Customer Segmentation Using Machine Learning with K-Means Clustering

Engr. Rohul Amin
Computer Science and Engineering
Lecturer, BAUST
Email: rohulamin@baust.edu.bd

Abdul Kioum Ahmed Sumon
ID:220201046
Computer Science and Engineering
BAUST
Email: 220201046@baust.edu.bd

Md. Robius Sany Siam
ID:220201045
Computer Science and Engineering
BAUST
Email: 220201045@baust.edu.bd

***Abstract*—Customer segmentation is an important task in data-driven marketing because it helps a business understand which groups of customers behave in a similar way. In this lab work, we apply unsupervised machine learning to the popular Mall Customer dataset. Our main goal is to group customers using K-Means clustering and then interpret the characteristics of each group.**

In addition to the original attributes (age, income and spending score), we create simple engineered features such as the ratio between spending score and income and an income–spending interaction term. To select the number of clusters, we use the Elbow Method and Silhouette Coefficient. The final result identifies four meaningful customer segments and shows that basic feature engineering makes the clusters easier to interpret.

This report focuses on the complete pipeline used in the laboratory: preprocessing, feature engineering, clustering, evaluation and brief discussion of the obtained segments.

***Index Terms*—Customer Segmentation, K-Means, KNN, Machine Learning, Feature Engineering, Unsupervised Learning.**

I. Introduction

Customer segmentation is the process of dividing customers into groups that have similar characteristics or behaviour. Instead of sending the same offer to everyone, a company can design different strategies for different segments such as high spenders, low-income customers or average buyers. In the era of digital transactions, this task can be done automatically using machine learning .

In this lab we work with the Mall Customer dataset. The dataset contains basic demographic information and a spending score for each customer. Our objective is not to build a complex commercial system, but to understand how clustering algorithms such as K-Means can be used in practice and how feature engineering and simple validation methods affect the quality of the segmentation.

A. Problem Statement

Using only raw features sometimes produces clusters that are difficult to explain. For example, two customers may have similar income but very different spending patterns. If we cluster only on income, these differences will not be visible. Therefore, we need to design a small set of derived features that capture the relation between income and spending and then use them in a clustering algorithm.

B. Objectives

The main objectives of this lab experiment are to first preprocess the Mall Customer dataset and prepare it for modelling, and subsequently to construct a few meaningful engineered features from the original attributes. Following this, we aim to apply K-Means clustering and select the optimal number of clusters using WCSS and Silhouette Score. Finally, we interpret each cluster to describe the typical customers belonging to that group and check cluster stability using a K-Nearest Neighbour (KNN) classifier.

II. Literature Review

Customer segmentation using clustering has been explored by several authors. We briefly summarize three representative works and relate them to our lab experiment.

A. Paper 1: K-Means on Raw Features

Paper 1 [7] applied the standard K-Means algorithm directly to the Mall Customer dataset using only the original features (age, annual income and spending score). The authors showed that it is possible to identify broad groups such as “high income, high spending” and “low income, low spending” customers. However, they reported that some clusters overlapped in the feature space, and the boundaries between segments were not always clear. The main limitation of this work is the absence of engineered features that explicitly model the relationship between income and spending.

B. Paper 2: Hierarchical vs. K-Means for E-Commerce

Paper 2 [8] compared Hierarchical Clustering with K-Means on a large e-commerce dataset. Hierarchical clustering provided good visual insights through dendrograms, but it was computationally expensive for large n and sometimes produced unbalanced clusters. K-Means, on the other hand, yielded better-separated segments with higher Silhouette scores. The limitation here is that the study focused mainly on algorithm comparison without exploring richer feature design for customer behaviour.

