Customer Segmentation Using Unsupervised Learning

AI Project(K24MCA18P)

Session (2024-25)

Submitted by

Piyush Jain (202410116100141)

Nancy Gupta (202410116100130)

Samay Verma (202410116100177)

Submitted in partial fulfilment of the Requirements for the Degree of

MASTER OF COMPUTER APPLICATION

Under the Supervision of

Dr. VIPIN KUMAR

Assistant Professor

Submitted to

DEPARTMENT OF COMPUTER APPLICATIONS
KIET Group of Institutions, Ghaziabad
Uttar Pradesh-201206
(April- 2025)

Table Of Contents

1. Introduction

1.1 Overview

Customer segmentation is the process of categorizing a customer base into distinct groups based on shared attributes such as demographics, behaviors, or purchasing habits. This project leverages unsupervised machine learning, specifically the K-Means clustering algorithm, to perform segmentation on a retail dataset. By analyzing features such as gender, age, annual income, and spending score, the system identifies natural groupings among customers, which can help businesses make informed, data-driven decisions.

1.2 Importance

In the modern business landscape, customer-centric strategies are essential for survival and growth. Understanding customer diversity allows businesses to:

- Personalize marketing efforts
- Optimize product recommendations
- Improve customer satisfaction
- Increase customer retention

  Customer segmentation plays a pivotal role in enabling these strategies by revealing hidden patterns in customer data.

1.3 Role of K-Means Algorithm

The K-Means algorithm is a popular unsupervised learning technique used for clustering data. It works by partitioning data points into *k* distinct clusters based on similarity. In the context of customer segmentation:

- It identifies clusters of customers with similar attributes.
- It helps visualize the distribution and characteristics of different customer groups.
- It enables targeted marketing by classifying customers into meaningful segments.

1.4 Objectives

The primary objectives of this project are:

- To implement customer segmentation using K-Means clustering.
- To preprocess and analyze customer demographic and behavioral data.
- To determine the optimal number of customer segments using the Elbow Method.
- To visualize the clusters using PCA and interpret group characteristics.
- To derive actionable insights that can support business decisions.

1.5 Significance

This project demonstrates the practical application of machine learning in real-world business scenarios. The ability to segment customers efficiently can:

- Enhance marketing ROI through precision targeting.
- Improve product development by focusing on customer needs.
- Assist in crafting personalized experiences, ultimately driving brand loyalty. By combining data science and business intelligence, this project highlights the transformative power of unsupervised learning in customer analytics.

2. Methodology

The methodology outlines the step-by-step approach used to perform customer segmentation using the K-Means clustering algorithm, a form of unsupervised learning. This process includes data collection, preprocessing, feature selection, scaling, model implementation, and visualization of results.

2.1 Data Collection

The dataset used in this project is the **Mall Customers Dataset**, which contains information about 200 customers of a mall. It includes the following features:

- CustomerID
- Gender
- Age
- Annual Income (in $1000s)
- Spending Score (1–100)

This dataset is ideal for segmentation because it captures both demographic and behavioral customer attributes.

2.2 Data Preprocessing

Before applying the clustering algorithm, the data undergoes a series of preprocessing steps:

- **Removing Irrelevant Columns**: The CustomerID column is excluded since it has no predictive power in clustering.
- **Encoding Categorical Data**: The Gender column, being categorical, is converted into numerical format using label encoding (Male = 0, Female = 1).
- **Handling Missing Values**: The dataset is checked for missing values. If any are found, appropriate handling methods such as imputation or removal are applied (in this dataset, there are no missing values).

2.3 Feature Selection

The features selected for clustering include:

- Gender (Encoded)
- Age
- Annual Income (k$)
- Spending Score (1–100)

These features are selected based on their relevance in capturing the diversity in customer behavior and purchasing capacity.

2.4 Feature Scaling

Since the features are on different scales (e.g., age vs. income), Standardization is applied using StandardScaler to ensure that each feature contributes equally to the distance calculations used by the K-Means algorithm. The features are transformed to have a mean of 0 and a standard deviation of 1.

2.5 Determining Optimal Number of Clusters

To find the best value of K (number of clusters), the Elbow Method is used:

- K-Means is run for values of K ranging from 1 to 10.
- The Within-Cluster Sum of Squares (WCSS) is calculated for each K.
- The point where the WCSS curve starts to flatten (elbow point) indicates the optimal number of clusters.

