Sentiment classification of movie reviews using decision trees and forests

2. Decision tree can be stopped early by adding threshold on the $entropy$ (or $information\ gain$). When $entropy(x, y) \leq \mu$ then make leaf with majority label.
   In this process number of nodes decreases and also the height of the tree. Performance of the increases because the decision trees are prone to overfit the training set which is reduced by early stopping.
   $\mu = 0.1$ means set has impurity of $1/70$ which can be ignored.
   Here is the experimental data for μ, $nodes$, $accuracy$

μ	$nodes$	$accuracy$ (%)
0.2	811	69.5
0.5	585	73.3
0.8	141	72.3
0.9	53	65.4

3. $Accuracy$ seems to be decreasing and $nodes$ are increasing for larger noise after adding noise in the data but changes were not significant for smaller noise in the data.

$noise$ %	$nodes$	$accuracy$ %	$noise$ %	$nodes$	$accuracy$ %
0.5	813	69.7	5.0	833	69.7
1.0	809	69.1	10	817	67.6

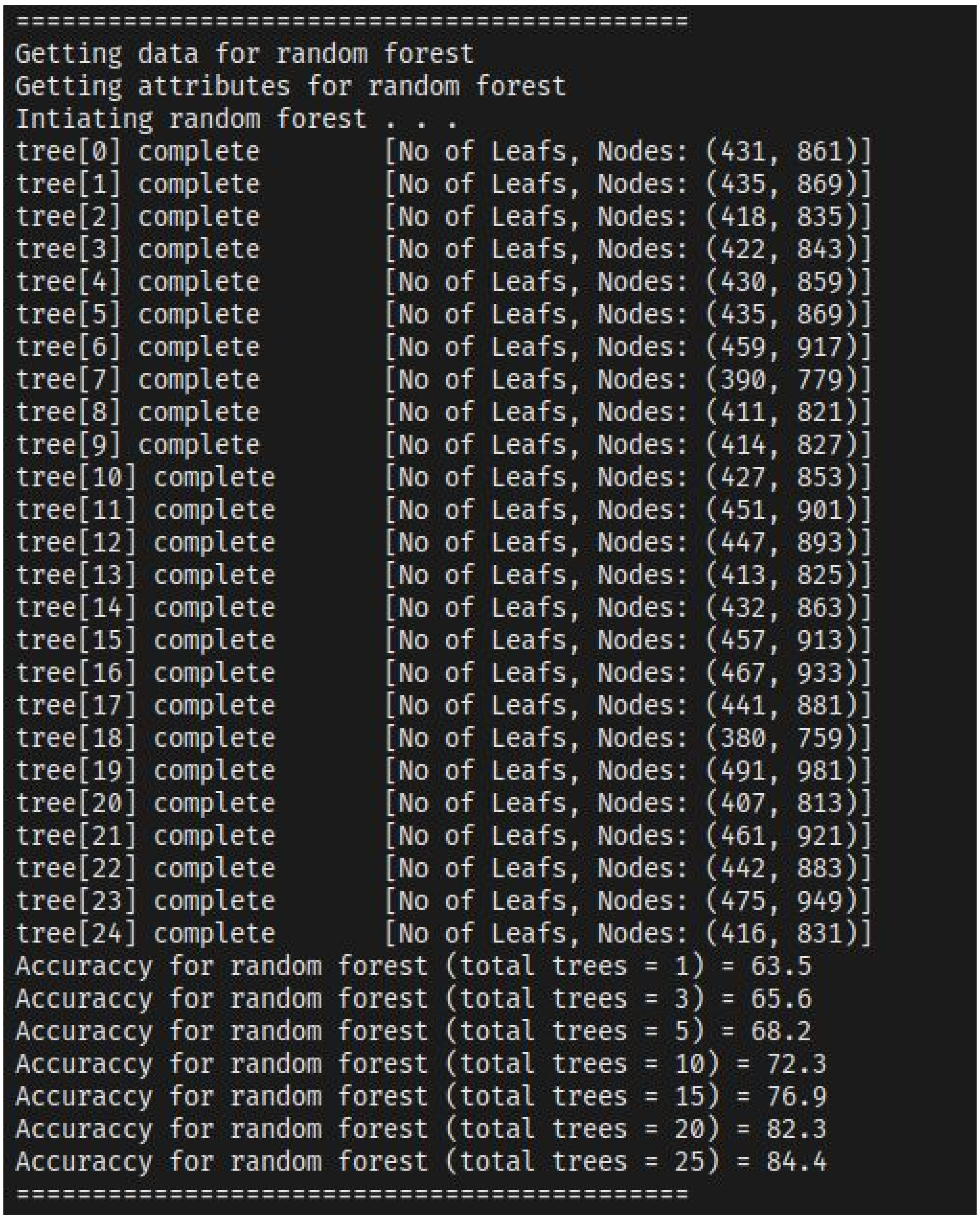
$*$ *nodes and accuracy at* 0.0 % *noise is* 811 *and* 69.5% *respectively*

