**P S G College of Technology**

**Dept of Applied Maths & Computational Sciences**

**VI Sem MSC(TCS) : Machine Learning Lab**

**Decision Trees**

1. Use the files zoo.csv (full data set with 101 records), zoo1.csv (training data set with 71 records) and zoo2.csv (testing data set with 30 records). Each record in these data sets has a number of characteristics that are used to classify zoological objects into a "type" (the last field) such as amphibian, bird, fish, reptile, etc.

a. Input the zoo1.csv to train the decision tree classifier and come up with a decision tree to classify a new record into one of the categories. Make sure you examine the data first and think about what field(s) to use for the classification scheme.

b. Use this to draw the full tree - how many leaves does it have? What is the classification accuracy on the training dataset?

c. Next, use the "fulltree" node and an analysis node to classify the records in the testing dataset, zoo2.csv.

d. What are the misclassified animals?

e. Suppose you wished to use a single level tree (i.e., 1R - just one attribute to classify) and you use the full data set (zoo.csv) to determine this. Which of the three attributes "milk", feathers" and "aquatic" yields the best results? Why do you think the results are so skewed in each case?