CS 6375

ASSIGNMENT \_\_\_2\_\_\_\_\_\_\_\_\_

Names of students in your group:

1. Nikita Sah – nxs170831
2. Ramya Ramineni – rsr170230

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.

* <https://pandas.pydata.org/>
* <http://www.onlamp.com/pub/a/python/2006/02/09/ai_decision_trees.html?page=1>
* <http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf>
* <https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/>

PROJECT REPORT

Assumptions:

1. The dataset will only be in .xlxs format.
2. During pruning when random function generates index for rootNode to be pruned it will be counted in the count of number of nodes to be pruned but the node will not be pruned.

For example: If pruning factor is 0.1 and total number of nodes in the tree is 250. Theoretically, it should prune (0.1 \* 250) – 25 nodes. But random number might generate the index for root node 20 times and we end up pruning only 5 nodes.

Learning:

Details about the implementation of ID3 algorithm and reduced pruning algorithm.

Python implementation of nested dictionary tree.

Best Results:

Pre-Pruned Accuracy

-------------------------------------

Number of training instances = 600

Number of training attributes = 21

Total number of nodes in the tree = 265.0

Number of leaf nodes in the tree = 133

Accuracy of the model on the training dataset = 100.0 %

Number of validation instances 2000

Number of validation attributes 21

Accuracy of the model on the validation dataset 74.3 %

Number of testing instances 2000

Number of testing attributes 21

Accuracy of the model on the testing dataset 75.14999999999999 %

Post-Pruned Accuracy

-------------------------------------

Number of training instances = 600

Number of training attributes = 21

Total number of nodes in the tree = 102.0

Number of leaf nodes in the tree = 52

Accuracy of the model on the training dataset = 77.5 %

Number of validation instances 2000

Number of validation attributes 21

Accuracy of the model on the validation dataset 67.95 %

Number of testing instances 2000

Number of testing attributes 21

Accuracy of the model on the testing dataset 67.60000000000001 %