CSCC11 Fall 2025 Project

1 Introduction

Welcome to the CSCC11 final project! The goal is to let you practice an end-to-end machine learning project and develop a deeper understanding of the machine learning models we covered in lecture by conducting a comparison study on a chosen regression or classification problem. It also provides an opportunity for you to explore an advanced model that is not covered in the lecture.

1.1 Learning Objectives

After finishing this project, you will

- gain hands-on experience in formulating a machine learning problem, and selecting or collecting appropriate datasets;
- familiarize the process of data cleaning, preprocessing, and feature engineering;
- implement, tune, and rigorously evaluate a suite of classic and advanced machine learning models;
- develop a deeper understanding of the strengths and weaknesses of different machine learning models;
- communicate effectively their methodology, results, and insights through a formal report and presentation.

1.2 Group Formation

This is a group project. You must form a team of four students. When the number of students in the class is not a multiple of four, we will accept a limited number of three-member groups. You will need to contact the instructor if you want to form a three-member group.

2 Project Requirements and Deliverables

Every group will be responsible for conducting an in-depth comparison study of at least four different models to solve either a regression or a classification problem.

2.1 Problem Formation

You will need to find a regression or classification problem to solve. The comparison study is to answer the research question: what are the advantages and limitations of different models

to solve the selected problem?

2.2 Data Collection, Proprocessing and Exploration

You will need to either collect your own data or find the data source to solve your problem. You can find datasets from many source. Kaggle, Huggingface, or UCI are three popular ones. You will need to clean the data and preprocess the data so that they are in the right format and representation for the learning models to use. You will need to explore the data, visualize them, feature engineer them before you apply a learning algorithm on them.

2.3 Model Requirements

Your comparison study must include

- at least three models covered in CSCC11 (e.g., Linear Regression, Basis Function Regression, Ridge Regression, KNN, Decision Trees, Random Forest, Class Conditional Models, Naive Bayes, Logistic Regression, Multi-Layer Perceptron, et. al.) and

- one *advanced* or *novel* model not covered in the lectures. This is the research component of your project. This could be an implementation of an advanced and well-known algorithm that performs better than those models covered in the lecture or a novel method you develop, such as a unique feature engineering approach, a custom model ensemble, or an experimental neural network architecture.

2.4 Deliverables

The final deliverable contains three items

- Group Signup (1 pts), due on Oct 3rd 22:00 EDT

- Proposal in a .pdf file (4 pts), due on Oct 31st 22:00 EDT

- A 10-minute presentation (5 pts), time to be arranged

- A .zip file containing a final report, well documented source code with build and run instructions (35 pts), due on Dec 2nd 22:00 EDT

2.4.1 Proposal

This is a 1-2 page short PDF file to discuss what your project topic is about and your intial plan of the models to be used in the comparison study. You should formulate the problem in the proposal, specify how to obtain the data to solve the problem, and what the tentative models to be used are. The document is meant to be short.

2.4.2 Final Project Report

The final project report a PDF document structured like a short academic paper. This report must detail your entire project, from problem formulation to your final analysis. The page limit is 3-5 pages. The marking scheme are as follows

- (10%) **Abstract**: Give an executive summary of your project and main findings
- (15%) **Introduction**: Clearly define the problem you are solving and the main research question you are investigating. What are the existing methodologies and what is your method. If you decide to implement an existing advanced method, then you should state why you decide to choose this particular advanced method.
- (15%) **Data Collection, Preprocessing, and Representation**: Present yoru data exploration results using figures or statistical summaries. Discuss the data cleaning and feature engineering you perform and why, how the data are represented in the models used, and how you split the data in the hyper parameter tuning.
- (36%) **Model Comparison Study**: Describe each model you use in the comparison study. What they are and why they are selected to solve the problem in your study. Detail the metrics you use to evaluate these models and the rational of choosing these metrics. Describe how your perform the hyperparameter tuning. What are the hyperparameters and what values you have tried during tuning. Provid evidence that your hyperparameter choices are reasonable.
- (6%) **Main Results**: Present your main results of the comparison study. What are the strengths and weaknesses of the models and what is your final selected model to solve the chosen regression or classification problem. You may want to use figures or tables or both to help the presentation fo the results.
- (3%) Future Work: You may also suggest future works to extend the study or open questions for further exploration.
- (15%) **Originality**: We would like to see how your comparison study differs from existing comparison study in the literature. How the way you apply the models to solve the selected problem differs from existing methods.

