Unsupervised classification

Machine Learning

Pablo Bande Sánchez Girón

1. Introduction

1.1. Problem

Customer Personality Analysis aims to understand a company's ideal customers and their unique needs, behaviours, and concerns. It helps tailor product offerings and marketing strategies to target specific customer segments effectively, optimise resource allocation, and enhance customer engagement and satisfaction. The objective is to implement three different unsupervised classification algorithms, one for each type **hierarchical, partitional** and **probabilistic**, and analyse the clusters and results obtained.

1.2. Dataset

The initial dataset initially consists of 2240 entries and 29 columns. Each one is described as follows:

- **ID:** Customer's unique identifier
- **Year_Birth:** Customer's birth year
- **Education:** Customer's education level
- **Marital_Status:** Customer's marital status
- **Income:** Customer's yearly household income
- **Kidhome:** Number of children in customer's household
- **Teenhome:** Number of teenagers in customer's household
- **Dt_Customer:** Date of customer's enrollment with the company
- **Recency:** Number of days since customer's last purchase
- **Complain:** 1 if the customer complained in the last 2 years, 0 otherwise
- **MntWines:** Amount spent on wine in last 2 years
- **MntFruits:** Amount spent on fruits in last 2 years
- **MntMeatProducts:** Amount spent on meat in last 2 years
- **MntFishProducts:** Amount spent on fish in last 2 years
- **MntSweetProducts:** Amount spent on sweets in last 2 years
- **MntGoldProds:** Amount spent on gold in last 2 years
- **NumDealsPurchases:** Number of purchases made with a discount
- **AcceptedCmp1:** 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- **AcceptedCmp2:** 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- **AcceptedCmp3:** 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- **AcceptedCmp4:** 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- **AcceptedCmp5:** 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- **Response:** 1 if customer accepted the offer in the last campaign, 0 otherwise
- **NumWebPurchases:** Number of purchases made through the company’s website
- **NumCatalogPurchases:** Number of purchases made using a catalogue
- **NumStorePurchases:** Number of purchases made directly in stores

- **NumWebVisitsMonth:** Number of visits to company's website in the last month

2. Methodology

2.1. Algorithms

1. Hierarchical: assumes that the data can be naturally grouped in a tree-like manner.
    a. **Agglomerative**: progressively merges individual data points or existing clusters into larger clusters based on a similarity measure and linkage criterion. In this experiment we try different linkages such as:
        i. Complete Linkage: computes the distance between two clusters as the maximum distance between any pair of samples from the two clusters. Tends to produce compact, spherical clusters, but it can be sensitive to outliers.
        ii. Ward Linkage: minimises the sum of squared differences within all clusters. It aims to minimise the variance within each cluster and tends to create clusters of similar sizes. It is less sensitive to outliers compared to complete linkage and often produces more balanced clusters.
        iii. Average Linkage: calculates the distance between two clusters as the average distance between all pairs of samples from the two clusters. It can be less influenced by outliers than complete linkage and can create more elongated clusters.
        iv. Single Linkage: considers the distance between two clusters as the minimum distance between any pair of samples from the two clusters. Sensitive to noise and outliers.

2. Nonhierarchical:
    a. **Partitional**: makes natural divisions of the data in a fixed number of clusters. Similar to supervised classification but without labels.
    b. **Probabilistic**: assumes that conditional densities of clusters have some known parametric distribution.

2.2. Dataset preprocessing

In order to work with a clean and adapted dataset for clustering, there were some steps needed in the preparation of the data:

- **Removed rows with null values**: During data cleaning, we identified rows in the dataset that contained missing values and removed them. Null values can introduce inconsistencies and affect the accuracy of analytical models, so removing them ensures data integrity.
- **Year_birth changed** into **Age**: In order to derive the age of each customer, we used the "Year_Birth" column, which indicates the birth year of each individual. By

subtracting the birth year from the current year, we calculated the age of each customer and replaced the "Year_Birth" column with an "Age" column. This transformation enables us to analyse the data based on the customers' ages rather than their birth years.

