CIND 119
Introduction to Big Data Analytics
Lab 3

1. Download **wine.csv** from your D2L course shell, under Course Materials, in the "Lab CSV Files" folder. Complete the following tasks:

    a. Read the file in SAS and display the contents using the **import** and **print** procedures.

    b. Develop a decision tree-based classification model using the **hpsplit** procedure of SAS.

    c. Navigate the contents of **Results View** by clicking on **HPSplit Wine Dataset**, and then by selecting **Model Assessment**. Examine the confusion matrix, fit statistics, and variable importance.

2. Download **mtcars.csv** from your D2L course shell, under Course Materials, in the "Lab CSV Files" folder. Complete the following tasks:

    a. Read the file in SAS and display the contents using **import** and **print** procedures.

    b. Develop a decision tree-based regression model using the **hpsplit** procedure of SAS. The model should **predict miles per gallon (mpg)** using the predictor variables.

    c. Navigate the contents of **Results View** by clicking on **HPSplit mpg regression tree** and then by selecting **Model Assessment**. Examine the fit statistics and variable importance.

Note: A tree regression is a decision tree model where the predicted outcome is a continuous variable. Similar to the Gini Index used in Module 4, we use *variance* for the tree regression as the split criterion. *Variance* is the same metric you learned in Module 1. You compute the variance of the outcomes of all data points assigned to a branch and the variable that produces the lowest variance is chosen as the next feature to be added to the tree.

In Lab 3, the mtcars.csv file has the outcome variable called "predict miles per gallon (mpg)" which is a continuous variable and hence, you should utilize a tree regression.

Reference: Agarwal, R. (2019, September 29). The simple math behind 3 decision tree splitting criterions. Medium.

End of CIND119 Lab 3

Ceni Babaoglu, PhD