**CSCI316 Big Data Mining Implementation and Techniques**

**Laboratory 4**

**Objective**

* To understand how to implement Decision Tree classifier *from scratch*

(Note: “Implementation from scratch” means “not relying on any pre-implemented machine learning libraries”, but libraries such as NumPy, Pandas and SciPy can be used.)

**Decision Tree**

Review Decision Tree theory and implementation technique in Lecture 4. Implement a Decision Tree classifier for the Mushroom dataset “agaricus-lepiota.data” (source: <https://archive.ics.uci.edu/ml/datasets/Mushroom> ).

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. The dataset has one binary class (i.e., the first column) and 22 attributes (i.e., all other columns) and contains 8124 records. Note that attribute #11 contains missing values (denoted by “?”). In this task, you can treat the missing values as a new category of the attribute.

Requirements

* The implementation contains a tree induction function and a classification function, plus any auxiliary functions you like.
* The tree induction function performs *multi-splits* based on the categorial values of features. It employs the *Shannon Entropy* (e.g., InfoGain) as the split criterion.
* Use 70% data for training and 30% data for testing. Compute the training and testing errors.