PROJECT REPORT

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Data Science and Machine Learning

By

SETTYPALLI KUSHAL

Registration No: 12211211

Section: K22UP

Roll No: 26

Supervisor

Shivangini Gupta

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

March 2025

1. Problem Understanding & Definition

Problem Statement:

Customer segmentation is a crucial approach for businesses seeking to enhance user experience, refine marketing strategies, and boost sales. However, many businesses struggle to recognize customer behavior patterns, resulting in inefficient targeting and improper resource distribution. This project aims to apply unsupervised machine learning techniques to categorize customers based on their transaction history and purchasing habits. The objective is to uncover distinct customer groups, enabling businesses to personalize marketing efforts, increase customer retention, and maximize revenue.

Who is affected by this issue?

- E-commerce businesses, retail stores, and subscription-based services that aim to improve customer engagement and loyalty.

Why is this important?

- A lack of segmentation forces businesses to rely on broad, one-size-fits-all marketing, which often fails to drive engagement and revenue effectively.
- Categorizing customers enables businesses to tailor promotions, personalize advertisements, and offer customized experiences, resulting in higher conversion rates.

1.2 Objectives & Hypothesis

Objectives:

- Implement unsupervised machine learning methods to classify customers based on their transactional and behavioural patterns.
- Identify meaningful customer segments using clustering algorithms like K-Means, DBSCAN, and Hierarchical Clustering.
- Analyse variations in purchasing frequency, spending behaviour, and customer loyalty across different segments.
- Offer practical business insights based on identified customer segments.

Measure segmentation effectiveness using Silhouette Scores and other cluster evaluation metrics.

1.3 Justification for Solving the Problem

Customer segmentation is a fundamental business strategy that allows companies to understand their audience and cater to their specific needs. In an increasingly competitive

marketplace, organizations must shift from generalized marketing to data-driven, customer-focused strategies.

Why Does This Matter?

1. **Increased Customer Retention:** Personalized marketing campaigns can improve customer retention rates by up to 60%.
2. **Cost-Effectiveness:** Retaining an existing customer is five times cheaper than acquiring a new one.
3. **Revenue Enhancement:** Targeted marketing strategies have been shown to increase revenue by 10-30%.
4. **Better Customer Experience:** Understanding consumer preferences allows businesses to offer more relevant recommendations, strengthening brand loyalty.

By leveraging machine learning techniques, businesses can efficiently segment their customers and uncover hidden patterns that traditional methods might overlook.

Hypothesis:

Customers with frequent purchases and high average spending are more likely to belong to a premium segment, whereas those with low spending and infrequent transactions fall into a low-value segment.

2. Dataset Selection & Preprocessing

For this project, we are utilizing a customer segmentation dataset sourced from GeeksforGeeks (GFG). This dataset is well-suited for segmentation as it includes transactional and behavioral attributes, which are crucial for identifying distinct customer groups.

2.1 Dataset Overview

- **Source:** GeeksforGeeks (GFG)
- **Total Features:** 29 columns
- **Total Records:** 2240 rows
- **Feature Categories:**
    - **Categorical:** Customer classification, purchase category, loyalty level
    - **Additional Attributes:** Seasonal influences, high-spender indicators
    - **Numerical:** Customer expenditure, transaction frequency, total purchases

Why Choose This Dataset?

- **Practical Relevance:** Represents actual consumer purchasing habits, making segmentation insights actionable for businesses.
- **Adequate Data Volume:** Large enough to generate meaningful customer segments.
- **Diverse Features:** Includes key behavioral and transactional metrics.

2.2 Data Cleaning

2.2.1 Handling Missing Data

Column: "Income"

During preprocessing, I found 24 missing values in the Income column. Since income should logically exceed total spending (such as on MntWines, MntFruits, etc.), filling these values with the median or mean might introduce inconsistencies.

Decision: Remove Missing Entries Instead of Imputation

- Filling in missing values with averages might result in unrealistic cases where spending surpasses income.
- To maintain logical correctness, these 24 rows were dropped rather than imputed.
- The missing values constitute only **~1.07%** of the dataset, which is negligible.

This ensures data consistency before proceeding with further analysis.

