Assignment#2

***Due Date (Hard!): October 30, 2017 11:55PM** **Total Marks: 10**

***Submit your assignment to the Course page at the CSE Eduserver**

***All submissions MUST be individual work and any form of copying, if found would lead to ZERO marks**

*** A ZIP file containing 2 PDF files may be uploaded with file name as YourFirstName_RollNo_Exp1.pdf and YourFirstName_RollNo_Exp2.pdf**

*** Use R package for all experiments**

1: Build the following classifiers using the cleaned dataset of your Assignment#1 submission.

a. Decision Tree with Gini index as the impurity measure

b. Decision Tree with Entropy as the impurity measure

c. Naïve Bayesian classifier

d. Artificial Neural Network – with and without hidden layers

You may construct the above classifiers with – **Case 1:** *A random sample of 80:20 for the Train set and Test set* – **Case 2:** *A 10-fold cross validation*

Show the confusion matrix and the values of accuracy, recall, precision and f-score.

Also plot the ROC curves for each of the above cases and find the Area Under ROC Curve (AUC) in all such cases. Which classifier performs the best with **Case 1** and **Case 2**?

2: Remove the class labels of the cleaned dataset of your Assignment#1 submission and apply the following clustering techniques:

Case A. K-means clustering with K=actual number of classes in your cleaned dataset

Case B. K-means clustering with K=actual number of classes in your cleaned dataset-1

Case C. K-means clustering with K=actual number of classes in your cleaned dataset+1

Case D. Density based clustering with the *radius parameter* = Average of the distance from the "centroid" to the furthest point in each of K-clusters obtained in **Case A** and the *minimum points parameter* = Minimum size of all the K-clusters obtained in **Case A**. Report the noise data returned by this method, if any.

Compare the Sum of Squared Error (SSE) values of **Case A, Case B and Case C.**

Report the Silhouette Coefficient values of individual clusters and also the whole set of clusters in **Case A** and **Case D**. Justify your views on comparing **Case A** and **Case D.**