Intelligent Data Analysis – Spring 2020

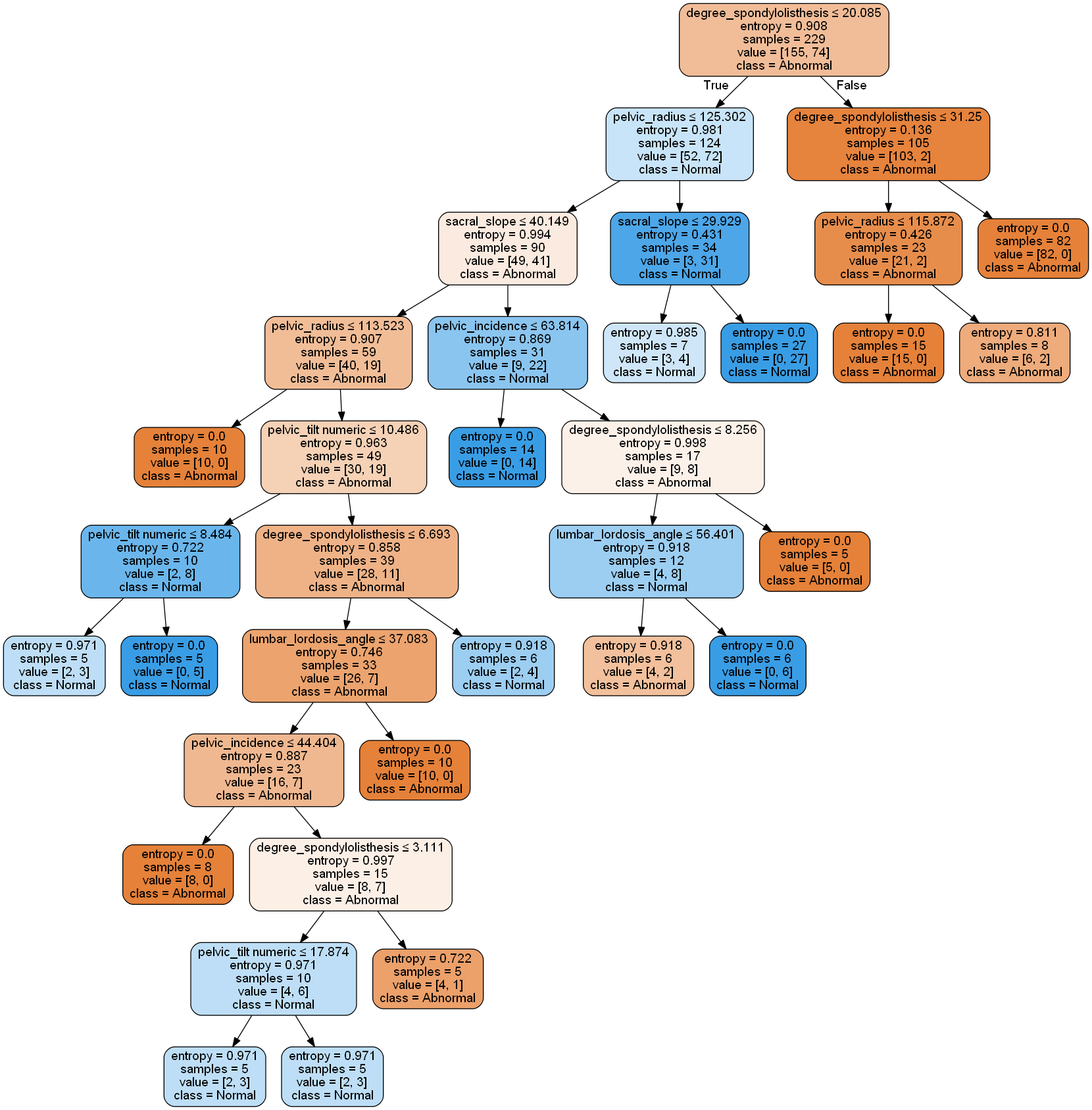
Homework #1

**Due Date: Feb. 5th, 2020, 10PM**

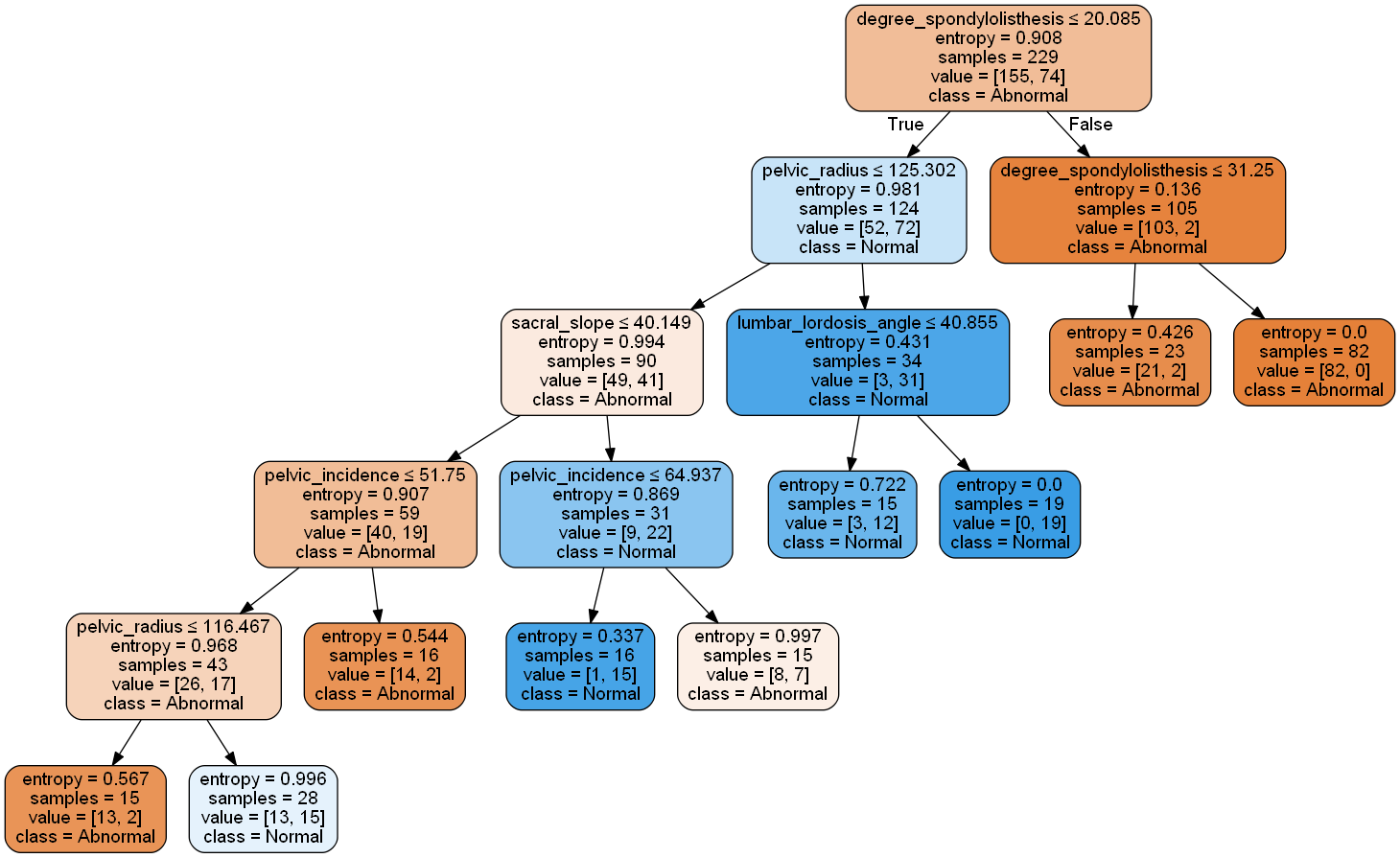
* Consider the two data files attached to this homework assignment. This data is taken from the UCI Machine Learning Repository and contains biomechanical features of some patients. The task is to predict whether the patient is normal or abnormal (call it Data2). The second dataset splits the abnormal group into two different diagnoses (call it Data3). Both datasets have same 310 feature vectors, six features, and a class column. Perform the following tasks with these datasets and submit the answers asked for in each task listed below.
* All your answers must be contained in a single pdf file. Upload this pdf file on Blackboard in response to this homework assignment.
* Use only one of the following two utilities: Scikit library function for decision trees or **fitctree** function of Matlab. Work done by any other toolbox will not be accepted.
* Each submission must be individual work of each student. Any plagiarism detected will be severely punished.