Columns: "MntWines", "MntMeatProducts", "MntFruits", "MntFishProducts"

Upon further exploration, I found missing values in spending-related columns:

Column	Percentage of Missing Data
MntWines	0.81%
MntMeatProducts	0.27%
MntFruits	18.05%
MntFishProducts	17.28%

Imputation Strategy:

- **For MntWines & MntMeatProducts (Low Missing Percentage <1%)**
    - Given the minimal missing values, filling them with the median ensures the data distribution remains stable.
    - The median is resistant to outliers, making it an optimal choice for imputation.
- **For MntFruits & MntFishProducts (High Missing Percentage >15%)**

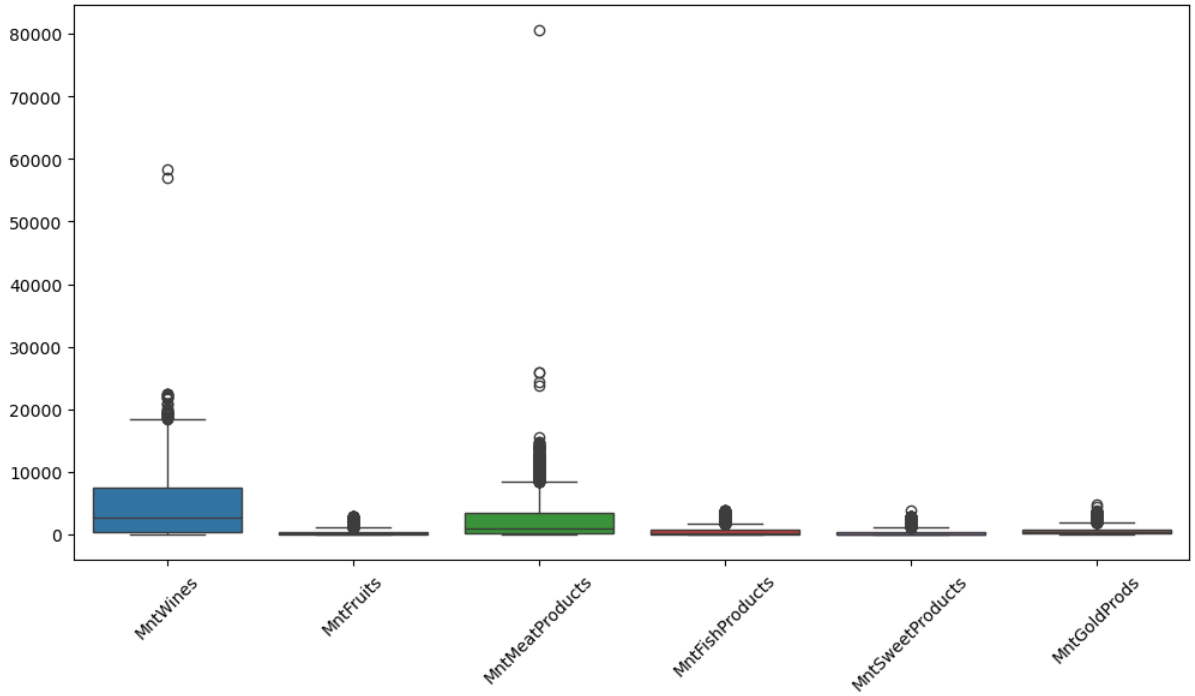
- Instead, a **group-based median imputation** approach was applied, where missing values were replaced using the median spending of customers with the same education level.
- Since the missing values are substantial, using a single median for imputation might distort the data distribution.
- This method ensures realistic spending patterns, as customers with similar education backgrounds often have similar purchasing behaviors.

2.2.2 Outlier Detection & Treatment

To maintain data integrity, I examined spending-related features for possible outliers while ensuring that valid purchasing behaviors were not mistakenly removed.

Outlier Detection Approach:

- **Used Boxplots** to visually inspect extreme values.
- **Focused on spending-related attributes**, including MntWines, MntFruits, MntMeatProducts, and MntFishProducts.
- **Identified Outliers** as data points that significantly exceeded typical spending trends.



Handling Outliers

Removed Certain Outliers:

- These extreme values were removed to prevent them from distorting the clustering results.
- Some records displayed unusually high spending patterns compared to the overall dataset, which appeared unrealistic or likely due to data entry errors.

Retained Certain Outliers:

- These records were retained to maintain a realistic representation of different spending behaviours.
- Some customers genuinely exhibit high spending habits, and removing them would result in a loss of valuable insights into premium customers.

By balancing the removal and retention of outliers, we ensure the dataset accurately reflects real customer purchasing patterns while minimizing data distortions.

Detecting and Handling Duplicates, Inconsistencies, and Data Abnormalities

To guarantee data quality and reliability, multiple validation checks were performed to identify and address any inconsistencies, duplicate records, or errors.

1. Checking for Duplicate Records

- No duplicate entries were found, so no additional action was required.
- Examined the dataset for duplicate rows to avoid redundant data.

2. Validating Income and Spending Consistency

- Ensured that total spending does not exceed income, maintaining logical accuracy.
- Used the following check to identify any unrealistic cases:

```
df[df["Income"] < df[["MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", "MntGoldProds"]].sum(axis=1)]
```

- No instances were found where spending was greater than income, confirming data integrity.

