National Institute of Technology Calicut
Department of Computer Science and Engineering
CS4038 Data Mining -Assignment 3

Instructions 14/ 03/ 2017

Run these questions in your selected/approved dataset.

Modifications will be suggested at the time of evaluation.

Reduce your data size to 30,000 tuples and perform the following operations.

1) Write a program to find the percentage of data filled under each attribute. Print the result with attribute name and the corresponding percentage value.

2) From Q1, for all percentage values less than 50 %, write a program to fill the missing values using Naive Bayesian Classification.

3) Discretize the numeric values using appropriate method. (Use domain knowledge and decide)

4) Standardize the numeric value attribute into a nearest integer value.

5) From Q4, if A-n is the first numeric attribute, write a program to form a binary tree using the first 99 tuples in A-n.

6) Using scikit learn Visualize the Data dispersion between A-3 (3^{rd} attribute) and A-7 (7^{th} attribute).