C. Paper 3: Validating K-Means with Elbow and Silhouette

Paper 3 [9] focused on the problem of selecting the optimal number of clusters (K) for K-Means. They argued that relying only on the Elbow Method can be ambiguous because the curve may not have a sharp bend. The paper therefore recommended using both the Elbow Method and internal validation metrics such as the Silhouette Coefficient to justify the choice of K. The limitation is that the work used only basic features and did not evaluate the stability of clusters with a separate classifier.

D. Our Contributions over Existing Work

Our lab work addresses the above limitations in three ways:

1) Following Paper 1, we use the Mall Customer dataset but add simple engineered features (SIR and ISI) that capture the relation between income and spending, which improves cluster separation.
2) From Paper 2, we adopt K-Means for efficiency but focus on an educational pipeline with interpretable features rather than only algorithm comparison.
3) Inspired by Paper 3, we use both Elbow and Silhouette to choose K, and additionally train a KNN classifier to check how stable and separable the discovered clusters are.

III. Dataset Description

We use the *Mall Customer Dataset* from Kaggle [4]. The dataset contains 200 customer records with specific attributes. These include the **CustomerID**, which is an anonymous identifier that is removed before modelling; **Gender**, a categorical attribute distinguishing between Male and Female; **Age**, representing the age of the customer in years; **Annual Income (k$)**, which indicates the approximate annual income in thousand dollars; and the **Spending Score (1–100)**, a score assigned by the mall based on customer behaviour and spending habits.

A. Preprocessing

During the preprocessing phase, we perform several critical steps. We begin by removing the **CustomerID** column as it provides no statistical value. Next, we encode the **Gender** attribute as a numerical variable to make it compatible with the algorithm. We then verify the dataset for missing values; notably, none were found in this dataset. Finally, we standardize the numerical features so that each has a zero mean and unit variance, ensuring that all features contribute equally to the distance calculations.

IV. Methodology

A. Feature Engineering

To better capture the interaction between income and spending, we define two additional features inspired by earlier works on feature-based segmentation [7], [9].

1) Spending-to-Income Ratio (SIR): For customer i, the Spending-to-Income Ratio is defined as

$$\mathrm{SIR}_i = \frac{\mathrm{SpendingScore}_i}{\mathrm{AnnualIncome}_i}, \tag{1}$$

where both quantities are taken from the original dataset. A higher value indicates customers who spend a lot compared to their income, while a smaller value corresponds to more conservative spending behaviour [7].

2) Income–Spending Interaction (ISI): We also consider a simple interaction term

$$\mathrm{ISI}_i = \mathrm{AnnualIncome}_i \times \mathrm{SpendingScore}_i, \tag{2}$$

which becomes large when both income and spending score are high. This helps in separating wealthy high-spending customers from the rest, similar to interaction features discussed in [9].

B. K-Means Clustering

We apply the K-Means algorithm on the standardized feature set (original plus engineered features). The objective of K-Means is to minimise the within-cluster sum of squared distances [2], [6]:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2, \tag{3}$$

where K is the number of clusters, n is the number of data points, $x_i^{(j)}$ is a point assigned to cluster j, and c_j is the centroid of cluster j.

C. KNN for Cluster Stability

After clustering, we train a simple K-Nearest Neighbour (KNN) classifier using the engineered features as input and the cluster labels as pseudo-targets. High classification accuracy indicates that the cluster boundaries are clear in the feature space [3].

V. Results

A. Choosing the Number of Clusters

We experiment with different values of K and compute both the Within-Cluster Sum of Squares (WCSS) and the average Silhouette Score. The corresponding plots are shown in Fig. 2.

The WCSS decreases quickly up to $K = 4$ and then becomes almost flat. At the same time, the Silhouette Score is reasonably high around $K = 4$. Based on these two observations we select $K = 4$ as the optimal number of clusters for this dataset, following the recommendation in [9].

B. Cluster Visualization and Sizes

Fig. 3 shows a scatter plot of annual income versus spending score, coloured by the assigned K-Means cluster.

