# Customer Segmentation Using Unsupervised Learning

**A PROJECT FILE**
**for**
**INTRODUCTION TO AI (AI201B)**
**Session (2024-25)**

**Submitted by**

**Kunal Prajapati**
**202410116100108**
**Doulat Biswal**
**202410116100070**
**Krishna**
**2024101161000103**
**Harsh Aggarwal**
**202410116100081**

**Submitted in partial fulfilment of the**
**Requirements for the Degree of**

**MASTER OF COMPUTER APPLICATION**

**Under the Supervision of**
**Mr. APOORV JAIN**
**Assistant Professor**

**Submitted to**

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

# TABLE OF CONTENT

# INTRODUCTION

## 1. Introduction

### 1.1 Background

Customer segmentation is the process of dividing a company's customer base into meaningful groups based on common characteristics. Traditionally done using domain knowledge and rules, AI now makes this process faster, more accurate, and scalable. In this project, we use **unsupervised learning** to perform customer segmentation using clustering algorithms.

### 1.2 Motivation

Manual segmentation of customers is slow, limited in accuracy, and doesn’t scale well with data. Unsupervised machine learning enables automatic discovery of hidden patterns in customer data. This leads to more personalized marketing, better customer satisfaction, and improved business strategy.

### 1.3 Objectives

- Apply unsupervised learning (KMeans Clustering) to segment customers.
- Discover patterns based on spending behavior and demographics.
- Visualize the customer clusters using dimensionality reduction techniques.
- Understand the characteristics of each cluster to aid marketing strategies.

### 1.4 Significance

This project demonstrates how unsupervised AI techniques like clustering can help organizations make data-driven decisions about marketing and customer relationship management. Segments discovered automatically can be used for targeted campaigns, loyalty programs, and improved service delivery.

# OVERVIEW

We used a dataset from Kaggle that contains detailed customer demographic information along with their transactional behaviour. The dataset includes features like **Income, Marital Status, Education**, and spending on various product categories such as wines, meat, fruits, and more.

To better analyse customer spending, we created a new feature called **Total Spending**, which sums up the expenditures on all product types. We then selected relevant features including **Income, Recency, Education level, family composition**, and **Total Spending** for clustering purposes.

Using **KMeans Clustering**, customers were automatically grouped into segments based on similarities in their data attributes. To visualize these segments, we applied **PCA (Principal Component Analysis)**, which reduces high-dimensional data to two components, allowing us to plot the customer segments clearly in a scatter plot.

This clustering approach helps businesses identify different customer profiles and tailor their services and marketing strategies accordingly

# METHODOLOGY

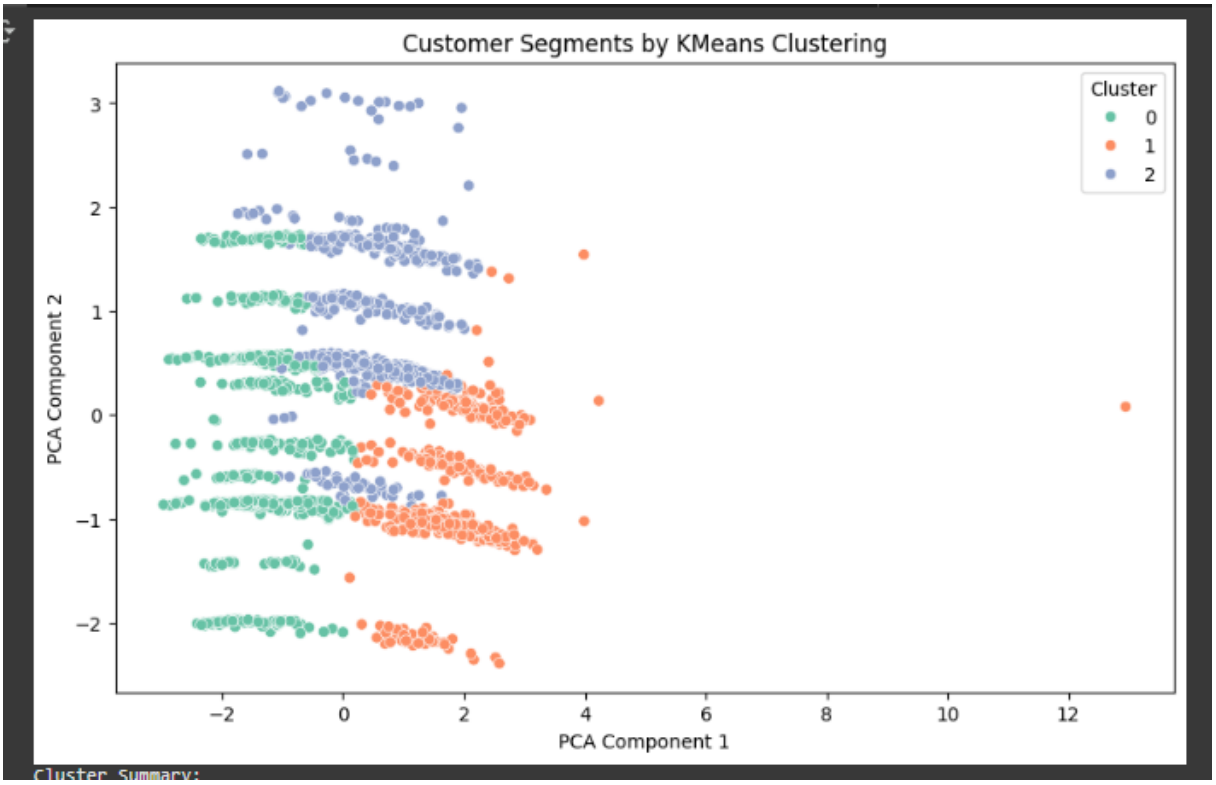
1. **Data Collection:**
    - Dataset from Kaggle: marketing_campaign.csv
2. **Preprocessing:**
    - Removed irrelevant columns like ID, date, and campaign cost.
    - Handled missing values.
    - Categorical variables (Education, Marital Status) were encoded into numbers.
3. **Feature Engineering:**
    - Created Total_Spend by summing spending across products.
    - Selected features like Income, Recency, Education, and Kid/Teen count for clustering.
4. **Feature Scaling:**
    - StandardScaler was used to normalize data before clustering.
5. **Clustering with KMeans:**
    - Used KMeans(n_clusters=3) to segment customers.
    - Each customer was assigned to a cluster label (0, 1, or 2).
6. **Visualization:**
    - PCA was applied to reduce data to 2 dimensions for visualization.
    - Scatterplot created to show how customers are distributed across clusters.
7. **Analysis:**
    - Mean values of features in each cluster were used to describe segment behavior.

