Overview of the project

In this project, we start by importing essential Python libraries like pandas, numpy, matplotlib, seaborn, and several machine learning modules from scikit-learn for preprocessing, clustering, and evaluation.

The dataset customer_data_advanced.csv is loaded, containing customer demographics and behavior data. Categorical features such as Gender, Profession, City, Marital Status, and Car Ownership are label-encoded to numerical values for machine learning compatibility.

After selecting key features like Age, Income, Spending Score, and others, the data is standardized using StandardScaler to ensure all features contribute equally.

We then apply multiple clustering algorithms: KMeans, Agglomerative Clustering, DBSCAN, and Gaussian Mixture Models (GMM) to segment customers into different groups.

The performance of KMeans, Agglomerative, and GMM clustering models is evaluated using the Silhouette Score, which measures how well each data point fits its assigned cluster.

For visualization, dimensionality reduction techniques PCA (Principal Component Analysis) and t-SNE are used to plot the high-dimensional data into 2D space, providing visual insights into cluster formation.

Cluster labels from KMeans, Agglomerative, and GMM are saved back into the dataframe, and the updated data is exported as a new CSV file customer_segments_labeled.csv.

In the advanced analysis step, we profile the KMeans clusters by calculating the mean values of features within each cluster, helping to understand customer groups' characteristics.

We also create visualizations such as a Spending Score vs Annual Income scatter plot and an Age Distribution box plot to explore how customer behaviors differ across clusters. Finally, a detailed cluster profile is saved as cluster_profiles.csv, summarizing the segmentation findings.

Customer Segmentation Using Machine Learning - Project Report

Abstract: This project focuses on segmenting customers based on demographic and behavioral features using advanced machine learning techniques. By leveraging clustering algorithms such as KMeans, Agglomerative Clustering, DBSCAN, and Gaussian Mixture Models (GMM), we aim to discover meaningful patterns among customers that can assist businesses in targeted marketing, product recommendations, and customer relationship management. Dimensionality reduction techniques like PCA and t-SNE are used for visualization, while cluster profiling provides deeper insights into each customer group.

Objectives:

- To preprocess and prepare the customer data for machine learning models.
- To apply different clustering algorithms and evaluate their performance.
- To visualize customer segments using PCA and t-SNE.
- To profile customer clusters based on their demographic and behavioral attributes.

Methodology: The project starts with importing libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn modules. The dataset customer_data_advanced.csv is loaded, and categorical features like Gender, Profession, City, Marital Status, and Car Ownership are label-encoded to convert them into numerical format. Selected features including Age, Income, Spending Score, etc., are then scaled using StandardScaler to standardize their ranges.

Multiple clustering algorithms are applied:

- **KMeans Clustering**: Used with 5 clusters.
- **Agglomerative Clustering**: Hierarchical clustering with 5 clusters.
- **DBSCAN**: Density-based clustering without a predefined number of clusters.
- **Gaussian Mixture Model (GMM)**: Probabilistic clustering assuming Gaussian distribution.

Each model's performance is evaluated using the Silhouette Score to identify the best segmentation. Dimensionality reduction is performed using PCA and t-SNE for 2D visualization of clusters. Clusters obtained from KMeans, Agglomerative, and GMM are saved into the dataset. Cluster profiling is conducted to understand the average characteristics of each group, and the final results are saved into separate CSV files.

Results:

- KMeans clustering achieved a high Silhouette Score, suggesting well-separated and dense clusters.
- PCA and t-SNE visualizations clearly show distinct clusters formed by KMeans.
- Cluster profiles reveal interesting patterns such as high-income, low-spending groups and young, high-spending groups.

- The final labeled dataset (customer_segments_labeled.csv) and cluster profiles (cluster_profiles.csv) were successfully generated.

Conclusion: This project demonstrates the effectiveness of machine learning techniques in segmenting customers into meaningful groups. Businesses can leverage these insights to create targeted strategies for each customer segment, ultimately improving customer satisfaction and increasing profitability. The visualizations and profiling further enhance the interpretability of the clustering results, making this approach highly practical for real-world applications.