to differentiate it from the final project report. Even better you can create a folder in your git repo containing the reports, name this folder `reports`.

Submission Guidelines

What to Submit:

1. **GitHub Repository** (Part A)
    - Ensure your instructor and TA have been added as collaborators
    - All code, notebooks, and preliminary results should be committed
    - README.md should be complete and professional
2. **PDF of your report** (Part C)
    - Compiled PDF to D2L
    - The PDF should include the seven sections described above

Peer Review Information

Next Week: Each student will review 2-3 peer proposals using a structured rubric. Your proposal should be:

- **Clear:** Understandable to your classmates
- **Complete:** All required sections with sufficient detail
- **Professional:** Publication-quality LaTeX formatting
- **Feasible:** Realistic scope for the semester timeline

The quality of your proposal directly impacts the usefulness of the peer feedback you'll receive.

Grading Rubric (49 points total)

Part A: GitHub Setup (9 points)

- Repository structure and organization: 6 points
- README documentation: 3 points

Part B: Preliminary Analysis (8 points)

- Data loading and documentation: 2 points
- Exploratory data analysis (4 visualizations): 4 points
- Baseline model implementation: 2 points

Part C: LaTeX Proposal (22 points)

Criteria	Excellent (4-5)	Good (3)	Fair (2)	Poor (1)	Score
Problem Definition & Motivation (4 pts)	Crystal clear problem statement with compelling motivation. Excellent literature context.	Clear problem with good motivation. Adequate context.	Somewhat vague problem. Limited motivation.	Unclear problem or weak motivation.	/4
Data Understanding (4 pts)	Comprehensive data analysis with insightful visualizations. Clear preprocessing plan.	Good data analysis with relevant visualizations. Basic preprocessing plan.	Limited data analysis. Few visualizations.	Minimal data understanding. No preprocessing plan.	/4

Criteria	Excellent (4-5)	Good (3)	Fair (2)	Poor (1)	Score
Methodology (5 pts)	Detailed, technically sound approach. Clear justification and architecture. Compares 3+ models.	Reasonable approach with justification. Compares 2-3 models.	Vague approach. Limited justification.	Technically flawed or unclear methodology.	/5
Evaluation & Timeline (4 pts)	Appropriate metrics with clear experimental design. Realistic, detailed timeline.	Reasonable metrics and design. Basic timeline.	Unclear metrics or evaluation. Vague timeline.	Inappropriate evaluation or unrealistic timeline.	/4
LaTeX Quality & Professionalism (3 pts)	Perfect template usage. Professional formatting. Clear figures/tables.	Good template usage. Minor formatting issues.	Some template deviations. Formatting needs work.	Poor template usage or unprofessional presentation.	/3

Deductions:

- -5 points: Not using the provided LaTeX template (unless pre-approved)
- -2 points: Missing GitHub repository link in proposal
- -2 points: Figures, tables, equations not properly referenced in text
- -1 point: Exceeds page limit (>6 pages of content)

File Naming Convention: (10 points)

- I am not kidding. Naming your file the correct way and a good commit message account for ~20% of the grade!
- PDF: `LastName_FirstName_CMSE492_ProjectProposal.pdf`
- Commit message for final GitHub push: "HW: Project Setup and Proposal - Complete"

Tips for Success

LaTeX/Overleaf Tips:

1. Start with the template early - don't wait until the last day to learn LaTeX
2. Compile frequently to catch errors early
3. Use the `\label{}` and `\ref{}` commands for figures, tables, and equations
4. Save your figures with descriptive names in a `figures/` folder in Overleaf
5. Ask for help!

GitHub Tips:

1. Commit early and often with descriptive messages

2. Don't commit large data files (>100MB) - use `.gitignore`
3. Test that your repository is accessible to collaborators
4. Keep your README updated as you make progress

Proposal Tips:

1. Be specific rather than vague - "I will use a neural network" is too general, what type of neural network?
2. Include preliminary results even if they're just baseline performance
3. Your peers will read this - make it clear and well-organized
4. Cite papers properly using BibTeX or the bibliography section in the template
5. Think about feasibility - can you actually complete this in the remaining weeks?
6. **Gantt chart:** Make it realistic! Account for other coursework, exams, and commitments. Include buffer time for debugging and unexpected issues.

Resources

Technical Writing:

- Geron's "Hands-On Machine Learning" Appendix A
- "How to Write a Good Paper": https://cs.stanford.edu/people/widom/paper-writing.html

---

Frequently Asked Questions

Q: What if I've never used LaTeX before?
A: That's expected! The provided template and instructions guide you through everything. Start early and use Overleaf's helpful error messages.

Q: Can I change the LaTeX template formatting?
A: No. You can only fill in content. Do not modify the document class or margins structure.

Q: What if my dataset is too large for GitHub?
A: Use `.gitignore` to exclude large files. Include a `data/README.md` explaining where to download the data instead.

Q: How detailed should my methodology be?
A: Detailed enough that a classmate could understand your approach. Include equations, architecture diagrams, and algorithmic descriptions.

Q: What if I don't have preliminary results yet?
A: At minimum, you should have EDA visualizations and a simple baseline. Even negative results are valuable to report.

Q: Can I use bullet points in my proposal?
A: Only where explicitly required (like listing requirements or itemizing features). The main narrative should be written in paragraph form with professional prose and smooth transitions between sections.

Q: How do I create a Gantt chart?
A: You can use Python (matplotlib-gantt, plotly) or even Excel/Google Sheets. Export it as a PNG or PDF and include it as a figure in your LaTeX document. The chart should span from now until December 8, 2025. A: Ask AI to create a Gantt chart for you. Even better ask AI to write code to produce the Gantt chart. Make sure that the text in the image is readable, color coordinated.

Q: Should my repository be public or private?
A: We recommend public to showcase your work professionally. If private, you must add the instructor and TA as collaborators.

Academic Integrity

- Your proposal must be your own work
- Properly cite all sources, datasets, and prior work
- You may discuss ideas with classmates, but write independently
- Code and visualizations must be your own (unless properly attributed)
- Plagiarism will result in zero points and academic misconduct reporting

---

Remember: This proposal is not just an assignment - it's the foundation for your entire semester project. Invest time in planning now to save yourself stress later. A well-thought-out proposal makes the rest of the project much smoother.

Good luck, and start early!