IE406 Applied Machine Learning:
Spring 2022

Assignment II
(Due Date: 11:59 PM on May 16, Monday)

A famous collection of data on whether a patient has diabetes, known as the Pima Indians dataset, and originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases can be found at Kaggle. Download this dataset from https://www.kaggle.com/kumargh/pimaindiansdiabetescsv. This dataset has a set of attributes of patients and a categorical variable telling whether the patient is diabetic or not. For several attributes in this dataset, a value of 0 may indicate a missing value of the variable. There is a total of 767 data points.

Build a simple naïve Bayes classifier to classify this dataset. You will use 20% of the data for testing and the other 80% for training. You should write this classifier yourself. Submit your Python code and your answers for the accuracy of the classifier on the 20% test data, where accuracy is the number of correct predictions as a fraction of total predictions. If you use Laplace smoothing, mention it in your answers.

Source: https://courses.engr.illinois.edu/cs498aml/