Assignment 1 (100 points)

Overview: Compare two decision tree induction methods on UCI datasets using (stratified) 10-fold cross-validation.

Classifiers:

1. C4.5 decision tree induction: use the Gain Ratio criterion for attribute selection. No pruning. No need to handle continuous attributes and missing data.
2. Random decision tree induction: select attributes randomly. No pruning. No need to handle continuous attributes and missing data.

   You may implement the decision tree induction algorithms by yourselves from scratch (not recommended), or by using APIs (Orange or Weka). It's perfectly fine if you do not implement anything and instead use existing tools/packages to perform the experiments.

   For cross-validation, you may either use a stratified version or a standard one. You may implement it by yourself or use existing packages. For your information, Weka has RandomTree and cross-validation built in, and has an option of no pruning.

Data sets:

From the UCI repository, http://archive.ics.uci.edu/ml/, choose as many as possible (**at least 10**) applicable datasets, i.e., the ones that are for the task of classification, with discreet attributes only, and no missing values (so that you do not need to handle continuous attributes and missing values).

If necessary, you may manually preprocess the datasets, e.g., change the class labels to integers of 1, 2, 3 … You may also change the format of the datasets so that your programs can read them properly.

Experiments:

For each dataset, use (stratified) 10-fold cross-validation and record the average tree height values as well as classification accuracy values.

Report:

Describe the details (e.g., implementation if any, software/packages used, parameter setting, etc.) of your work/experiments. Include all raw experiment results, and several screenshots as supporting evidence.

Use a table to present your experiment results. The table should include the following information: name of the dataset, number of instances, number of attributes, number of classes, C4.5 average height, C4.5 average accuracy, Random average height, and Random average accuracy.

Observe and report the tree heights and accuracy performance of C4.5 and Random. Would C4.5 generate shorter trees than Random on average? Which is more accurate on average? Provide a summative analysis on your experiment results to support your answers. There are no correct answers to these questions.

Submission: Submit a single zip file to TRACS including your code (if any) and report.