1. (30) Take Data2 and split it into randomly selected 230 training instances and remaining 80 as test instances. Create decision trees using the training set and the “minimum records per leaf node” values of 5, 15, 25, 40 and 50.
   1. Show the trees for all the five cases of “min record per node” values. Comment on what you see in a comparative analysis of the five trees. Just reporting the numbers is not enough; you must try to give an explanation of any differences in decision tree structures. Which of these five trees would you prefer to use and why?

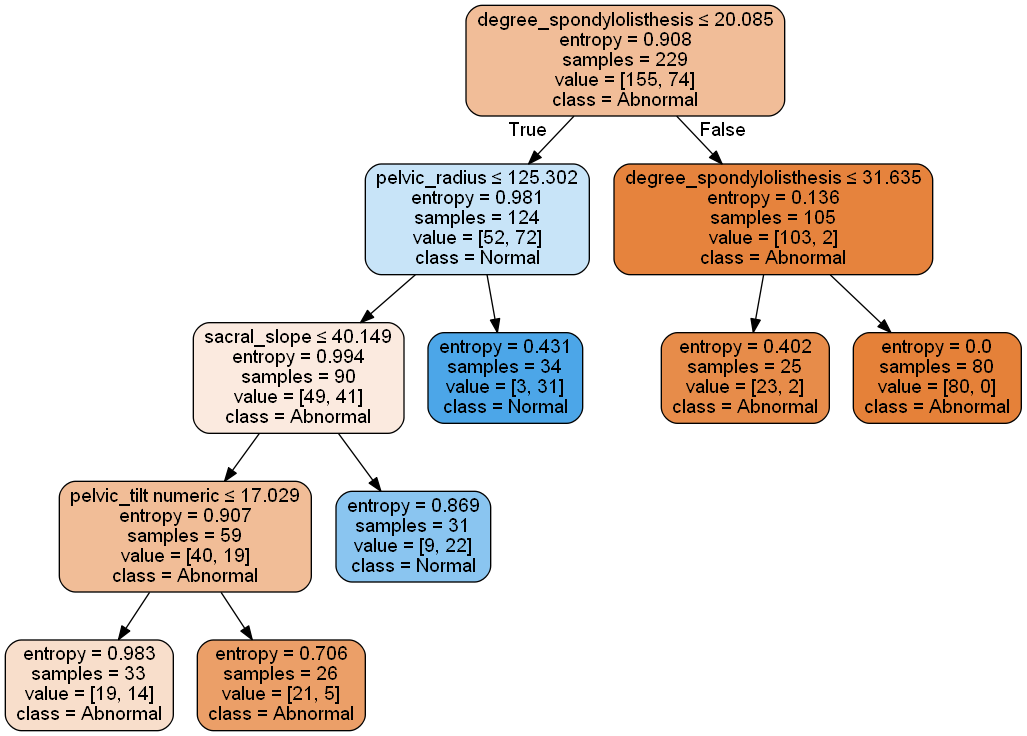
**DATA2\_5 min record per node**



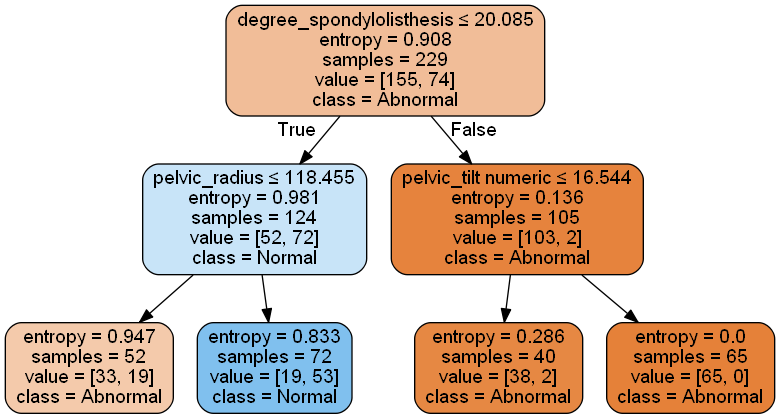
**DATA2\_15 min record per node**



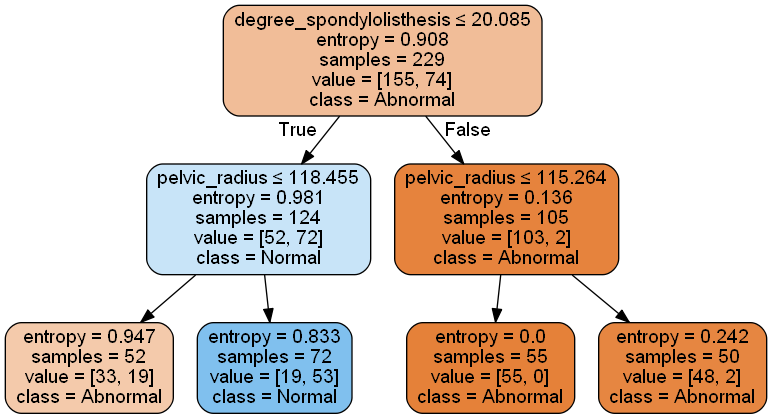
**DATA2\_25 min record per node**



**DATA2\_40 min record per node**

****

**DATA2\_50 min record per node**



The way I set decision tree is using DecisionTreeClassifier(criterion="entropy",min\_samples\_leaf=minleaf), by setting the min\_samples\_leaf variable in the function of DecisionTreeClassifier can decided the minimum records per leaf node. It can control the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min\_samples\_leaf training samples in each of the left and right branches.

As the number of records per leaf node increase, the depth of the tree decreases, and the gain start also increases. What’s more, after increasing the node number to a certain value, the tree structure will decrease. In conclusion, as the depth of a tree increases, the impurity per node will decrease until a certain value.

I prefer to choose the tree with 25 minimum nodes, it has the highest accuracy 0.8875. Also, it has the highest precision and recall.

