Unsupervised Classification for Customer Segmentation

Lucía Sánchez Bella

December, 2024

1 Introduction

In today's competitive business environment, understanding customer behavior is essential for tailoring products and services to meet customer needs effectively. Customer Analysis enables businesses to analyze and segment their customer base, identifying patterns and behaviors that inform decision-making. By grouping customers into distinct segments, companies can better allocate resources, optimize marketing strategies, and enhance customer satisfaction.

2 Problem Description

This study will analyze a dataset [1] containing 2240 entries and 28 variables, which provides information about customer demographics, purchasing habits, and interactions with promotional campaigns.

The variables include:

- **Demographics:** ID, Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome.
- **Company Relationship:** Dt_Customer, Recency.
- **Product Spending:** MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds.
- **Promotions and Campaigns:** AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response.
- **Purchases:** NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, NumDealsPurchases

Further information about the variables can be found in [1].

3 Methodology

This section presents the methodology employed, which comprises several essential components: exploratory data analysis, feature engineering, data preprocessing and model implementation.

3.1 Exploratory Data Analysis

The exploratory data analysis in this study comprises the essential steps for understanding the dataset's structure, and detecting any irregularities.

Visual representations have shown that the Marital_Status and Education columns contain categories with few entries that could be grouped for simplicity. Additionally, 24 null values were found in the Income column, representing just 1% of the data, which can be removed without impacting the analysis.

3.2 Feature Engineering

To enhance the quality of the data and simplify the clustering analysis, several feature engineering steps were applied to transform, and reduce the dimensions of the dataset. These steps include:

- **Client Tenure:** Created a new feature, `Days Customer`, based on the *Dt Customer* column.
- **Marital Status:** The `Marital_Status` column was simplified into two categories: 1 for individuals in a relationship (*Married, Together*) and 0 for those not in a relationship (*Divorced, Widow, Alone, YOLO, Absurd, Single*).
- **Education:** The `Education` column was consolidated into two categories: higher education levels (*PhD, 2nd Cycle, Graduation, Master*) were grouped as `1`, and *Basic* was grouped as `0`.
- **Children in the Household:** A new feature `Kids` was created by adding the `Kidhome` and `Teenhome` columns.
- **Total Expenses:** Added spending across product categories into a new `Expenses` feature.
- **Campaign Engagement:** Created `TotalAcceptedCmp`by summing all accepted campaigns.

- **Age:** Calculated the customer's age based on the `Year Birth` column.

3.3 Data Preprocessing

To prepare the dataset for clustering analysis, several preprocessing steps were applied. Rows with missing values were removed, as they represented a minimal proportion of the data and did not significantly impact the dataset. Outliers in numerical features were identified and eliminated using the interquartile range (IQR) to avoid distortion in the clustering process. Numerical features were scaled to a uniform range between 0 and 1, ensuring that variables with larger scales did not dominate the analysis. Finally, categorical variables had already been transformed into binary values during feature engineering, so no further encoding was required.

3.4 Model Implementation

This section outlines the implementation of different clustering techniques used to analyze and group data based on similarity. These techniques include Hierarchical Clustering, Partitional Clustering, and Probabilistic Clustering, all executed using Python's *sklearn* clustering library [2].

3.4.1 Hierarchical Clustering

In the context of hierarchical clustering, agglomerative clustering was implemented, where clusters are formed by iteratively merging the closest pairs. As the choice of method and distance metric can impact the cluster structure, several combinations have been employed.

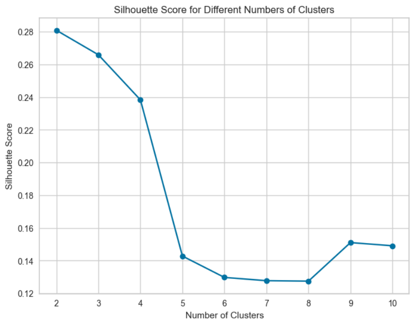


Figure 1: Silhouette Score for case 1

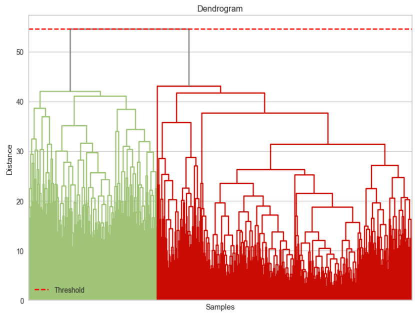


Figure 2: Dendogram for case 1

To determine the optimal number of clusters, the Silhouette Score was calculated for several numbers of clusters, followed by a dendrogram to provide an initial view of the hierarchical structures.