- **Created a new column with all spent:** To obtain a comprehensive measure of customers' overall spending, we aggregated the amounts spent on various product categories. We combined the columns "MntWines," "MntFruits," "MntMeatProducts," "MntFishProducts," "MntSweetProducts," and "MntGoldProds" into a single column called "AllSpent." This consolidation allows us to analyze the total expenditure of customers across multiple product categories.
- **Removed 'Z_CostContact'** and **'Z_Revenue'** columns: After careful consideration, we decided to exclude the "Z_CostContact" and "Z_Revenue" columns from our analysis. These columns were likely irrelevant for our purposes.
- **Transformed** 'Education', 'Marital_Status' and 'DT_education' **categorical variable into numerical** using **LabelEncoding**: The "DT_education" column in the dataset initially contained categorical values representing customers' educational backgrounds.
- Standardized some of the columns using StandardScaler.

Finally we ended with 2216 entries and 28 columns.

2.3. Principal Component Analysis

Performing dimensionality reduction on the selected features is essential in order to effectively handle a high number of features. When the feature count increases, the task becomes more challenging. A significant portion of these features tends to exhibit correlation, resulting in redundancy. Consequently, to address this issue, I will employ a dimensionality reduction technique: PCA.

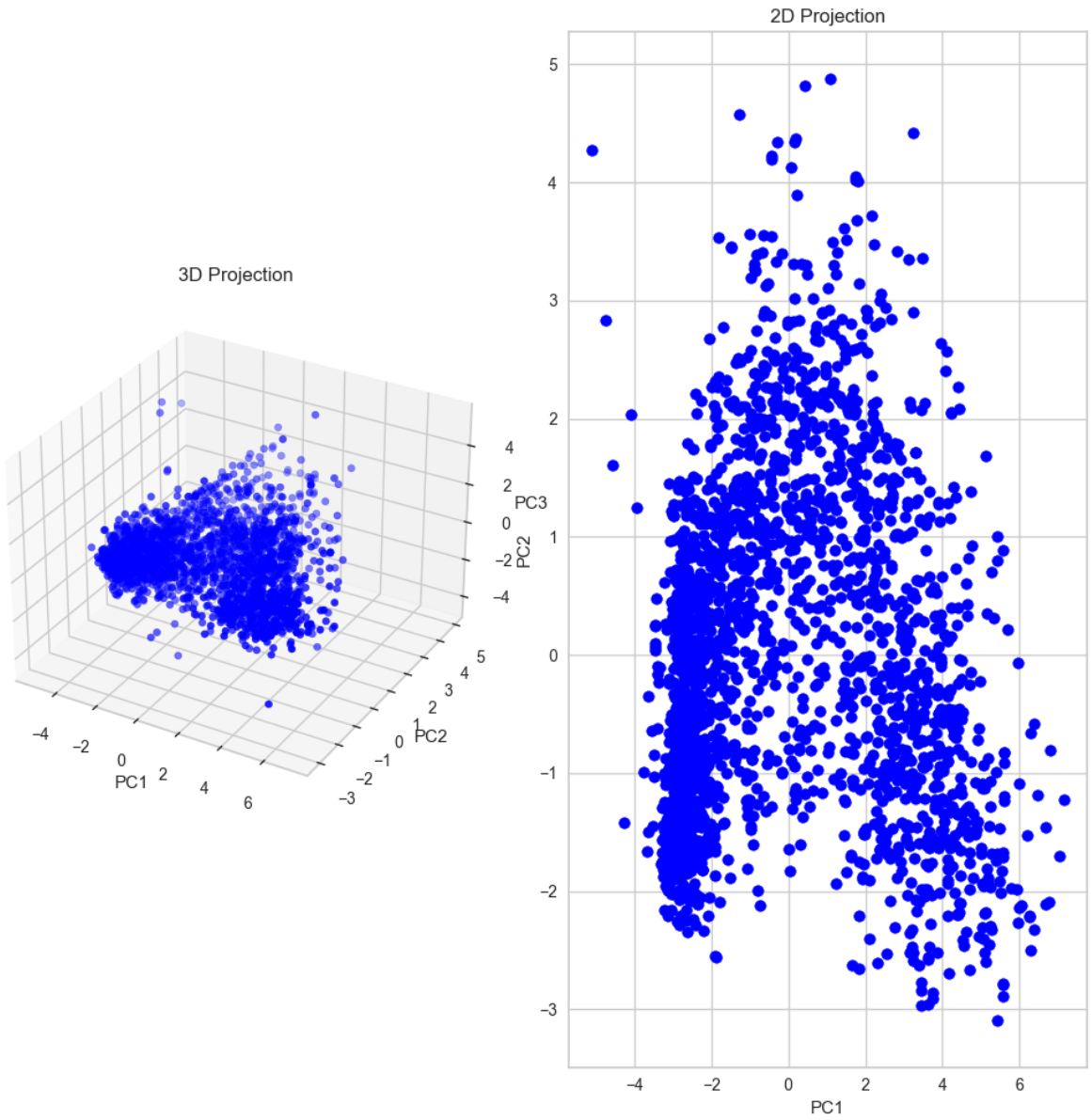

Image 1: PCA 2D and 3D projection.

2.4 Hyperparameter estimation

In order to find the optimal evaluating the optimal number of clusters for each value, we plotted the within-cluster sum of squares (WCSS) against different values of K. The WCSS is a measure of how compact the data points are within each cluster. The results obtained are the following:

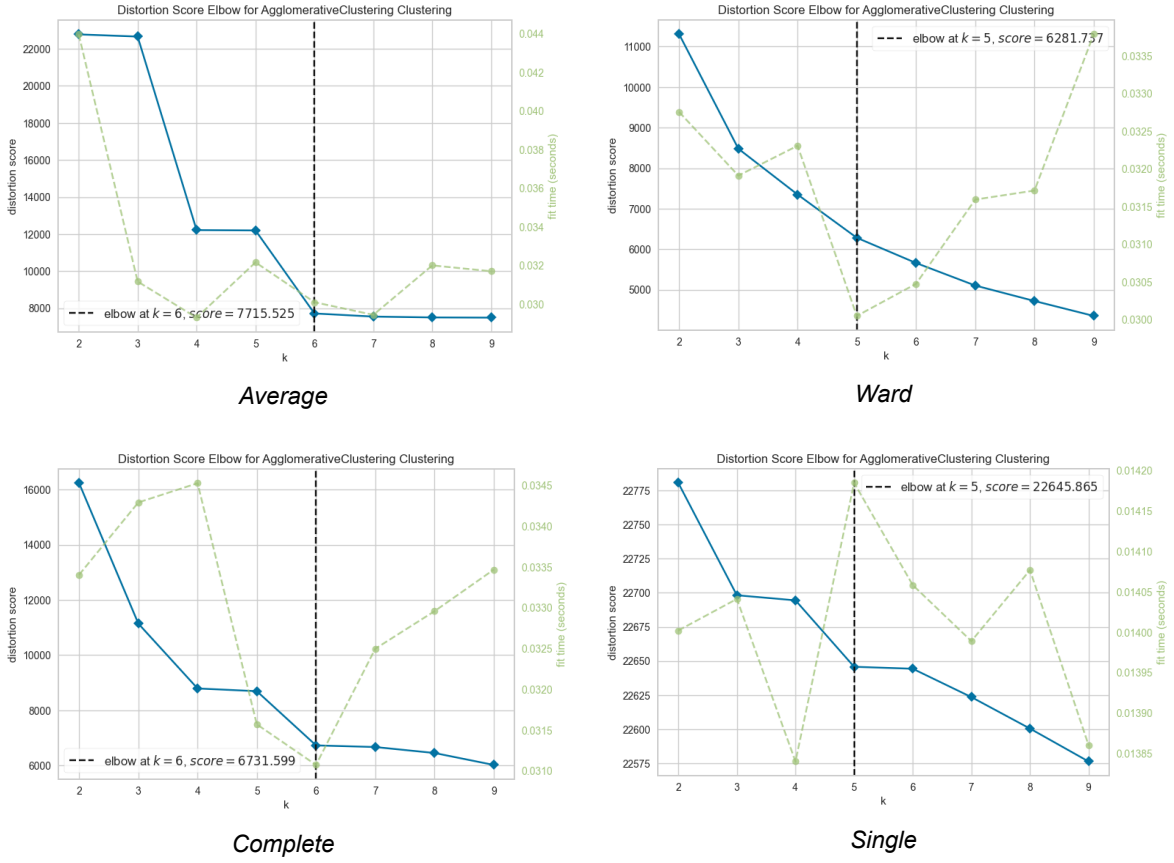

Image 2: Agglomerative k evaluation for average, ward, complete and single linkage

