



भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 360: Machine Learning Lab
Practice Assignment 5

Instructions: This is only for practice. Complete it by 12:00 PM today. Your completion will be reviewed by the Teaching Assistants.

1. Download [Synthetic dataset](#) containing mixed attribute types: numeric (age, income), categorical (occupation, region), and ordinal (education level, satisfaction). Write a program to do the followings:
 - (a) Read the dataset.
 - (b) Perform Normalization on the numeric attributes.
 - (c) Encode the categorical and ordinal attributes (using label encoding or one-hot encoding).
 - (d) Cluster the dataset into different groups of persons(ID) using K-means clustering method (**without using in-built Python library/ packages**) Consider the following for this:
 - Number of initial centroids, K : 2 to 8.
Hint: Apply K-means clustering, separately for all the given K -values.
 - Stopping criteria: No changes in the cluster assignment.
 - **Distance measures:** For numeric attribute: Euclidean distance, For categorical attribute: Hamming distance, For ordinal attributes: Manhattan distance (after encoding).
 - (e) Report and visualize the obtained final clusters with respect to different K , and write the result analysis.
 - (f) Determine the optimal cluster using [Silhouette score](#) for the different K -values, without using in-built Python library/ packages.
 - (g) Plot the silhouette scores to identify the optimal number of clusters.