

CS 6375

ASSIGNMENT 2 (Part II)

Names of students in your group:

Muhammad Arsalan Malik (mxm162431)

Md Shihabul Islam (mxi170330)

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Lecture Notes

ID3-Example ppt

Python Documentation - <https://docs.python.org>

REPORT

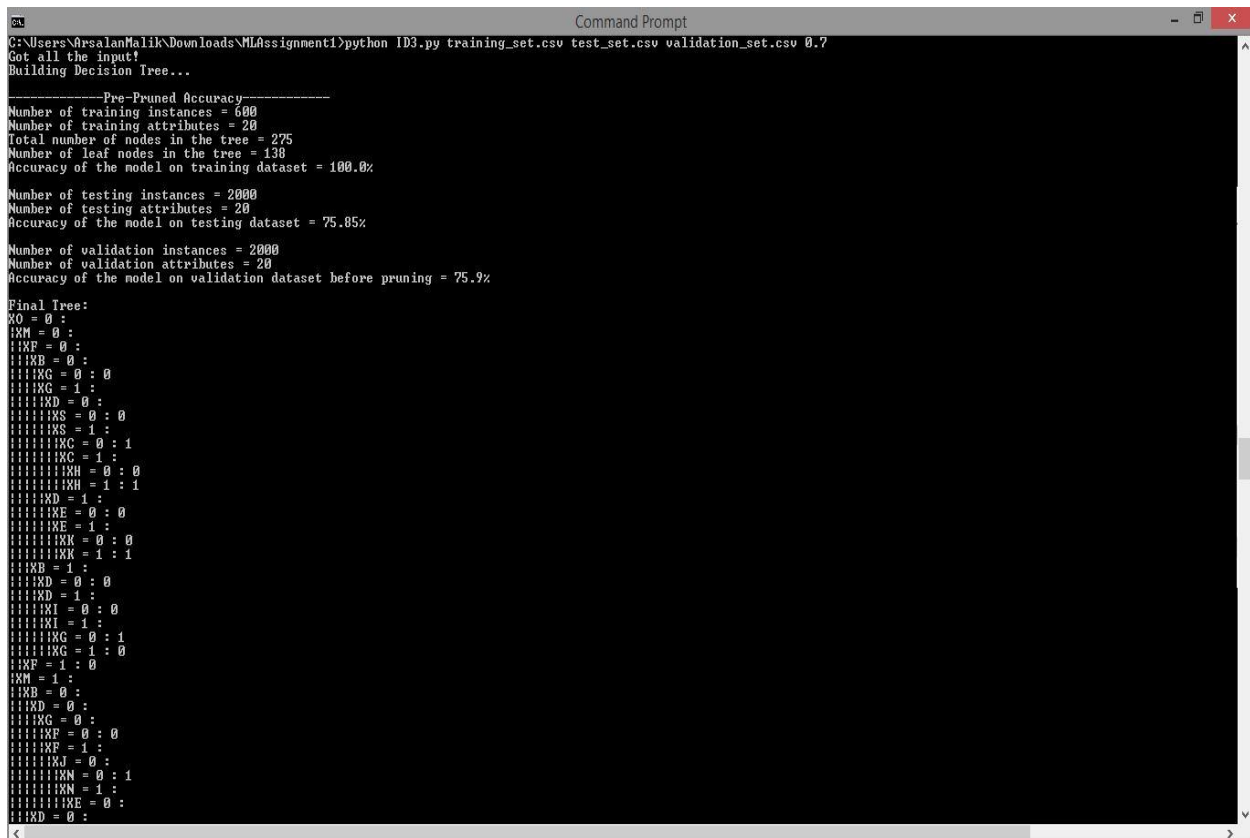
Assignment-2 (ii)

Assumptions:

- The attribute values in the data are assumed to be only Boolean values.
- No handling of missing data or blank columns is done.
- The program when run from the command line accepts 5 arguments in which first argument is the name of the main file, i.e., "ID3.py", the next three arguments are the absolute paths of the training, test and validation data sets respectively and last one is the pruning factor.
- Pruning of the nodes is done randomly.
- All the nodes, except root, are considered when selecting nodes for pruning.
- All the assumptions that are mentioned in the instruction's document for the assignment are followed.

