# Assignment-1 Part-2(Report)


**Team Members**

1) Rajan Jhaveri rjj160330

2) Varun Dani vxd162230


## Accomplishment, Learning and Best Results

The implemented ID3 algorithm does binary classification of the data provided to it and predicts the output label of the target class to predict which class a given instance belongs to and thus calculating the accuracy with which it predicted the class. The assumption here is that the data given as input will be binary and the labels to be associated with it will also be binary.

By implementing the ID3 algorithm we learnt the supervised machine learning method of **classification** and thus got clarity on how we train the machine with respect to the task (in this case classification given to it) and see the improved performance it gives by getting trained on more data. We also learned that if we try to go very deep down the tree using the training data, we get a great training accuracy and it seems pretty good at first but turns out, it is actually overfitting because we create a rule for each instance presented in the training set. Thus we have kept ***85%*** purity of the node as the threshold and turns out we can actually get better results on validation and testing data.

We used Depth First Traversal for printing the tree and for searching(predicting class labels for validation and testing instances) the tree we follow the path based on the value of the splitting attribute. By convention, we have used that for a value 0 of the attribute, we search the left subtree and the right subtree when the value of splitting attribute is 1.

It was a great task as we got to know how the ID3 Algorithm works in the background.

We got the best results on Data_Set1 with pruning factor 0.15 and on Data_Set2 with pruning factor 0.12

## Assumptions :

- ✓ We have assumed that we go deep down the tree till we get a node that is 85% pure. If the user wants to have more accuracy on the training data and wants to go deeper in the tree, the user can pass optional $5^{th}$ argument which will be considered as the purity required on each node.
- ✓ The nodes to be deleted during pruning should be in the lower half of the tree i.e. the node to be deleted should be at a level greater than or equal to half the value of total depth of the tree. We try to make sure that we don't prune a large part of the tree.
- ✓ The number of nodes to be deleted is not calculated for all the nodes to be deleted at the same time but a new number is generated only after one node is pruned. This way we can avoid a situation where in a node to be pruned was already pruned because it's parent was pruned.
- ✓ The max value of Pruning Factor can be 0.3 because it makes no sense to randomly prune more than 0.3 part of the tree because it would be effectively prune 0.5 to 0.6 part of the

tree approximately because all the subtrees of a node being pruned are also considered to be pruned.

**Screenshots:**

Taking inputs from the user along with the arguments

```
java ID3 training_set.csv validation_set.csv test_set.csv 0.05
```

