CSE 477: Data Mining

Assignment 02 (Exploring a Dataset)

For this assignment, you need to explore a dataset and find some insights about the dataset.

The dataset **adult.data** is posted in your course folder. bit.ly/ewucse477

The semantics of this dataset is also given in another file – **adult.names**

You must submit a report (hardcopy) and your code file in the following link:
http://bit.ly/ewucse477codesubmit

Deadline of report submission is: **30 May 2019 by 12:00 PM**

The report must contain the answer of these questions:

- How many records are there?
- How many features are there?
- How many features are continuous, and how many are nominal?
- For the continuous features, what are the average, median, maximum, and minimum values? What is the standard deviation?
- For the continuous features, use appropriate plotting tool to make 2-dimensional scatter plots of two features at a time. What relationships can you find? *You must present at least 2 scatter plots; one with positive correlation and another with negative correlation.*
- Find dissimilarity between two records. You must find the dissimilarity between record 1 and record XXX where XXX is the last three-digit of your id. *If last three digit of your id is 001 then you must compare record 002 and 001.*