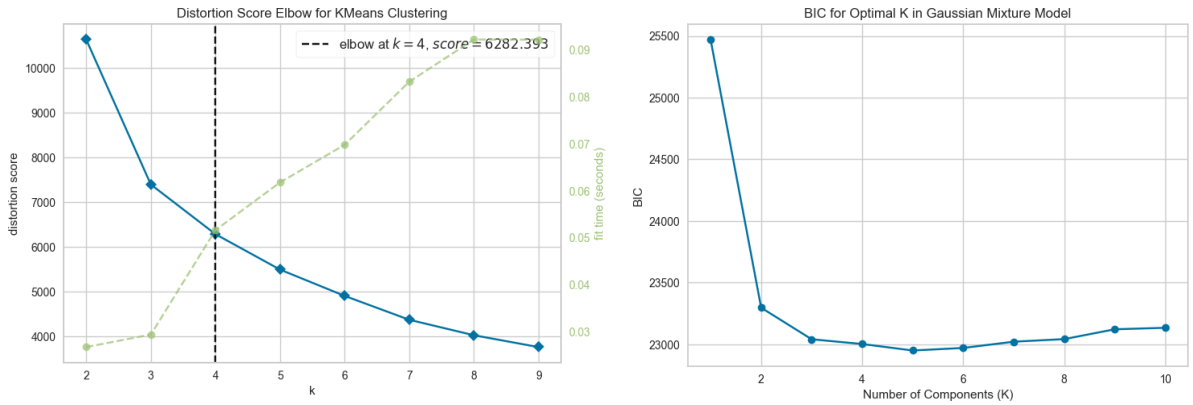

Image 3: K evaluation for KMeans and GMM

For the Gaussian Mixture model we had to use another library which does not indicate the exact optimal K value, so we performed the analysis with 2 and 3 values.

3. Results

In this section we will compare the results obtained for each model and talk about the best one for each algorithm.

1. Agglomerative

a. Average

Although the number of clusters seems to be three, there are actually six groups. Perhaps the methodology used for calculating the optimal K value was not ideal. Also, outliers affect negatively on this kind of clustering so this might have also been a factor in the failure of the grouping.

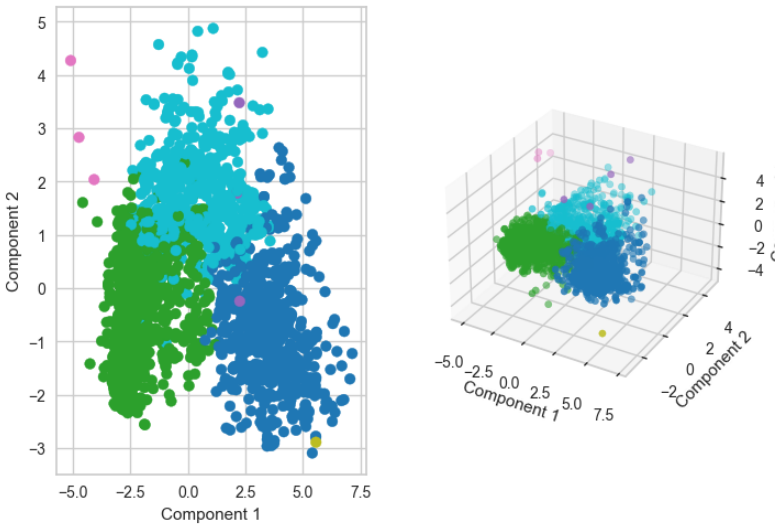

Image 4: Agglomerative with average linkage visual representation of the clusters

b. Ward

Opposed to the average linkage, here we can clearly differentiate the five different clusters in the data. As it was mentioned before, it tends to create clusters of similar sizes and is less sensitive to outliers. This can be observed in the resultant graphical representation.

