# Programming Project 3: Selective Choice

Deadline: March 09, 2025

**Task Overview.** In this project, you will need to complete the following tasks:

1. Find and clean a new large dataset.
2. Build and train various ML models.
3. Report the training performance of your models.

The goal of this project is to give you a lot of freedom so that you can have the flexibility to do cool things with your existing machine learning skill set while still learning new tools.

## 1 Task 1: Find and Clean a New Dataset (25 points)

You are required to find and clean a new *large* dataset with more than 10,000 data samples; each sample has at least 10 features. This dataset cannot be the ones that you have used in previous projects. You are allowed to source datasets from the following places:

- Kaggle: `https://www.kaggle.com/datasets`
- University of California – Irvine's ML Repository: `https://archive.ics.uci.edu`
- Any dataset with the permission of the instructor.

You will need to divide your dataset into training, validation, and test sets. You will use the training set to train your model, the validation set to tune/select hyper-parameters, and the test set for the final evaluation of your trained models.

The output of this task include: (i) a notebook file named *proj3_data_preprocess.ipynb* which contains your codes; and (ii) the original datasets (or link to the datasets).

## 2 Task 2: Build and Train Various ML Models (65 points)

Your task is to build, train, and evaluate various machine learning models based on your new datasets. These machine learning models include:

1. Linear regression (for a regression task) or logistic classification (for a classification task).
2. Support vector machine

3. Decision trees

4. Random forests

5. Neural networks

In particular, you will need to apply at least two different kernel tricks for linear regression and support vector machines. The goal is to study how different kernel tricks will influence the performance of these models.

Finally, for neural nets, recall that in Project 2, a `random_seed` is given in advance, allowing you to obtain obtain identical results for each time you re-run your program. Different `random_seed` values will likely result in different learning results. Therefore, for Project 3, you will need to train your neural nets with at least five to ten different random seeds. The performance of your neural nets will be averaged over these ten random seeds.

The output of this task is a notebook file named *proj3_machine_learning.ipynb* which contains your codes.

**Important notes.** For this task, you can use Python libraries (e.g., sci-kit learn, pytorch, etc) instead of implementing the above methods from scratch.

## 3 Task 3: Write Report (10 points)

Finally, you will write a report on your dataset and the results of your model performance. Your report should include the following results:

1. A description of your dataset and where you find it with the link to download the dataset.

2. A description of the task you are trying to accomplish with the data via machine learning techniques.

3. A description of any pre-processing you did to the dataset.

4. A summary of prediction accuracy of all the evaluated machine learning methods in training, validation, and test sets.

5. Write 4-5 sentences summarizing your observations regarding these results, with a comment on the under-fitting/over-fitting/convergence performance of your models.

Save your report in a file named *proj3_report.pdf*.

## 4 Submission

You will need to submit three files: (i) *proj3_data_preprocess.ipynb*; (ii) *proj3_machine_learning.ipynb*; and (iii) *proj3_report.pdf* on Canvas.