Anmol Harsh 18CS10005
Ayudh Saxena 18CS10007

REPORT

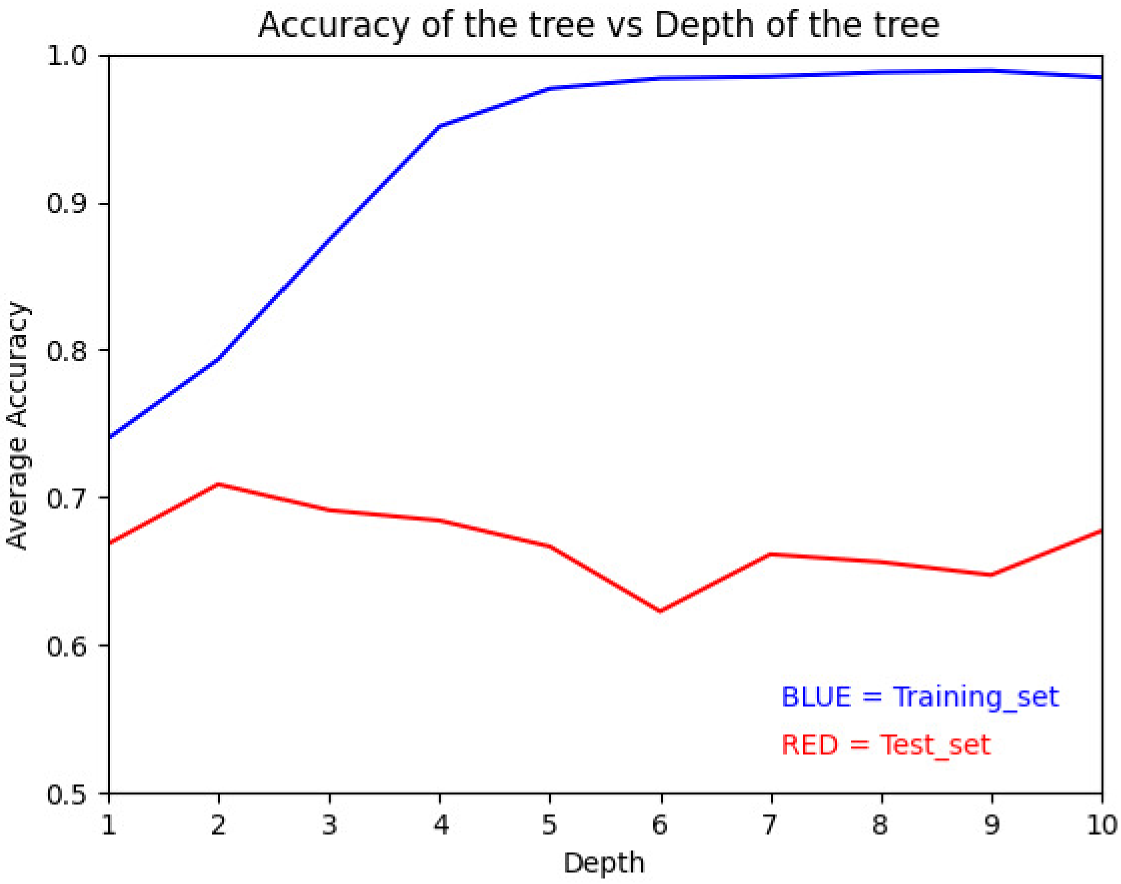
Assignment 1 : Decision Trees

- ## Procedure
  - Converted the .data file to .csv and read the data using the Python csv library.
  - Defined classes for tree nodes, attribute set etc.
  - Implemented the ID3 algorithm using information gain heuristic.
    - Tackled examples with missing attribute values by assigning the value which is most common in the examples having its target classification.
  - Implemented a best_depth function that takes in a dataset and returns a tree with highest test accuracy.
    - Iterated over possible depths (in this case, 10)
    - For each depth, shuffled the data set and then divided it in the ratio 60:20:20.
      - 1st part : Training set to train the tree,
      - 2nd part : Validation set which is used in pruning the tree
      - 3rd part : Test set to test the accuracy of the tree constructed
  - Used Reduced Error Pruning to prune the tree
    - Tested nodes from bottom to up (leaf to root)
    - If pruning the node, increases the accuracy on the validation set (obtained from the best_depth function), then the node is replaced with a leaf node having most common classification at that node.
  - Used the matplotlib library to plot the graph between Depth vs Accuracy on both test-set as well as training set
  - Implemented a function to write a graphviz file (.gv) for printing the tree.
    - Non-leaf nodes are colored purple.
    - Leaf Nodes are colored red/green
      - Red :- Non-recurrence event
      - Green :- Recurrence event

- Results

  - Step II : Finding the best depth

Depth	Average accuracy on test-set (%)	Accuracy on test-set of the best tree (%)	Accuracy on training set (%)
1	66.84	78.94	72.02
2	70.87	80.70	76.57
3	69.12	77.19	83.91
4	68.42	77.19	93.70
5	66.66	75.43	97.02
6	62.28	71.92	97.02
7	66.14	73.68	97.02
8	65.61	75.43	97.02
9	64.73	73.68	97.02
10	67.71	70.87	97.02



- Step III : Pruning the tree

  - The tree with maximum accuracy of of depth 2 is chosen for pruning
  - **Three** nodes are pruned.

Depth of the node pruned	Attribute name	Accuracy on validation set before pruning (%)	Accuracy after on validation set after pruning (%)
1	Tumor-size	70.87	71.92
1	Breast-quad	71.92	73.68
1	Tumor-size	73.68	75.43

  - Accuracy increases from 70.87% to 75.43%.