Assignment 1: Data Mining

Due: Oct. 26

The attached dataset is provided by the Golden Biotech Company of Jasper (GBCJ). Each record of the dataset represents a Petri dish kept in the Chemistry Lab of GBCJ. Each record is composed of four attributes as follows:

Attribute A1 represents the number of different microscopic residue of an organic element on the dish. Attribute A2 represents the length (in micron) of the tallest residue in the dish and attribute A3 represents the hue saturation of the largest residue after being exposed to phosphoric acid. The last attribute is Class that represents some grouping of organic elements.

The authorities of the GBCJ want you to write a program to:

1- Discretize the attribute A1 using Entropy-Based discretization

When either the condition "a" or condition "b" is true for a partition, then that partition stops splitting:

a- The number of distinct classes within a partition is 1.

b- The ratio of the minimum to maximum frequencies among the distinct values for the attribute Class in the partition is <0.5 and the number of distinct values within the attribute of Class in the partition is Floor(n/2), where n is the number of distinct values in the original dataset.

(**Note:** Your program must work for any attribute that is made-up of integer values)

2- Discretize the attribute A2 using Segmentation by Natural Partitioning

To get the attribute values between 5 to 95 percentiles, simply (i) sort the data in ascending order and (ii) keep values from Floor(n*0.05) to Floor(n*0.95), where n is the number of values in the dataset.

(**Note:** Your program must work for any attribute that is made-up of continuous numbers)

3- Calculate the correlation between A1 and A3 and remove A3, if correlation is >0.6 or correlation is <-0.6

(**Note:** Your program must work for any two numeric attributes).

4- (Optional) Apply Principal Component Analysis on the dataset, convert it into a new dataset, and save the new dataset.

5- **You are not permitted to use any existing software to complete this assignment.**