Printing the decision tree

```
Pre Pruned Decision Tree

  XO = 0
  |   XM = 0
  |   |   XF = 0
  |   |   |   XB = 0
  |   |   |   |   XG = 0 : 0
  |   |   |   |   XG = 1
  |   |   |   |   |   XD = 0
  |   |   |   |   |   |   XS = 0 : 0
  |   |   |   |   |   |   XS = 1 : 1
  |   |   |   |   |   XD = 1 : 0
  |   |   |   XB = 1 : 0
  |   |   XF = 1 : 0
  |   XM = 1
  |   |   XB = 0
  |   |   |   XD = 0
  |   |   |   |   XG = 0
  |   |   |   |   |   XF = 0 : 0
  |   |   |   |   |   XF = 1
  |   |   |   |   |   |   XJ = 0
  |   |   |   |   |   |   |   XN = 0 : 1
  |   |   |   |   |   |   |   XN = 1
  |   |   |   |   |   |   |   |   XE = 0
  |   |   |   |   |   |   |   |   |   XK = 0 : 0
  |   |   |   |   |   |   |   |   |   XK = 1 : 1
  |   |   |   |   |   |   |   |   XE = 1 : 0
  |   |   |   |   |   |   XJ = 1 : 1
  |   |   |   |   XG = 1 : 1
  |   |   |   XD = 1
  |   |   |   |   XC = 0
  |   |   |   |   |   XF = 0
  |   |   |   |   |   |   XG = 0 : 0
  |   |   |   |   |   |   XG = 1
  |   |   |   |   |   |   |   XP = 0
  |   |   |   |   |   |   |   |   XS = 0 : 0
  |   |   |   |   |   |   |   |   XS = 1 : 1
  |   |   |   |   |   |   |   XP = 1 : 0
  |   |   |   |   |   XF = 1
  |   |   |   |   |   |   XJ = 0 : 1
```

```
|   |   |   |   |   |  XJ = 1
|   |   |   |   |   |  |  XE = 0
|   |   |   |   |   |  |  |  XG = 0
|   |   |   |   |   |  |  |  |  XI = 0 : 1
|   |   |   |   |   |  |  |  |  XI = 1 : 0
|   |   |   |   |   |  |  |  XG = 1 : 0
|   |   |   |   |   |  |  XE = 1 : 1
|   |   |   |  XC = 1 : 0
|   |   XB = 1
|   |   |  XI = 0 : 0
|   |   |  XI = 1
|   |   |   |  XC = 0
|   |   |   |  |  XK = 0
|   |   |   |  |  |  XP = 0 : 1
|   |   |   |  |  |  XP = 1
|   |   |   |  |  |  |  XS = 0
|   |   |   |  |  |  |  |  XG = 0 : 1
|   |   |   |  |  |  |  |  XG = 1
|   |   |   |  |  |  |  |  |  XF = 0 : 0
|   |   |   |  |  |  |  |  |  XF = 1 : 1
|   |   |   |  |  |  |  |  XS = 1 : 0
|   |   |   |  |  |  XK = 1 : 0
|   |   |   |  |  XC = 1 : 0
XO = 1
|   XI = 0
|   |   XM = 0
|   |   |  XQ = 0
|   |   |   |  XF = 0
|   |   |   |  |  XH = 0
|   |   |   |  |  |  XB = 0 : 0
|   |   |   |  |  |  XB = 1
|   |   |   |  |  |  |  XC = 0 : 1
|   |   |   |  |  |  |  XC = 1 : 0
|   |   |   |  |  XH = 1 : 1
|   |   |   |  XF = 1 : 0
