Major Course Output
Machine Learning Application to Real-World Data

Release Date: January 12, 2026
Due Date: March 21, 2026 (Saturday), 8:00 AM

Note: This is the graduate school version of this major course output.

---

1 Project Description

In this project, students will apply the machine learning concepts and techniques they have learned in the course to a real-world dataset. The goal is to develop a comprehensive understanding of the end-to-end process of formulating a supervised machine learning task, selecting appropriate machine learning models, training and evaluating these models, and interpreting the results. Through this project, the students should be able to demonstrate the following learning outcomes:

- **L01.** Define and differentiate the various machine learning techniques and approaches.
- **LO2.** Assess the impact of the design and parameters of various machine learning models in the context of a machine learning task
- **LO3.** Apply machine learning approaches on various tasks on real-world datasets, and effectively communicating findings and results

2 Groupings

The project must be accomplished in pairs of 2. If the class size is not divisible by 2, the instructor is responsible for balancing in the number of students in each group, but no group can have more than 2 members. Students may pick their own group members.

3 Dataset and Project Proposal

For the first phase of this project, students must propose a project that involves the application of **supervised machine learning** techniques to a real-world dataset. The following criteria must be observed for the proposal:

1. The dataset must be interesting to explore on a personal level, i.e., **the task must be something that the students care about**, not just selected out of convenience or compliance. It must also **have some level of social relevance or significance**, if possible in the Philippine context.
2. **The task must relatively unexplored**. Avoid selecting tasks that have already been extensively studied, either in academic works or in blogs, tutorial and other online resources. While it is not required for the task to be completely novel, students must avoid simply replicating existing works. Toy datasets are not allowed.
3. **There must be a concrete and feasible plan to acquire a dataset to train the models for the task**. It is highly recommended to pick tasks where there are already publicly avaiable datasets. If the students plan to collect their own data, they must show that their plan is feasible within the timeframe of the project.

We suggest looking for datasets in dataset-focused academic journals such as Data in Brief and Scientific Data. You are not required to use datasets from these journals, however. Online data repositories such as Kaggle, UCI Machine Learning Repository are strongly discouraged, unless you can justify the proposal according to the criteria above.

Each group must prepare the following information for their proposed project:

1. The supervised machine learning task to be performed, in the format: **"Predict Y based on X"**, where Y is the target variable to be predicted, and X is the target list of features to be used for prediction.
2. The dataset source or plan on how the dataset will be collected
3. Short justification (1–2 sentences) on why the task is interesting to the group members

All groups must seek the approval of the instructor for their proposed task and the corresponding dataset as early as possible. All groups must have an approved dataset already on or before January 30, 2026 (Friday), 11:59 PM. Please follow your instructor's directions on how to submit your proposal for approval.

4 Machine Learning Project Implementation

Once the proposal has been approved, the group may proceed to implement the machine learning project. The following is a general outline of the steps that must be performed:

1. **Data Preparation.** Collect the dataset and clean the data.
2. **Exploratory Data Analysis (EDA).** Analyze the dataset to understand its characteristics, distributions, and relationships between features. Visualizations and statistical summaries must be generated for a better understanding of the data.
3. **Data Preprocessing** Perform any necessary preprocessing steps to prepare the data for machine learning. Perform feature selection if needed.
4. **Model Selection and Training.** Select appropriate supervised machine learning models for the task. Train the models using the prepared dataset. Ensure proper training-validation-test splits to avoid data leakage.
5. **Error Analysis and Model Tuning.** Analyze the errors made by the models and make necessary adjustments to improve model performance.
6. **Model Evaluation.** Evaluate the performance of the trained models using appropriate metrics. Compare the performance of different models and explain the possible reasons why models performed well or poorly.

Please note that the steps outlined above is just a guide. You may adjust the steps as needed based on how your project progresses. For example, you may find that additional data preprocessing is needed after performing EDA, or you may need to revisit model selection after performing error analysis. Please do not treat the steps above as a rigid sequence that must be followed exactly.

For this project, you are required to make an effort to produce the best possible model for your task. This may entail exploring multiple types of models and comparing the best results. You must eventually identify the best performing model/s, justified by the experiments you conducted.

You may use any machine learning libraries or frameworks in your implementation, but you must be prepared to explain how the models work.

5 Deliverables

The following are the deliverables for this project:

5.1 Jupyter Notebook

The entire implementation of the machine learning project must be documented in a Jupyter Notebook. The Notebook must include both code cells needed to replicate all the outputs of the project, as well as markdown cells that provide explanations on the different steps taken in the project.

5.2 Conference Paper

A conference paper, following the IEEE Conference Paper Template, must be prepared to summarize the project. The paper must be a maximum of 6 pages long (including figures and references). The paper must include the following sections:

1. Introduction and Related Work
2. Dataset Description
3. Methodology
4. Experiments and Results
5. Discussion
6. Conclusion
7. References

Write the conference paper as if you are submitting it to a machine learning conference, not as if you are writing a report for a class. Projects that have potential can be further developed into actual conference submissions after the course.

6 Submission

All deliverables must be via AnimoSpace before March 21, 2026 (Saturday), 8:00 AM. The submission must include two files: `.zip` file containing the Jupyter Notebook and all supporting files, and a `.pdf` file of the conference paper. If the dataset files are small enough (less than 100 MB), you may include them in the `.zip` file. Otherwise, please provide a link to a public repository where the dataset can be downloaded.

7 Oral Exam

Each group will be required to attend an oral exam, where the instructor will ask questions about the project. The oral exam is a **major part of the grade**. Groups who submit complete projects, but are unable to answer questions about their process will receive major penalties in the grade and, in extreme cases where ownership of the work could not be established, will be treated as academic dishonesty resulting in a grade of 0.0 for the course. Details for the oral exam will be announced by the invidividual instructors.

8 Generative AI Use Policy

You are allowed to use AI tools for aid in concept understanding, exploration, and brainstorming. You are **not** allowed to use AI tools to directly generate code or text that will be used in your submission. If you do use AI tools, you must adhere to the following guidelines:

1. You must declare the use of such tools by specifying the tool that was used, and a **detailed description** of how the tool was used.
2. You must validate any AI response through your own understanding of the concepts or through your own research.
3. You must be able to articulate the thought processes, rationales, and implementation details of your work, and through this you must be able to show that human agency was maintained even if AI was used in augmenting the process.

9 Rubrics for Grading

Refer to the AnimoSpace course page for the detailed rubrics for grading.