University of Ottawa

School of Electrical Engineering and Computer Science

CSI5155 - Fall 2024

Assignment 1 – Supervised Learning

Total marks - 100

Instruction:

1. This is an individual assignment. Submit your assignment using uOttawa's BrightSpace before the due date.
2. For the implementation, you should either upload your code on BrightSpace or provide a link to a GitHub repository. Note that if you choose to use GitHub, the date and time of the last change to your repository should be **before** the assignment deadline.
3. Use Scikit-Learn to complete the assignment. You are also encouraged to use the Morningstar Cluster.
4. We will post a schedule for you to demonstrate your assignment to the Teaching Assistant (TA). These will be individual, one-on-one demonstrations.

Description

This assignment considers a dataset collected to assess an individual's risk of drug use and abuse based on numerous factors, including psychological, social, individual, environmental, and economic factors, which are also associated with several personality traits.

The dataset can be found at this link:
https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

This dataset contains a row identifier, 12 features describing the user data, and 18 classification problems related to using 18 different drugs.

For each drug, it indicates whether a person has 'never used', 'used over a decade ago', 'used in the last decade', 'used in the last year', 'used in the last month', 'used in the last week', or 'used in the last day'.

In this assignment, we will convert the multi-class problems into binary classification tasks. To do so, we merge the categories 'used over a decade ago' and 'never used' into the '**non-user**' group and place the other five categories ('used in last decade', 'used in last year-month-week-day') into a '**user**' group.

Furthermore, we will study only two classification tasks: individuals who enjoy chocolate and those who take magic mushrooms. That is, we want to build accurate models to typify individuals who use chocolate or magic mushrooms. We will not consider the use of other drugs at this time.

To accomplish this task, you should construct two separate datasets:
- The first dataset contains features 2 through 13 and the label 'Choc' (feature 20).
- The second dataset contains features 2 through 13 and the label 'Mushroom' (feature 29).

The following paper describes the data, as well as the data quantification process used to create the dataset: https://link.springer.com/book/10.1007/978-3-030-10442-9
(or https://arxiv.org/abs/1506.06297).

A. Supervised learning [30 marks]

Complete the following steps.

Classification task for Chocolate users.

a. Learning: Transform the data into a binary classification task, import the data into your machine learning environment and construct models using the following six (6) algorithms: a single decision tree (DT), a random forest (RF) learner, a support vector machine (SVM), a gradient boosting (GB) ensemble, a multi-layer perceptron (MLP), and a k-nearest neighbour (k-NN) classifier. You should use the standard holdout (train-then-test) evaluation method and perform parameter tuning to obtain the "best" model.
b. Evaluation: Display the confusion matrices corresponding to the final six (6) models and calculate the recalls and precisions against the test data.
c. Evaluation: Draw a figure to show the ROC Curves for the six models.

Classification task for Magic Mushroom users.

a. Learning: Transform the data into a binary classification task, import the data into your machine learning environment and construct models using the following six (6) algorithms: a single decision tree (DT), a random forest (RF) learner, a support vector machine (SVM), a gradient boosting (GB) ensemble, a multi-layer perceptron (MLP), and a k-nearest neighbour (k-NN) classifier. You should use the standard holdout (train-then-test) evaluation method and perform parameter tuning to obtain the "best" model.
b. Evaluation: Display the confusion matrices corresponding to the final six (6) models and calculate the recalls and precisions against the test data.
c. Evaluation: Draw a figure to show the ROC Curves for the six models.

B. Class imbalance [40 marks]

The two datasets are imbalanced because one of the labels is more common, which can affect the learning results. In particular, the chocolate dataset is highly skewed, while the magic mushrooms dataset is somewhat more balanced.

B1. Data balancing for 'Chocolate'.

Apply two (2) data-centric approaches of your own choice to address the class imbalance:

i. Use an algorithm to undersample the majority class. Next, employ the six algorithms from Section A on the undersampled dataset to construct six (6) new models.
ii. Apply another algorithm (such as SMOTE) to oversample the minority class. Next, employ the six (6) algorithms from Section A on the oversampled dataset to construct six (6) new models. (Reference to SMOTE: https://arxiv.org/abs/1106.1813)
iii. Rebalance the data by combining undersampling and oversampling. That is, you should use both methods together and determine the sampling percentage of each one to get a more balanced dataset. Next, employ the six (6) algorithms from Section A above against the rebalanced dataset to construct six (6) new models.

For both (i) and (ii), complete the following steps:

a. Evaluation: Display the confusion matrices corresponding to the three (3) sets of six (6) models and calculate the recalls and precisions against the same test data you used in Section A.
b. Evaluation: Draw three (3) figures to show the three (3) ROC Curves for the six (6) models.

B2. Data balancing for 'Mushrooms'.

Apply the same two (2) data-centric approaches you used for Chocolate to rebalance this data.

i. Use an algorithm to undersample the majority class. Next, employ the six algorithms from Section A on the undersampled dataset to construct six (6) new models.
ii. Apply another algorithm (such as SMOTE) to oversample the minority class. Next, employ the six (6) algorithms from Section A on the oversampled dataset to construct six (6) new models. (Reference to SMOTE: https://arxiv.org/abs/1106.1813)
iii. Rebalance the data by combining undersampling and oversampling. That is, you should use both methods together and determine the sampling percentage of each one to get a more balanced dataset. Next, employ the six (6) algorithms from Section A above against the rebalanced dataset to construct six (6) new models.

For both (i) and (ii), complete the following steps:

a. Evaluation: Display the confusion matrices corresponding to the three (3) sets of six (6) models and calculate the recalls and precisions against the same test data you used in Section A.
b. Evaluation: Draw three (3) figures to show the three (3) ROC Curves for the six (6) models.

Report: Create a PDF document containing all the Confusion Matrices, Recalls, Precisions, and ROC Curves. Label all your tables and figures clearly.

C. Discussion, Synthesis and Summary [30 marks]

Submit a 400-word to 500-word summary discussing the results you obtained and the lessons you learned when analyzing this data.

- Your answer should focus on the behaviour of the algorithms, the results obtained, and the impact of rebalancing.
- Your answer should also highlight the differences between the models constructed against the two datasets and the differences between the rebalancing processes and results for these two datasets.