School of Information Technologies and Engineering, ADA University

CSCI4734 – Machine Learning

Summer 2025

Course Project

Project Proposal Submission Deadline: June 15, 2025, 23:59

Archive File Submission Deadline: July 2, 2025, 23:59

Overview

You get 20 points of your total grade for the Machine Learning course based on a project that you are expected to complete by the end of the term. The project is intended to start you in the directions of machine learning research or applications of machine learning techniques to solve real-world problems. The project will require you to do fair shares of coding, reporting, and presenting.

Grading Components

Project Proposal (10%)

You are expected to pick an area of interest that you deem viable to apply ML algorithms to. These areas include, but are not limited to, life sciences, finance, physical sciences, theoretical machine learning, computer vision, natural language processing, music, commerce, etc.

Feel free to drop an email or visit the office to discuss and brainstorm ideas.

A good project idea is one that you are excited about. Feel free to be ambitious yet realistic (building AGI is not going to fit into a few weeks, maybe you can give it a shot for SDP). You should be aware that training deep learning models on big data can be very time and compute consuming, so make sure you have the necessary resources if you are planning to work with big data and deep learning. Alternatively, if you are already working on an ML research or industry project, then you may already have a project idea. However, you must make sure that your project leverages the ideas and concepts discussed in the course. Replicating the results of a research paper is also a good project idea if supplemented with new contributions so that it is not just a duplicate of previously done work.

Once you have identified your project idea, submit a proposal by the set deadline. You can find the proposal template on the contents section of the Blackboard course page. You can add pointers to a relevant dataset and an example of prior research on the topic. The proposal is mainly intended to make sure you decide on a project topic and get feedback. If your proposal follows the instructions and the project seems to have been thought out with a reasonable plan, you should do well on the proposal.

It can be useful to look up existing research on topics relevant to yours. You can use the following resources for researching related work:

- https://paperswithcode.com/
- https://dl.acm.org/search/advanced
- https://scholar.google.com/
- https://arxiv.org/search/advanced
- https://dblp.org/

To get familiar with the lifecycle of a machine learning project, you are recommended to check the following resource:

Book: Aurelien, G. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 3rd edition, Chapter 2: End-to-End Machine Learning Project.* 2022.

Accompanying notebook: https://github.com/ageron/handson-ml2/blob/master/02_end_to_end_machine_learning_project.ipynb

EDA and Data Preprocessing (20%)

An important aspect of conducting an ML project is the identification of relevant data sources. If the datasets need considerable preprocessing to suit proper modeling, or if you are planning to collect the data yourself, make sure that you have conducted a cost-benefit analysis as it may take a lot of your time, and it constitutes a small (yet important) part of the expected work.
Datasets should contain enough examples to get statistically significant results.
Use of non-public datasets is allowed; however, you must make sure that you are in possession of the required permissions to use it and share it with the instructor if asked to do so.
Some places to check out:

- https://www.kaggle.com/datasets
- https://datasetsearch.research.google.com/
- https://github.com/awesomedata/awesome-public-datasets
- https://sebastianraschka.com/blog/2021/ml-dl-datasets.html
- https://archive.ics.uci.edu/ml/index.php
- https://registry.opendata.aws/
- https://dataportals.org/
- https://www.reddit.com/r/datasets/top/?t=all
- https://www.reddit.com/r/datasets/comments/shbzwe/is_there_a_master_list_of_places_to_look_for/

For this part of the project, you are expected to conduct an Exploratory Data Analysis and preprocess the data to prepare it for modeling.
EDA serves the purpose of getting initial insights from the data and often includes but is not limited to the following components:

- Visualization (histogram, scatter plot, bar chart, violin plot, box plot, map, etc.)
- Missing values detection
- Outlier detection
- Five-number summary
- Data type identification (categorical, numerical, ordinal)
- Correlation analysis

Data preprocessing serves the purpose of preparing the data for machine leaning algorithms to be applied on. It often includes but is not limited to the following components:

- Handling outliers
- Handling missing values
- Data cleaning
- Scaling
- Data normalization
- Feature removal
- Feature engineering (can also be a part of modeling phase)

Some of the most popular EDA and data preprocessing tools are *pandas*, *matplotlib*, *seaborn*, and *scikit-learn*.

Modeling (40%)

As the most important stage of the project, this part requires you to select, build, train, fine-tune and evaluate machine learning model(s) on the prepared data. You are expected to apply some of the following techniques:

- Linear regression
- Logistic regression
- Decision trees
- Ensemble learning
- K-means
- DBSCAN

- Naïve Bayes
- Principal Component Analysis
- Support Vector Machines
- Deep Neural Networks

The algorithm(s) should be applied in such a way that the desirable working result is achieved at the end. Non-working prototypes will cause a significant reduction in the grade.

This phase of the project may have the following steps (does not have to be in the given order):

- Selecting algorithm(s)
- Applying algorithm(s)
- Hyperparameter tuning (you can use grid search)
- Feature selection
- Evaluation
- Error analysis
- Comparative analysis

MLOps (10%)

Choose an MLOps tool to track your experiments. Some examples are:

- https://www.mlflow.org/
- https://dvc.org/
- https://kedro.org/
- https://pycaret.org/
- https://wandb.ai/
- https://metaflow.org/
- https://www.tensorflow.org/tensorboard

Report (10%)

You are expected to produce a final project report that includes following sections:

- Problem formulation
- Discussion of related works (optional)
- EDA and data preprocessing
- Modeling
- Experiments
- Discussion of results

If you have been advised by another instructor/professor for this project, make sure you fully acknowledge their contributions. The final report will be judged based on the clarity, discussion of the key points, and interpretation of the results.

You can find the report template on the Blackboard course page.

Presentation (10%)

Each team is expected to present and discuss their project and be prepared to answer relevant questions. A time slot of 10 minutes will be scheduled for each presentation including question answering session.

The presentation should follow the same content as the project report and be limited to at most 10 slides. The team is also expected to demonstrate the working program. Each project member's role must be clearly indicated in the presentation. Each team member is expected to be aware of the whole scope of the project. The team must be able to clearly and effectively explain the work that has been done, including context, methods, and results.

Submission

Please follow the instructions below when you make your submissions to the Blackboard System:

- The language of choice for the project is Python. Consider using Jupyter Lab for your own convenience.

- One of the team members is expected to submit a single compressed archive file with all relevant files. It must include the report, the presentation slides, and the source code.
- Name the archive file according to the template (all capital):
- CSCI4734_2025U_30045_PROJECT_<TEAMNUMBER>

- You can make 3 submissions before the deadline.
- The latest submission will be considered for grading.
- For one day delay, your grade will be deducted 25%. No submission is accepted after one day delay.
- Exact mirrors of existing project solutions with no substantial technical contribution on your part will be graded with zero.
- You may lose up to half of the grade if existing analysis/ML solutions for the same dataset can be found on the Web. It is in your best interest to refrain from popular datasets.
- Upon detection, free riders will lose a considerable number of points and will be graded zero if no understanding whatsoever is displayed.
- Furthermore, any team whose solution raises question(s), will be asked for some explanations.