**Assignment - 1**

**Big Data Analytics** **(CSOE17)**

**Weightage- 20%**

**Dataset Description:**

The wine data set consist of total 1599 tuples. Training dataset consist of 1119 tuples and testing dataset consist of 480 tuples. Total number of attributes are 12.

**Implement the following tasks:**

1. Linear regression to detect/estimate the value of “quality” attributes of a wine in the dataset.

**Input**: Training dataset and testing dataset

**Output**: Graphical representation of fitting a line over your training data and sum of square error calculation for test dataset.

1. Use the following classifier to detect the quality of wine. As all the attributes of the data sets have numerical values you can take “quality” attribute of the dataset as class label. In Quality attribute the values greater than or equal to 7 can be considered as “good “ quality and the quality value less than 7 can be considered as “bad” quality. So now you have two class problem.
   1. Logistic Regression
   2. Linear Regression as a classifier
   3. SVM
   4. Naïve Bayesian
2. Compare all the classifier based on Accuracy, Precision, Recall, F-measure, Sensitivity, and Specificity and discuss the result

**Input**: Test results from task 2

**Output**: values of each measures with respect to each classifiers

1. Find the correlation between attributes and apply PCA (Discuss about the correlation with respect to wine dataset). While applying PCA leave attribute “quality” as it is. So you have total 11 attributes. All the classifiers (a to d) should be applied again over the newly constructed dataset (dataset constructed using PCA) for the following number of attributes and evaluated using measures given in (3)
   1. 7 (i.e. from PCA we get 11 attribute so take first 7 attributes of the newly constructed dataset, call this dataset as **redwine\_7\_training and redwine\_7\_testing**)
   2. 4 (i.e. from PCA we get 11 attribute so take first 4 attributes of the newly constructed dataset, call this dataset as **redwine\_4\_training and redwine\_4\_testing**)

**Input**: {**redwine\_7\_training and redwine\_7\_testing, redwine\_4\_training and redwine\_4\_testing**}

**Output**: values of each measures with respect to each classifiers and each dataset (both redwine\_7 and redwine\_4)

So what do you observe when taking 7 attribute and when you are taking 4 attributes what impact does the dimensionality have in various classifiers (based on evaluation measures).

**Note**: While using PCA for construction of new data from the existing data, PCA should be applied over both training and testing data.

**Deliverables:**

1. Every task should be implemented in separated code files. Since dataset and tasks are predefined do not prompt for user input anywhere in the program. When executing each program file for the given task, specified outputs should be generated automatically (without any user intervention).
2. You should make a report which consist of discussion regarding each tasks about your findings and observations after the experiments/coding. (It can consist of graphs and any other diagrams). It should also include the screenshots of your each task execution i.e. output of each tasks.
3. All the materials should be zipped in a single folder and uploaded. Name of the folder should be your roll number and assignment number, for example – “11100898\_Assignment1.zip”

**Important** : Your **submission deadline is 8th October 2020.** The implementation of code should be in python and if you want to use any other language first get my permission.