BIG DATA HW2

1. The KDDCUP99 database, which is described in the attached file, has two kddcup.data.gz training file and a corrected.gz test file. Using RANDOM FOREST in SPARK, we want to obtain the ordering accuracy and runtime on four cores for the five main classes of DOS, R2L, U2R, probing and NORMAL in the following cases:

a: Draw two graphs of accuracy and runtime for the tree number 10-20-30-40-50 and the maximum tree depth of 10.

b: Repeat the best answer to Section A with a maximum depth of tree 20 with GINI or ENTROPY criteria.

c: Provide the following metric values for the best answer: Precision, Recall, F-measure, True Positive Rate, False Positive Rate.

2. The MNIST database, which is described in the appendix, is used to train and test a decision tree in SPARK:

a: Draw the accuracy graph by varying the depth of the tree from k = 3 to8.

b: Change the MaxBins parameter between 4, 8, 16, and 32 with the optimal tree depth. Plot the output accuracy graph according to this parameter and explain the changes in the accuracy of the results.