Assignment #11

Data Mining — Due: December 10, 2018

Note: Consider only a classification problem. That is, there is a variable which indicates classes. The location of the class variable is not fixed. Make your program to handle more than two classes. You can assume that values of the class variable are integers starting with 1. Assume your data has both numerical and categorical variables. Further assume that the categorical variables are coded as integers starting with 1. You may use one of two (R or Python) language for this assignment.

1. Prompt the user to type in the filename of the training data.
2. Prompt the user to enter the locations of the categorical variables and the class variables.
3. Prompt the user to enter the filename of the test dataset. (Assume the column location of the class variable is as same as that of the training dataset.)
4. Perform Bagging depending on the choice by the user.
5. For Bagging Ensemble method, use (1) decision trees with depth 2 and (2) decision trees with depth 4 as the classifier and 51 bootstraps as the number of re-sampled data.

   Only the output of the **test data** is necessary for this assignment. The output file for classification generated by the program must look like:

```
  (1)  Tree with depth 2
ID, Actual class, tree-depth2 pred
-----------------------------
1, 1, 1
2, 2, 2
3, 1, 1
(continue)

Confusion Matrix (tree-depth2)
----------------------------------
                   Predicted Class
                      1       2
Actual      1       239      14
Class       2        12     153

Model Summary (tree-depth2)
-----------------------------
Overall accuracy = .793

  (2)  Tree with depth 4

ID, Actual class, tree-depth4 pred
-----------------------------
1, 1, 1
2, 2, 2
3, 1, 1
(continue)

Confusion Matrix (tree-depth4)
----------------------------------
                   Predicted Class
                      1       2
Actual      1       239      14
Class       2        12     153

Model Summary (tree-depth4)
-----------------------------
Overall accuracy = .793
```