PES
UNIVERSITY
CELEBRATING 50 YEARS

ML Lab Week 13 Clustering Lab Instructions

1. Objective

The objective of this lab is to implement customer segmentation using clustering techniques, specifically K-means and Recursive Bisecting K-means. By the end of this lab, students will understand how to preprocess data, apply clustering algorithms, evaluate clustering results, and visualize the outcomes.

2. Core Concepts

Introduction to Clustering

Clustering is an unsupervised machine learning technique used to group similar data points together based on their features. The main goal of clustering is to identify inherent structures within the data without prior knowledge of labels.

Types of Clustering

Clustering can be broadly categorized into several types:

- **Partitioning Clustering:** This approach divides the dataset into distinct non-overlapping subsets (clusters). Each data point belongs to exactly one cluster. K-means is a popular partitioning clustering algorithm.
- **Hierarchical Clustering:** This method builds a hierarchy of clusters either by a bottom-up approach (agglomerative) or a top-down approach (divisive). Recursive Bisecting K-means is a variant that recursively splits clusters into subclusters.
- **Density-Based Clustering:** This technique groups together data points that are closely packed together, marking as outliers points that lie alone in low-density regions. DBSCAN is a well-known density-based clustering algorithm.

In this lab, we will focus on **K-means and Recursive Bisecting K-means clustering** methods.

3. Files Provided

boilerplate.ipynb :A Jupyter Notebook file containing the skeleton code for the lab.

4. Lab Procedure

1. **Download and Setup:** Download the boilerplate.ipynb file and open it in your Jupyter Notebook environment.
2. **Review the Code:** Read through the entire notebook to understand the structure, the datasets being used, and the helper functions provided.
3. **Complete the Code:** Your main task is to fill in the sections marked with TODO.
4. **Execute the Notebook:** Run all the cells in the notebook from top to bottom. Ensure that all plots are generated and visible in the output.
5. **Analyze the Results:** Observe the performance metrics and the visualizations.
6. **Prepare Deliverables:** Once you have completed the notebook and answered the questions, prepare the two required files for submission as detailed in the next section.

5. Deliverables

You are required to submit **two separate files**. Please follow the naming conventions and content requirements carefully.

Deliverable 1: Completed Jupyter Notebook

This file demonstrates your coding work and the results you obtained.

File Naming Convention:
Rename your completed notebook to
SRN_SECTION.ipynb .

Example: PES1UG23CS123_A.ipynb

Content Requirements:

1. All TODO sections in the notebook must be filled with the correct and functional code.
2. **SRN Identification:** Your SRN must replace the <PES1UG23CSXXX> placeholder if it appears.
3. **Complete Output:** The notebook must be submitted with all cell outputs saved and visible. Do not clear the output before submitting.

Deliverable 2: Lab Report

This document contains your analysis and interpretation of the results.

File Format:
PDF Document.

File Naming Convention:
Name your report SRN_SECTION_Report.pdf .