Typically, the elbow occurs around K=5 for this dataset.

2.6 Applying K-Means Clustering

Once the optimal number of clusters is selected:

- The K-Means algorithm is applied with the chosen K value.
- Each data point is assigned to the cluster whose centroid is closest in Euclidean distance.
- The final output is a label for each customer indicating their cluster group.

2.7 Dimensionality Reduction using PCA

To visualize the clusters in a 2D space:

- **Principal Component Analysis (PCA)** is applied to reduce the 4-dimensional data into 2 principal components.
- PCA helps retain the maximum variance while reducing complexity, allowing us to plot customer segments clearly.

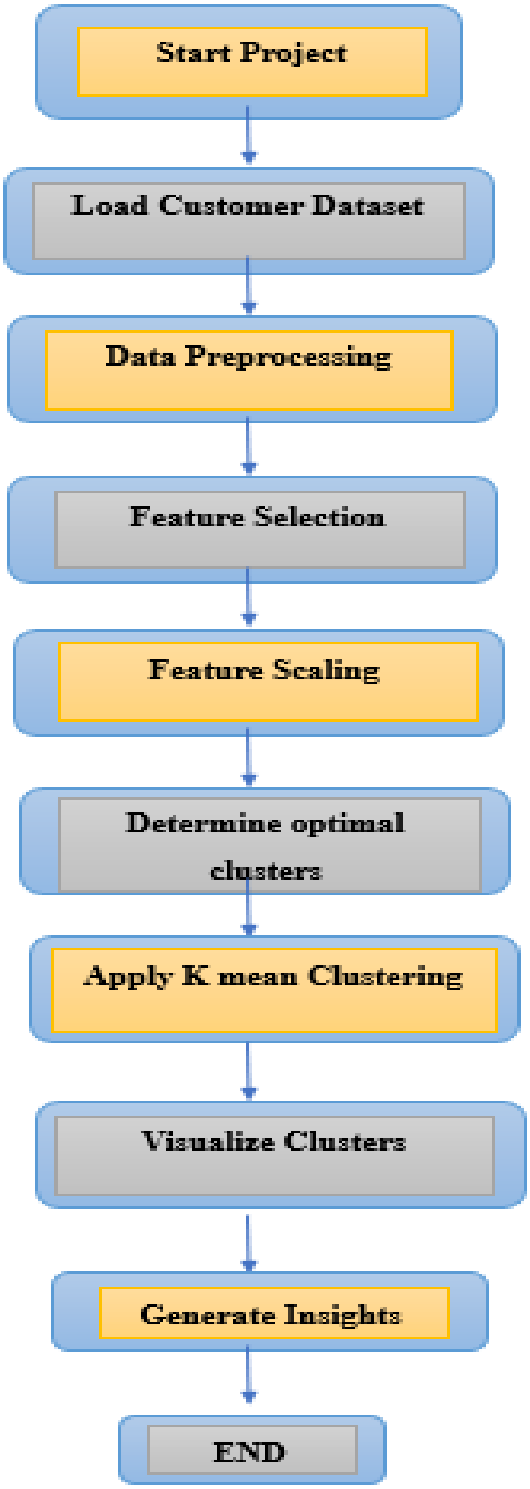
2.8 Visualization and Interpretation

- A scatter plot is generated using the two PCA components.
- Points are colored according to their cluster assignments.
- This visualization helps identify the distribution, density, and separation of the clusters.
- Statistical summaries for each cluster (mean age, income, spending score) are used to interpret the characteristics of each customer group.

2.9 Tools and Technologies Used

- **Python** for coding and analysis
- **Pandas** for data manipulation
- **Matplotlib & Seaborn** for visualization
- **Scikit-learn** for machine learning and clustering
- **Google Colab** as the development environment

3. Flowchart : Customer Segmentation using Unsupervised Learning



4. Code Implementation

4.1 Installing Required Libraries

To ensure all necessary packages are available, we install seaborn for advanced visualizations. Most other libraries come pre-installed in Google Colab.

```
# Step 1: Install Required Libraries (if needed)
!pip install seaborn --quiet
```

4.2 Importing Necessary Libraries

These libraries are used for data manipulation (numpy, pandas), visualization (matplotlib, seaborn), machine learning (KMeans), preprocessing (StandardScaler), and dimensionality reduction (PCA).

```
# Step 2: Import Required Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

4.3 Loading the Dataset

The dataset used is the "Mall Customers" dataset, which contains customer demographic and spending behavior information. It is loaded from a local path in Colab.

