COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Machine Learning, I

CS 5805

Final Term Project- Proposal

The final term project for **CS 5805: Machine Learning I** consists of three phases:

1. Proposal
2. Formal final report
3. Presentation

We are now commencing the first phase of the project, which involves dataset selection and the preparation of a written proposal.

Dataset Selection Criteria

The selected dataset must meet the following requirements:

- Select an interesting, real-world dataset with clear applications in industry.
- The dataset must be multivariate and contain at least 50,000 observations. If you identify a particularly compelling dataset with fewer than 50,000 samples, please consult with me.
- The dataset must include both numerical and categorical variables, with at least two variables of each type. Feature engineering may be applied to generate additional variables if necessary.
- The dataset must originate from a non-classified, publicly available source.
- Dataset allocation will follow on a *first-come, first-served* basis.

Proposal Requirements

You are required to submit a written proposal (minimum one full A4 page) that demonstrates how the selected dataset satisfies the above criteria. The proposal must also address the following questions:

- **Regression Analysis**[10 points]
  - Which feature is selected as the dependent variable, and which features are the independent variables?
  - Is feature engineering required? Is encoding required? Justify your answers.
- **Clustering and Classification**[10 points]
  - Which feature is selected as the dependent variable, and which features are the independent variables?
  - Is feature engineering required? Is encoding required? Provide justification.
  - Is the task a binary or multi-class classification problem? Provide justification.

- **Association Rule Mining**[10 points]
    - Which feature is selected as the dependent variable, and which features are the independent variables?
    - How will association rule mining provide meaningful insights for the selected dataset? Provide justification.

Dataset Resources

Several publicly available sources of datasets include:

- Kaggle
- UCI Machine Learning Repository
- Google Dataset Search
- Analytics India Magazine – Top 10 Public Datasets

Submission Guidelines

- The deadline for submitting the proposal and selected dataset is **Sunday, September 28, 2025**.
- Submit the proposal as a **PDF document** before the deadline.
- Upload the dataset file (Excel, CSV, JSON, etc.) to **Canvas** under *Term Project – Proposal*.
- Complete the following shared Excel sheet with the details of your dataset before the deadline: Shared Dataset Registration Sheet

Please ensure that you sign in with your Virginia Tech email address to access and complete the dataset registration sheet.