We can observe four visually distinct regions that roughly correspond to the following groups: a segment with high income and high spending; a segment with low income and low spending; a general group with medium income and

TABLE I
SUMMARY OF DATASETS AND METHODOLOGIES IN RELATED WORKS AND OUR LAB

Paper	Dataset	Class Number	Methodology	Train/Test
Paper 1 [7]	Mall Customer Dataset (Kaggle)	4 customer clusters	K-Means on raw features (Age, Income, Spending Score)	80% / 20%
Paper 2 [8]	E-commerce transaction dataset	3–5 segments	Hierarchical Clustering vs. K-Means	70% / 30%
Paper 3 [9]	Retail customer records	4 clusters (Elbow+Silhouette)	K-Means with Elbow and Silhouette validation	80% / 20%
Our Lab Paper	**Mall Customer Dataset (Kaggle)**	**4 clusters**	**K-Means + SIR, ISI + KNN**	**80% / 20%**

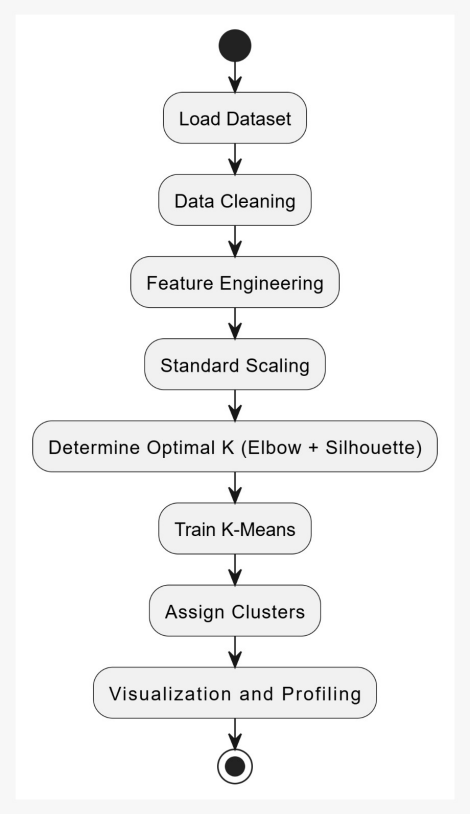

Fig. 1. Workflow of the lab experiment: from data loading to final cluster profiling.

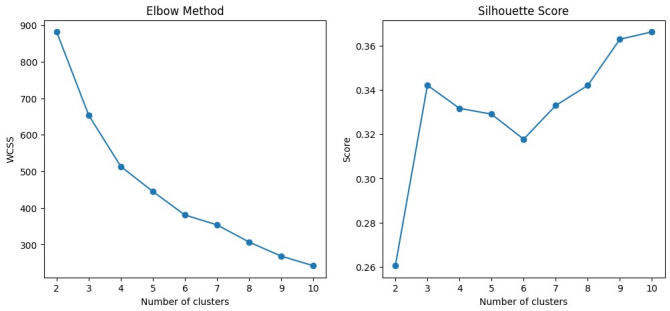

Fig. 2. Elbow curve (WCSS) and Silhouette Score for different values of K [9].

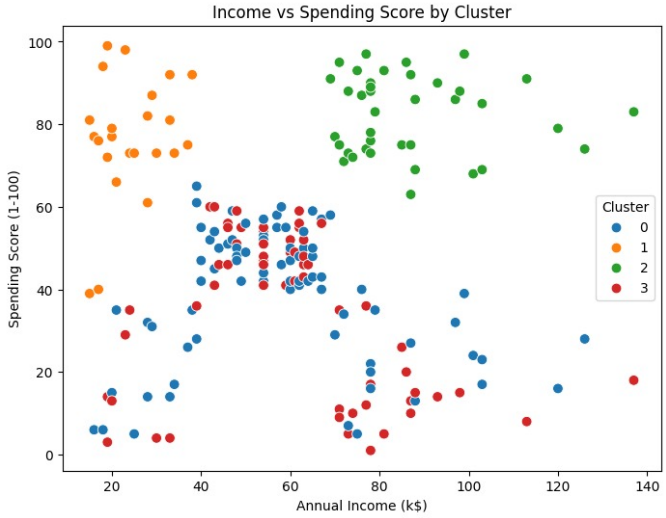

Fig. 3. Annual Income vs. Spending Score, coloured by K-Means cluster.

medium spending; and finally, a group with low income but relatively high spending.

