Unsupervised classification

Machine Learning

Pablo Bande Sánchez Girón

Introduction

Algorithms:

1. Hierarchical: Agglomerative
2. Partitional: KMeans
3. Probabilistic: Gaussian Mixture Model

Agglomerative linkage:

1. Complete
2. Ward
3. Average
4. Single

Goal: Group data and analyze

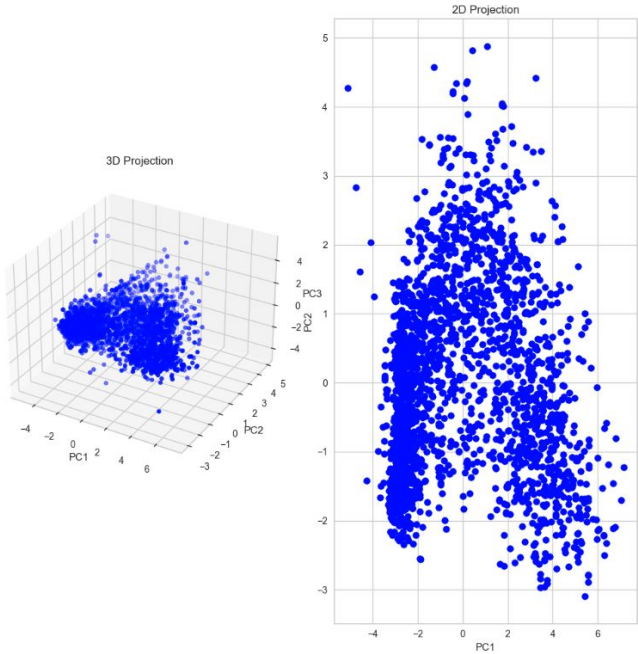
Dataset

The dataset is about **Customer Personality Analysis** and initially consisted of 2240 entries and 29 columns.

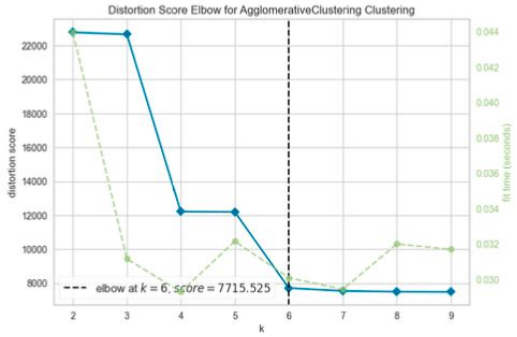
Preprocessing:

- Numerical to categorical using Label encoding
- Removed null values
- Converted Year_Birth into Age
- New column unifying all spent by the user and removed irrelevant columns
- Standardized
- Applied Principal Component Analysis

