Lab 1
EC9630: Machine Learning
Chapter 3: Bayesian Pattern Classification
Duration: 3 Hours
Introducing Machine learning datasets

In this lab you will learn KNN classification algorithm.

1. Load the breast cancer dataset (Breast Cancer Wisconsin (Diagnostic) Data Set) from *scikit-learn* datasets module.
2. Laern the data. Find the,
   - shape of the data.
   - Sample count per class.
   - Features in the dataset.
   - Other properties of the dataset.
3. Divide your data into two sets, feature values **(X)** and target values **(y)**.
4. Now fit a nearest neighbor model with 5 nearest neighbors.
   - Fit that model on the whole data (train the model on whole data).
   - Test your model on the same data (No test, train split)
   - Print the score of your model.
   - Now divide the whole data into 80% train and 20% test.
   - Train the same model (5 nearest neighbors) on training data and test your trained model on the unseen test data.
   - Print the current score and compare it with the old one.
5. Do the following experiment on the split data. Change the number of neighbors from 1 to 8 and see how training accuracy and testing accuracy are changing with number of neighbors.
6. Plot both the accuracy values on the same graph and see the changes.

7. Select the best number of neighbors from the above graph.
8. Fit a nearest neighbor model with that value and print the score.
9. Calculate the root mean squared value of your model.
10. Plot a confusion matrix and interpret the result.

At the end of this lab, you have to submit a lab report. It will be marked for 100 Marks.