```
# Step 3: Load Dataset (Mall Customers Dataset from GitHub)

data = pd.read_csv("/content/Mall_Customers.csv")
data.head()
```

4.4 Exploratory Data Analysis (EDA)

Here, we examine the structure, data types, and summary statistics of the dataset to understand its contents and identify potential issues like missing values.

```
# Step 4: Explore the Data
print("\nDataset Info:")
print(data.info())

print("\nSummary Statistics:")
print(data.describe())
```

4.5 Data Preprocessing

The categorical column "Genre" (gender) is converted into numerical format using label encoding to make it compatible with machine learning algorithms.

```
# Step 5: Encode Gender Column
data['Genre'] = data['Genre'].map({'Male': 0, 'Female': 1})
```

4.6 Feature Selection

We select relevant features for clustering. These include gender, age, income, and spending score, which provide a good basis for customer segmentation.

```
# Step 6: Feature Selection
X = data[['Genre', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)']]
```

4.7 Feature Scaling

StandardScaler is applied to normalize the feature values so that each feature contributes equally to the distance calculations used by K-Means.

```
# Optional: Scale the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

4.8 Finding the Optimal Number of Clusters (Elbow Method)

The Elbow Method helps determine the optimal number of clusters (k) by plotting the Within-Cluster Sum of Squares (WCSS). The "elbow point" in the curve indicates the best k.

```
# Step 7: Find Optimal Number of Clusters using Elbow Method
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)

# Plot the Elbow Curve
plt.figure(figsize=(8, 5))
plt.plot(range(1, 11), wcss, 'bo-')
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

4.9 Applying K-Means Clustering

K-Means clustering is applied with the chosen number of clusters (e.g., 5). Each customer is assigned to one of the clusters.

```
# Step 8: Apply KMeans Clustering
kmeans = KMeans(n_clusters=5, init='k-means++', random_state=42)
clusters = kmeans.fit_predict(X_scaled)
data['Cluster'] = clusters
```

4.10 Visualizing Clusters Using PCA

PCA (Principal Component Analysis) reduces the dimensionality of the dataset to 2 components for easier visualization. A scatter plot shows how customers are grouped into different clusters.

