Customer Segmentation

I. Problem Statement

Customer segmentation is the process of dividing a company's customer base into distinct groups or segments based on certain characteristics or behaviors. The goal is to identify meaningful patterns and differences among customers to better understand their needs, preferences, and behaviors. By segmenting customers, businesses can tailor their marketing strategies, product offerings, and customer experiences to target each segment more effectively.

Unsupervised learning techniques are commonly used to solve customer segmentation problems because they allowed me to discover patterns and structures in the data without the need for pre-labeled or labeled examples.

II. Dataset

To solve this problem, I used Credit Card Dataset for Clustering dataset from Kaggle. The provided dataset is a customer-level dataset containing information on the credit card usage behavior of approximately 9000 active credit card holders over the past 6 months. It includes 18 behavioral variables for each customer. The dataset provides valuable information for customer segmentation and developing marketing strategies based on customer behavior and preferences.

Behavioral Variables	Description
CUST_ID	Identification of the credit card holder (Categorical)
BALANCE	The balance amount left in the account to make purchases
BALANCE_FREQUENCY	The frequency at which the balance is updated (score between 0 and 1)
PURCHASES	The total amount of purchases made from the account
ONEOFF_PURCHASES	The maximum purchase amount made in one transaction
INSTALLMENTS_PURCHASES	The amount of purchases made in installments
CASH_ADVANCE	The cash in advance given by the user
PURCHASES_FREQUENCY	The frequency of purchases being made (score between 0 and 1)
ONEOFFPURCHASESFREQUENCY	The frequency of one-off purchases (score between 0 and 1)
PURCHASESINSTALLMENTSFREQUENCY	The frequency of purchases made in installments (score between 0 and 1)
CASHADVANCEFREQUENCY	The frequency of cash in advance being paid (score between 0 and 1)
CASHADVANCETRX	The number of transactions made with "Cash in Advance"
PURCHASES_TRX	The number of purchase transactions made
CREDIT_LIMIT	The credit card limit for the user
PAYMENTS	The amount of payment made by the user
MINIMUM_PAYMENTS	The minimum amount of payments made by the user
PRCFULLPAYMENT	The percentage of the full payment paid by the user
TENURE	The tenure of credit card service for the user

III. DESIGN & IMPLEMENTATION

Performed customer segmentation using K-means clustering with the dataset provided and then determined the optimal number of clusters using the elbow method. By employing the K-means algorithm, I divided the customers into distinct groups based on their characteristics. To evaluate the performance of the clustering model, I utilized a Decision Tree Classifier and computed evaluation metrics such as the confusion matrix, classification report, and accuracy score. These metrics allowed me to assess how accurately the model predicted the clusters for unseen data.

To perform clustering, the K-means algorithm is employed. The data is standardized using the StandardScaler, and the optimal number of clusters (k) is determined using the elbow method. The data is also subjected to dimensionality reduction using Principal Component Analysis (PCA) to reduce it to two dimensions for better visualization. With the chosen optimal number of clusters, K-means clustering is performed on the standardized data. The resulting clusters are visualized using a scatter plot, where each cluster is represented by a different color. Cluster centers are added to the plot to indicate the centroid of each cluster. Next, the count of samples in each cluster is visualized using a countplot to understand the distribution of customers across clusters. The dataset is then segregated into four separate DataFrames based on the cluster assignments. Each DataFrame corresponds to a specific cluster, allowing for further analysis and understanding of customer characteristics within each cluster.

To evaluate the clustering model, the dataset is split into training and testing sets. A Decision Tree Classifier (ID3) is trained on the training data and used to predict the clusters for the testing data. Evaluation metrics such as the confusion matrix, classification report, and accuracy score are calculated to assess the performance of the model in predicting the clusters for unseen data. Overall, this approach enabled me to segment customers effectively and gain insights into their preferences and behaviours.

By combining K-means clustering and the ID3 model, you can assess how well the customer clusters derived from K-means can be classified by the decision tree model. This approach allows you to explore relationships and patterns within the clusters and evaluate the effectiveness of the clustering solution using the ID3 model as a classification tool.