Course: CSL7620- Machine Learning

Course project - Semester I

CUSTOMER SEGMENTATION USING CLUSTERING

(based on their purchase history and demographics)

Submitted By:
Layaa Vishwakarma (M25CSA017)
Prapti Halder (M25CSA022)
Suparni Maitra (M25CSA029)

TABLE OF CONTENTS

1. AIM OF THE PROJECT

This project aims to perform customer segmentation using various unsupervised machine learning clustering algorithms, to identify distinct and meaningful groups within a customer dataset.

The objective is to leverage various clustering techniques like **DBSCAN, Hierarchical Clustering and K-Means Clustering**, to partition customers based on their characteristics and behaviours. This will provide insights into the customer base, enabling businesses to tailor marketing strategies, product development and customer service efforts more effectively.

To ensure a robust segmentation, the performance of each clustering method is evaluated using three primary metrics- **Separation/Spread Ratio, the Calinski-Harabasz (CH) Index and the Davies-Bouldin Index (DBI).** The final outcomes are compared to determine the most optimal and thoughtful segmentation approach for the given dataset.

2. WORKFLOW OF THE PROJECT

(a) Data Preprocessing:

The raw dataset was first cleaned to ensure consistency and reliability.

- Handling missing values
- Encoding categorical data
- Dimensionality Reduction

(b) Data Clustering:

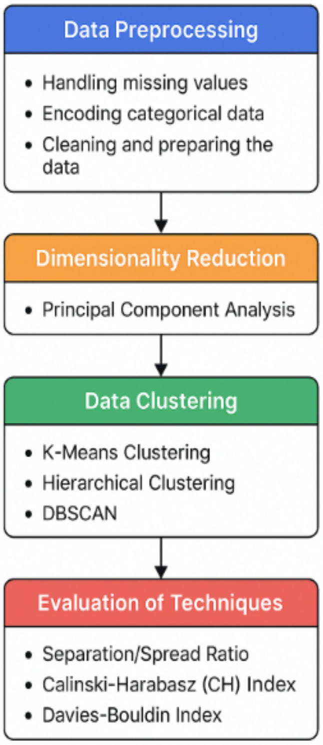
Three main clustering algorithms were applied to segment the customers based on similar behavioral patterns:

- K-Means Clustering
- Hierarchical Clustering
- DBSCAN

(c) Evaluation of Techniques:

To assess the clustering performance of the tree algorithms ,several evaluation metrics were used.

- Separation/Spread Ratio
- Calinski-Harabasz (CH) Index
- Davies-Bouldin Index (DBI)


Data Preprocessing
• Handling missing values
• Encoding categorical data
• Cleaning and preparing the data
Dimensionality Reduction
• Principal Component Analysis
Data Clustering
• K-Means Clustering
• Hierarchical Clustering
• DBSCAN
Evaluation of Techniques
• Separation/Spread Ratio
• Calinski-Harabasz (CH) Index
• Davies-Bouldin Index

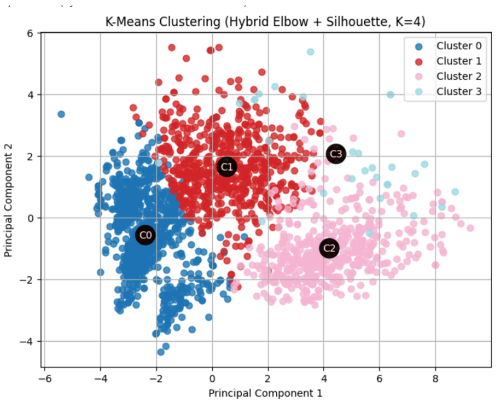
3. CLUSTERING METHODS USED

(A) K-Means Clustering:

The algorithm followed two main iterative steps:

- Assignment step: Each data point was assigned to its nearest centroid using Euclidean distance.

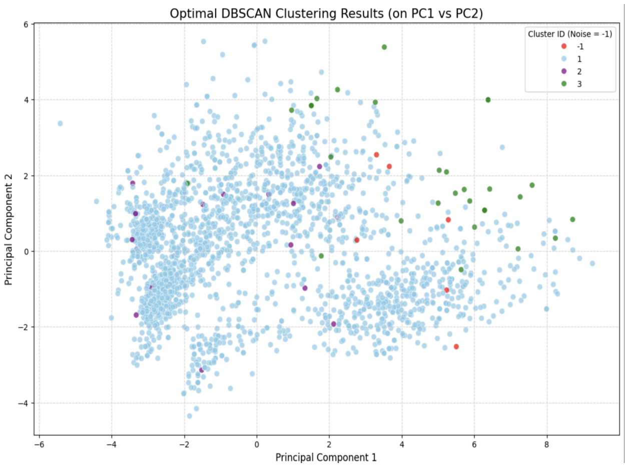
- Update step: Each centroid was recalculated as the mean of all points belonging to that cluster. The process continued until the centroid shift was less than the tolerance (1e-4) or maximum iterations (300) were reached.
- The inertia (Within-Cluster Sum of Squares) was computed to measure cluster compactness.
- The optimal K was chosen as the average of Elbow K and Silhouette K for a balanced selection. The final K-Means model was trained using the optimal K obtained from the hybrid method.
- The algorithm assigned each record a cluster label (0, 1, 2, …, K-1).
- The final centroids represented the centers of the clusters in the PCA-transformed feature space.
- Clusters were visualized using a 2-D scatter plot formed by the first two PCA components.
- Each centroid was labeled (C0, C1, C2, …) directly on the plot for clear interpretation.
- The visualization demonstrated how data points were grouped based on their similarities.