2.4.3 Presentation

This meant to be a concise, engaging presentation summarizing your project's key findings for the class. You are encouraged to present the project together as a team.

2.4.4 Peer Evaluation

Your group project grade will be affected by how much contribution you have made to the teamwork. A peer evaluation form will be circulated among group members to let every

group member evaluate the contributes from other group members and provide constructive anonymous feedbacks. There will be a contribution factor that gets applied to the original group work mark. How do we calculate the contribution factor will be communicated to the class by the time the proposal is due.

3 Project Ideas

You are encouraged to find your own project ideas. Here are two possible projects, one for regression and one for classfication comparison studies by using the Toronto crime data. There are other project ideas using this dataset that you may want to explore on your own.

3.1 Problem Formulation

We want to predict the urban safety in Toronto. This can be formulated either as a regression problem or classification problem. Let's first look at the data we have access to.

3.1.1 Data

You will find the **Major Crime Indicators (MCI)** dataset from the City of Toronto's Open Data Portal at `https://data.torontopolice.on.ca/pages/major-crime-indicators`. This rich, spatio-temporal dataset contains records of major crime incidents from 2014 to the present, offering a realistic machine learning problem.

The key features from the dataset are

- `neighbourhood_158`: Spatial location (158 distinct neighbourhoods in Toronto)
- `report_date`: Temporal information (when the incident was reported)
- `occ_date`: Temporal information (when the incident ocurred)
- `offence`: Crime category (Assault, Break and Enter, Theft, Auto Theft, Robbery)
- `long_wgs84`, `lat_wgs84`: Geographic coordinates
- Additional contextual features available in the dataset

3.1.2 The Neighbourhood Safety Index

Instead of working directly with raw crime counts, your first task is to construct a meaningful target variable: the Neighbourhood Safety Index (NSI).

1. **Assign Severity Weights:** Define and justify a severity scale for different crime categories. For example:

- Assault = 5
- Break & Enter = 3
- Auto Theft = 2
- Theft = 1
- Robbery = 4

You should research and justify your chosen weights in your proposal and final report.

2. **Calculate Weighted Score:** For each of the 158 neighbourhoods in a given month, calculate:

$$\texttt{TotalCrimeScore} = \sum_i (\texttt{count}_i \times \texttt{weight}_i) \ , \tag{1}$$

where i ranges over crime categories.

3. **Normalize and Invert:** Scale the score to a 0-1 range and invert it:

$$\texttt{Neighbourhood SafetyIndex} = 1 - \frac{\texttt{TotalCrimeScore} - \texttt{MinScore}}{\texttt{MaxScore} - \texttt{MinScore}} \ .$$

A score near 1.0 indicates higher safety, while scores near 0.0 indicate lower safety.

3.1.3 Prediction Task

You will build models to predict neighbourhood safety one month into the future. This forecasting horizon is long enough to be practically useful for urban planning and resource allocation, while short enough to be tractable with the available data.

Accurate safety predictions can inform police resource allocation, community intervention programs, urban planning decisions, and public awareness and transparency.

3.2 Regression Project Stream - Predicting Safety Index

The goal is to predict the Neighborhood Safety Index for each neighborhood one month into the future. You want to select at least three classic machine learning regression models covered in the lecture. You want to pick the evaluation metrics. You want to define and select proper features for this project.

3.3 Classification Project Stream - Predict Crime Type

The goal is to predict crime category given features (location, time, premises) of a crime incident. You want to select at least three classic machine learning classification models

covered in the lectrure. You also need to select the evaluation metrics. This is a multi-class classification problem.