|   |   |   XQ = 1
|   |   |   |  XJ = 0
|   |   |   |  |  XN = 0
|   |   |   |  |  |  XP = 0 : 1
|   |   |   |  |  |  XP = 1
|   |   |   |  |  |  |  XB = 0
|   |   |   |  |  |  |  |  XF = 0 : 0
|   |   |   |  |  |  |  |  XF = 1 : 1
|   |   |   |  |  |  |  XB = 1 : 0
|   |   |   |  |  XN = 1 : 0
|   |   |   |  XJ = 1 : 1
|   |   XM = 1
|   |   |  XQ = 0
|   |   |   |  XF = 0
|   |   |   |  |  XL = 0
|   |   |   |  |  |  XC = 0 : 1
|   |   |   |  |  |  XC = 1
|   |   |   |  |  |  |  XH = 0 : 1
|   |   |   |  |  |  |  XH = 1
|   |   |   |  |  |  |  |  XU = 0
|   |   |   |  |  |  |  |  |  XB = 0
|   |   |   |  |  |  |  |  |  |  XD = 0 : 1
|   |   |   |  |  |  |  |  |  |  XD = 1 : 0
|   |   |   |  |  |  |  |  |  XB = 1 : 0
|   |   |   |  |  |  |  |  XU = 1 : 1
```

```
|   |   |   |   |  XL = 1
|   |   |   |   |  |  XC = 0
|   |   |   |   |  |  |  XB = 0 : 0
|   |   |   |   |  |  |  XB = 1
|   |   |   |   |  |  |  |  XP = 0 : 0
|   |   |   |   |  |  |  |  XP = 1 : 1
|   |   |   |   |  |  XC = 1 : 1
|   |   |   |  XF = 1 : 0
|   |   |  XQ = 1 : 0
|   XI = 1
|   |   XT = 0
|   |   |  XH = 0
|   |   |   |  XP = 0
|   |   |   |  |  XF = 0 : 0
|   |   |   |  |  XF = 1
|   |   |   |  |  |  XQ = 0
|   |   |   |  |  |  |  XK = 0 : 1
|   |   |   |  |  |  |  XK = 1
|   |   |   |  |  |  |  |  XC = 0 : 0
|   |   |   |  |  |  |  |  XC = 1 : 1
|   |   |   |  |  |  XQ = 1
|   |   |   |  |  |  |  XK = 0 : 0
|   |   |   |  |  |  |  XK = 1 : 1
|   |   |   |  XP = 1
|   |   |   |  |  XS = 0
|   |   |   |  |  |  XD = 0
|   |   |   |  |  |  |  XC = 0
|   |   |   |  |  |  |  |  XJ = 0
|   |   |   |  |  |  |  |  |  XN = 0 : 0
|   |   |   |  |  |  |  |  |  XN = 1 : 1
|   |   |   |  |  |  |  |  XJ = 1
|   |   |   |  |  |  |  |  |  XB = 0 : 0
|   |   |   |  |  |  |  |  |  XB = 1
|   |   |   |  |  |  |  |  |  |  XG = 0 : 1
|   |   |   |  |  |  |  |  |  |  XG = 1 : 0
|   |   |   |  |  |  |  XC = 1 : 0
|   |   |   |  |  |  XD = 1
|   |   |   |  |  |  |  XM = 0
|   |   |   |  |  |  |  |  XC = 0 : 1
|   |   |   |  |  |  |  |  XC = 1 : 0
|   |   |   |  |  |  |  XM = 1 : 1
|   |   |   |  |  XS = 1 : 1
|   |   |   |  XH = 1
|   |   |   |  |  XJ = 0
|   |   |   |  |  |  XC = 0
|   |   |   |  |  |  |  XN = 0 : 1
|   |   |   |  |  |  |  XN = 1
|   |   |   |  |  |  |  |  XF = 0