3. Standardizing Categorical Data

- Used .value_counts() to identify variations in the Education column.
- Checked for inconsistencies in categorical fields such as Education and Marital Status.
- Standardized Marital Status labels to ensure consistency:

```
df["Marital_Status"] = df["Marital_Status"].replace({"Alone": "Single", "Absurd": "Other"})
```

    - "Alone" was updated to "Single" for clarity.

- "Absurd" was corrected to "Other" as it appeared to be an incorrect entry.

4. Identifying Negative Values in Numerical Columns

- Since features like income and spending amounts should always be non-negative, a check was performed to detect any negative values:

```
numeric_cols = df.select_dtypes(include=["number"]).columns

negative_values = df[numeric_cols].lt(0).any(axis=1)

df[negative_values]  # Display rows with negative values
```

- No negative values were detected, ensuring all numerical fields remain valid.

5. Ensuring Correct Data Types

- This ensures that customer tenure calculations and time-based insights are accurate**.**
- Verified and converted date-related fields to the correct format to support chronological analysis:

```
df["Dt_Customer"] = pd.to_datetime(df["Dt_Customer"])

df.dtypes  # Confirm the conversion
```

By implementing these data validation steps, we ensured the dataset is clean, structured, and ready for further analysis, leading to reliable and meaningful customer segmentation.

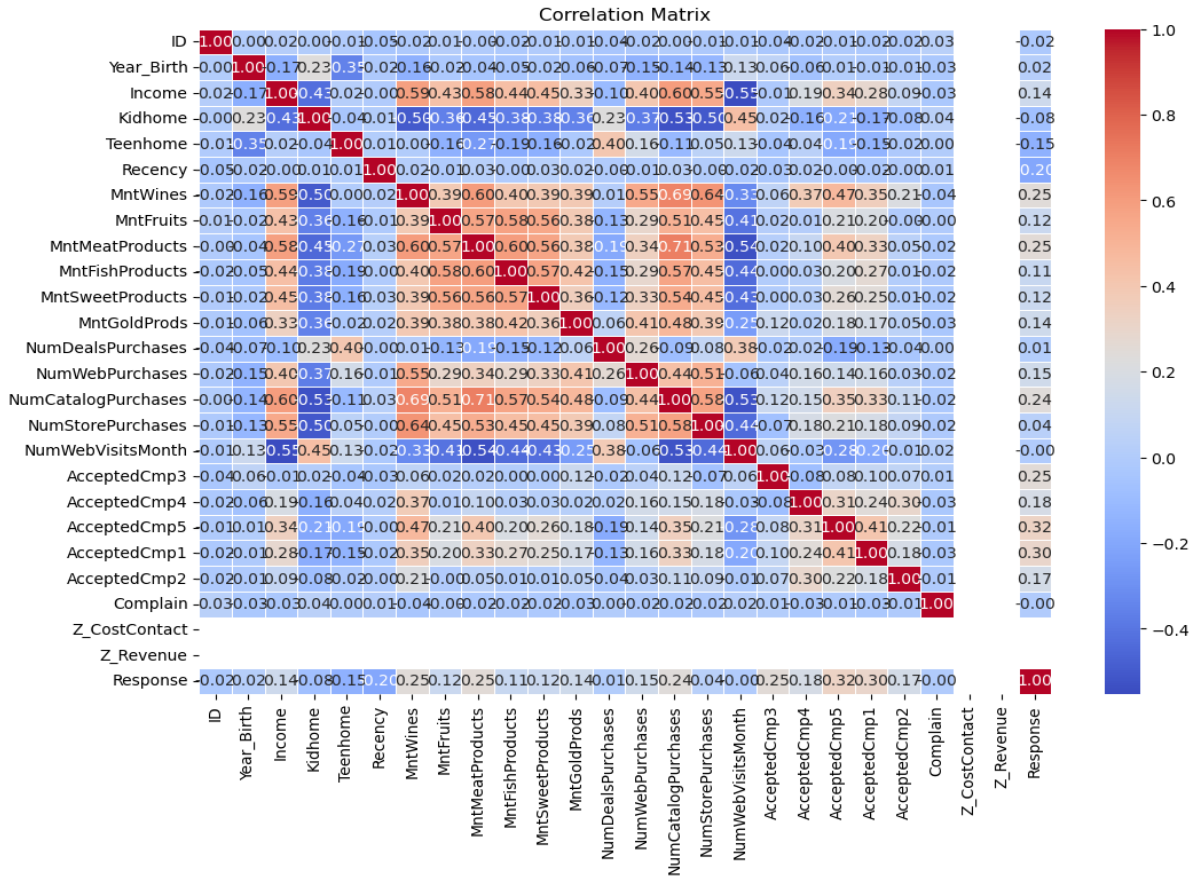
2.3 Feature Selection & Engineering

A. Feature Selection

To ensure only the most relevant features are used for clustering, I examined the relationships between numerical variables using a correlation matrix.