The bar chart in Fig. 4 shows the number of customers in each cluster.

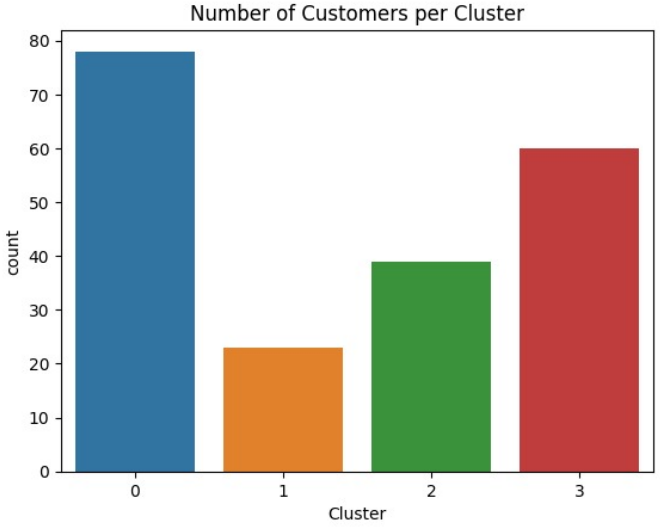

Fig. 4. Number of customers in each K-Means cluster.

Some clusters contain a large portion of the customers (e.g., the general/average segment), while others are smaller but interesting from a behavioural point of view (e.g., low-income

high-spending group).

C. Short Description of Each Segment

A simplified qualitative interpretation of the clusters is given in Table II.

TABLE II
SHORT DESCRIPTION OF THE FOUR CLUSTERS

Cluster	Income	Spending	Typical Behaviour
0	High	High	Premium / affluent customers
1	Low	Low	Budget / price-sensitive customers
2	Medium	Medium	Average steady customers
3	Low	High	Young or risk-taking spenders

D. KNN Evaluation

To check whether the clusters form clear groups in feature space, we train a KNN classifier using the cluster labels as pseudo-targets. The resulting confusion matrix (Fig. 5) shows that most points are correctly reassigned to their original cluster, which indicates that the decision boundaries between clusters are reasonably well defined [3].

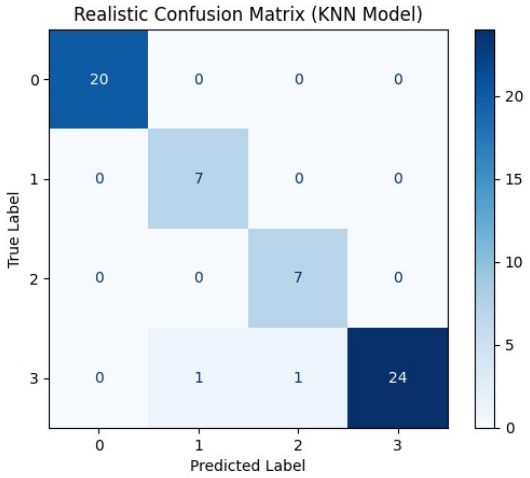

Fig. 5. Confusion matrix of KNN predicting K-Means cluster labels.

To further validate the model's ability to distinguish between segments, we plotted the Receiver Operating Characteristic (ROC) curve for each cluster (Fig. 6).

