University of Ottawa
School of Electrical Engineering and Computer Science
CSI5155 - Fall 2025

Assignment 1 – Supervised Learning

Total marks: 100
Due date: 3 October, 2025 (12:59pm ET)

Instructions:
1. This is an **individual** assignment. Submit your assignment (source code, output, and written report) using uOttawa's BrightSpace before the due date.
2. Use Scikit-Learn to complete the assignment.
3. In addition to the source code and written report, all students are expected to demonstrate and explain their projects during a time slot the teaching assistant will schedule after submission. This demonstration will be used in conjunction with the submitted assignment for assignment evaluation.
4. Please refer to the AI Policy and Late Policy on the course syllabus.

Description
This assignment considers the Customer Personality Analysis dataset, which contains a detailed analysis of a company's ideal customers, at https://www.kaggle.com/imakash3011/customer-personality-analysis/version/1

"*Customer personality analysis helps a business modify its product based on its target customers from different customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then sell the product only on that particular segment.*"

In the original task, machine learning engineers focused on customer segmentation and clustering. However, in this assignment, the target feature is "complain," and we aim to construct classifiers to predict this behaviour. That is, we wish to learn a model to distinguish between **satisfied customers** and **those who complained** in the past. This is a binary classification task.

A. Supervised learning [40 marks] Complete the following tasks.
1. Learning: Import the data into your machine learning environment and conduct feature engineering. Next, construct models using the following six (6) algorithms:
    a. logistic regression (LR)
    b. single decision tree (DT)
    c. support vector machine (SVM)
    d. k-nearest neighbour (k-NN)
    e. random forest (RF) learner
    f. gradient boosting (GB) ensemble *[Note: we will talk about RF and GB in class on Oct.1, but you can complete the assignment before then using scikit learn.]*

You should use the five-fold cross-validation evaluation method and perform parameter tuning to get the “best” model.

2. Evaluation: Show the confusion matrices corresponding to the six (6) models and output the recalls and precisions.

3. Visualization: Generate a figure to show the ROC curves for the six models.

B. Class imbalance [40 marks]
The dataset is *imbalanced* in that one of the labels occurs more frequently; this may affect the results of learning. Complete the following tasks.

1. Learning and Rebalancing: Apply two (2) data-centric approaches of your own choice to address this issue, namely:
    a. Use one algorithm for *undersampling* the majority class. Next, employ the six abovementioned algorithms with the undersampled dataset to construct six (6) new models.
    b. Apply another algorithm (such as SMOTE) for *oversampling* the minority class. Employ the six (6) algorithms mentioned above with the oversampled dataset to construct six (6) new models.

2. Evaluation: For the models in both (a) and (b), show the confusion matrices corresponding to each model and calculate the recalls and precisions.

3. Visualization: For the models in both (a) and (b), generate a figure to show the ROC Curves for the six models.

C. Reporting [20 marks]
Submit a **400 to 500** words written summary, discussing the results you obtained and the lessons you learned when analysing this data, focusing on the behaviour of the algorithms, the results obtained, and the impact of sampling.

Tips:

- There are many methods in scikit-learn to help with tasks such as generating a train-test split, conducting parameter optimization with cross-validation, and evaluating classifier performance:
    - E.g. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
    - E.g. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
- You may find this package useful for dealing with class imbalance:
    - https://imbalanced-learn.org/stable/