```
# Step 10: View Clustered Data
print("\nClustered Data Sample:")
print(data.groupby('Cluster').mean())
```

5: Output Explanation

This chapter provides a detailed explanation of the outputs generated during the implementation of the customer segmentation model. Each output corresponds to a key step in the clustering process, revealing important insights into the structure and behavior of customer data.

5.1 Elbow Curve Plot

Output: A line plot showing the Within-Cluster Sum of Squares (WCSS) for different values of *k* (number of clusters).

Explanation:

- The Elbow Method helps identify the optimal number of clusters.
- In the plot, the WCSS decreases as the number of clusters increases.
- The "elbow point" (where the curve bends) usually represents the best choice of clusters. In our case, this occurs at **k = 5**.
- This means 5 is the ideal number of customer segments that balance accuracy and simplicity.

5.2 Cluster Assignment

Output: Each customer is labeled with a **Cluster ID** ranging from 0 to 4.

Explanation:

- The K-Means algorithm assigns each customer to one of the 5 clusters based on similarity in features such as age, gender, income, and spending score.
- These cluster labels are stored in a new column Cluster in the dataset.
- This assignment helps categorize customers into distinct groups for targeted marketing and personalized strategies.

5.3 PCA-Based Cluster Visualization

Output: A 2D scatter plot displaying customers grouped by clusters using PCA (Principal Component Analysis).

Explanation:

- The data is reduced to two principal components for easy visualization.
- Each color in the scatter plot represents a different customer cluster.
- Clearly separated clusters show that the segmentation has effectively grouped similar customers together.
- This visual proof of distinct groupings validates the success of the clustering algorithm.

5.4 Cluster Profile Summary

Output: A table showing the **average values** of features (e.g., Age, Income, Spending Score) for each cluster.

Explanation:

- This aggregated data helps interpret each customer segment.
- For example:
    - **Cluster 0** might represent young high spenders.
    - **Cluster 1** could be older customers with low spending scores.
    - **Cluster 3** might include middle-aged customers with average income and moderate spending.
- Such segmentation enables businesses to design tailored strategies for each group.

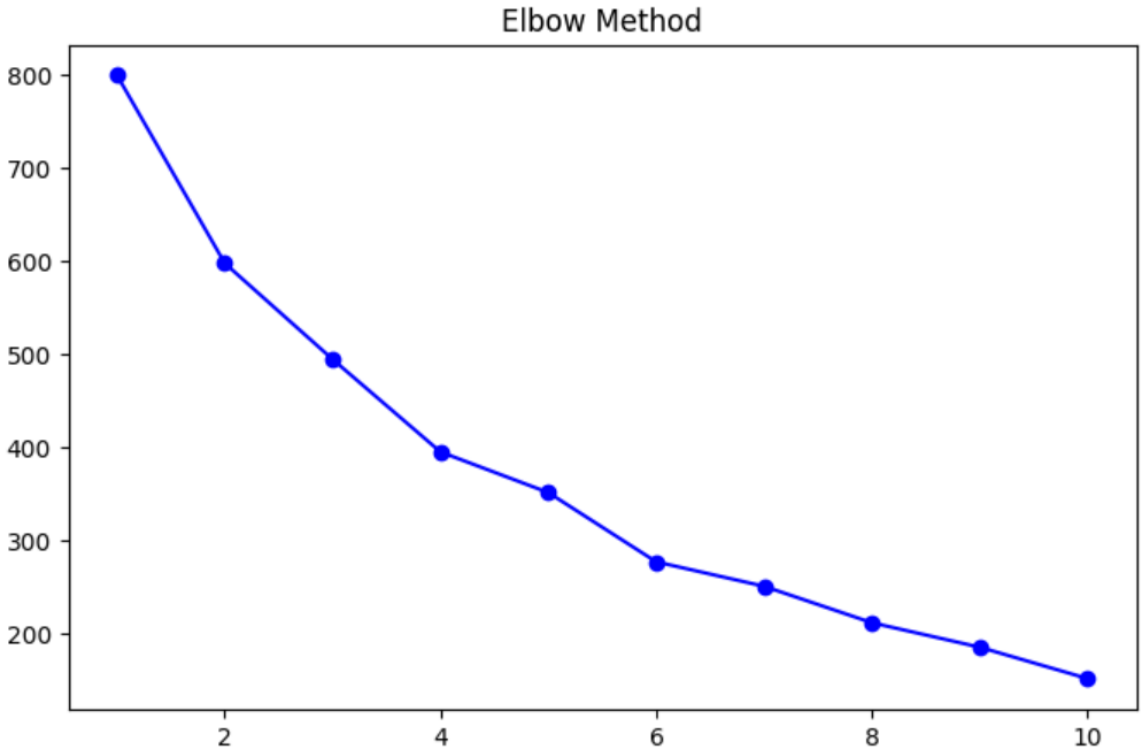
5.5 Key Observations from Output

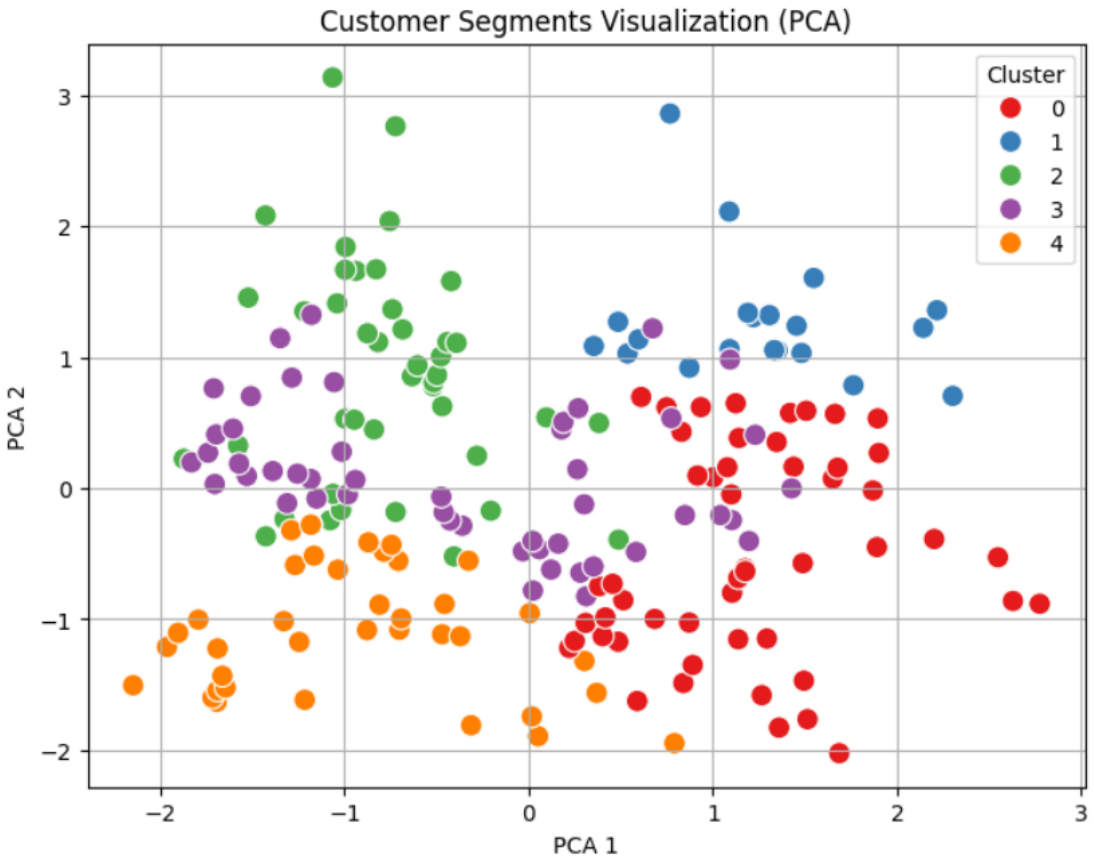
- The algorithm successfully divides customers into meaningful and interpretable segments.
- Each cluster has unique characteristics that can be aligned with specific business actions.
- PCA visualization confirms that the clustering is well-separated and effective.

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Genre                   200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None

Summary Statistics:
       CustomerID         Age  Annual Income (k$)  Spending Score (1-100)
count  200.000000  200.000000          200.000000              200.000000
mean   100.500000   38.850000           60.560000               50.200000
std     57.879185   13.969007           26.264721               25.823522
min      1.000000   18.000000           15.000000                1.000000
25%     50.750000   28.750000           41.500000               34.750000
50%    100.500000   36.000000           61.500000               50.000000
75%    150.250000   49.000000           78.000000               73.000000
max    200.000000   70.000000          137.000000               99.000000
```