Results:

Screenshots of one run of the program with pruning factor = 0.7



```
Command Prompt
C:\Users\ArsalanMalik\Downloads\MLAssignment1>python ID3.py training_set.csv test_set.csv validation_set.csv 0.7
Got all the input!
Building Decision Tree...

-----Pre-Pruned Accuracy-----
Number of training instances = 600
Number of training attributes = 20
Total number of nodes in the tree = 275
Number of leaf nodes in the tree = 138
Accuracy of the model on training dataset = 100.0%

Number of testing instances = 2000
Number of testing attributes = 20
Accuracy of the model on testing dataset = 75.85%

Number of validation instances = 2000
Number of validation attributes = 20
Accuracy of the model on validation dataset before pruning = 75.9%

Final Tree:
X0 = 0 :
  X1M = 0 :
    X1F = 0 :
      X1G = 0 :
        X1G = 0 : 0
        X1G = 1 :
          X1D = 0 :
            X1S = 0 : 0
            X1S = 1 :
              X1C = 0 : 1
              X1C = 1 :
                X1H = 0 : 0
                X1H = 1 : 1
              X1D = 1 :
                X1E = 0 : 0
                X1E = 1 :
                  X1K = 0 : 0
                  X1K = 1 : 1
                X1B = 1 :
                  X1D = 0 : 0
                  X1D = 1 :
                    X1I = 0 : 0
                    X1I = 1 :
                      X1G = 0 : 1
                      X1G = 1 : 0
                    X1F = 1 : 0
                    X1M = 1 :
                      X1B = 0 :
                        X1D = 0 :
                          X1G = 0 :
                            X1F = 0 : 0
                            X1F = 1 :
                              X1J = 0 :
                                X1M = 0 : 1
                                X1M = 1 :
                                  X1E = 0 :
                                    X1D = 0 :
```

```
Command Prompt
:::IXM = 1 :
:::IXE = 0 :
:::IXK = 0 : 0
:::IXK = 1 : 1
:::IXE = 1 : 0
:::IXJ = 1 :
:::IXC = 0 :
:::IXT = 0 :
:::IXL = 0 :
:::IXE = 0 :
:::IXI = 0 : 0
:::IXI = 1 : 1
:::IXE = 1 : 1
:::IXL = 1 : 0
:::IXT = 1 : 1
:::IXC = 1 : 1
:::IXG = 1 :
:::IXU = 0 : 1
:::IXU = 1 :
:::IXI = 0 : 0
:::IXI = 1 : 1
:::IXD = 1 :
:::IXC = 0 :
:::IXP = 0 :
:::IXG = 0 : 0
:::IXG = 1 :
:::IXP = 0 :
:::IXS = 0 : 0
:::IXS = 1 : 1
:::IXD = 1 : 0
:::IXF = 1 :
:::IXJ = 0 : 1
:::IXJ = 1 :
:::IXE = 0 :
:::IXG = 0 :
:::IXI = 0 : 1
:::IXI = 1 : 0
:::IXG = 1 : 0
:::IXE = 1 :
:::IXT = 0 :
:::IXG = 0 : 1
:::IXG = 1 : 0
:::IXT = 1 : 1
:::IXC = 1 : 0
:::IXB = 1 :
:::IXI = 0 : 0
:::IXI = 1 :
:::IXC = 0 :
:::IXK = 0 :
:::IXP = 0 : 1
:::IXP = 1 :
:::IXS = 0 :
:::IXG = 0 : 1
:::IXG = 1 :
:::IXF = 0 : 0
:::IXF = 1 : 1
:::IXS = 1 : 0
```

```
Command Prompt
:::IXK = 1 : 0
:::IXC = 1 : 0
XO = 1 :
IXI = 0 :
IXM = 0 :
IXQ = 0 :
IXF = 0 :
IXH = 0 :
IXB = 0 : 0
IXB = 1 :
IXC = 0 : 1
IXC = 1 : 0
IXH = 1 : 1
IXF = 1 : 0
IXQ = 1 :
IXJ = 0 :
IXM = 0 :
IXP = 0 : 1
IXP = 1 :
IXD = 0 :
IXF = 0 : 0
IXF = 1 : 1