First, in Case 1 the **Complete** linkage method with the **Manhattan** distance metric was used. The Complete linkage method considers the maximum distance between any two points from different clusters, ensuring that all points within a cluster are closely connected, and the Manhattan distance metric, which computes the sum of absolute differences across dimensions, is effective for handling datasets with a large number of variables [3]. As shown in Figure 1, the optimal number of clusters is 2, with a Silhouette score of 0.28.

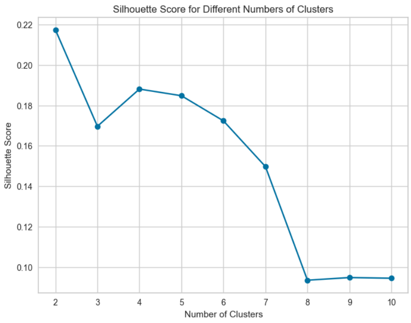


Figure 3: Silhouette Score for case 2

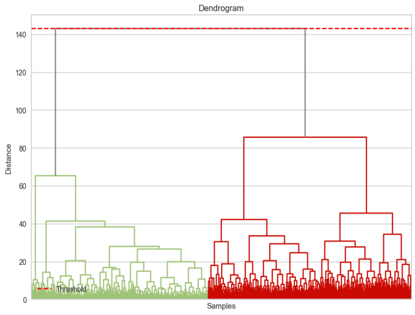


Figure 4: Dendogram for case 2

Then, in Case 2 the **Ward** linkage method with the **Euclidean** distance metric was applied. The Ward method minimizes the variance within each cluster by merging the two clusters that result in the smallest increase in total within-cluster variance [3]. As shown in Figure 3, the optimal number of clusters remains at 2. However, the Silhouette score has slightly decreased to 0.219, indicating a minor reduction in clustering quality.

Finally, in Case 3 the **Complete** linkage method with the **Canberra** distance metric was employed. The Canberra distance is a weighted metric that calculates the sum of absolute differences between two points, scaled by the sum of their absolute values for each dimension [3]. As shown in Figure 5, the optimal number of clusters remains 2, with a Silhouette score of 0.27, which is very similar to the score obtained in Case 1.

Moreover, Figures 2, 4, and 6 illustrate how the choice of method and distance metric affects the criteria used for separating clusters. These differences lead to distinct cluster structures, even when the optimal number of clusters remains the same. Specifically, the number of samples and the hierarchical distribution within each cluster vary depending on the method used in each case.

