CIND 119: Introduction to Big Data Analytics
Assignment 1 (15% of the final grade)
Supervised Learning Using SAS

1. Download the breast-cancer-dataset.csv from your D2L Assignment 1 link. Complete the following tasks (5 points):

    a. Read the file in SAS and display the contents using the import and print procedures. (1 point)

    b. Develop a decision tree-based classification model using the hpsplit procedure of SAS. (2 points)

    c. Navigate the contents of Results View by clicking on HPSplit breast-cancer-dataset, and then by selecting Model Assessment. Examine the confusion matrix, fit statistics, and variable importance. (2 points)

2. Using the confusion matrix, compute the following assessment metrics accuracy, recall, and precision (see lecture for formulas). (5 points)

    Condition for marks: 3 points for accuracy, 1 point for precision, and 1 point for recall.

3. Change the grow algorithm to “gini” and recompute the metrics from question 2. Does entropy build a more accurate classifier or gini? (5 points)

Reference: UCI Machine Learning Repository [Breast-cancer dataset]

End of CIND119 Assignment 1