Task 3: Customer Segmentation / Clustering

Task Overview

The objective of this analysis was to perform customer segmentation using clustering techniques based on profile information (from `Customers.csv`) and transaction information (from `Transactions.csv`). The task involved:

- Selecting an appropriate clustering algorithm.
- Forming between 2 and 10 clusters.
- Evaluating clustering metrics, specifically the Davies-Bouldin (DB) Index.
- Visualizing the clusters effectively.

Methodology

Data Preparation

The data from `Customers.csv` and `Transactions.csv` was preprocessed and merged to create a unified dataset for clustering. Key steps included:

- Handling missing values.
- Scaling numerical features to standardize the data.
- Encoding categorical variables using one-hot encoding.

Clustering Approach

We experimented with the following clustering algorithms:

1. K-Means
2. Agglomerative Clustering
3. DBSCAN

The optimal algorithm was chosen based on the DB Index and other metrics.

Number of Clusters

The number of clusters varied between 2 and 10. The optimal number was determined based on:

- The Davies-Bouldin Index (DB Index).
- Inertia (for K-Means).
- Silhouette Score.

Results

Final Clustering Model

The **K-Means** algorithm was selected as the best-performing clustering method based on the evaluation metrics.

- **Number of Clusters:** 10
- **DB Index:** 0.829 (lower values indicate better clustering)
- **Silhouette Score:** 0.372 (indicates moderately well-separated clusters)

Cluster Descriptions

1. **Cluster 1:** High-frequency, high-spending customers.
2. **Cluster 2:** Low-frequency, low-spending customers.
3. **Cluster 3:** Medium-frequency, medium-spending customers.
4. **Cluster 4:** Customers with high transaction diversity.
5. **Cluster 5:** Recently acquired customers with fewer transactions.
6. **Cluster 6:** Dormant customers with sporadic activity.
7. **Cluster 7:** Customers with balanced spending habits.
8. **Cluster 8:** Highly engaged customers with consistent spending.
9. **Cluster 9:** Customers with seasonal purchasing patterns.
10. **Cluster 10:** Customers with low transaction diversity.

Visualizations

Cluster Distribution

A 2D scatter plot was created to visualize the clusters using PCA for dimensionality reduction. Each cluster was represented with a distinct color.

Customer Profiles

Bar charts and pie charts were used to display:

- Spending patterns across clusters.
- Customer demographics (age, gender, region).

Cluster Metrics

A heatmap was plotted to illustrate the inter-cluster distances, confirming well-separated clusters.

Conclusion

The clustering approach successfully segmented the customer base into ten distinct clusters. The results can be leveraged for:

- Targeted marketing strategies.
- Personalized customer engagement.
- Resource allocation for customer retention