**Name – Bbiswabasu Roy**

**Roll – 19EC30008**

**Assignment – 3**

**Preprocessing data:**

The features in the dataset can be broken into 3 categories:

* Binary categorical features
* Non-binary categorical features
* Numerical features

Among them numerical features can be left as it is because they represent scalar values of features. Binary categorical features can also be left as they are because similarities between 0 and 0 is more than that between 0 and 1 and similarities between 1 and 1 is more than that between 1 and 0.

However, the non-binary categorical values cannot be used directly to build the model and hence one hot encoding was applied to convert them to binary categorical features. For example, the feature “Education” takes values {0, 1, 2, 3} and after applying one hot encoding 4 binary features were added namely “Education\_0”, “Education\_1”, “Education\_2”, “Education\_3” and the column “Education” was removed. If the “Education” column had value of 2, then “Education\_2” was assigned value 1 while others were assigned value 0. Same principle was applied on other non-binary categorical features as well.

Also, the ID column in the original dataset was useless and hence removed.

Finally, normalization was applied on all columns to scale the values appropriately using the formula:

Where, is the value in j-th column of i-th row.

**Clustering Algorithm:**