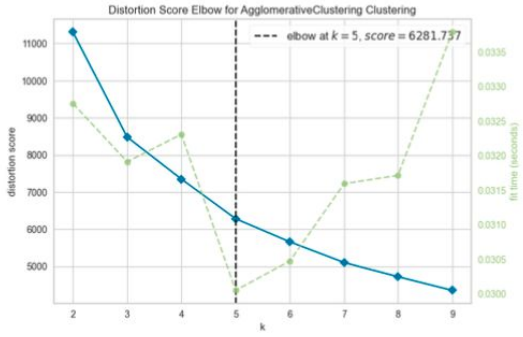
Result:
2216 entries and 28 columns



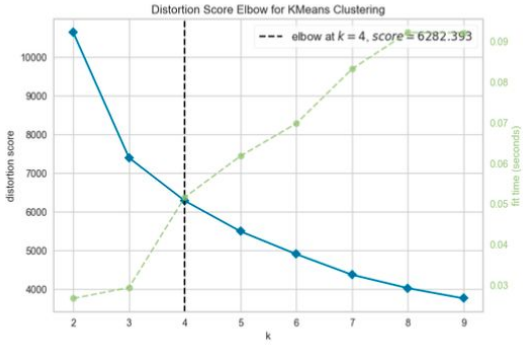
Hyperparameter estimation: K

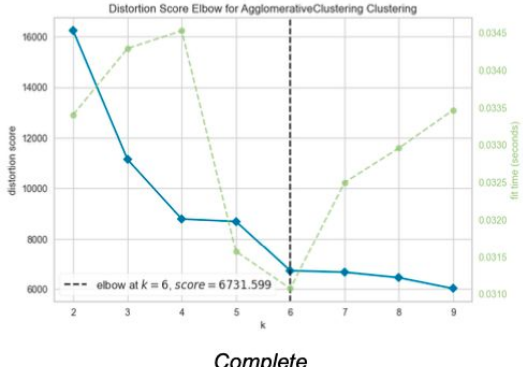


Average

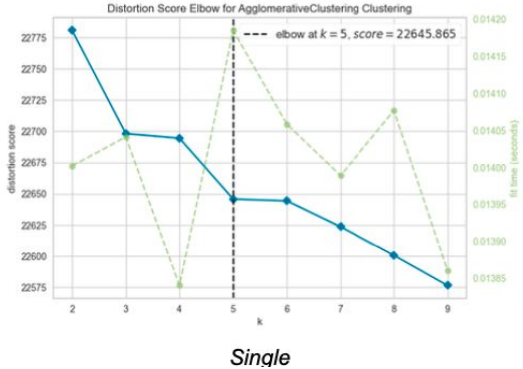


Ward

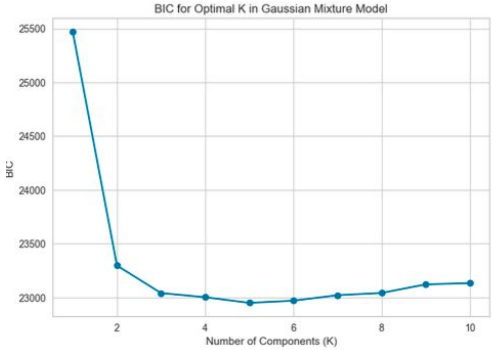




Complete

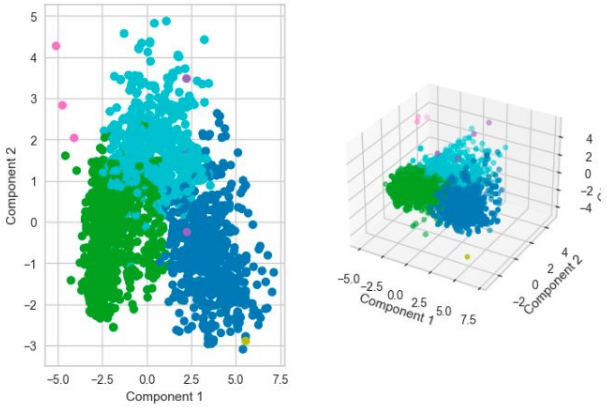


Single

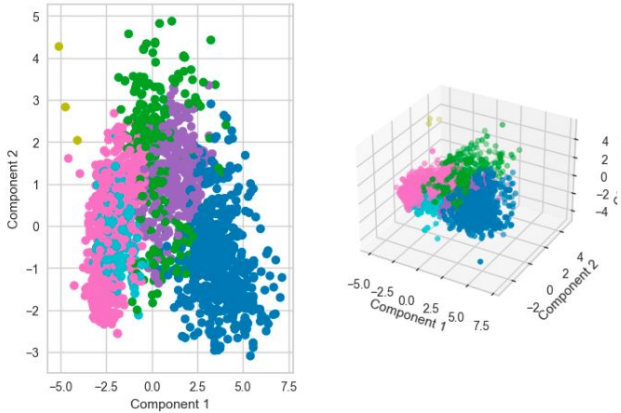


Results: Agglomerative

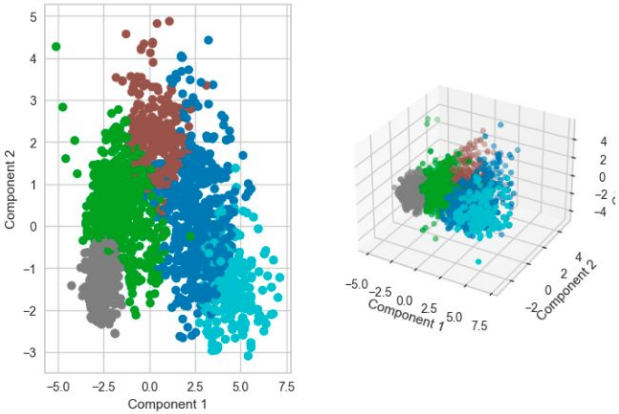
Average



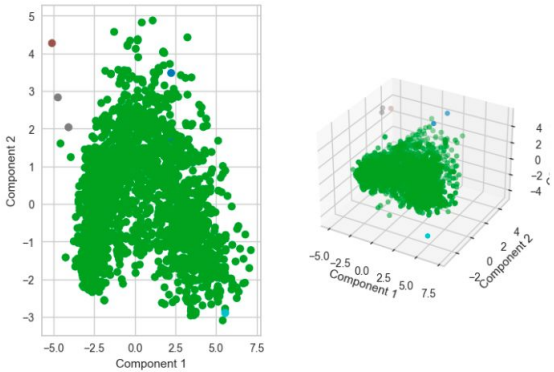