|   |   |   |  |  |  |  |  |  XG = 0 : 1
|   |   |   |  |  |  |  |  |  XG = 1 : 0
|   |   |   |  |  |  |  XF = 1 : 0
|   |   |   |  |  XC = 1
|   |   |   |  |  |  XM = 0 : 0
|   |   |   |  |  |  XM = 1
|   |   |   |  |  |  |  XF = 0
|   |   |   |  |  |  |  |  XR = 0 : 1
|   |   |   |  |  |  |  |  XR = 1 : 0
|   |   |   |  |  |  |  XF = 1 : 1
|   |   |   |  XJ = 1 : 1
|   |   XT = 1
```

```
|   |   |  XS = 0
|   |   |   |  XQ = 0
|   |   |   |   |  XK = 0
|   |   |   |   |   |  XC = 0
|   |   |   |   |   |   |  XR = 0 : 1
|   |   |   |   |   |   |  XR = 1
|   |   |   |   |   |   |   |  XB = 0
|   |   |   |   |   |   |   |   |  XD = 0 : 0
|   |   |   |   |   |   |   |   |  XD = 1 : 1
|   |   |   |   |   |   |   |  XB = 1 : 0
|   |   |   |   |   |   |  XC = 1 : 1
|   |   |   |   |   |  XK = 1
|   |   |   |   |   |   |  XD = 0
|   |   |   |   |   |   |   |  XF = 0 : 0
|   |   |   |   |   |   |   |  XF = 1 : 1
|   |   |   |   |   |   |  XD = 1 : 0
|   |   |   |   |  XQ = 1
|   |   |   |   |   |  XM = 0
|   |   |   |   |   |   |  XN = 0
|   |   |   |   |   |   |   |  XU = 0 : 1
|   |   |   |   |   |   |   |  XU = 1 : 0
|   |   |   |   |   |   |  XN = 1 : 1
|   |   |   |   |   |  XM = 1 : 0
|   |   |   |  XS = 1
|   |   |   |   |  XL = 0
|   |   |   |   |   |  XD = 0
|   |   |   |   |   |   |  XU = 0 : 1
|   |   |   |   |   |   |  XU = 1
|   |   |   |   |   |   |   |  XB = 0
|   |   |   |   |   |   |   |   |  XE = 0 : 1
|   |   |   |   |   |   |   |   |  XE = 1
|   |   |   |   |   |   |   |   |   |  XC = 0 : 1
|   |   |   |   |   |   |   |   |   |  XC = 1 : 0
|   |   |   |   |   |   |   |  XB = 1
|   |   |   |   |   |   |   |   |  XG = 0
|   |   |   |   |   |   |   |   |   |  XH = 0 : 0
|   |   |   |   |   |   |   |   |   |  XH = 1 : 1
|   |   |   |   |   |   |   |   |  XG = 1 : 0
|   |   |   |   |   |   |  XD = 1 : 0
|   |   |   |   |  XL = 1
|   |   |   |   |   |  XH = 0
|   |   |   |   |   |   |  XD = 0
|   |   |   |   |   |   |   |  XQ = 0 : 0
|   |   |   |   |   |   |   |  XQ = 1
|   |   |   |   |   |   |   |   |  XB = 0 : 0
|   |   |   |   |   |   |   |   |  XB = 1 : 1
|   |   |   |   |   |   |  XD = 1
|   |   |   |   |   |   |   |  XB = 0 : 1
|   |   |   |   |   |   |   |  XB = 1 : 0
|   |   |   |   |   |  XH = 1 : 0
```

