DMG Assignment 4

Aim: Perform clustering.

Deadline: 30 November 2020

Instructions:

1. Mention all assumptions if any in the report.
2. Report and code in .py should be submitted in the classroom in a zip folder with the name ‘A4_RollNumber1_RollNumber2’.
3. You are free to use any library or data processing techniques.
4. Include one runner function in code which takes test_X.csv as input and produces result.csv. All preprocessing to be done on data before applying the model should be present in the runner function.
5. Some students will be randomly picked for a demo of assignment 4. So write the code on your own, make sure you don’t cheat. If you can't answer the questions during your demo, 50% of your marks will be deducted.
6. A single team member will submit on the google classroom and will mention the contributions of each member in the report.

The following should be included in the Report:

1. Explain your methodology: approach and reason clearly in the report.
2. Visualize skewness of data before and after preprocessing (if done any).
3. Add all data analysis steps which you have performed on the dataset.
4. Make a section “Learning”, which describes your learning in doing this assignment.
5. Report Centroid/representative object/prototype of each cluster.
6. Visualize your clusters. (You can use lesser data points/ dimensions for visualizations).
7. Compare your cluster distribution with the below true label count.

In one of the cluster method variations, you will make 7 clusters and compare your results with the below data, You are free to increase or decrease clusters in other variations and report your insights.

Cluster_1 - 540, Cluster_2 - 542, Cluster_3 - 743, Cluster_4 - 540, Cluster_5 - 540, Cluster_6 - 675, Cluster_7 - 540,

The evaluation will be based on:

- The report containing your various cluster methods and hyperparameters tried. You have to try at least 3 variations (Kmeans, Kmeans++, any other of your choice).