✓ 14s completed at 5:38 PM






```
Clustered Data Sample:
         CustomerID     Genre        Age  Annual Income (k$)  \
Cluster
0         65.333333  0.490196  56.470588           46.098039
1        159.500000  0.000000  39.500000           85.150000
2        100.809524  0.000000  28.690476           60.904762
3        151.510204  1.000000  37.897959           82.122449
4         50.526316  1.000000  27.315789           38.842105

         Spending Score (1-100)
Cluster
0                     39.313725
1                     14.050000
2                     70.238095
3                     54.448980
4                     56.210526
```

6. Conclusion

This project successfully demonstrates the practical application of unsupervised learning techniques—specifically the **K-Means clustering algorithm**—for customer segmentation. By analyzing key features such as gender, age, annual income, and spending score, we were able to identify distinct customer groups with similar purchasing behavior and demographic traits.
Through preprocessing and scaling of the data, determining the optimal number of clusters using the **Elbow Method**, and visualizing the results using **PCA**, the project showcased a full machine learning pipeline from raw data to insightful business segmentation.
The five identified clusters reflect real-world customer diversity, offering a valuable framework for:

- Targeted marketing campaigns
- Personalized customer experiences
- Improved customer relationship management
- Strategic decision-making based on data insights

Key Takeaways:

- Unsupervised learning can uncover hidden patterns in customer behavior without requiring labeled data.
- K-Means clustering is effective, fast, and interpretable for customer analytics tasks.
- Visualizations such as PCA help validate the clustering and make insights accessible to non-technical stakeholders.

In conclusion, this project emphasizes how machine learning, when integrated with business intelligence, can transform raw customer data into actionable strategies that enhance both **customer satisfaction** and **business performance**.