## Pre-pruned accuracy (Summary and results)

```
Pre-Pruned Accuracy
-----------------------------------

Number of training instances = 600
Number of training attributes = 20
Total number of nodes in the tree = 209
Number of leaf nodes in the tree =105
Accuracy of the model on the training dataset = 96


Number of validation instances = 2000
Number of validation attributes = 20
Accuracy of the model on the validation dataset before pruning = 75.95


Number of testing instances = 2000
Number of testing attributes = 20
Accuracy of the model on the testing dataset = 76.25
```

## Pruned Tree

```
Pruned Nodes =  96 114 58 156 19 89 147 125 137 9


Post Pruned Decision Tree

  XO = 0
  |   XM = 0
  |   |   XF = 0
  |   |   |   XB = 0
  |   |   |   |   XG = 0 : 0
  |   |   |   |   XG = 1
  |   |   |   |   |   XD = 0
  |   |   |   |   |   |   XS = 0 : 0
  |   |   |   |   |   |   XS = 1 : 1
  |   |   |   |   |   XD = 1 : 0
  |   |   |   XB = 1 : 0
  |   |   XF = 1 : 0
  |   XM = 1
  |   |   XB = 0
  |   |   |   XD = 0
  |   |   |   |   XG = 0
  |   |   |   |   |   XF = 0 : 0
  |   |   |   |   |   XF = 1 : 1
  |   |   |   |   XG = 1 : 1
  |   |   |   XD = 1
  |   |   |   |   XC = 0
  |   |   |   |   |   XF = 0
  |   |   |   |   |   |   XG = 0 : 0
  |   |   |   |   |   |   XG = 1
```

```
|   |   |   |   |   |   |  XP = 0
|   |   |   |   |   |   |   |  XS = 0 : 0
|   |   |   |   |   |   |   |  XS = 1 : 1
|   |   |   |   |   |   |  XP = 1 : 0
|   |   |   |   |   |  XF = 1
|   |   |   |   |   |   |  XJ = 0 : 1
|   |   |   |   |   |   |  XJ = 1
|   |   |   |   |   |   |   |  XE = 0
|   |   |   |   |   |   |   |   |  XG = 0
|   |   |   |   |   |   |   |   |   |  XI = 0 : 1
|   |   |   |   |   |   |   |   |   |  XI = 1 : 0
|   |   |   |   |   |   |   |   |  XG = 1 : 0
|   |   |   |   |   |   |   |  XE = 1 : 1
|   |   |   |   |  XC = 1 : 0
|   |   |  XB = 1
|   |   |   |  XI = 0 : 0
|   |   |   |  XI = 1
|   |   |   |   |  XC = 0
|   |   |   |   |   |  XK = 0
|   |   |   |   |   |   |  XP = 0 : 1
|   |   |   |   |   |   |  XP = 1
|   |   |   |   |   |   |   |  XS = 0
|   |   |   |   |   |   |   |   |  XG = 0 : 1
|   |   |   |   |   |   |   |   |  XG = 1
|   |   |   |   |   |   |   |   |   |  XF = 0 : 0
|   |   |   |   |   |   |   |   |   |  XF = 1 : 1
|   |   |   |   |   |   |   |  XS = 1 : 0
|   |   |   |   |   |  XK = 1 : 0
|   |   |   |   |  XC = 1 : 0
XO = 1
|   XI = 0
|   |   XM = 0
|   |   |  XQ = 0
|   |   |   |  XF = 0
|   |   |   |   |  XH = 0
|   |   |   |   |   |  XB = 0 : 0
|   |   |   |   |   |  XB = 1
|   |   |   |   |   |   |  XC = 0 : 1
|   |   |   |   |   |   |  XC = 1 : 0
|   |   |   |   |   |  XH = 1 : 1
|   |   |   |   |  XF = 1 : 0
|   |   |   |  XQ = 1
|   |   |   |   |  XJ = 0
|   |   |   |   |   |  XN = 0
|   |   |   |   |   |   |  XP = 0 : 1
|   |   |   |   |   |   |  XP = 1
|   |   |   |   |   |   |   |  XB = 0
|   |   |   |   |   |   |   |   |  XF = 0 : 0
|   |   |   |   |   |   |   |   |  XF = 1 : 1
|   |   |   |   |   |   |   |  XB = 1 : 0
|   |   |   |   |   |  XN = 1 : 0
|   |   |   |   |  XJ = 1 : 1
|   |   |  XM = 1
|   |   |   |  XQ = 0
|   |   |   |   |  XF = 0
|   |   |   |   |   |  XL = 0 : 1
|   |   |   |   |   |  XL = 1
|   |   |   |   |   |   |  XC = 0
|   |   |   |   |   |   |   |  XB = 0 : 0
|   |   |   |   |   |   |   |  XB = 1
```

```
|   |   |   |   |   |   |   |  XP = 0 : 0
|   |   |   |   |   |   |   |  XP = 1 : 1
|   |   |   |   |   |   |  XC = 1 : 1
