Customer Segmentation using KMeans Clustering

1. Merging Datasets and Aggregating Transaction Data

- **Objective:** Combine customer profile data with their respective transaction data.
- **Steps:**
    - **Merging Data:** The customer data and transaction data are merged based on a common identifier (e.g., customer ID). This ensures that each customer has corresponding transaction information available for analysis.
    - **Aggregating Transaction Data:** For each customer, transaction data is aggregated to obtain key metrics such as the total number of transactions, total spending, frequency of purchases, or any other relevant metrics. This aggregation helps in summarizing the customer's behavior.
    - **Combined Data:** The aggregated transaction data is then joined with the customer profile data to create a comprehensive dataset that includes both demographic and behavioral features.

2. Feature Selection and Standardization

- **Objective:** Prepare the data for clustering by selecting relevant features and ensuring uniform scaling.
- **Steps:**
    - **Feature Selection:** Choose the most relevant features that reflect the customer's behavior or profile, such as age, gender, total spending, purchase frequency, etc. This step ensures that only meaningful data is included in the clustering process.
    - **Standardization:** Standardize the selected features using techniques like Min-Max Scaling or Standard Scaler. This is crucial for KMeans since it is sensitive to the scale of the data. Standardization ensures that all features contribute equally to the distance metric used by KMeans.

3. Clustering Using KMeans

- **Objective:** Apply KMeans clustering to segment customers into distinct groups based on their behavior.
- **Steps:**
    - **KMeans Algorithm:** The KMeans algorithm is applied to the standardized dataset to segment customers into 5 distinct clusters. KMeans works by grouping data points into clusters based on the nearest mean of the cluster, iterating until convergence.
    - **Choosing K (Number of Clusters):** The number of clusters is set to 5, based on domain knowledge or experimentation. The optimal number of clusters can also be determined using techniques like the Elbow Method or Silhouette Score.

4. Evaluating Clusters Quality

- **Objective:** Assess the quality of the clustering results.
- **Steps:**

- **Davies-Bouldin Index:** This metric measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower Davies-Bouldin Index indicates better separation between clusters.
- **Silhouette Score:** This score evaluates how similar a point is to its own cluster compared to other clusters. A higher silhouette score (closer to 1) indicates better-defined clusters.

5. Visualization of Clusters

- **Objective:** Visualize the customer segments to better understand the clustering results.
- **Steps:**
  - **Principal Component Analysis (PCA):** PCA is used to reduce the dimensionality of the dataset while preserving the variance. This step helps in projecting high-dimensional data into 2 or 3 dimensions for easy visualization.
  - **Plotting the Clusters:** After applying PCA, the clusters are visualized on a 2D plot (if using 2D PCA). Each point represents a customer, and colors are used to distinguish between clusters. This visual representation helps in understanding the separation and distribution of customer segments.

6. Saving the Segmented Data

- **Objective:** Save the results of the clustering for future analysis or deployment.
- **Steps:**
  - **Save to CSV:** The final dataset, which includes customer profile data along with their respective cluster labels, is saved to a CSV file. This allows for further analysis, reporting, or integration with other systems.

Conclusion

In this process, we successfully segmented customers into distinct groups based on their behavior, which can be leveraged for targeted marketing, personalized offers, or improving customer services. The evaluation of the clustering results ensures the quality and meaningfulness of the segments, and the visualization aids in understanding the separation between the customer groups. By saving the segmented data, we ensure that the results are accessible for further use.