(B) DBSCAN Clustering:

For DBSCAN Clustering, the major steps done in the codeflow are -

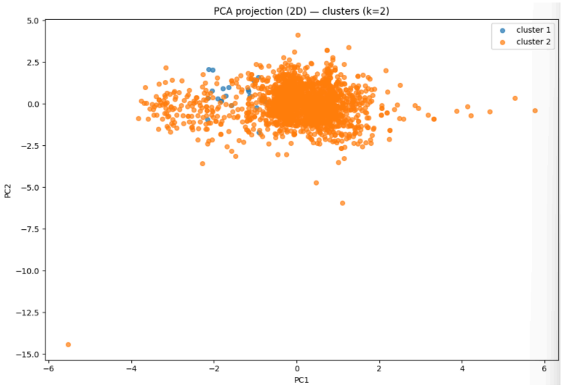
- This section focuses on finding an optimal value for eps, a key parameter in the DBSCAN algorithm that defines the radius of a neighborhood. It employs the k-distance graph method, shown in the image given.
- To find the best-performing combination of eps and minpts, the notebook performs a grid search. It tests a range of eps values (from 5.0 to 8.5) and minpts values (20, 30, 40, 50, 60).
- Using the optimal parameters, the final clustering results were found, and then visualized using a scatter plot of the first two principal components (PC1 vs. PC2).

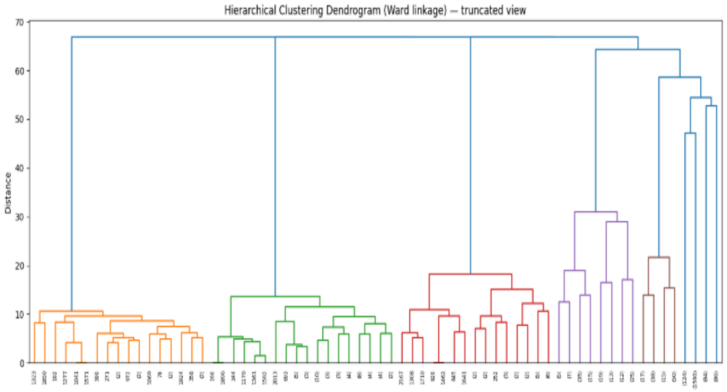


(c) Hierarchical Clustering:

- A pairwise Euclidean distance matrix was computed between all samples. Using Ward's linkage method, clusters were iteratively merged to minimize the total within cluster variance.

- A dendrogram was plotted to visualize the merging process. By cutting the dendrogram at a certain level the number of clusters k is obtained.
- Several values of k were tested, and the k value with the highest silhouette score was chosen as the optimal number of clusters.
- To visualize the clusters, a PCA projection to two dimensions was plotted, showing the separation between clusters.





4. TECHNIQUES USED TO EVALUATE THE CLUSTERING METHODS

(a) Separation/Spread Ratio: This metric evaluates the quality of clustering results by comparing how well-separated the clusters are from each other versus how compact each cluster is. The following steps show how it was implemented -

- Filtering noise (since it would wrongly affect the results).
- Calculating Average Cluster Spread or Compactness (lower is better).
- Calculating Average Cluster Separation (higher is better).
- The function then returns the ratio of average separation to average spread (higher value implies better clustering).

(b) Calinski-Harabasz (CH) Index: The CH index measures the ratio between the between-cluster dispersion and the within-cluster dispersion. This shows how distinct and compact the clusters are. A higher score indicates the clusters are dense and well separated.

- For each clustering method, the Calinski-Harabasz score was computed using calinski_harabasz_score(X,labels) from sklearn.metrics.
- the method with the highest score was considered to have produced the most distinct and well-defined clusters.

(c) Davies-Bouldin Index (DBI): It measures how well clusters are separated from each other and how compact the members of each cluster are.

- **Cluster Centroid Calculation:** For each cluster, the centroid was computed as the mean of all data points within that cluster, along with the Intra-cluster Scatter, Inter cluster distance, Cluster similarity and DBI aggregation.
- Lower DBI values (Hierarchical and DBSCAN) indicate that these algorithms produced tighter and better-separated clusters, while higher DBI value (K-Means) suggests that the clusters were less distinct and possibly overlapping.

5. OBSERVATIONS

	K-Means	**Hierarchical**	**DBSCAN**
Separation to Spread Ratio	1.1649	2.4407	2.3936
CH - Index	86.743	97.278	78.566
DBI	3.431	0.819	1.009

- The logical calculation of **'Separation to Spread Ratio'** signifies that the best clustering result was given by hierarchical clustering, with the highest score of 2.4407. DBSCAN, being a close competitor, scored 2.3936, with K-Means producing the lowest result.
- The **Calinski–Harabasz (CH) Index** values indicate that the Hierarchical Clustering method achieved the highest score (97.27), followed by K-Means (86.74) and DBSCAN (78.56). Since a higher CH score signifies better clusters, hierarchical clustering produced the most distinct and cohesive customer groups.
- The **Davies–Bouldin Index (DBI)** results also support this finding. Hierarchical Clustering again performed best with the lowest DBI (0.8194), while DBSCAN achieved a slightly higher value (1.0093) and K-Means performed the worst (3.4313). A lower DBI corresponds to tighter clusters with less overlap.
- Overall, the results are consistent across all internal validation measures, demonstrating the robustness of hierarchical clustering for this dataset.

6. CONCLUSION

- Based on the Separation/Spread ratio, CH Index and DBI the **Hierarchical Clustering** algorithm provided the most effective segmentation of customers, forming compact, well-separated, and interpretable clusters.
- DBSCAN, while performing better than K-Means, though capable of identifying arbitrary-shaped clusters and outliers, performed slightly less effectively on this dataset due to its reliance on density parameters.
- K-Means produced the least satisfiable result showing higher intra-cluster variation, suggesting that it may be sensitive to cluster initialization or data distribution.