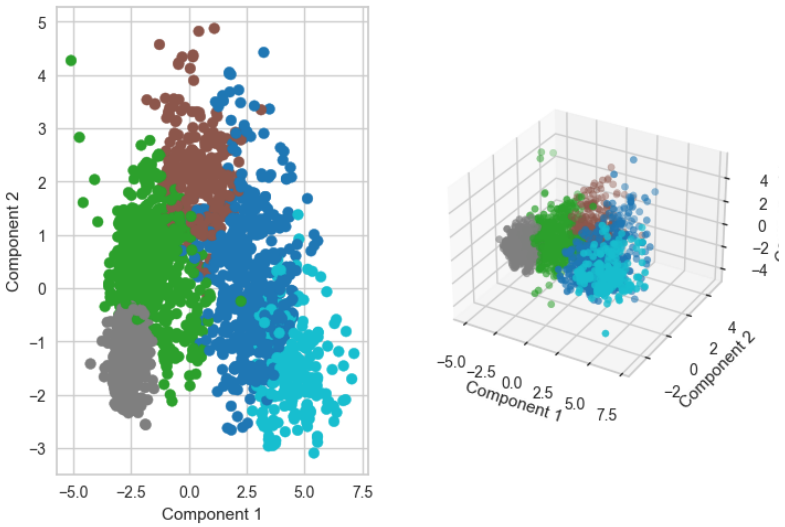

Image 5: Agglomerative with ward linkage visual representation of the clusters

c. Complete

This is another example of bad clustering, even though complete linkage tends to produce compact and spherical shapes, it is not the case. This might also be because of the outliers.

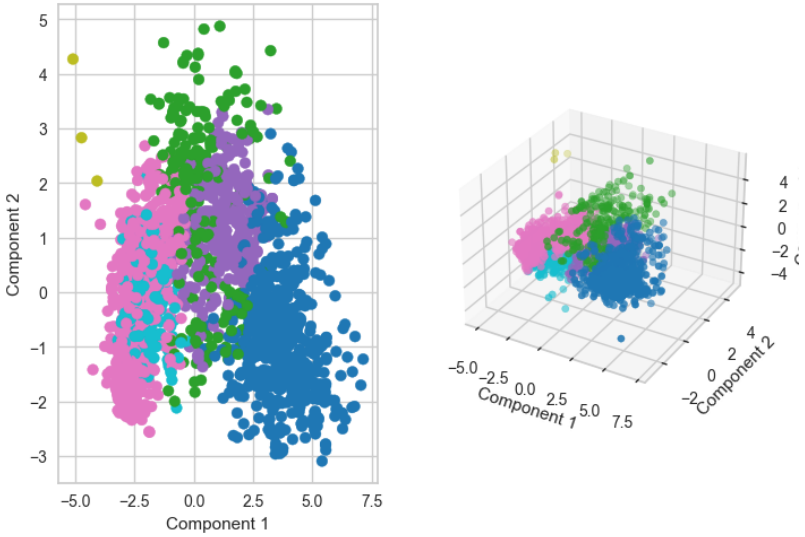

Image 6: Agglomerative with complete linkage visual representation of the clusters

d. Single

This one is sensitive to noise and outliers and can be seen in the way it basically found one big dispersed cluster in green and assigned far points to the rest of the clusters to find.

