ASSIGNMENT 2: CLASSIFICATION

WEKA documentation is available on Canvas under *Software and Tutorials*.

A. (60 points) Classification Learning

Use the following 5 learning schemes (OneR to IBk), to analyze any 5 datasets from the `UCI-arff-datasets.zip` file on Canvas. For test options, first choose "Use training set", and then choose "Percentage Split" using default 66% percentage split. Finally use 10-fold cross-validation.

Report the 3 error rates on each dataset (training set, 66% split and cross-validation) using each of the following machine learning algorithms. Explain in a few lines what you find interesting about each of the 5 learning schemes after assessing the performance on your selected datasets.

- OneR or Logistic Regression
- Naive Bayes
- J4.8
- PRISM (you need to discretise numeric attributes when you use PRISM)
- IBk (you can tweak k for different error rates)

B. (10 points) Using the datasets explain in your own words why training error should not be the sole measure of model accuracy?

C. (10 points) Pick any 2 datasets from your 5 datasets. Explain and compare the performance of J4.8, PRISM and Naïve Bayes on the two datasets using the confusion matrix produced by each of these algorithms.

D. (20 points) Provide lift charts from WEKA for models learned on any 3 datasets where the predictive accuracy of the model learned by the algorithm is better than flipping a coin. So you will need to provide a total of 3 lift charts, one for each dataset. Explain in 1-2 sentences the reason for selecting each model and presenting the lift chart.