DATA MINING ASSIGNMENT 6
(Decision Trees)

Debarati Das

1PI13CS052

ANSWERS

1. **What is the difference (if any) between the two decision trees?**

The tree obtained from BFTree is binary while the other one is not Binary.
It can be observed that the BF Tree yields more leaf nodes and has a greater size than the other J48 tree.

2. **Why is there a difference?**

Standard algorithms such as ID3, C4.5 and CART for the top-down induction of decision trees expand nodes in depth-first order in each step using the divide-and-conquer strategy.The selection attribute in C4.5 is based on entropy gain in tree grown phase. A fixed order is used to expand nodes (normally, left to right) these decision trees. In Best-first decision trees the selection of best split is based on boosting algorithms which is used to expand nodes in best-first order instead of a fixed order. This algorithm uses the both the gain and gini index in calculating the best node in tree grown phase of the decision tree. This method adds the ”best” split node to the tree in each step. The best node is the node that maximally reduces impurity among all nodes available for splitting (i. e. not labeled as terminal nodes). Although this results in the same fully-grown tree as standard depth-first expansion, it enables us to investigate new tree pruning methods that use cross-validation to select the

number of expansions.

3. **Take one example that you create on your own and explain how each decision tree will be used to predict the class for your example**

Say we take a customer database with age, income,student, credit rating and buys computer as attributes. First we need to find an attribute to split on. This can be found by obtaining the Info(D) or Entropy of the Database and then the Info(Attribute). The gain can be calculated which is the difference between Info(D)-Info(Attribute). Whichever attribute has the highest Gain will be chosen as the Attribute to Split on. When the split yields a pure set, the decision making can stop.