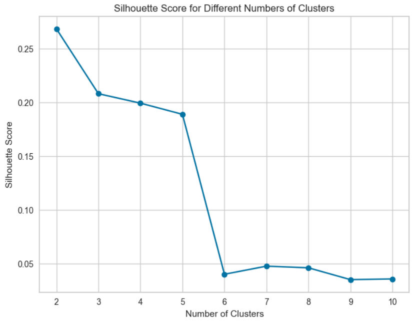


Figure 5: Silhouette Score for case 3

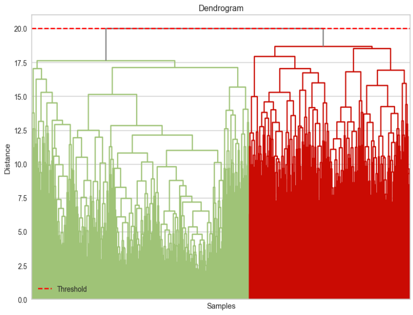


Figure 6: Dendogram for case 3

3.4.2 Partitional Clustering

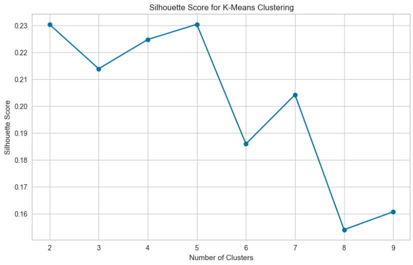


Figure 7: Silhouette Score for K-means

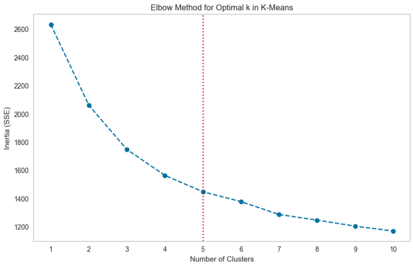


Figure 8: Elbow for k-means

Regarding Partitional Clustering, the K-Means algorithm was implemented, which partitions the data into k clusters by assigning each data point to the nearest cluster center and then recalculating the cluster centroids based on the assigned points. In order to determine the optimal number of clusters, both the Silhouette Score and the Elbow method were used. As indicated in Figure 7, the optimal number of clusters is 5, which aligns with the location of the elbow point represented in Figure 8.

3.4.3 Probabilistic clustering

Finally, the selected probabilistic method was the Gaussian Mixture Model (GMM), implemented using the Expectation-Maximization (EM) algorithm. This model assumes that the data is generated from a mixture of multiple Gaussian distributions and uses the EM algorithm to iteratively estimate the parameters of these distributions to best fit the data.

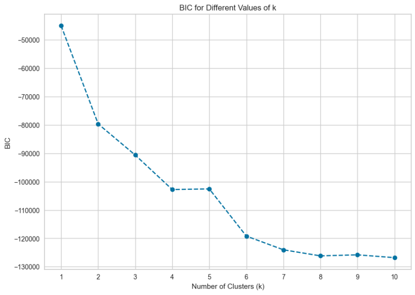


Figure 9: BIC for GMM

To determine the number of clusters, the BIC (Bayesian Information Criterion) coefficient was used. From Figure 9, it can be concluded that the optimal number of clusters, balancing model complexity and BIC values, is 6.

4 Results

4.1 Hierarchical Clustering

The Silhouette Plots from Figures 10, 11, and 12 reveal the performance of hierarchical clustering in three different cases. In Case 1, the clustering quality is moderate, with Cluster 1 showing good cohesion, while Cluster 2 exhibits several points with negative silhouette values, indicating some misclassification or poor separation from Cluster 1. However, Cases 2 and 3 show worse performance for Cluster 2, with a significant number of negative silhouette values, especially in Case 2. This suggests that the clustering in Case 2 and Case 3 is of lower quality compared to Case 1, as was previously mentioned.

Furthermore, Figures 16, 17, and 18 suggest that different criteria have been employed to separate the clusters in each case, as each Decision Tree emphasizes different attributes for the clustering. However, the analysis across the three cases reveals how the clusters overlap consistently in all cases, highlighting the shared patterns and similarities despite the use of varied separation criteria in the trees.