IXB = 1 : 0
IXN = 1 : 0
IXJ = 1 :
IXL = 0 :
IXU = 0 : 0
IXU = 1 :
IXK = 0 :
IXU = 0 : 1
IXU = 1 : 0
IXK = 1 : 1
IXL = 1 : 1
IXM = 1 :
IXQ = 0 :
IXF = 0 :
IXL = 0 :
IXC = 0 : 1
IXC = 1 :
IXH = 0 : 1
IXH = 1 :
IXU = 0 :
IXB = 0 :
IXD = 0 : 1
IXD = 1 : 0
IXB = 1 : 0
IXU = 1 : 1
IXL = 1 :
IXC = 0 :
IXB = 0 : 0
IXB = 1 :
IXP = 0 : 0
IXP = 1 : 1
IXC = 1 : 1
IXF = 1 : 0
IXQ = 1 : 0
IXS = 1 : 0
```

```
Command Prompt
[XTI = 1 :
[XTI = 0 :
[IXH = 0 :
[IXP = 0 :
[IXP = 0 : 0
[IXP = 1 :
[IXQ = 0 :
[IXK = 0 : 1
[IXK = 1 :
[IXC = 0 : 0
[IXC = 1 : 1
[IXQ = 1 :
[IXK = 0 : 0
[IXK = 1 : 1
[IXP = 1 :
[IXS = 0 :
[IXD = 0 :
[IXC = 0 :
[IXJ = 0 :
[IXM = 0 : 0
[IXM = 1 : 1
[IXJ = 1 :
[IXD = 0 : 0
[IXB = 1 :
[IXG = 0 : 1
[IXG = 1 : 0
[IXC = 1 : 0
[IXD = 1 :
[IXM = 0 :
[IXC = 0 : 1
[IXC = 1 : 0
[IXM = 1 : 1
[IXS = 1 : 1
[IXH = 1 :
[IXJ = 0 :
[IXC = 0 :
[IXM = 0 : 1
[IXM = 1 :
[IXP = 0 :
[IXG = 0 : 1
[IXG = 1 : 0
[IXF = 1 : 0
[IXC = 1 :
[IXM = 0 : 0
[IXM = 1 :
[IXP = 0 :
[IXR = 0 : 1
[IXR = 1 : 0
[IXF = 1 : 1
[IXJ = 1 :
[IXS = 0 : 1
[IXS = 1 :
[IXG = 0 :
[IXB = 0 : 0
[IXB = 1 :
[IXD = 0 : 1
[IXS = 1 : 0
[IXS = 1 : 0
```

```
Command Prompt
[IXC = 0 : 1
[IXC = 1 :
[IXD = 0 : 1
[IXD = 1 : 0
[XTI = 1 :
[IXS = 0 :
[IXQ = 0 :
[IXK = 0 :
[IXC = 0 :
[IXH = 0 :
[IXH = 0 :
[IXE = 0 : 0
[IXE = 1 : 1
[IXH = 1 : 1
[IXB = 1 :
[IXB = 0 :
[IXD = 0 : 0
[IXD = 1 : 1
[IXB = 1 : 0
[IXC = 1 : 1
[IXK = 1 :
[IXD = 0 :
[IXP = 0 : 0
[IXP = 1 : 1
[IXD = 1 : 0
[IXQ = 1 :
[IXM = 0 :
[IXM = 0 :
[IXM = 0 : 1
[IXU = 1 : 0
[IXM = 1 :
[IXP = 0 :
[IXB = 0 :
[IXF = 0 : 0
[IXF = 1 : 1
[IXB = 1 : 1
[IXP = 1 : 1
[IXM = 1 : 0
[IXS = 1 :
[IXL = 0 :
[IXD = 0 :
[IXU = 0 : 1
[IXU = 1 :
[IXB = 0 :
[IXE = 0 : 1
[IXE = 1 :
[IXC = 0 : 1
[IXC = 1 : 0
[IXB = 1 :
[IXG = 0 :
[IXM = 0 : 0
[IXM = 1 : 1
[IXG = 1 : 0
[IXD = 1 :
[IXG = 0 : 0
[IXG = 1 : 1
[IXS = 1 : 0
```

```
Command Prompt

|||||XG = 1 : 1
|||||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-----
Number of training instances = 600
Number of training attributes = 20
Total number of nodes in the tree = 177
Number of leaf nodes in the tree = 89
Accuracy of the model on training dataset = 85.83333333333333%

Number of testing instances = 2000
Number of testing attributes = 20
Accuracy of the model on testing dataset = 68.35%

Number of validation instances = 2000
Number of validation attributes = 20
Accuracy of the model on validation dataset after pruning = 71.5%

Pruned Tree:
XO = 0 :
|XN = 0 :
| |XN = 0 :
| | |XG = 0 : 0
| | |XG = 1 :
| | |XD = 0 :
| | | |XS = 0 : 0
| | | |XS = 1 :
| | | |XG = 0 : 1
| | | |XG = 1 :
| | | |XN = 0 : 0
| | | |XN = 1 : 1
| | | |XD = 1 :
| | | |XE = 0 : 0
| | | |XE = 1 :
| | | |XK = 0 : 0
| | | |XK = 1 : 1
| | | |XB = 1 :
| | | |XD = 0 : 0
| | | |XD = 1 :
| | | |XI = 0 : 0
| | | |XI = 1 :
| | | |XG = 0 : 1
| | | |XG = 1 : 0
| | | |XF = 1 : 0
| | | |XN = 1 :
| | | |XS = 1 : 0
```

```
Command Prompt

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

