K-Means Clustering and Support Vector Machine (SVM)

K-Means Clustering

The Marketing Analytics Division of a retail company aims to segment its customers based on their purchasing behavior to improve targeted promotions and loyalty programs. You are provided with the Wholesale Customers dataset, which includes annual spending patterns of clients in various product categories.

Dataset: wholesale_customers.csv

Column Description

- Channel: Type of customer (e.g., Horeca/Hotel/Restaurant/Café or Retail).
- Region: Geographical region of the customer.
- Fresh: Annual spending on fresh products.
- Milk: Annual spending on milk products.
- Grocery: Annual spending on grocery items.
- Frozen: Annual spending on frozen products.
- Detergents_Paper: Annual spending on detergents and paper products.
- Delicassen: Annual spending on delicatessen (specialty) products.

Your Task:

Your task is to apply K-Means Clustering to group customers into distinct segments based on their spending patterns.

- What to Do:

1. Load and explore the dataset, handling any missing or inconsistent values.
2. Normalize numerical columns to ensure fair clustering.
3. Use the Elbow method and Silhouette Score to determine the optimal number of clusters (k).
4. Apply K-Means and visualize clusters using PCA (2D).
5. Interpret each cluster – identify customer types (e.g., high spenders, budget customers).

Allowed Libraries:

• Numpy
• Pandas
• Matplotlib
• Scikit-learn

Support Vector Machine (SVM)

The National Socioeconomic Research Institute (NSRI) is studying income inequality trends and aims to predict whether a person earns more than $50,000 per year based on demographic and employment features. You are provided with the Adult Income dataset, which contains census information such as age, education, occupation, and work hours.

Dataset: adult.csv

Column Description

- age: Age of the individual.
- workclass: Type of employment (e.g., Private, Self-emp, Govt).
- education: Highest education level attained.
- education-num: Numeric representation of education level.
- marital-status: Marital status of the individual.
- occupation: Type of occupation.
- relationship: Relationship status (e.g., Husband, Wife).
- race: Race of the individual.
- sex: Gender of the individual.
- capital-gain: Capital gains from investments.
- capital-loss: Capital losses from investments.
- hours-per-week: Average working hours per week.
- native-country: Country of origin.
- income: Target variable (>50K = 1, <=50K = 0).

Your Task:

Your task is to build an SVM classifier to predict whether a person earns more than $50K annually.

- What to Do:

6. Load and preprocess the dataset (handle missing values marked as '?').
7. Encode categorical variables using LabelEncoder or OneHotEncoder.
8. Normalize numerical features for optimal model performance.
9. Split the dataset into training (80%) and testing (20%) sets.
10. Train an SVM classifier with linear, polynomial, and RBF kernels.
11. Evaluate model performance using accuracy, confusion matrix, and F1-score.
12. Visualize decision boundaries using PCA for dimensionality reduction.

Allowed Libraries:

• Numpy
• Pandas
• Matplotlib
• Scikit-learn