**Assignment on Clustering**

[Submission deadline: **11:59 pm,** **28th August 2021**]

1. Describe a problem related to your organisation where clustering technique can be applied.
2. Identify feature variables and justify the reasons for selecting these to form the clusters.
3. What are the benefits that could be achieved by forming the clusters?
4. Collect relevant data and form the clusters. You may use k-means / hierarchical clustering technique. The sample python codes of both the techniques are provided to you for your easy reference.

**Solution to Assignment**

Name: Chinmay Harkawat  
Organization: Sardar Vallabhbhai National Institute of Technology (SVNIT) Surat  
Role: Student (2nd Year, Mechanical)  
Contact: [u20me157@med.svnit.ac.in](mailto:u20me157@med.svnit.ac.in) [[LinkedIn](linkedin.com/in/chinmay-harkawat-45290a1b4)] Phone: +91 9619750652

**Introduction**

During these troubled times the news of the vaccine trial completion was a relief to many. However, one of the major challenge faced by organizations and governments worldwide was to distribute these vaccines efficiently, safely and quickly.

Keeping that requirement in mind I tried to define the problem statement as follows.

**Problem Statement:** Vaccine distribution in Maharashtra [(a)](#Q1).

**Proposed Solution**: Using statistical data [[1](https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Maharashtra#Statistics)], any region having cases more than 50,000 is identified and then according to their location in terms of latitude and longitude they are classified into clusters. This classification ensures vaccine is distributed in an optimized manner [(c)](#Q3).

|  |  |
| --- | --- |
| **Feature Variables** [(b)](#Q2) | |
| Variable Name | Function |
| Latitude | Location Specification |
| Longitude | Location Specification |

Data Collection:

|  |  |
| --- | --- |
| Cases Data | [Link](https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Maharashtra#Statistics) |
| CSV File Used | [Link](https://github.com/janeka1122/AI_ML_SCM/blob/main/Assignment%201/Assignment1_Dataset.csv) |
| Final Repo | [Link](https://github.com/janeka1122/AI_ML_SCM/tree/main/Assignment%201) |
| Note: The final repository contains 2 solutions Assignment1\_Try1 and Assignment1\_Try2  The reason is explained below. | |

The 1st attempt at the problem used only 10 data points. However, that did not give a satisfactory elbow graph.   
The 2nd attempt used 28 data points which gave a much more typical / defined elbow graph.

This observation is concurrent to that one mentioned in the lecture where it was stated that clustering was used for large datasets.