Accuracy: 0.8625

Per-class precision score: [0.92156863 0.75862069]

recall: [0.87037037 0.84615385]

Accuracy: 0.85

Per-class precision score: [0.92 0.73333333]

recall: [0.85185185 0.84615385]

Accuracy: 0.8875

Per-class precision score: [0.90909091 0.84 ]

recall: [0.92592593 0.80769231]

Accuracy: 0.825

Per-class precision score: [0.87037037 0.73076923]

recall: [0.87037037 0.73076923]

Accuracy: 0.825

Per-class precision score: [0.87037037 0.73076923]

recall: [0.87037037 0.73076923]







* 1. For each of the five decision trees compute and report the accuracy, precision (for each class), and recall (for each class) values. Compare and comment on these values and show these values on a plot. Give your reasons for the observed trends/differences.

1. (30) Repeat the same tasks as done in Question-1 above for Data3 (Now the decision tree has three classes to work with). In addition to reporting results for parts (1a) and (1b) comment on the comparison of results obtained for (1a) and (2a) and also for (1b) and (2b).
2. (30) Take Data2 for this question. Use “1” to denote “Abnormal” and “0” to denote “Normal) class labels. Compute and report the correlations between each attribute (column) and the class label column. Drop the column that has the highest magnitude of the correlation value (It could be positive or negative). Now repeat that you did for Question number1 above.
   1. Repeat the task (1a) for the truncated dataset.
   2. Repeat the task (1b) for the truncated dataset.
   3. Compare and contrast the results of (1a) with those of (2a), and the results of (1b) with those of (2b).
3. (10) These 10 points are for good organization and presentation of results in your submission.