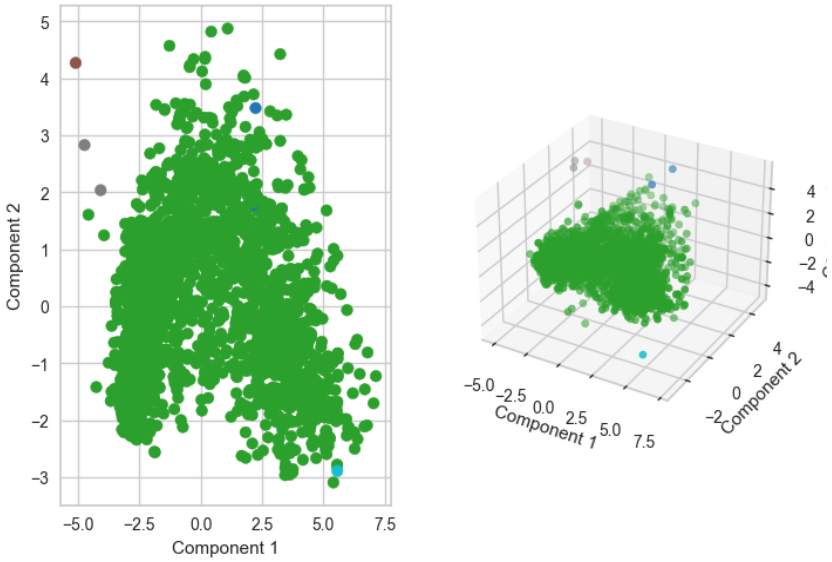

Image 7: Agglomerative with single linkage visual representation of the clusters

2. KMeans

As we can observe in image 8, there are four well differentiated clusters. While the red and dark blue ones are more compact, the other two have bigger size and are more expanded, taking some of the outliers. In general, KMeans performed well with this data.

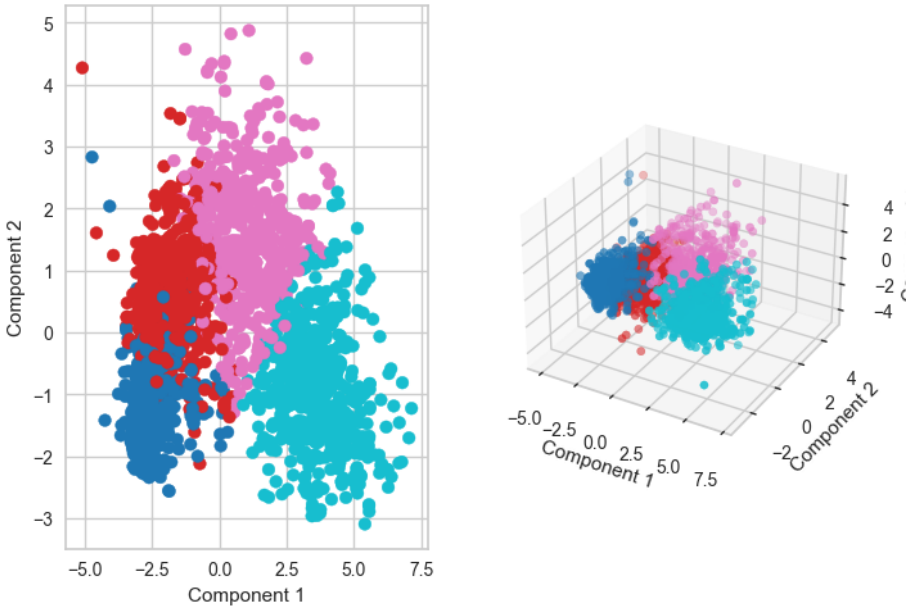

Image 8: KMeans visual representation of the clusters

3. Gaussian Mixture Model

Finally, GMM was executed with two different values of K: 2 and 3 due to the possible interpretation of the elbow method. In the first one we can see how the clusters are not very similar, one is completely compact while the second one is dispersed through the data, keeping the distribution to be clearly differentiated from the other. The second execution shows three clusters, here we can appreciate more similar groups in shape and size compared to the previous analysis. We could conclude that three numbers of clusters is better than two for this algorithm.

