1. **Breast Cancer Diagnosis**[1]

   The goal of this Lab is to diagnose breast cancer using the features extrracted from digital images of Fine Needle Aspirates (FNA) of a breast mass. You are provided a sample of images of benign and malignant cases, just to become familiar with the original images. You will work on numerical features extracted from those images. You will work with the diagnostic data set, WDBC. There are 32 attributes in the data set: the first attribute is a patient ID, the second is diagnosis (B for Benign, M for Malignant). The other 30 attributes are the features that you will work with to build a diagnosis tool for breast cancer.

   Ten real-valued features are calculated for each nucleus in the digital image of the FNA of a breast mass.[2] They are:

   - radius (mean of distances from center to points on the perimeter)
   - texture (standard deviation of gray-scale values)
   - perimeter
   - area
   - smoothness (local variation in radius lengths)
   - compactness (perimeter^2 / area - 1.0)
   - concavity (severity of concave portions of the contour)
   - concave points (number of concave portions of the contour)
   - symmetry
   - fractal dimension "coastline approximation" - 1)

   Ther mean, standard deviation, and the mean of three largest values for each image has been computed, to represent each image using 3×10 features. There are 569 instances in the data set, 212 of which are malignant, and 357 are benign.

   (a) Download the WDBC data from: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)` .

   (b) Choose the first 30 malignant cases and the first 50 benign cases in the data set as the test set and the rest as the training set.

   (c) Binary Classification Using Logistic Regression[3]

      i. Depict scatter plots of the features in your training set in a scatter matrix. (See p. 129 of the textbook).

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

[2] For more details see: `https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis`.

[3] Some logistic regression packages have a built-in $\mathscr{L}_2$ regularization. To remove the effect of $\mathscr{L}_2$ regularization, set $\lambda = 0$ or set the budget $C \to \infty$ (i.e. a very large value).

ii. Use logistic regression[4] to solve the binary classification problem. Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets.

iii. Calculate the p-values for your logistic regression parameters and prune those variables that are not statistically significant. Refit a logistic regression model using your pruned set of features.[5] Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets.

iv. Do your classes seem to be well-separated to cause instability in calculating logistic regression parameters?

2. **Breast Cancer Prognosis**

The goal of this Lab is to determine the prognosis of breast cancer patients using the features extrracted from digital images of Fine Needle Aspirates (FNA) of a breast mass. You will work with the prognostic data set, WPBC. There are 34 attributes in the data set: the first attribute is a patient ID, the second is an outcome variable that shows whether the cancer recurred after two years or not (N for Non-recurrent, R for Recurrent), the third variable is also an income variable that shows the time to recurrence. There are 30 attributes that are similar to those you used in the diagnostic data set. Additionally, the diameter of the excised tumor in centimeters and the number of positive axillary lymph nodes are also given in the data set.

Important Note: Time to recurrence (third attribute) should *not* be used for classification, otherwise, you will be able to perfectly classify!

There are 198 instances in the data set, 151 of which are nonrecurrent, and 47 are recurrent.

(a) Download the WPBC data from: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.

(b) Select the first 130 non-recurrent cases and the first 37 recurrent cases as your training set. Add record #197 in the data set to your training set as well.

(c) There are four instances in your training set that are missing the lymph node feature (denoted as ?). This is not a very severe issue, so replace the missing features with the median of the lymph node feature in *your training set*.

(d) Binary Classification Using Naïve Bayes' Classifiers

i. Solve the problem using a Naïve Bayes' classifier. Use Gaussian class conditional distributions. Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets.

[4]If you encountered instability of the logistic regression problem because of linearly separable classes, modify the Max-Iter parameter in logistic regression to stop the algorithm immaturely and prevent from its instability.

[5]R calculates the p-values for logistic regression automatically. One way of calculating them in Python is to call R within Python. There are other ways to obtain the p-values as well.

ii. This data set is rather imbalanced. Balance your data set using SMOTE, by downsampling the common class in the training set to 90 instances and upsampling the uncommon class to 90 instances. Use $k = 5$ nearest neighbors in SMOTE. Remember not to change the balance of the test set. Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets. Does SMOTE help?

iii. (Extra Credit, 10 points) Repeat 2(d)i and 2(d)ii using multinomial priors.[6]

(e) (Extra Credit, 15 points) Solve the regression problem of estimating time to recurrence (third attribute) using the next 32 attributes. You can use KNN regression. To do it in a principled way, select 20% of data points each class in your training set to choose the best $k \in 1, 2, \ldots, 20$, and the rest 80% as the *new training set*. Report your MSE on the test set using the k you found and the whole training set (not only the new training set!). For simplicity, use Euclidean Distance. Repeat this process when you apply SMOTE to your new training set to only upsample the rare class and make the data completely balanced. Does SMOTE help in reducing the MSE?

[6]We briefly covered them in the lecture without using the term multinomial. Research what they mean.