Assignment 2: Data Mining

Due: Nov. 18

The attached two pairs of training and test sets are provided by the Design and Fashion Company of Atlanta (DFCA). Each record of the datasets represents a product, and it is composed of ordering venue of the product (Venue), color of the product (Color), Model of the product (Model), category of the product (Category), location that product was manufactured (Location), weight of the product (Weight), variety of the product (Variety), material used in building the product (Material), and volume of the order received from the clients (Volume).

All the attributes for the files of "Assignment 2--Training set for ID3" and "Assignment 2--Test set for ID3" are discretized by the DFCA authorities. Two attributes of Weight and Material are not discretized in the files of "Assignment 2--Training set for Bayes" and "Assignment 2--Test set for Bayes" and they must be treated as attributes with continuous values. The attribute Volume serves as the class attribute in both pairs of training and test sets.

The authorities of the DFCA want you to write a program to give them a choice of using either ID3 or Naïve Bayesian classifier. Your program must ask the user to select/enter the methodology.

The details for use of each methodology are as follows:

1- For ID3

a) Use the "Assignment 2--Training set for ID3" to extract rules and test the quality of the extracted rules against the "Assignment 2—Test set for ID3".
b) You need to apply both pre-pruning and post-pruning methods during the creation of the decision tree. Consider the file, let us call it F1, for a leaf generated from one of the unique values of a given attribute. Let c1 be the count of records with the dominant class in F1. The parameter α is the *mixture ratio* and it is defined as $\alpha = c1/|F1|$. If α is less than a threshold, T1, then F1 will not be split further and the class for all the records in F1 is considered to be the same as the dominant class.

   The value of T1 is decided by the user. In other words, your program must ask the user to select/enter a threshold value for the mixture ratio.
c) For post-pruning use the simple criteria provided in the class presentation:

   If $(N-M)/Q \leq gK$, then the subtree can be removed.

   N is the number of records in the Test set that are correctly classified by the rules extracted from the tree before removal of a subtree.

   M is the number of records in the Test set that are correctly classified by the rules extracted from the tree after removal of the subtree.

   g is a parameter, and $0 < g \leq 0.015$

   K is the total number of branches in the subtree

The value of g is decided by the user. In other words, your program must ask the user to select/enter a value for of g that is within the valid range.

2- Naïve Bayesian classifier

Use the "Assignment 2--Training set for Bayes" as the training set and classify the records of the "Assignment 2--Test set for Bayes"

3- Calculate Accuracy, Error Rate, Sensitivity, Specificity, and Precision for the selected classifier in reference to the corresponding test set.

4- **You must develop your own codes and you are not permitted to use any existing software developed by another entity.**