CSAL4243 – Assignment 2

You can use any programming language to perform the experiments. There is no restriction of using only Python or Matlab or C/C++ for the task(s).

Task 1:
The program you developed for Naïve Bays Classifier (NBC) in Assignment-1 is specific for the given simple dataset of 15 samples only.

You are now required to:

Make the program code of the NBC generic so that it can accept any matrix of dataset (in CSV file) as long as the columns represent **features** and last column is the **outcome**. There can be any number of samples and you have to take care of the m-estimate also.

You will be using the following sample datasets:

1. Iris Flower - https://archive.ics.uci.edu/ml/datasets/Iris
2. Thyroid Disease - https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease
3. Breast Cancer - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Instructions for use of datasets:

- Handle the problems of missing values.
- Shuffle the dataset before using in training or testing.
- Use 90% of the dataset for training and 10% for testing.
- Convert the '*.data' files downloaded from UCI Repository to '*.csv'.

Marking Scheme:

1. Read and load data into respective variables. [3]
2. Design functions to Calculate Probabilities and Likelihoods. [3]
3. Ask user for input data and predict outcome. [3]
4. Design a GUI for the program. [1]

Total: [10]