```
Command Prompt

|||||IXD = 1 :
|||||IXM = 0 :
|||||IXC = 0 : 1
|||||IXC = 1 : 0
|||||IXM = 1 : 1
|||||IXS = 1 : 1
|||||IXH = 1 :
|||||IXJ = 0 :
|||||IXC = 0 :
|||||IXM = 0 : 1
|||||IXM = 1 :
|||||IXF = 0 :
|||||IXG = 0 : 1
|||||IXG = 1 : 0
|||||IXF = 1 : 0
|||||IXC = 1 :
|||||IXM = 0 : 0
|||||IXM = 1 :
|||||IXF = 0 :
|||||IXR = 0 : 1
|||||IXR = 1 : 0
|||||IXF = 1 : 1
|||||IXJ = 1 :
|||||IXS = 0 : 1
|||||IXS = 1 :
|||||IXG = 0 :
|||||IXD = 0 : 0
|||||IXD = 1 :
|||||IXD = 0 : 1
|||||IXD = 1 :
|||||IXE = 0 : 0
|||||IXE = 1 : 1
|||||IXG = 1 :
|||||IXC = 0 : 1
|||||IXC = 1 :
|||||IXD = 0 : 1
|||||IXD = 1 : 0
|||||IXT = 1 :
|||||IXS = 0 :
|||||IXQ = 0 :
|||||IXK = 0 :
|||||IXC = 0 :
|||||IXR = 0 :
|||||IXH = 0 :
|||||IXE = 0 : 0
|||||IXE = 1 : 1
|||||IXH = 1 : 1
|||||IXD = 1 : 0
|||||IXC = 1 : 1
|||||IXK = 1 :
|||||IXD = 0 :
|||||IXF = 0 : 0
|||||IXF = 1 : 1
|||||IXD = 1 : 0
|||||IXQ = 1 :
|||||IXM = 0 :
|||||IXS = 1 : 0
```

```
Command Prompt

|||||IXS = 0 :
|||||IXQ = 0 :
|||||IXK = 0 :
|||||IXC = 0 :
|||||IXR = 0 :
|||||IXH = 0 :
|||||IXE = 0 : 0
|||||IXE = 1 : 1
|||||IXH = 1 : 1
|||||IXR = 1 : 0
|||||IXC = 1 : 1
|||||IXK = 1 :
|||||IXD = 0 :
|||||IXF = 0 : 0
|||||IXF = 1 : 1
|||||IXD = 1 : 0
|||||IXQ = 1 :
|||||IXM = 0 :
|||||IXM = 0 :
|||||IXU = 0 : 1
|||||IXU = 1 : 0
|||||IXM = 1 :
|||||IXP = 0 :
|||||IXB = 0 :
|||||IXE = 0 : 0
|||||IXF = 1 : 1
|||||IXB = 1 : 1
|||||IXP = 1 : 1
|||||IXM = 1 : 0
|||||IXS = 1 :
|||||IXL = 0 :
|||||IXD = 0 :
|||||IXU = 0 : 1
|||||IXU = 1 :
|||||IXB = 0 :
|||||IXE = 0 : 1
|||||IXE = 1 :
|||||IXC = 0 : 1
|||||IXC = 1 : 0
|||||IXD = 1 :
|||||IXG = 0 :
|||||IXH = 0 : 0
|||||IXH = 1 : 1
|||||IXG = 1 : 0
|||||IXD = 1 :
|||||IXG = 0 : 0
|||||IXG = 1 : 1
|||||IXL = 1 :
|||||IXH = 0 :
|||||IXD = 0 :
|||||IXQ = 0 : 0
|||||IXQ = 1 :
|||||IXB = 0 : 0
|||||IXB = 1 : 1
|||||IXD = 1 : 1
|||||IXH = 1 : 0
```

Conclusion:

In the first run with pruning factor = 0.70, the accuracy of the training data before pruning is 100.00% with testing accuracy of 75.85% and validation set accuracy is 75.9% which clearly indicates that the model is suffering from “overfitting”. After pruning the training accuracy decreases to 85.83% with testing accuracy of 68.35%. There is a little decrease in the training data accuracy, which indicates that the nodes removed clearly were the reason for the model to overfit.

For pruning factor = 0.8, the training data accuracy changed to 79.33% and the accuracy of the testing data decreased very little, i.e., 73.6%, and validation accuracy became 74.2% which shows that the nodes pruned were very significant in overfitting the model.

For pruning factor = 0.90, the training accuracy decreased to 95.6%, while the testing data accuracy also decreased a little to 74.85% and validation set accuracy reduced to 75.55%.

As pruning factor increases, the test accuracy increases up to a limit and then it maintains almost constant value. i.e., as pruning factor is further increased after certain limit, there is no significant change in the accuracy of the test set and training set.