**Big Data Analytics**

**Programming Assignments #4A and #4B**

For this project you will build decision trees in Spark environment using the APIs and your own program and compare the results. Perform the following tasks and submit the output as mentioned in each task. You will use CloudEra Spark release ([LINK-for-Info](https://www.cloudera.com/documentation/spark2/latest/topics/spark2.html)) for performing these tasks. Install a version of Spark on your Laptop and perform these tasks. Feel free to use any one of the permissible languages.

Problem #1 is to be submitted as Assignment #4A and is due by 9PM, April 8th, 2019.

Problem #2 is to be submitted as Assignment #4B and is due by 9PM, April 12th, 2019.

All submissions must be in the form of a single pdf file containing all information needed to be submitted.

1. (40) Use the MLLib API of Spark to construct a decision tree for the Breast Cancer Diagnostic data ([Data-Link1](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)) (we call it dataset1), available from the UC-Irvine ML repository. Select appropriate parameters to generate only a 3-level deep decision tree. Submit the following.
   1. Your program code.
   2. The choice of parameters and attribute selection metric (Gini index, info gain, etc.) used.
   3. Any assumptions made.
   4. Validation and Train/Test Strategy used.
   5. Decision tree Obtained.
   6. Performance shown by the confusion matrix.
2. (60) Now use your own code to build a decision tree in Spark. Model your algorithm based on the homework assignment you did for designing the decision tree learning algorithm. Use excactly the same parameter choices as used in (1.) above.
   1. Submit the same items (1.a-1.f) as for the question above.
   2. Reproduce the results from the 1.e an d 1.f from the previous question and compare with the outputs obtained by your algorithm.