Project Documentation: Customer Clustering and Classification Using KMeans

---

Title:

Customer Clustering and Classification Using KMeans

Introduction:

The objective of this project is to analyze customer data from a banking dataset and apply KMeans clustering to identify distinct customer segments. Additionally, the project evaluates the effectiveness of the clustering in predicting the target variable `y`, which indicates whether a customer subscribed to a term deposit.

Data Description:

- **Dataset**: `bank-additional-full.csv`
- **Source**: UCI Machine Learning Repository
- **Delimiter**: Semicolon (`;`)
- **Features**: Includes demographic information, previous contact details, and various other attributes.
- **Target Variable**: `y` (indicates subscription to a term deposit: 'yes' or 'no').

Initial Data Exploration:

- **Shape**: Displays the dimensions of the dataset.
- **Null Values**: Checks for missing values in the dataset.
- **Data Types**: Provides information on the data types of each column.

Steps Involved:

1. **Reading the Dataset:**
    - Load the dataset into a Pandas DataFrame.
2. **Data Exploration:**
    - Display the first 5 rows.
    - Check the shape of the dataset.
    - Examine null values.

    - Retrieve information about the dataset.
    - Analyze the distribution of the target variable `y`.
3. **Data Visualization:**
    - Plot a count plot of the target variable `y`.
    - Create a histogram plot for the `age` feature.
4. **Data Preprocessing:**
    - Drop the target variable `y` for clustering.
    - Convert categorical features into dummy variables using one-hot encoding.
    - Standardize the features using `StandardScaler`.
5. **Clustering with KMeans:**
    - Apply KMeans clustering to the standardized data.
    - Set the number of clusters to 2.
    - Predict cluster assignments and append them to the original DataFrame.
6. **Evaluation:**
    - Compare the clustering results with the actual target variable `y`.
    - Convert `y` to binary format for comparison.
    - Compute and print the accuracy of clustering compared to the target variable.

Methods and Methodology:

1. **Standardization:**
    - Standardize the dataset to ensure all features are on the same scale, which is essential for distance-based algorithms like KMeans.
2. **KMeans Clustering:**
    - Apply KMeans clustering to segment customers into two clusters based on their attributes.
    - KMeans is used to find the best clusters by minimizing the within-cluster variance.
3. **Accuracy Evaluation:**
    - Evaluate the clustering effectiveness by comparing the cluster labels with the actual target values.
    - Calculate accuracy scores for both possible interpretations of the cluster labels.

Future Work:

- **Improvement in Clustering:**
    - Experiment with different numbers of clusters and other clustering algorithms (e.g., DBSCAN, Agglomerative Clustering) to improve segmentation.
- **Feature Engineering:**
    - Explore additional feature engineering techniques to enhance clustering performance.
- **Model Evaluation:**
    - Conduct further validation using metrics such as silhouette score and Davies-Bouldin index.
- **Integration:**

    - Integrate the clustering results with a recommendation system for targeted marketing strategies.

Conclusion:

The project demonstrates how KMeans clustering can be used to segment customers and evaluate the clustering results against the actual subscription status. Standardizing the data and converting categorical features into dummy variables are crucial preprocessing steps for effective clustering. The accuracy of clustering in predicting customer subscription status provides insights into the potential effectiveness of the segmentation approach.

Results:

- **Accuracy Scores:**
    - The accuracy of clustering compared to the target variable `y` is provided for both interpretations of cluster labels.

Summary:

This project involved preprocessing a customer dataset, applying KMeans clustering to identify segments, and evaluating the clustering effectiveness. By standardizing the data and applying clustering, the project offers valuable insights into customer segmentation and its alignment with customer subscription behavior. Future work will focus on refining the clustering approach and exploring additional features for improved results