Complete



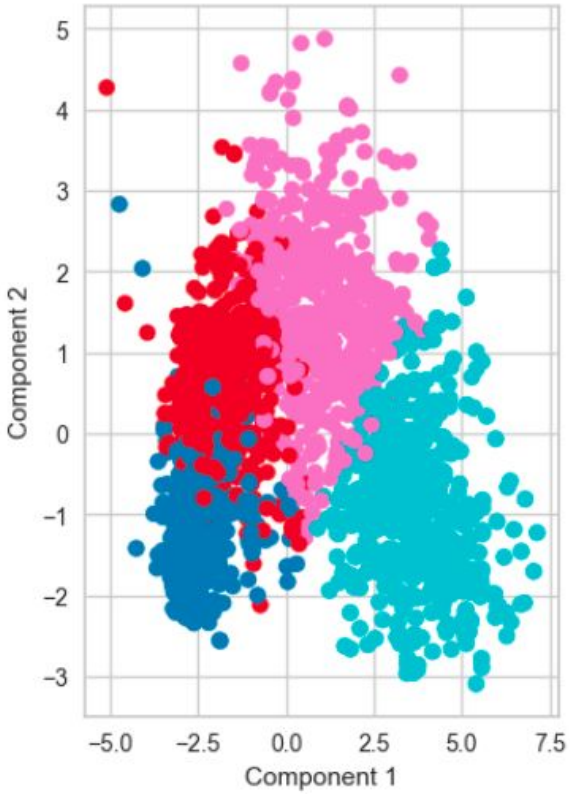
Ward

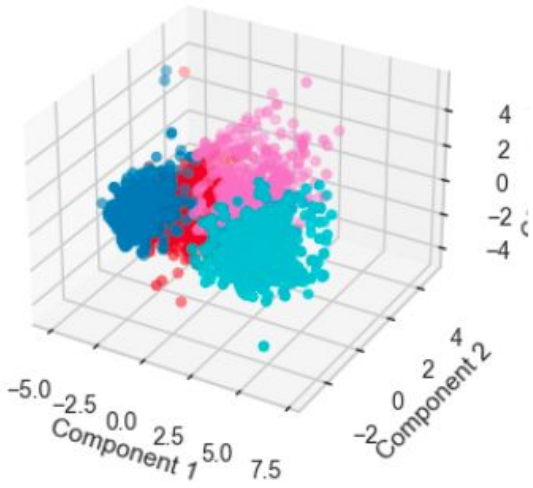


Single

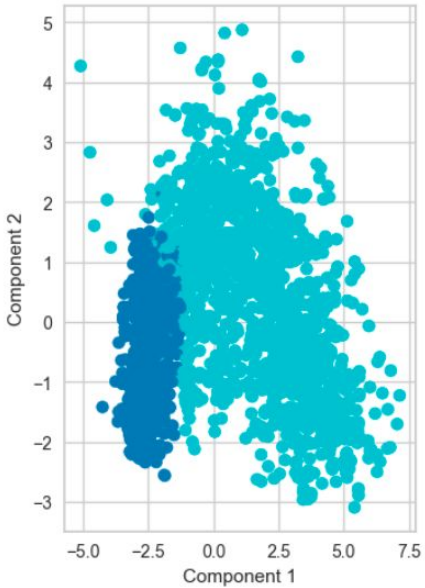


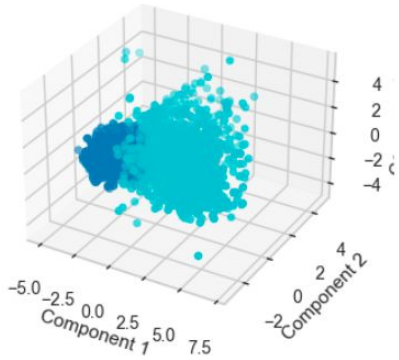
Results: KMeans

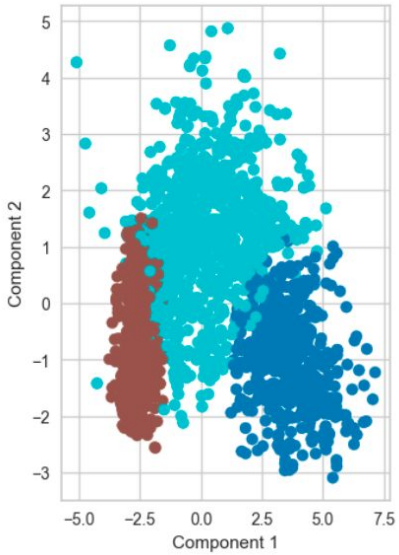


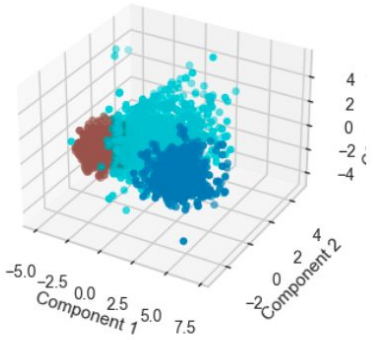


Results: Gaussian Mixture Model









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. 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.

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