Clustering Analysis

This project applies clustering algorithms to the **Adult dataset** from the UCI Machine Learning Repository. The goal is to group individuals based on demographic and work-related attributes using **K-Means** and **Hierarchical Clustering**. The analysis includes preprocessing steps, clustering, evaluation, and visualization of results. Internal metrics like the **Silhouette Score** are used to evaluate clustering quality.

Dataset and Loading

The dataset, which contains 48,842 records, is loaded directly from the UCI repository. It includes features such as age, work class, education, hours worked per week, and income level. Missing values (represented as ?) are removed during the loading step, and columns are labeled with meaningful names for better understanding.

Data Preprocessing

- **Handling Categorical Data**: Since clustering algorithms require numerical input, categorical features (e.g., workclass, occupation, native_country) are converted to numerical values using **Label Encoding**.
- **Normalization**: Numerical features (e.g., age, hours_per_week) are normalized using **StandardScaler** to ensure all attributes contribute equally to the clustering process. This step is crucial because clustering algorithms are sensitive to differences in feature scales.

K-Means Clustering

- The **K-Means** algorithm is applied with 2 clusters (as income levels are binary: <=50K and >50K).
- The model identifies cluster centroids and assigns each data point to the nearest cluster.
- The **Silhouette Score** is calculated to evaluate how well the data points fit within their clusters. A higher score indicates better-defined clusters.

Visualization of K-Means Clusters

- To visualize the clusters in two dimensions, **Principal Component Analysis (PCA)** is applied to reduce the dataset's dimensionality. The resulting two components are plotted, with points colored according to their cluster assignments. This visualization helps interpret how the algorithm groups data in a simplified space.

Hierarchical Clustering

- **Agglomerative Clustering** is used to form clusters hierarchically by merging similar data points.
- The **Ward linkage method** is chosen to minimize the variance within each cluster during the merging process.

- A **Dendrogram** is plotted to visualize the hierarchical structure of the clusters, showing how clusters are formed at different levels of granularity

Evaluation

- Both clustering methods are evaluated using the **Silhouette Score**, which measures the cohesion and separation of clusters. This metric helps compare the quality of clustering between K-Means and Hierarchical Clustering.

Key Takeaways

- **K-Means Clustering**: Effective for creating distinct clusters and provides numerical centroids for interpretation.
- **Hierarchical Clustering**: Offers a visual representation of cluster formation through a dendrogram, allowing flexibility in selecting the number of clusters.
- Both methods yield similar results in terms of cluster quality but differ in interpretability and computational complexity.

How to Run

- Ensure you have Python installed along with the required libraries (pandas, numpy, scikit-learn, matplotlib, seaborn, scipy).
- Load the dataset from the UCI repository or any equivalent .csv file.
- Run the script in a Jupyter notebook, Google Colab, or your local Python environment to see clustering results and visualizations.