BDA564 End-to-end Big Data Analytics
ASSIGNMENT#1
TERM#3 (2017-18)

Deadline: July 8, 2018 (Sunday, Midnight, 23:59)

General Instructions:

- Your assignment is mainly about to understand and develop the prepared model (Sample 5: Train, Test, Evaluate for Binary Classification: Adult Dataset) on Azure Machine Learning Studio.
- There are 7 questions in this assignment.
- Each question is worth 10 points for a total of 70 points.
- There is no deadline extension. If you do so, your assignment will be graded out of 35 points.
- There is no tolerance for plagiarism. You should be very careful about this and you must write every single sentence in your responses with your own words.

Part-A: Please explain the dataset in terms of its general importance, size, features, and main use. (10 points)

Part-B: Please explain the preprocessing (data preparation) stage briefly. What happens in the *clean missing data* and *select column in dataset*. Why are these steps relevant for the data processing? (10 points)

Part-C: Please explain the algorithm (Two-class Boosted Decision Tree) used in this model very briefly. Why do you think that this algorithm might have been chosen? (10 points)

Part-D: Please comment on the trained model. What is the output? In the visualization part of train model, what are those trees? (10 points)

Part-E: What are your comments related to the outputs of score model and evaluate model? You need to refer to the confusion matrix, accuracy, precision, recall. (10 points)

Part-F: Please construct a parallel pipeline beginning from split module with the same algorithm and test different parameters including maximum number of leaves, learning rate, and number of trees constructed. (10 points)

Part-G: Please suggest an alternative algorithm and provide a third pipeline for its execution. You are free to choose any algorithm but it should be a classification algorithm for two classes. (10 points)