CMP4336 – Introduction to Data Mining

Homework 1

Deadline: April 4, 2018 till 23:59 (strict deadline, no extension!)

The dataset given in the following link consists of 285 instances and 18 attributes.

https://archive.ics.uci.edu/ml/datasets/University

Write a program that performs the following tasks on the above-mentioned dataset:

1) Replace the missing values using one of the methods we have discussed in the lecture hour.
2) Calculate the mean, standard deviation, mode, and skewness of all numerical attributes and report them.
3) Find the most frequent value of each categorical variable.
4) Plot the probability distribution of both continuous and random variables.
5) Using "academic-emphasis" attribute as the class variable, plot the scatter plots of each pair of attributes.
6) Compute the distance matrix using Euclidean distance.
7) Compute the distance matrix using Mahalonobis distance.
8) Choose one of the discretization methods we have discussed in the lecture and discretize all numerical attributes using that method.

Guidelines

1. Use Python, R, or MATLAB.
2. Do not use the built-in functions for statistical parameter computations, write the functions yourself.
3. Submit a single pdf file which includes the response and required output for each of the tasks given above and the source code you have written.
4. Submission will be made through itslearning, NOT e-mail.