4. There was significant changes in $accuracy$, $nodes$ and $height$. Number of $nodes$ were reduced to $1/4th$ and $height$ of the tree reduced to $half$ of previous height. $Test\ accuracy$ of the tree increased from 69.5% to 75.1%. $Training\ accuracy$ of the tree decreased from 89.3% to 82.8%.

5. Prediction $accuracy$ of the test set increased with number of trees in random forest. Here is the experimental data.

$no.\ of\ trees$	$accuracy$ %	$no.\ of\ trees$	$accuracy$ %
1	63.5	15	76.9
3	65.6	20	82.3
5	68.2	25	84.4
10	72.3		

As seen in above data $accuracy$ is increasing linearly with $no.\ of\ trees$ ($upto$ 20). After 20 relation is not longer linear.



```
============================================
Getting data for random forest
Getting attributes for random forest
Intiating random forest . . .
tree[0] complete        [No of Leafs, Nodes: (431, 861)]
tree[1] complete        [No of Leafs, Nodes: (435, 869)]
tree[2] complete        [No of Leafs, Nodes: (418, 835)]
tree[3] complete        [No of Leafs, Nodes: (422, 843)]
tree[4] complete        [No of Leafs, Nodes: (430, 859)]
tree[5] complete        [No of Leafs, Nodes: (435, 869)]
tree[6] complete        [No of Leafs, Nodes: (459, 917)]
tree[7] complete        [No of Leafs, Nodes: (390, 779)]
tree[8] complete        [No of Leafs, Nodes: (411, 821)]
tree[9] complete        [No of Leafs, Nodes: (414, 827)]
tree[10] complete       [No of Leafs, Nodes: (427, 853)]
tree[11] complete       [No of Leafs, Nodes: (451, 901)]
tree[12] complete       [No of Leafs, Nodes: (447, 893)]
tree[13] complete       [No of Leafs, Nodes: (413, 825)]
tree[14] complete       [No of Leafs, Nodes: (432, 863)]
tree[15] complete       [No of Leafs, Nodes: (457, 913)]
tree[16] complete       [No of Leafs, Nodes: (467, 933)]
tree[17] complete       [No of Leafs, Nodes: (441, 881)]
tree[18] complete       [No of Leafs, Nodes: (380, 759)]
tree[19] complete       [No of Leafs, Nodes: (491, 981)]
tree[20] complete       [No of Leafs, Nodes: (407, 813)]
tree[21] complete       [No of Leafs, Nodes: (461, 921)]
tree[22] complete       [No of Leafs, Nodes: (442, 883)]
tree[23] complete       [No of Leafs, Nodes: (475, 949)]
tree[24] complete       [No of Leafs, Nodes: (416, 831)]
Accuraccy for random forest (total trees = 1) = 63.5
Accuraccy for random forest (total trees = 3) = 65.6
Accuraccy for random forest (total trees = 5) = 68.2
Accuraccy for random forest (total trees = 10) = 72.3
Accuraccy for random forest (total trees = 15) = 76.9
Accuraccy for random forest (total trees = 20) = 82.3
Accuraccy for random forest (total trees = 25) = 84.4
============================================
```