Assignment 2

Machine Learning (**CS564**)

Date:- 05-Oct-2020 **Deadline:-12-Oct-2020**

Instructions:

1. You are supposed to write code from scratch for filter methods but you can use libraries for pre-processing and post-processing.
2. The assignments should be completed and uploaded before the deadline.
3. Markings will be based on the correctness and soundness of the outputs. Marks will be deducted in case of plagiarism.
4. Proper indentation and appropriate comments are mandatory.
5. You should zip all the required files and name the zip file as ***roll_no*.zip**, eg. **1701cs11.zip.**
6. Make necessary assumptions if required. For further clarification, you can write an email: cs5642020@gmail.com
7. Upload your assignment (the zip file) with the following link: https://www.dropbox.com/request/crkfxYPdsgL0nhjgNb1P

DATASET :

Cancer Gene Expression Data:

1.Leukemia
2.DLBCL
3.Lung

Testing Size: 20%

You are supposed to do data preprocessing as per the problem requirement.

PROBLEM:

The study of gene expression of cells and tissue is one of the major ways for discovery in medicines. The main challenge of such gene data is high input dimensionality, heterogeneity in the data with very low sample size. To overcome this, gene subset selection/Feature Selection has become a crucial and essential step.

A. Apply the three filter methods(**1. Mutual Info[f1] 2. F Classif[f2] and 3.T-Test[f3]**) on the three datasets to get important features(N: No. of selected feature is not restricted but

it should be less than 20% of toat features). Now, use KNN(C1) and SVM(C2) for Classification. Report Accuray, F-Score, and Confusion Matrix.

B. Select the most important N/3 features from each of these three filter methods(f1,f2,f3).

$$F = \{ f1 \cup f2 \cup f3)$$

Classify the test data with the Classifiers(C1 and C2) and Compare them with the above result.

C. Now apply feature selection in a cascaded manner and Classify with C1 and C2.

a. F1(N features) → F2(2N/3 features out of selected features from F1) → F3(N/3 features out of selected features from F2)
b. F2 → F3 → F1
c. F3 → F1 → F2

D. Classify the test data using wrapper methods(Sequential Forward Search and Sequential Backward Search) with N features.

Present a comparison table of the obtained results with a summary of your observations.