CS 484: Introduction to Machine Learning

Fall Semester 2023 Assignment 5

# Question 1 (100 points)

The Center for Machine Learning and Intelligent Systems at the University of California, Irvine manages the Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>). We will use two of the datasets in the repository for analyses, namely, the **WineQuality\_Train.csv** for training and the **WineQuality\_Test.csv** for testing.

The categorical target variable is *quality\_grp*. It has two categories, namely, 0 and 1. The Event category is 1. The input features are *alcohol*, *citric\_acid*, *free\_sulfur\_dioxide*, *residual\_sugar*, and *sulphates*. These five input features are considered interval variables.

We will train two models. One is a classification tree, and another is a binary logistic regression.

The classification tree has the following specifications.

* The Splitting Criterion is Entropy.
* The maximum tree depth is five.
* The initial random state value is 20230101 for the classification tree.

The binary logistic regression has the following specifications.

* The model must include the Intercept term.
* Use the All-Possible Subset method to determine the model with the lowest Akaike Information Criterion.

After we train these two models, we will compare them using a suite of model performance metrics and charts.

1. (20 points) What are the Root Average Squared Error values of both models for both training and testing partitions?
2. (20 points) What are the Area Under Curve values of both models for both training and testing partitions?
3. (10 points) Generate the Receiver Operating Characteristic curve for both models on the training partition. Please put the two curves in the same chart frame. Don’t forget to add the diagonal reference line.
4. (10 points) Generate the Precision and Recall chart for both models on the training partition. Please put the two curves in the same chart frame. Don’t forget to add the No-Skills line to the chart.
5. (10 points) What is the threshold for the Event probability based on the F1 Score from the training partition? Please calculate the thresholds of both models.
6. (10 points) Using the F1 Score threshold, what are the Misclassification Rates of both models when evaluated only on the testing partition?
7. (10 points) Generate the Cumulative Gain and Lift table for both models using the predicted Event probabilities from the testing partition. Which model has the highest Lift value in Decile 1?
8. (10 points) Based on all the above model performance metrics and charts, which model will you pick as the Champion model?