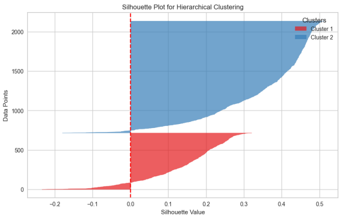


Figure 10: Case 1

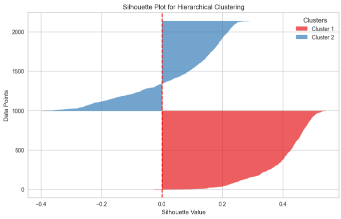


Figure 11: Case 2

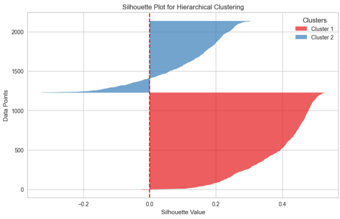


Figure 12: Case 3

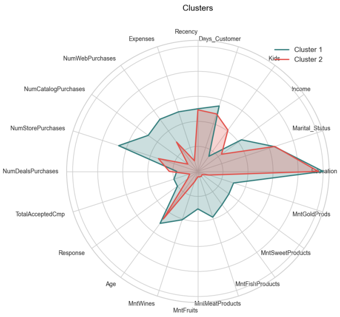


Figure 13: Case 1

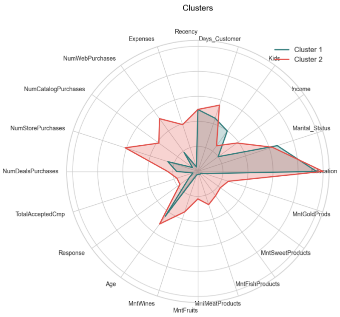


Figure 14: Case 2

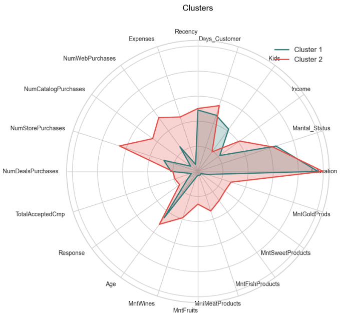


Figure 15: Case 3

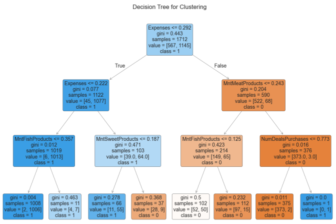

Decision Tree for Clustering
Expenses <= 0.292 gini = 0.443 samples = 1712 value = [567, 1145] class = 1
True
False
Expenses <= 0.222 gini = 0.077 samples = 1122 value = [45, 1077] class = 1
MntMeatProducts <= 0.243 gini = 0.204 samples = 590 value = [522, 68] class = 0
MntFishProducts <= 0.357 gini = 0.012 samples = 1019 value = [6, 1013] class = 1
MntSweetProducts <= 0.187 gini = 0.471 samples = 103 value = [39.0, 64.0] class = 1
MntFishProducts <= 0.125 gini = 0.423 samples = 214 value = [149, 65] class = 0
NumDealsPurchases <= 0.773 gini = 0.016 samples = 376 value = [373.0, 3.0] class = 0
gini = 0.004 samples = 1008 value = [2, 1006] class = 1
gini = 0.463 samples = 11 value = [4, 7] class = 1
gini = 0.278 samples = 66 value = [11, 55] class = 1
gini = 0.368 samples = 37 value = [28, 9] class = 0
gini = 0.5 samples = 102 value = [52, 50] class = 0
gini = 0.232 samples = 112 value = [97, 15] class = 0
gini = 0.011 samples = 375 value = [373, 2] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1

Figure 16: Case 1

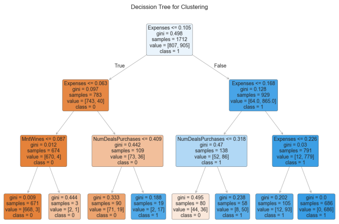

Decission Tree for Clustering
Expenses <= 0.105 gini = 0.498 samples = 1712 value = [807, 905] class = 1
True
False
Expenses <= 0.063 gini = 0.097 samples = 783 value = [743, 40] class = 0
Expenses <= 0.168 gini = 0.128 samples = 929 value = [64.0, 865.0] class = 1
MntWines <= 0.087 gini = 0.012 samples = 674 value = [670, 4] class = 0
NumDealsPurchases <= 0.409 gini = 0.442 samples = 109 value = [73, 36] class = 0
NumDealsPurchases <= 0.318 gini = 0.47 samples = 138 value = [52, 86] class = 1
Expenses <= 0.226 gini = 0.03 samples = 791 value = [12, 779] class = 1
gini = 0.009 samples = 671 value = [668, 3] class = 0
gini = 0.444 samples = 3 value = [2, 1] class = 0
gini = 0.333 samples = 90 value = [71, 19] class = 0
gini = 0.188 samples = 19 value = [2, 17] class = 1
gini = 0.495 samples = 80 value = [44, 36] class = 0
gini = 0.238 samples = 58 value = [8, 50] class = 1
gini = 0.202 samples = 105 value = [12, 93] class = 1
gini = 0.0 samples = 686 value = [0, 686] class = 1