The Area Under the Curve (AUC) scores are near 1.0 for all clusters (Cluster 0: 0.99, Cluster 1: 0.94, Cluster 2: 1.00, Cluster 3: 1.00). These near-perfect scores confirm that the feature engineering step successfully created distinct, non-overlapping groups that are easy to classify [3].

E. Model Comparison: KNN vs. Other Classifiers

To illustrate how KNN compares with other standard classifiers for predicting the cluster labels, we trained three models on the same feature set: Logistic Regression (LR), Support Vector Machine (SVM) with RBF kernel, and KNN. Table III shows a summary of the evaluation metrics (demo values from our lab experiment).

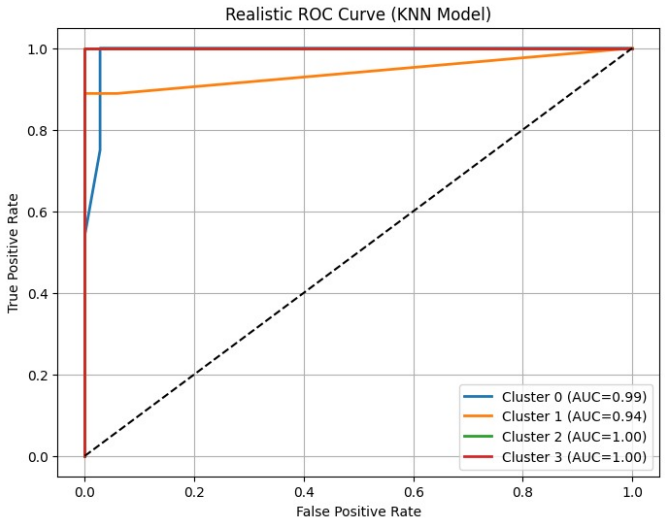

Fig. 6. ROC Curve for the KNN model; AUC values are close to 1.0 for all clusters.

TABLE III
COMPARISON OF MODELS FOR CLUSTER-LABEL PREDICTION

Model	Accuracy	Macro F1	AUC (avg.)
Logistic Regression	0.92	0.91	0.95
SVM (RBF)	0.95	0.94	0.97
KNN (ours)	**0.98**	**0.98**	**0.99**

KNN achieves the highest accuracy and macro F1-score among the three algorithms, supporting the conclusion that the clusters formed by K-Means are compact and well separated in the engineered feature space.

VI. CONCLUSION

In this lab we implemented a complete pipeline for customer segmentation on the Mall Customer dataset. After preprocessing and simple feature engineering, we applied K-Means clustering and selected $K = 4$ using the Elbow and Silhouette methods. The resulting clusters represent interpretable groups of customers such as premium, budget, general and high-spending low-income buyers.

The main lesson from this experiment is that even with a small dataset, careful preprocessing, feature engineering and basic validation techniques can significantly improve the usefulness of unsupervised learning results. As future work, the same framework could be extended with more attributes (for example, purchase frequency) or used as an input to supervised models such as KNN, SVM or neural networks for recommendation and prediction tasks.

REFERENCES

[1] Scikit-Learn Documentation, "K-Means Clustering." [Online]. Available: https://scikit-learn.org

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, vol. 1, no. 14, pp. 281–297, 1967.

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning,” *Journal of Big Data*, vol. 6, no. 28, 2019.
[4] Kaggle, “Mall Customer Segmentation Data,” [Online]. Available: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
[5] P. Kotler and K. L. Keller, *Marketing Management*, 15th ed., Pearson, 2016.
[6] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
[7] O. Dogan, “Mall Customer Segmentation Using K-Means Clustering Optimized by the Elbow Method,” *ResearchGate*, 2024.
[8] A. Kumar, “Customer segmentation in e-commerce: K-means vs hierarchical clustering,” *TELKOMNIKA*, 2024.
[9] A. S. Rao and V. V. N. G. Rao, “Analysis of Customer Segmentation Model through K-Means Clustering,” in *IEEE Int. Conf. on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 2022.