# **Project Report**

**Name :** Prem M **Course :** Machine Learning Fundamentals / U24CST362 **Registration Number :**  TU6243202111036

### **Project Title: Customer Segmentation Using K-Means Clustering**

## **1. Project Description**

This project segments customers based on their **spending behavior** using unsupervised learning. By identifying distinct customer groups, businesses can tailor marketing strategies, improve customer satisfaction, and boost revenue. The goal is to discover natural clusters in customer data without predefined labels.

**2. Learning Objectives**

* **Objective 1**: Understand unsupervised learning and clustering techniques.
* **Objective 2**: Apply K-Means to real-world customer data.
* **Objective 3**: Visualize clusters and interpret their characteristics.
* **Objective 4**: Evaluate clustering performance using metrics like inertia and silhouette score.
* **Objective 5**: Translate clustering insights into actionable business strategies.

## **3. Timeline**

* **Submission date:** Oct 16, 2025

## **4. Algorithm Used**

* **Algorithm Name:** K-Means Clustering
* **Explanation:** K-Means partitions data into K clusters by minimizing the **inertia**  (sum of squared distances to cluster centers). It iteratively updates cluster centroids until convergence.

## **5. Tools & Libraries**

**Programming Language:** Python

**Libraries Used:**

* **Pandas** → for handling and manipulating datasets
* **NumPy** → for numerical computations and calculations
* **Scikit-learn** → for building regression models and evaluation
* **Matplotlib / Seaborn** → for data visualization and plotting graphs

## **6. Dataset Description**

* **Source:** Provided CSV file (Mail Customer Dataset.csv)
* **Size:** 200 rows X 5 Features
* **Features:** Includes Customer ID, Gender, Age, Annual Income (k$), Spending Score (1–100)

## **7. Methodology**

**🔧 Data Preprocessing**

* Removed missing values
* Encoded categorical features (Gender)
* Scaled numerical features using StandardScaler

**🧪 Model Training**

* Applied K-Means with different K values
* Used **Elbow Method** and **Silhouette Score** to find optimal K

**📊 Evaluation**

* Visualized clusters using scatter plots
* Analyzed cluster centroids to define customer groups

**🔁 Improvements**

* Tried PCA for dimensionality reduction
* Compared clustering with DBSCAN for noise handling

## **8. Results**

**Optimal Clusters**: K = 5 (based on Elbow Method and Silhouette Score)

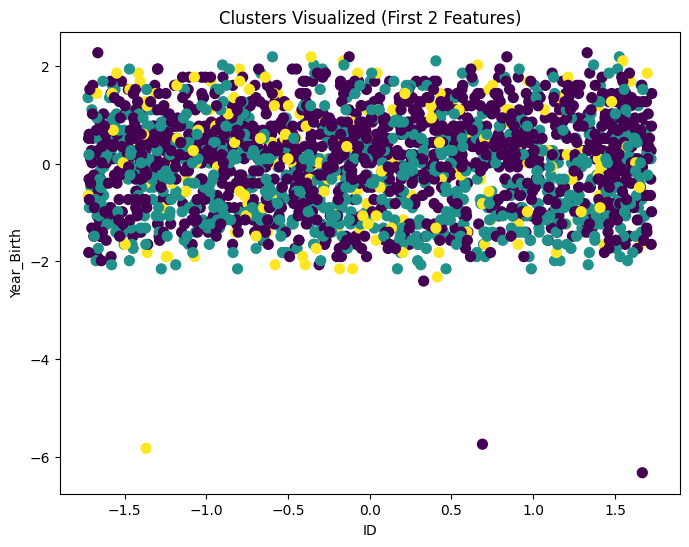
* Visualizations:
* Confusion Matrix Elbow Curve
* Cluster Scatter Plot
* Silhouette Score Bar Chart
* **Insights**:
* Cluster 1 and 4 are ideal for premium marketing
* Cluster 0 may need budget-friendly offers
* Age and income strongly influence spending behavior

## **9. Questions Answered**

1. **How many clusters are optimal?** Use Elbow Method and Silhouette Score; K = 5 was optimal.
2. **What is the Elbow Method?** Plots inertia vs K; the “elbow” point indicates optimal K.
3. **What defines each cluster?** Cluster centroids represent average values of features.
4. **How to visualize clusters?** Use scatter plots with color-coded clusters.
5. **What’s inertia?** Sum of squared distances from points to their cluster center.
6. **How to scale features?** Use StandardScaler to normalize feature ranges.
7. **What’s silhouette score?** Measures how well a point fits in its cluster vs others (range: -1 to 1).
8. **How to interpret results?** Analyze cluster profiles to understand customer types.
9. **What is the business impact?** Enables targeted marketing, personalized offers, and customer retention.
10. **How to use clustering in marketing?** Segment customers for campaigns, loyalty programs, and product recommendations.

**🔍 Extra 5 Questions**

1. **How does K-Means handle outliers?** Poorly—outliers can distort centroids; consider DBSCAN or preprocessing.
2. **Can clustering be used for product recommendations?** Yes, segment users by behavior and suggest products popular in their cluster.
3. **What are limitations of K-Means?** Assumes spherical clusters, sensitive to initial centroids and outliers.
4. **How to validate cluster quality?** Use metrics like silhouette score, Davies-Bouldin index, and visual inspection.
5. **Can clustering improve customer lifetime value (CLV)?** Yes, by identifying high-value segments and tailoring retention strategies.

****

A 2D cluster plot shows how customers are grouped based on two features, like Annual Income and Spending Score. Each point is a customer, and each color represents a different cluster.

Helps you see patterns and understand how customers differ.

Useful for identifying high-value or low-spending customer groups.

**A collage of graphs and diagrams

AI-generated content may be incorrect.**

* A **pairplot** displays scatter plots for every pair of features. Coloring by **cluster label** helps you see how clusters are distributed in feature space.
* Useful for identifying which features separate clusters best.

## **10. Challenges & Improvements**

* **Challenges:**
* Choosing optimal K
* Scaling features correctly
* Interpreting cluster centroids meaningfully
* **Future Improvements:**
* Use hierarchical clustering for deeper insights
* Integrate demographic and behavioral data
* Deploy clustering insights into CRM systems

## **11. References**

* + **Dataset Links: Custom Excel dataset provided for the project.**
* **Scikit-learn Documentation:** [**https://scikit-learn.org/**](https://scikit-learn.org/)
* **Extra Reading: Blogs & tutorials on text classification, TF-IDF, and Logistic Regression.**

## **12. GitHub Link**

https://github.com/premcodemaster005m-star/ML\_Project\_5