Figure 17: Case 2

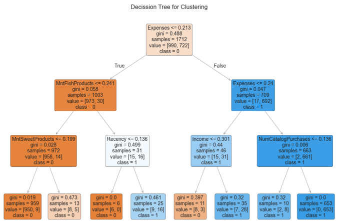
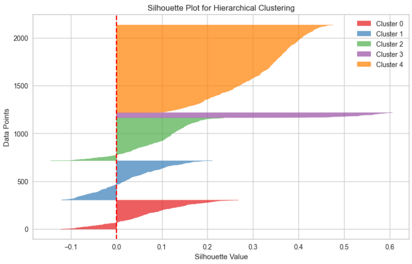

Decission Tree for Clustering
Expenses <= 0.213 gini = 0.488 samples = 1712 value = [990, 722] class = 0
True
False
MntFishProducts <= 0.241 gini = 0.058 samples = 1003 value = [973, 30] class = 0
Expenses <= 0.24 gini = 0.047 samples = 709 value = [17, 692] class = 1
MntSweetProducts <= 0.199 gini = 0.028 samples = 972 value = [958, 14] class = 0
Recency <= 0.136 gini = 0.499 samples = 31 value = [15, 16] class = 1
Income <= 0.301 gini = 0.44 samples = 46 value = [15, 31] class = 1
NumCatalogPurchases <= 0.136 gini = 0.006 samples = 663 value = [2, 661] class = 1
gini = 0.019 samples = 959 value = [950, 9] class = 0
gini = 0.473 samples = 13 value = [8, 5] class = 0
gini = 0.0 samples = 6 value = [6, 0] class = 0
gini = 0.461 samples = 25 value = [9, 16] class = 1
gini = 0.397 samples = 11 value = [8, 3] class = 0
gini = 0.32 samples = 35 value = [7, 28] class = 1
gini = 0.32 samples = 10 value = [2, 8] class = 1
gini = 0.0 samples = 653 value = [0, 653] class = 1

Figure 18: Case 3

In both cases, the two clusters represent contrasting customer profiles, with their roles reversed. In Case 1 (Figure 13), Cluster 1 consists of higher-income, highly educated individuals with smaller families who are highly engaged shoppers, spending significantly across all product categories and responding moderately to marketing campaigns, while Cluster 2 includes lower-income individuals with larger families who are less engaged, with minimal spending and low marketing responsiveness. Conversely, in Cases 2 (Figure 14) and 3 (Figure 15), Cluster 1 mirrors the previous Cluster 2, and Cluster 2 mirrors the previous Cluster 1. The only notable difference is that in Case 1, Cluster 1 shows slightly higher catalog purchases and total expenses compared to Case 2.

4.2 Partitional Clustering

The K-Means model focuses on distinguishing clusters based on education, expenses, deals purchased and meat products purchased (Figure 21).



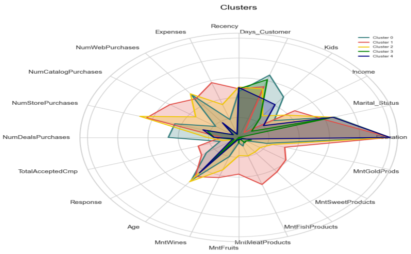


Figure 19: Silhouette for K-Means

Figure 20: Clusters for K-Means

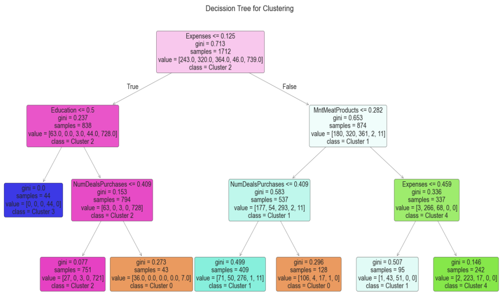


Figure 21: Decision Tree for K-Means

The Silhouette plot in Figure 19 shows a clear separation between the fourth and fifth clusters. However, the remaining clusters exhibit significantly poorer results, with some negative values, particularly in the second cluster. These observations align with the Gini values from the decision tree in Figure 21, which suggest distorted boundaries for the second cluster. This is reflected in the Gini score of 0.5 for both leaves of Class 1.

