Intelligent Data Analysis – Fall 2018
Homework #1
Due Date: Sept. 17th, 2018, 9PM

- 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 **fitctree** function of Matlab (or any other tool box that you are familiar with). If you are not using Matlab then include in your submission the name of the toolbox that you use and an example of commands issued for generating the decision trees and other outputs. Do "help fitctree" within matlab to learn about this matlab function.

- 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 210 training instances and remaining 100 as test instance. Create decision trees using the training set and the "minimum records per leaf node" values of 3, 8, 12, 30, and 50.
    a. Show the trees for all the five cases of min record 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 the changes observed. Which of these five trees would you prefer to use and why?
    b. For each of the five decision trees compute and report the accuracy, precision, and recall values. Comment on the comparison of these values and show these values on a plot. Give your reasons for the observed trends/differences.
2. (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 (a) and (b) comment on the comparison of results obtained for (1a) and (2a) and also for (1b) and (2b). Give your analysis for the differences in results. Label this answer as 2c in your submission.
3. (30) Take Data2 for this question. Partition each column into four sets of equal width of values. Assign these intervals as values 0, 1, 2, and 3 and replace each value in the original data by its corresponding interval number.
    a. Show the boundaries for each interval for each attribute.
    b. Learn a decision tree with this transformed data and compute performance parameters in the same way as done for 1b and 2b.
    c. Compare the performance metric as obtained in 1b with those obtained here in 3b. Explain the differences in performance and give your intuitive reasons why these differences are observed.
4. (10) These 10 points are for good organization and presentation of results in your submission.