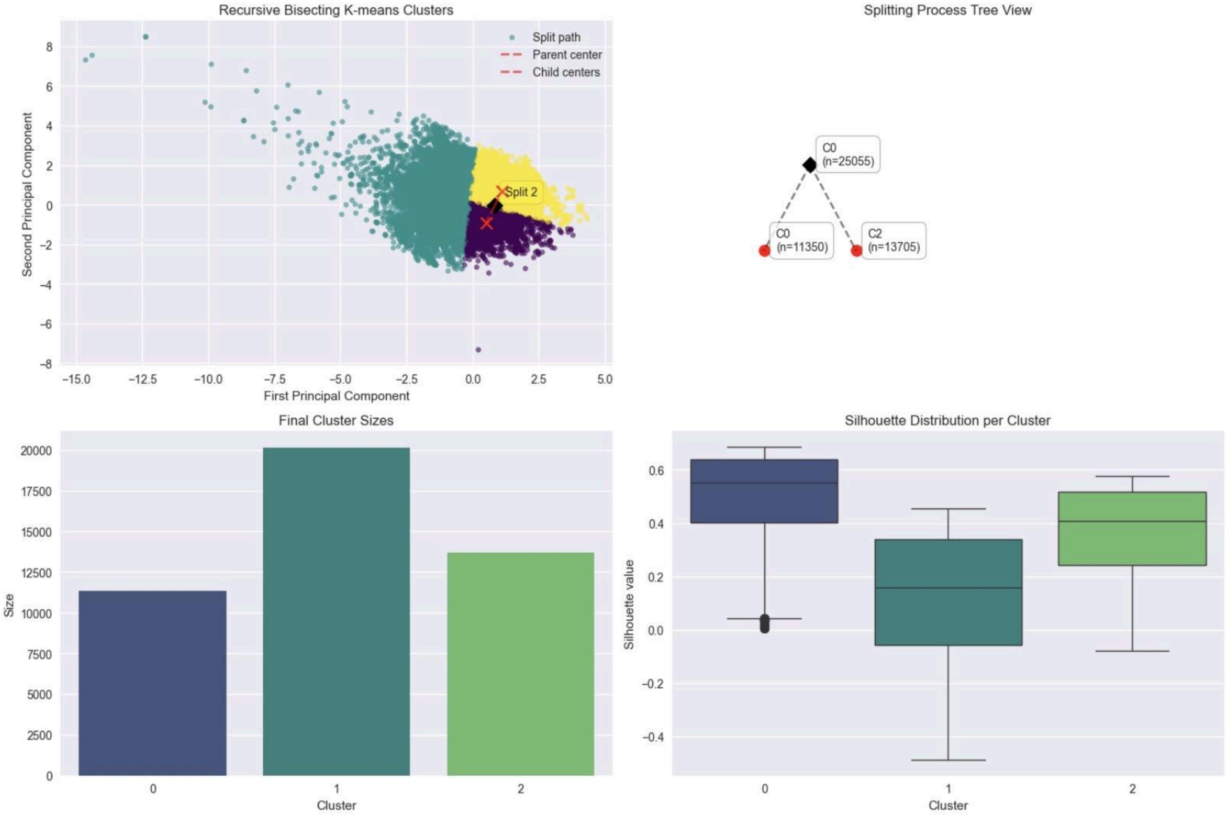
Example: PES1UG23CS123_A_Report.pdf

Content Requirements:

1. **Cover Page:** Include your fullname, SRN, and section.
2. **Analysis Questions:** Provide clear and concise answers to all 8 analysis questions from the notebook. The questions are divided into three sections:
   1. Dimensionality Justification:
      Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset? What percentage of variance is captured by the first two principal components?
   2. Optimal Clusters:
      Looking at both the elbow curve and silhouette scores, what is the optimal number of clusters for this dataset? Justify your answer using both metrics.
   3. Cluster Characteristics:
      Analyze the size distribution of clusters in both K-means and Bisecting K-means. Why do you think some clusters are larger than others? What might this tell us about

the customer segments?

4. Algorithm Comparison:
   Compare the silhouette scores between K-means and Recursive Bisecting K-means. Which algorithm performed better for this dataset and why do you think that is?
5. Business Insights:
   Based on the clustering results in the PCA space, what insights can you draw about customer segmentation that might be valuable for the bank's marketing strategy?
6. Visual Pattern Recognition:
   In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?



3. **Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of** 4 screenshots, divided as

   1. Feature Correaltion matrix for the dataset
   2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA
   3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means
   4. K-means Clustering Results with Centroids Visible (Scatter

Plot)

K-means Cluster Sizes (Bar Plot)

Silhouette distribution per cluster for K-means (Box Plot)

6. Submission

Submit the two files (SRN_SECTION.ipynb and SRN_SECTION_Report.pdf) through the designated google form before the deadline. Please ensure to follow the naming conventions of the submission files.