CSCI 6600/7350
Homework 5 (Due midnight, Saturday Nov 14th)

In this homework, we empirically study feature selection/dimensionality reduction approaches using the iris data and

1. Apply the following feature selection methods to pick the top 2 features for the iris data (https://archive.ics.uci.edu/ml/datasets/iris):
    a. Signal to noise ratio
    b. Two tailed t-test
    c. Relief, assuming the two randomly selected objects are:
        i. 4.3,3.0,1.1,0.1,Iris-setosa. (line 14)
        ii. 6.2,2.2,4.5,1.5,Iris-versicolor (line 69)

2. Apply the PCA and LDA dimensionality reduction methods on the following two data sets:
    a. biodegradation data (https://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation#)
    b. glass data (https://archive.ics.uci.edu/ml/datasets/glass+identification)

    c. For each data set, split the data with 80% training and 20% testing, and use scikit learn programs (program templates can be downloaded from the course web site) to answer the following questions:

        - For PCA, what is the optimal number of principle component to use such that the classification accuracy on the data with reduced dimension will be the highest? Show the classification accuracy vs. number of principle components plot and derive your answer.

        - Does PCA improve the classification accuracy obtained from the original data?

        - For LDA, is it able to improve the classification accuracy with reduced dimension? What is the smallest number of Linear Discriminators needed to obtain the highest classification accuracy on this data? Show accuracy vs. the number of LDs to justify your answer.