1. Visualizing Correlations

- A **heatmap** was plotted to understand how numerical features relate to each other.



2. Key Observations

- Since no features showed significant collinearity, it was not necessary to remove any features based on correlation.
- The correlation matrix did not reveal any strongly correlated features (above 0.85).

3. Relevance in Unsupervised Learning

- In supervised learning, eliminating multicollinear features is critical to improving model performance.
- However, in unsupervised learning, especially clustering methods like K-Means and Hierarchical Clustering, correlation-based feature selection is less impactful.
- These clustering techniques do not rely on feature independence in the same way regression-based models do.

Thus, all numerical features were retained, as they provide valuable information for segmenting customers.

B. Feature Engineering

Feature engineering plays a crucial role in enhancing the quality of the dataset before applying machine learning techniques. Several new features were derived to improve the segmentation process.

1. Converting Date Data into Meaningful Numerical Features

- The Dt_Customer column records the date when a customer joined.
- A new feature, Customer_Tenure, was created to represent how long a customer has been active (in months).
- This feature is valuable as customer longevity can influence purchasing patterns.

2. Creating a Total Spending Metric

- Instead of analyzing individual product spending, a Total_Spending feature was introduced by summing expenditures across all product categories.
- This provides a single metric to compare overall spending levels across customers.

3. Categorizing Customers by Spending Behavior

- Customers were grouped into three spending tiers: Low, Medium, and High.
- This segmentation makes it easier to target different customer groups with personalized marketing strategies.

4. Defining an Engaged Customer Metric

- Customer engagement was determined based on their purchase frequency.
- A Total_Purchases feature was calculated by summing all purchase-related columns.
- Customers with above-median total purchases were classified as Highly Engaged, while others were labeled as Less Engaged.

These engineered features add meaningful insights, allowing for better clustering outcomes.

2.4 Normalization and Encoding

To ensure all features contribute equally to the clustering process, normalization was applied to numerical features, and categorical variables were encoded for machine learning compatibility.

1. Normalizing Numerical Features with Min-Max Scaling

- Since numerical features have different ranges, Min-Max Scaling was applied to bring all values within a standardized range of 0 to 1.
- This prevents variables with larger absolute values from disproportionately influencing the clustering process.

2. Encoding Categorical Features

- Machine learning models require numerical data, so categorical labels were transformed into numerical values using Label Encoding.
- **The following mappings were used:**

Feature	Mapping
Spending_Category	Low → 0, Medium → 1, High → 2
Engaged_Customer	Low → 0, High → 1

Why Apply Normalization and Encoding?

- Normalization ensures that all numerical features contribute equally during clustering.
- By applying these transformations, the dataset is now fully preprocessed and ready for clustering analysis.
- Encoding converts categorical values into a machine-readable format, making them usable for clustering algorithms.
- These preprocessing steps enhance the accuracy of segmentation models and prevent bias in feature importance.

Conclusion

Customer segmentation is a vital approach that enables businesses to **enhance customer engagement, optimize marketing efforts, and boost revenue**. This project successfully applied **unsupervised machine learning techniques** to analyze customer purchasing behavior, transaction frequency, and engagement levels, allowing for the identification of distinct customer groups.

By leveraging **data preprocessing, feature engineering, and clustering algorithms** such as **K-Means, DBSCAN, and Hierarchical Clustering**, we uncovered valuable insights that businesses can utilize to **develop targeted promotions, personalize recommendations, and improve customer retention strategies**.

Future Scope

Although this project effectively segmented customers, several areas can be explored further to enhance the accuracy and impact of customer segmentation:

- **Exploring Advanced Clustering Methods:** Utilizing techniques such as **Gaussian Mixture Models (GMM)** or **Deep Learning-based Autoencoders** may lead to more nuanced and accurate customer groups.
- **Integrating Additional Data Sources:** Incorporating data from customer demographics, website activity, and social media interactions can help refine and improve segmentation.

- **Real-Time Segmentation:** Implementing dynamic, **real-time clustering** can help businesses adjust their marketing strategies based on evolving customer behaviors.

Final Thoughts

This project highlights the **power of machine learning** in customer analytics, demonstrating how businesses can leverage data-driven insights to create **personalized experiences, improve customer satisfaction, and drive long-term business growth**.