# CODE

```
# 1. Import required libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

# 2. Load dataset
df = pd.read_csv('/content/marketing_campaign.csv', sep='\t')

# 3. Data Cleaning
df.drop(['ID', 'Dt_Customer', 'Z_CostContact', 'Z_Revenue'], axis=1, inplace=True)
df.dropna(inplace=True)

# 4. Encode categorical variables
le = LabelEncoder()
for col in ['Education', 'Marital_Status']:
    df[col] = le.fit_transform(df[col])

# 5. Create Total Spending feature
df['Total_Spend'] = df[['MntWines', 'MntFruits', 'MntMeatProducts',
                        'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']].sum(axis=1)

# 6. Select features for clustering
features = ['Income', 'Recency', 'Education', 'Marital_Status', 'Kidhome',
            'Teenhome', 'Total_Spend']
X = df[features]

# 7. Feature Scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 8. KMeans Clustering
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(X_scaled)

# 9. Visualize clusters using PCA (2D)
pca = PCA(n_components=2)
pca_result = pca.fit_transform(X_scaled)
df['PCA1'] = pca_result[:, 0]
```

```
df['PCA2'] = pca_result[:, 1]

plt.figure(figsize=(10, 6))
sns.scatterplot(x='PCA1', y='PCA2', hue='Cluster', data=df, palette='Set2')
plt.title('Customer Segments by KMeans Clustering')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.show()

# 10. Analyze cluster characteristics
cluster_summary = df.groupby('Cluster')[features].mean()
print("Cluster Summary:\n", cluster_summary)
```

# OUTPUT




```
Cluster Summary:
                Income    Recency  Education  Marital_Status   Kidhome  \
Cluster
0        34438.377551  48.788776   2.222449        3.702041  0.910204
1        77690.590994  49.814259   2.454034        3.776735  0.041276
2        57782.665718  48.716927   2.587482        3.721195  0.092461

           Teenhome  Total_Spend
Cluster
0        0.362245   137.813265
1        0.035647  1387.679174
2        1.061166   669.401138
```

# CONCLUSION

This project implemented customer segmentation using **unsupervised learning** with the KMeans algorithm. It successfully grouped customers into segments based on demographic and spending data.

The PCA visualization helped to observe how the clusters are separated, and the analysis of feature averages in each cluster provided business insights.

This unsupervised approach eliminates the need for labeled training data, making it ideal for real-world business applications where class labels are not predefined.

In the future, more advanced clustering techniques like DBSCAN or hierarchical clustering can be explored for better accuracy and scalability. This project highlights the power of AI in enhancing customer understanding and decision-making.

.