# Customer Segmentation Using Supervised 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 a key task in marketing that involves dividing a customer base into distinct groups that exhibit similar behaviors or characteristics. It helps organizations tailor their products, services, and marketing strategies. With the rise of artificial intelligence, customer segmentation can be achieved effectively using machine learning techniques such as supervised learning.

### 1.2 Motivation

Understanding customer behavior is crucial for businesses to thrive in competitive markets. Segmenting customers based on behavior and spending helps identify high-value customers and create personalized campaigns. Manual segmentation is inefficient and lacks accuracy; hence, AI-based segmentation offers a smarter, faster alternative.

### 1.3 Objectives

- Implement a supervised learning model to classify customers into segments.
- Analyze customer data to identify patterns based on spending and demographics.
- Automate segmentation for smarter marketing and business strategies.

### 1.4 Significance

This project showcases the practical application of supervised learning in a business scenario. By automating segmentation, companies can better allocate resources, boost customer satisfaction, and increase return on investment.

# OVERVIEW

The project uses a marketing campaign dataset that includes details such as customer income, education, age, marital status, and spending on various products. The goal is to classify customers into segments (Low, Medium, High) based on their total spending.

The dataset is preprocessed to remove unnecessary columns, handle missing values, and convert categorical data into numerical format. The total amount spent by each customer is calculated, and based on this value, the customers are divided into three segments using quantile-based binning.

A supervised machine learning model, specifically a Decision Tree Classifier, is trained on the features to predict customer segments. The model is evaluated for accuracy and analyzed to understand which features influence the segmentation the most.

# METHODOLOGY

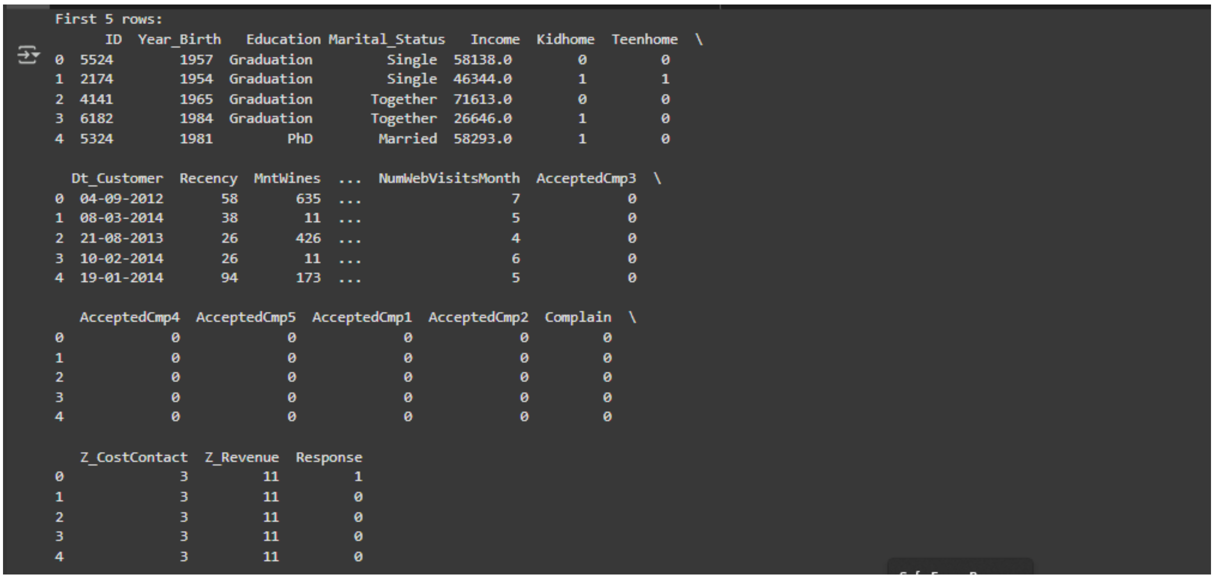
1. **Data Collection:**
   - Dataset sourced from Kaggle (marketing_campaign.csv).
2. **Data Preprocessing:**
   - Removed irrelevant columns like IDs and date fields.
   - Handled missing values and encoded categorical variables (Education, Marital Status).
3. **Target Variable Creation:**
   - Total spending = sum of spending on wine, fruits, meat, fish, sweets, and gold products.
   - Customers categorized into three segments: Low, Medium, High based on spending.
4. **Model Selection:**
   - Supervised learning model used: **Decision Tree Classifier**.
   - Features selected: Age, Income, Education, Marital Status, Kid/Teen count, and various product spends.
5. **Training and Testing:**
   - Dataset split into training and testing sets (70/30).
   - Model trained on the training data.
6. **Evaluation:**
   - Accuracy score and classification report used to evaluate performance.
   - Feature importance visualized using bar chart.

