ML Assignment on K-Means Clustering

**Name**: Dhananjay Siddhanath Jagtap  
**Roll No**: 282057  
**Batch**: B3

# Problem Statement

This assignment involves applying K-Means Clustering, an unsupervised machine learning algorithm, on the Mall Customers dataset. The steps include:

1. Data Pre-processing: Apply techniques such as label encoding and data transformation (if necessary).  
2. Data Preparation: Split the dataset into training and testing sets (Train-Test Split).  
3. Apply K-Means Clustering algorithm to group the customers.  
4. Model Evaluation: Analyze and visualize the clusters.  
5. Apply Cross-Validation and evaluate the stability of clusters.

# Objective

To help students understand and practically apply clustering using the K-Means Clustering algorithm, and interpret the results to derive meaningful insights from customer data.

# Theory

What is K-Means Clustering?

K-Means Clustering is an unsupervised machine learning algorithm used to identify groups (clusters) in unlabeled data. It helps discover patterns by organizing data into K distinct non-overlapping subgroups (clusters), where each data point belongs to the cluster with the nearest mean (centroid).

Characteristics:  
- Unsupervised: No labeled output; the algorithm finds structure in data by itself.  
- Centroid-based: Each cluster is defined by the mean of the points within it.  
- Iterative Process: Keeps refining cluster assignments until convergence.

# How the K-Means Algorithm Works

1. Step 1: Choose the number of clusters K.  
2. Step 2: Initialize K random centroids (these can be random or selected from the data points).  
3. Step 3: Assign each data point to the nearest centroid (forming K clusters).  
4. Step 4: Compute new centroids as the mean of all data points assigned to each cluster.  
5. Step 5: Repeat steps 3 and 4 until the centroids no longer change or changes are minimal (convergence).  
6. Step 6: Output final clusters and centroids.

# Use Cases of K-Means Clustering

- Customer Segmentation: Segment customers based on spending behavior, income, or demographics.  
- Market Basket Analysis: Group products based on purchase patterns.  
- Inventory Categorization: Group items by performance metrics.  
- Behavioral Analytics: Group users based on activity or interaction history.

# Implementation Summary

- Preprocessed the dataset using label encoding and scaling where needed.  
- Applied KMeans algorithm to cluster customers based on Annual Income and Spending Score.  
- Used the Elbow Method to find the optimal value of K.  
- Evaluated cluster performance visually using scatter plots with centroids marked.  
- Applied cross-validation (or silhouette score) to verify the stability and quality of the clusters.

# Key Observations

- The dataset revealed natural groupings of customers with high spending scores and income levels.  
- K-Means helped discover targetable customer segments for marketing, loyalty programs, and personalized services.  
- Customers with low spending scores despite high income can be targeted with engagement strategies.

# Conclusion

In this assignment, we effectively demonstrated how to implement and evaluate K-Means Clustering on real-world customer data. The algorithm grouped customers into meaningful clusters that can provide valuable business insights. Through visualization and evaluation, we validated that our clustering strategy was appropriate and useful for segmenting the dataset.

This hands-on exercise emphasized:  
- The importance of proper data preprocessing  
- Choosing the right number of clusters (K)  
- Evaluating clustering performance both visually and quantitatively

The knowledge gained from this task will be beneficial for solving real-world problems involving customer segmentation, behavioral analysis, and pattern discovery in large datasets.