Lab Report on

K Means Clustering

Submitted to

Department of Computer Science and Engineering

Nepal Engineering College

in Partial Fulfillment of the

Requirements for the Degree of B.E. in Computer

Submitted By

Dipson Thapa (020-313)

Date: *28/07/2024*

K Means Clustering

Introduction:

K-Means Clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct groups, or clusters. The algorithm works by initializing a set of cluster centroids, then iteratively refining these centroids by assigning each data point to the nearest centroid and recalculating the centroids as the mean of all points assigned to them. This process continues until the centroids stabilize and the clusters are formed. K-Means Clustering is widely applied in various fields such as market segmentation, image compression, and pattern recognition to uncover underlying patterns and structure within the data.

Dataset:

The dataset used for this K-Means Clustering analysis is the "Mall Customer Segmentation Data" which contains information about the spending behavior of customers in a mall. You can access it here.

Link: https://www.kaggle.com/datasets/sahirmaharajj/country-health-trends-dataset

Notebook name: K Means Clustering_020313.ipynb

Libraries Used:

1. **Pandas**:

   - Utilized for data manipulation and analysis. It offers data structures like DataFrames that are ideal for handling tabular data.

2. **Pathlib**:

   - A standard library for handling filesystem paths in a more readable and efficient way.

3. **numpy**:

   - A fundamental package for numerical computations in Python, providing support for arrays and matrices along with a collection of mathematical functions to operate on these data structures.