# CODE

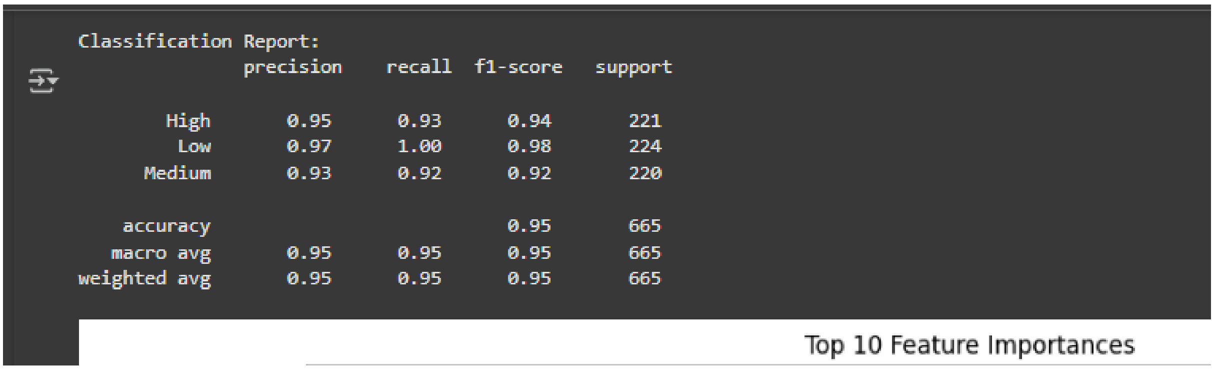
```
# Step 1: Import libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report

# Step 2: Load the dataset
df = pd.read_csv('/content/marketing_campaign.csv', sep='\t')  # Adjust path if needed
print("First 5 rows:\n", df.head())

# Step 3: Drop unnecessary columns (like ID, date fields if not needed)
df.drop(['ID', 'Dt_Customer', 'Z_CostContact', 'Z_Revenue'], axis=1, inplace=True)

# Step 4: Handle missing values
df.dropna(inplace=True)

# Step 5: Encode categorical variables
categorical_cols = ['Education', 'Marital_Status']
le = LabelEncoder()
for col in categorical_cols:
    df[col] = le.fit_transform(df[col])

# Step 6: Create a target variable
# Let's define 'High-Value Customers' based on total spend (can be adjusted)
df['Total_Spend'] = df[['MntWines', 'MntFruits', 'MntMeatProducts',
                        'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']].sum(axis=1)
df['Segment'] = pd.qcut(df['Total_Spend'], q=3, labels=['Low', 'Medium', 'High'])  # 3
categories

# Step 7: Define features and target
X = df.drop(['Segment', 'Total_Spend'], axis=1)
y = df['Segment']

# Step 8: Train/Test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Step 9: Train Decision Tree Classifier
model = DecisionTreeClassifier()
```

```
model.fit(X_train, y_train)

# Step 10: Evaluate
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))

# Optional: Feature Importance Visualization
plt.figure(figsize=(12, 6))
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(10).plot(kind='barh')
plt.title("Top 10 Feature Importances")
plt.show()
```