|   |   |   |   |  XF = 1 : 0
|   |   |   |  XQ = 1 : 0
|   |  XI = 1
|   |   |  XT = 0
|   |   |   |  XH = 0
|   |   |   |   |  XP = 0
|   |   |   |   |   |  XF = 0 : 0
|   |   |   |   |   |  XF = 1 : 1
|   |   |   |   |  XP = 1
|   |   |   |   |   |  XS = 0
|   |   |   |   |   |   |  XD = 0 : 0
|   |   |   |   |   |   |  XD = 1
|   |   |   |   |   |   |   |  XM = 0 : 1
|   |   |   |   |   |   |   |  XM = 1 : 1
|   |   |   |   |   |  XS = 1 : 1
|   |   |   |  XH = 1
|   |   |   |   |  XJ = 0
|   |   |   |   |   |  XC = 0
|   |   |   |   |   |   |  XN = 0 : 1
|   |   |   |   |   |   |  XN = 1
|   |   |   |   |   |   |   |  XF = 0 : 1
|   |   |   |   |   |   |   |  XF = 1 : 0
|   |   |   |   |   |  XC = 1
|   |   |   |   |   |   |  XM = 0 : 0
|   |   |   |   |   |   |  XM = 1
|   |   |   |   |   |   |   |  XF = 0
|   |   |   |   |   |   |   |   |  XR = 0 : 1
|   |   |   |   |   |   |   |   |  XR = 1 : 0
|   |   |   |   |   |   |   |  XF = 1 : 1
|   |   |   |   |  XJ = 1 : 1
|   |   |  XT = 1
|   |   |   |  XS = 0
|   |   |   |   |  XQ = 0
|   |   |   |   |   |  XK = 0
|   |   |   |   |   |   |  XC = 0
|   |   |   |   |   |   |   |  XR = 0 : 1
|   |   |   |   |   |   |   |  XR = 1
|   |   |   |   |   |   |   |   |  XB = 0
|   |   |   |   |   |   |   |   |   |  XD = 0 : 0
|   |   |   |   |   |   |   |   |   |  XD = 1 : 1
|   |   |   |   |   |   |   |   |  XB = 1 : 0
|   |   |   |   |   |   |  XC = 1 : 1
|   |   |   |   |   |  XK = 1
|   |   |   |   |   |   |  XD = 0
|   |   |   |   |   |   |   |  XF = 0 : 0
|   |   |   |   |   |   |   |  XF = 1 : 1
|   |   |   |   |   |  XD = 1 : 0
|   |   |   |   |  XQ = 1
|   |   |   |   |   |  XM = 0
|   |   |   |   |   |   |  XN = 0
|   |   |   |   |   |   |   |  XU = 0 : 1
|   |   |   |   |   |   |   |  XU = 1 : 0
|   |   |   |   |   |   |  XN = 1 : 1
|   |   |   |   |   |  XM = 1 : 0
|   |   |   |  XS = 1
|   |   |   |   |  XL = 0
|   |   |   |   |  XD = 0
```

```
|   |   |   |   |   |   XU = 0 : 1
|   |   |   |   |   |   XU = 1
|   |   |   |   |   |   |   XB = 0
|   |   |   |   |   |   |   |   XE = 0 : 1
|   |   |   |   |   |   |   |   XE = 1
|   |   |   |   |   |   |   |   |   XC = 0 : 1
|   |   |   |   |   |   |   |   |   XC = 1 : 0
|   |   |   |   |   |   |   XB = 1
|   |   |   |   |   |   |   |   XG = 0
|   |   |   |   |   |   |   |   |   XH = 0 : 0
|   |   |   |   |   |   |   |   |   XH = 1 : 1
|   |   |   |   |   |   |   |   XG = 1 : 0
|   |   |   |   |   XD = 1 : 0
|   |   |   |   XL = 1
|   |   |   |   |   XH = 0
|   |   |   |   |   |   XD = 0
|   |   |   |   |   |   |   XQ = 0 : 0
|   |   |   |   |   |   |   XQ = 1
|   |   |   |   |   |   |   |   XB = 0 : 0
|   |   |   |   |   |   |   |   XB = 1 : 1
|   |   |   |   |   |   XD = 1
|   |   |   |   |   |   |   XB = 0 : 1
|   |   |   |   |   |   |   XB = 1 : 0
|   |   |   |   |   XH = 1 : 0
```

## Post pruned accuracy (Summary and results)

```
Post-Pruned Accuracy
------------------------------------

Number of training instances = 600
Number of training attributes = 20
Total number of nodes in the tree = 169
Number of leaf nodes in the tree =85
Accuracy of the model on the training dataset = 93


Number of validation instances = 2000
Number of validation attributes = 20
Accuracy of the model on the validation dataset After pruning = 77.0


Number of testing instances = 2000
Number of testing attributes = 20
Accuracy of the model on the testing dataset = 76.9
```