You can use sklearn and Pandas or any other Python package for this lab.

1. **Breast Cancer Coimbra Data Set**[1]

   Clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls. There are 9 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

   (a) Load the `dataR2.csv` file from the Dropbox folder. It contains a slightly modified version of Breast Cancer Coimbra Data Set from: `https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra`. (5 pts)

   (b) Exploratory data analysis:

      i. Make a scatterplot matrix of the features in the dataset. Use color to show Classes 0 and 1.[2] (10 pts)

      ii. Select the first 40 rows of Class 0 and the first 48 rows of Class 1 as the training set and the rest of the data as the test set. (10 pts)

   (c) Classification using KNN on Breast Cancer Coimbra Data Set

      i. Use sklearn's KNN method to classify training data for $k \in \{88, 87, 86, \ldots, 3, 2, 1\}$ (i.e. in reverse order). Use majority polling and Euclidean distance. Plot training misclassification error rate (the percentage of training data that are misclassified) for each k. (25 pts)

      ii. Use sklearn's KNN method to classify test data for $k \in \{88, 87, 86, \ldots, 3, 2, 1\}$ (i.e. in reverse order), using the model you developed in 1(c)i.[3] Use majority polling and Euclidean distance. Plot test misclassification error rate (the percentage of training data that are misclassified) for each k on the same plot as the training misclassification error rate. Make sure to plot in reverse order of k. (25 pts)

      iii. Which k^* is the most suitable k among those values? (5 pts)

   Let us further explore some variants of KNN.

   (d) Replace the Euclidean metric with the Minkowski distance[4] with parameter p and calculate the test error for all 440 choices of $k \in \{88, 87, 86, \ldots, 3, 2, 1\}$ and $p \in \{1, 2, 3, 4, 5\}$. Find the pair (k^*, p^*) for which the misclassification error rate on the test set is the smallest and report it. (20 pts)

[1] This Lab is assigned in October in Fall semesters, which is Breast Cancer Awareness Month. Please help in raising awareness about breast cancer. `https://www.nationalbreastcancer.org/breast-cancer-awareness-month`

[2] See the example `https://plotly.com/python/splom/`

[3] Attention: DO NOT use test data to train a new model. You must predict the label of test data using training data only, and compare the labels predicted by training data with the actual labels of the test data to calculate the test misclassification error. Otherwise, there is no point in using training data!

[4] You can use sklearn.neighbors.DistanceMetric.

2. **Extra Credit (Weighted KNN)** (20 pts)

   (a) The majority polling decision can be replaced by weighted decision, in which the weight of each point in voting is *inversely proportional* to its distance from the query/test data point. In this case, closer neighbors of a test point will have a greater influence than neighbors which are further away. Calculate the test error for all 440 choices of $k \in \{88, 87, 86, \ldots, 3, 2, 1\}$ and $p \in \{1, 2, 3, 4, 5\}$. Find the pair (k^*, p^*) for which the misclassification error rate on the test set is the smallest and report it..

Note: You must submit Extra Credit along with your main homework. If you submit Extra Credit Late, your whole homework will be late.