**Assignment 2**

**The following has to be done using Bayesian learning (Naïve Bayes classifier):**

1) Randomly divide the data into 80% for training and 20% for testing. Apply the following:

1. Handle the missing values in both train and test set. [5]
2. Encode categorical variables using appropriate encoding method (in-built function allowed). [5]
3. After completing step (a) and (b), compute 5-fold cross validation on the training set (normalisation of data is allowed, if required). Print the final test accuracy. [10]

2) Apply PCA (select number of components by preserving 95% of total variance) on the processed data from step (1).

1. Plot the graph for PCA (in-built function allowed for PCA and visualisation). [20]
2. Use the features extracted from PCA to train your model. Compute 5-fold cross validation on the training set (normalisation of data is allowed, if required). Print the final test accuracy. [10]

3) Using the processed data from step (1), apply the following:

1. A feature value is considered as an outlier if its value is greater than mean + 3 x standard deviation. A sample having maximum such outlier features must be dropped. [5]
2. Using the sequential backward selection method, remove features. [15]
3. Print the final set of features formed. [5]
4. Compute 5-fold cross validation on the training set (normalisation of data is allowed if required). Print the final test accuracy. [5]

4) Report and results. [20]

**Dataset Description:**

Use Train\_E.csv as data for this assignment. The “life\_expectancy” column will be used as labels

**Submission Guidelines:**

Implementation has to be done in Python. **No function for Naïve Bayes classifier should be used.** Provide a report on your study with proper description. Keep your codes and report along with your results in a single folder. Submit the compressed .zip file following groupno\_asgn2.zip naming convention. For example, 5\_asgn2.zip for Group no 5.