# OUTPUT

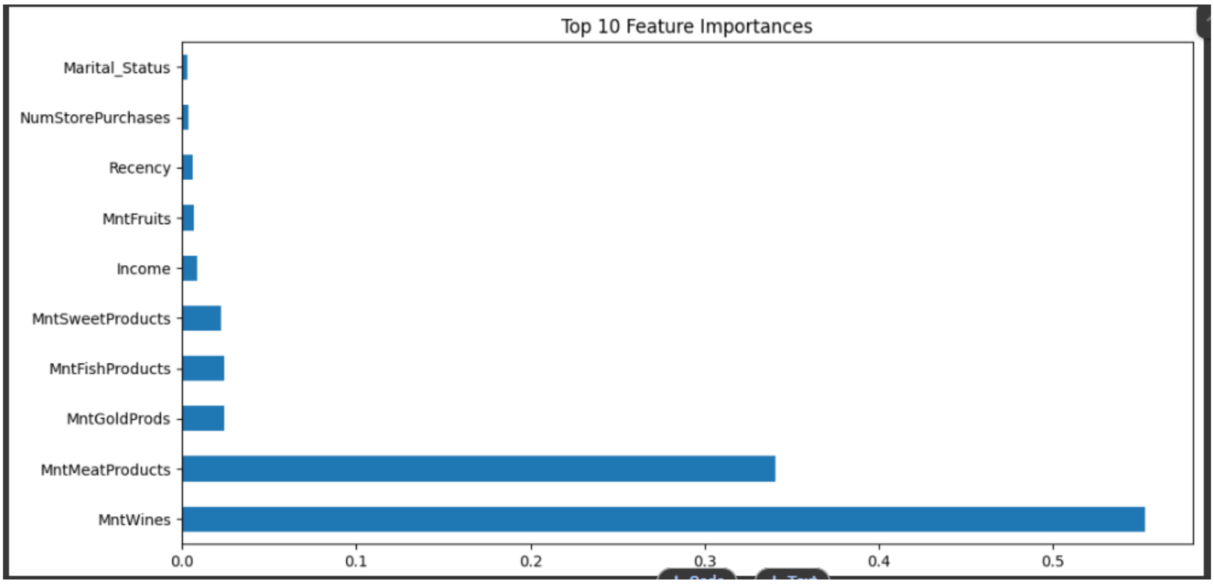


```
First 5 rows:
     ID  Year_Birth   Education Marital_Status   Income  Kidhome  Teenhome  \
0  5524        1957  Graduation         Single  58138.0        0         0   
1  2174        1954  Graduation         Single  46344.0        1         1   
2  4141        1965  Graduation       Together  71613.0        0         0   
3  6182        1984  Graduation       Together  26646.0        1         0   
4  5324        1981         PhD        Married  58293.0        1         0   

  Dt_Customer  Recency  MntWines  ...  NumWebVisitsMonth  AcceptedCmp3  \
0  04-09-2012       58       635  ...                  7             0   
1  08-03-2014       38        11  ...                  5             0   
2  21-08-2013       26       426  ...                  4             0   
3  10-02-2014       26        11  ...                  6             0   
4  19-01-2014       94       173  ...                  5             0   

   AcceptedCmp4  AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  \
0             0             0             0             0         0   
1             0             0             0             0         0   
2             0             0             0             0         0   
3             0             0             0             0         0   
4             0             0             0             0         0   

   Z_CostContact  Z_Revenue  Response  
0              3         11         1  
1              3         11         0  
2              3         11         0  
3              3         11         0  
4              3         11         0  
```



```
Classification Report:
              precision    recall  f1-score   support

        High       0.95      0.93      0.94       221
         Low       0.97      1.00      0.98       224
      Medium       0.93      0.92      0.92       220

    accuracy                           0.95       665
   macro avg       0.95      0.95      0.95       665
weighted avg       0.95      0.95      0.95       665
```

Top 10 Feature Importances



Top 10 Feature Importances

# CONCLUSION

This project successfully implements customer segmentation using supervised learning. The use of a Decision Tree classifier enabled accurate classification of customers into Low, Medium, and High-value segments based on spending patterns and demographic features.

The approach demonstrated how AI and machine learning can be applied to real-world business problems like marketing personalization and campaign optimization. With feature importance analysis, businesses can also understand which customer attributes most influence purchasing behavior.

Future improvements could involve using ensemble methods like Random Forests or Gradient Boosting to improve accuracy, or exploring deep learning for larger datasets. The integration of automated pipelines can further streamline segmentation in enterprise settings.

In conclusion, this project showcases how AI can empower smarter decision-making in marketing through efficient and automated customer segmentation.

.