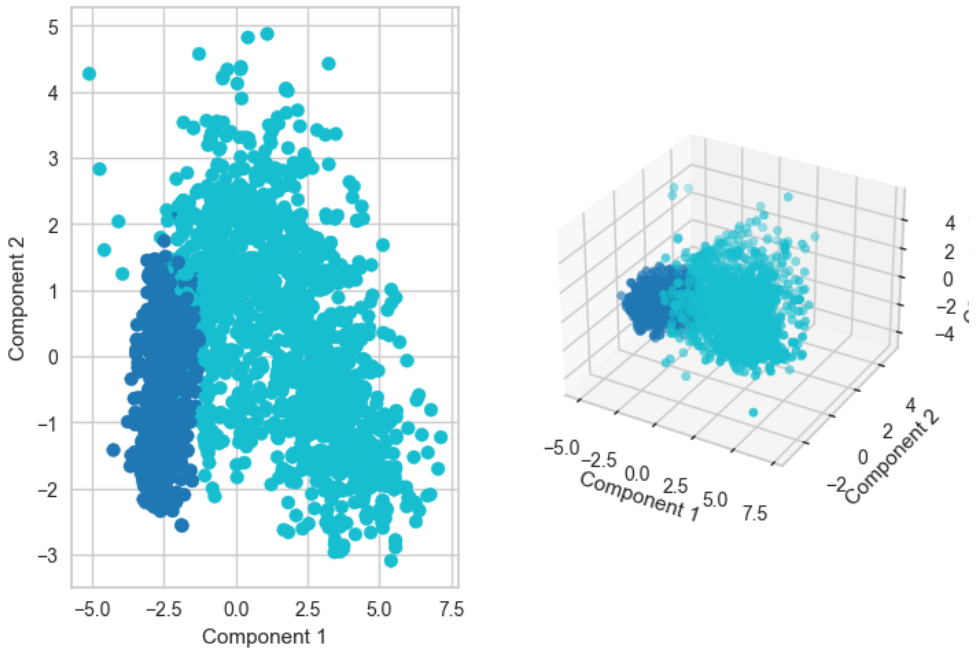

Image 9: GMM k = 2 visual representation of the clusters

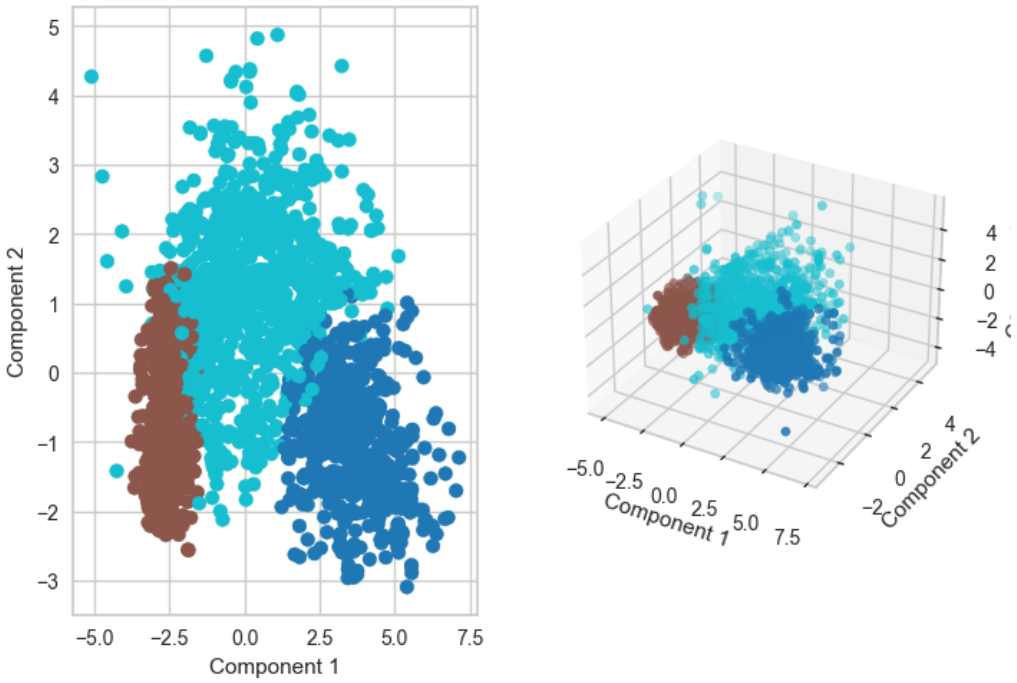

Image 10: GMM k = 3 visual representation of the clusters

4. Conclusion

Non-hierarchical clustering methods generally resulted in more compact and distinguishable clusters compared to hierarchical clustering, except for agglomerative clustering with ward linkage. The presence of noise and outliers in the data negatively affected the performance of clustering methods. Based on the obtained results, it was concluded that the data is best split into three or four distinct groups.

For future work, it is recommended to focus on preprocessing the data to address the issue of noise and outliers. Implementing effective techniques, such as outlier detection and noise reduction methods will enhance the performance of clustering algorithms obtaining more compact and distinguishable clusters, facilitating better insights and interpretations of the data.

5. References

[1] Akash Patel. Analysis of company's ideal customers. Data retrieved from Kaggle,https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis. 2021.

[2] Pedro Larrañaga, Concha Bielza. Unsupervised classification