# CIVE 7150 - Data-Driven Decision Support for Civil and Environmental Engineering

Homework 4

Due date: April 1, 2024

## Instructions

- Submit a PDF writeup named "LASTNAME_HW4.pdf" on Canvas. Please answer all the questions in the PDF. If a problem asks for graphs or other results, please extract the results from the code and include them in the report. The PDF report should have complete answers to all the questions.
- Please use Jupyter notebooks in Python and include comments in the code if needed.

**Datasets:** The datasets for this assignment are:

- The SPAMBASE dataset from the UCI repository is available at: https://archive.ics.uci.edu/ml/datasets/spambase.

  The first 57 columns are features counting word frequencies (see documentation at https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names). The last column indicates 1 for the SPAM class and 0 for the HAM class.

- The Mushroom dataset from the UCI repository is available at: https://archive.ics.uci.edu/ml/datasets/Mushroom

  This dataset consists of 22 categorical attributes, and the prediction task is whether the mushroom is edible or poisonous. The class label is the first column in the dataset.

## Problem 1 [Decision Trees] - 25 points

(a) Use an existing package to train a decision tree on the SPAMBASE training data, without pruning. Use the information gain splitting criteria. Compute the training and the testing error, accuracy, F1 score, and AUC and report these metrics. Write down some observations about the training and testing metrics.

(b) Change the splitting criteria to use the Gini index and report the same metrics. Compare them to the information gain metrics.

(c) Implement a pruning criteria that sets an upper bound on the maximum depth of the tree. Generate a graph that plots the training and testing error as a function of the tree depth on the SPAMBASE data. Please explain your observations. What is the optimal depth of the tree that you would recommend based on this analysis?

## Problem 2 [Random Forest Ensemble] - 25 points

(a) Use an existing package to train a Random Forest ensemble with 10, 50, 100, and 500 decision trees on the SPAMBASE dataset. Report accuracy, F1 score, and AUC on both the training and testing sets for $T \in \{10, 50, 100, 500\}$. How do the metrics change as $T$ increases?

(b) Compare the metrics obtained for Random Forest with the Decision Tree metrics obtained in Problem 1. Write down some observations.

(c) Compute the variable importance for each feature and include a plot.

## Problem 3 [AdaBoost Ensemble] - 25 points

(a) Use an existing package to train an AdaBoost ensemble with 10, 50, 100, and 500 base classifiers on the SPAMBASE dataset. Use a decision tree classifier as the base classification model. Report accuracy, F1 score, and AUC on both the training and testing sets.

(b) Compare AdaBoost with the Random Forest ensemble for 10, 50, 100, and 500 base learners by looking at various metrics on the training and testing sets.

(c) Plot the ROC curves for the decision tree model, Random Forest with 100 trees, and AdaBoost with 100 trees.

## Problem 4 [Naive Bayes classifier] - 25 points

In this problem you will implement your own Naive Bayes classifier and you will compare it with a package implementation. You will use the Mushroom dataset for this problem. Split the dataset into 75% for training and 25% for testing.

(a) Train the Naive Bayes classifier. Compute the prior probabilities for the *Edible* and *Poisonous* classes from the training data. For each feature $X_i$ in the dataset compute the probabilities $P[X_i = x|Y =$

*Edible*], and $P[X_i = x|Y =$ *Poisonous*$]$ from the training data. Use the Laplace smoothing method when computing these probabilities.

The Naive Bayes classifier stores these prior and conditional probabilities.

(b) For each point in the testing set, estimate the probability that it belongs to the *Edible* and *Poisonous* classes. Use the Naive Bayes classifier probabilities computed in part (a).

(c) Compute accuracy, precision, recall, and F1 score for your Naive Bayes classifier on the testing data.

(d) Compare the results obtained by your implementation with those obtained with a Naive Bayes package (trained on the same dataset). Use several metrics, including accuracy, precision, recall, and F1 score. Are the results similar or different?