Moreover, the five clusters reveal distinct profiles. Cluster 0 consists of highly engaged individuals with a moderate income level, showing a strong tendency for online purchases and deal interactions. Cluster 1 represents higher-income, well-educated individuals who are highly engaged shoppers, spending heavily on products like wine, fruit, and meat, and showing a strong response to marketing efforts. Cluster 2, similar to Cluster 1 but with slightly less engagement, still shows notable spending, particularly on wine and meat, though their marketing response is lower. Cluster 3, on the other hand, consists of customers with a lower income, very low spending on products and minimal interaction with deals, stores, or catalog purchases. Lastly, Cluster 4 captures a group with low engagement in spending, particularly on wine, fruits, and other products, with a focus on online purchases but very low total spending.

4.3 Probabilistic Clustering

Lastly, the Gaussian Mixture Model (GMM) provided a very clear separation of the clusters, with the Gini impurity reaching 0 for the leaves of the Decision Tree, except for class 4 (Figure 22), indicating that the clusters are perfectly differentiated. The GMM's clustering approach highlighted three key factors: marital status, expenses, and response to the last promotion.

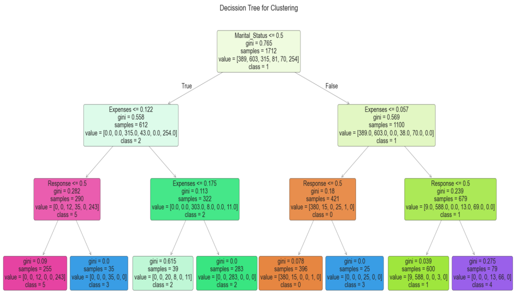


Figure 22: Decision Tree for GMM

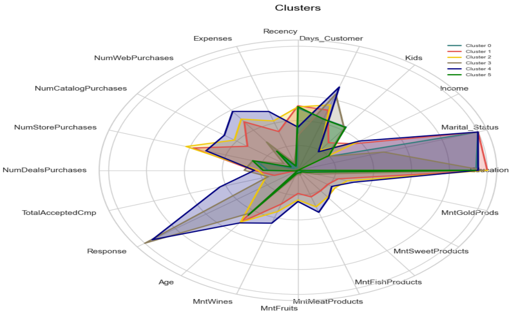


Figure 23: Clusters for GMM

Cluster 0 consists of low-income individuals with moderate family sizes. They have minimal engagement, spending very little across all product categories and showing low participation in store or catalog purchases. Cluster 1 represents moderate-income customers who actively shop through catalogs and in stores, with notable spending on wines, meats sweet products, and gold items. Cluster 2 includes higher-income, highly engaged customers who are the biggest spenders, particularly on wines, meat, and fish, and who favor catalog and online purchases. Cluster 3 is made up of lower-income individuals with modest spending habits, primarily on wines and sweet products, but they stand out for their high responsiveness to marketing campaigns. Cluster 4 represents affluent, highly engaged customers who are significant spenders across all product categories, especially wines and meat, with a strong preference for web and catalog purchases and the highest acceptance of marketing campaigns. Lastly, Cluster 5 mirrors Cluster 1, with the primary difference being Marital Status, where Cluster 5 represents single individuals.

5 Conclusion

These results demonstrate that different methods and models apply varying criteria for cluster separation. Hierarchical Clustering showed that different metrics produced identical results but with varying quality, indicating that the chosen distance measures did not significantly impact the final cluster formation. The probabilistic clustering approach, however, yielded the best separation overall, highlighting its effectiveness in distinguishing between customer groups. Despite using various criteria and methods, all models resulted in similar clusters, typically revealing two predominant patterns: one group of wealthy, active individuals and another of lower-income, less engaged individuals. These consistent findings underscore the importance of certain attributes in clustering, while also pointing out that different models may emphasize distinct nuances, yet ultimately capture similar underlying structures in the data.

References

[1] A. Patel, "Customer personality analysis," 2021. Available: `https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data`.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, 2011.

[3] P. S. University, "Stat 555: Hierarchical clustering - agglomerative methods," 2018.