Final Project

This assignment is for candidates enrolled in our Programs in Analytics and Statistical Studies (PASS), and replaces Assignment 4. Others are welcome to take it, and view model answers and compare their answers with it, but grades will be provided only to those registered in the above program.

What is required to be done?

1) Perform all the analysis required in R.
2) You are supposed to submit your analysis along with the codes involved and the interpretations for each question.
3) You are expected to write a report of the analysis including interpretations of the results obtained.

Data set: ALL Data set

Packages: Biobase, DMwR, randomForest.

About the data set:

The data set is from the study on acute lymphoblastic leukemia.

The data consists of microarray samples from 128 individuals carry the disease acute lymphoblastic leukemia. There are two different types of tumors among these samples as T-cell ALL (33 samples) and B-cell ALL (95 samples).

Our interest is to study B-cell ALL samples.

Question to be worked on:

Use the data of the case study used in the lesson 4 for a small experiment of comparing several variants of kNN and random forests, using leave one out cross validation. Regards the problem of having too many features on the original data, simply choose 30 genes randomly. Having this filtered data set with 30 random genes, compare a few variants of random forests (varying for instance the parameter that controls how many trees make up the ensemble), with a few other variants of k-Nearest Neighbours (varying the number of neighbours). For the comparison use the LOOCV routines available in the book R package and explained in the chapter supporting this lesson. Check and comment the obtained results, in particular in comparison with the results shown in the book, to observe the impact of the random feature selection you have used.