CS 5393 – Introduction to Machine Learning
Fall 2025

PROJECT OVERVIEW

In this project, you will explore a dataset of your choice and use the dataset to build a model or models that accomplish one or more machine learning tasks.

Dataset Requirements

Choose a dataset that is already formatted as tabular data (rows and columns). If the data requires significant preprocessing, prepare it before analysis. Your dataset must meet the following criteria:

1. **Size**: The dataset must contain at least **3,000 rows**.
2. **Data Type**: It should not include image or free text-based features.
3. **Restrictions**: Do **not** use commonly used datasets such as **Iris, Digits, Penguins, or Titanic**.

Machine Learning Task

You can choose **one** or **combine multiple** machine learning tasks from the following:

- **Classification**
- **Regression**
- **Dimensionality Reduction**
- **Clustering**
- **Anomaly Detection**

PROJECT PROPOSAL GUIDELINES

Your **project proposal** must be written in a **Jupyter Notebook** and should include the following sections:

1) Project Title

Provide a clear and descriptive title for your project.

2) Author(s)

List your name(s) and email address(es).

3) Objectives

State the overall goal of your project. Use bullet points to list specific objectives.

4) Significance of the Project

Provide detailed answers to the following:

4.1) Dataset Description

- What is the dataset about?
- How many records (rows) and attributes (columns) does the dataset have?
- What are the types, meanings, and sizes of each attribute?
- What is the overall dataset size (in bytes)?

- Where did you obtain the dataset? (Include references.)
- Include a sample of the dataset records for reference.

4.2) Machine Learning Task Explanation

- What machine learning task(s) will you perform?
- Why is this task important? (Explain its real-world applications and relevance.)
- How will you use the dataset to accomplish the machine learning task?

5) Timetable

Provide a **detailed plan** outlining the to-do list you will follow. Include:

- A clear breakdown of activities and expected deliverables. You can break down the activities based on the machine learning pipeline we discussed in the class.
- Start and end dates for each task

6) References

List all sources in proper citation format. Sort references alphabetically by the last name of the first author. Each reference should include:

- Author(s)
- Title of the article/book
- Journal/conference/book name
- Publisher (for books)
- Month and year of publication
- Page numbers (if applicable)
- For websites, include the **access date**

SUBMISSION & GRADING

Submission Deadline: Submit all deliverables to the class website on Canvas by **11:59 PM** before the due dates.

This project accounts for **25% of your final grade**. You may complete it individually or in a team of **two**.

Items	Due Date	Grade
Project Proposal	**Saturday, Oct 11, 2025**	5%
Project Progress Report	Saturday, Nov 8, 2025	5 %
Project Presentation	Tuesday, Dec 9, 2025	5 %
Project Final Report	Tuesday, Dec 9, 2025	10%

Submit your proposal as a **ZIP file** containing the following:

- **Jupyter Notebook (.ipynb) file**
- **Rendered HTML version** of the notebook
- **Additional files** (e.g., figures or datasets used in your notebook)

USEFUL LINKS

MACHINE LEARNING DATASET REPOSITORY

- https://archive.ics.uci.edu
- https://www.kaggle.com/datasets
- https://openml.org/search?type=data&sort=runs&status=active

WRITING REPORT USING MARKDOWN IN JUPYTER NOTEBOOK

- https://www.datacamp.com/tutorial/markdown-in-jupyter-notebook
- https://www.ibm.com/docs/en/watson-studio-local/1.2.3?topic=notebooks-markdown-jupyter-cheatsheet