Final Project Report

Name: R13250028
Repository: GitHub Link (https://github.com/rita1015/bigdata_R13250028)

Objective

The goal of this project is to classify particle accelerator datasets into appropriate groups using unsupervised learning. According to the guideline, if the dataset has n dimensions, we must group it into 4n - 1 clusters. The evaluation metric is the Fowlkes-Mallows Index (FMI), which measures the similarity between predicted and hidden true clusters.

Methodology

- Algorithm Used: K-Means Clustering

We applied the K-Means clustering algorithm as our main unsupervised learning method. This method was executed using the scikit-learn library with n_clusters = 4n - 1, and n_init = 10 to increase stability.

- ✓ Public Dataset: 4 features →15 clusters
- ✓ Private Dataset: 6 features → 23 clusters

- Why K-Means is Suitable

1. Efficiency: K-Means is computationally efficient for medium-sized datasets and works well when the number of clusters is known.
2. Scalability: It handles large amounts of data efficiently.
3. Cluster Geometry: Suitable when clusters are roughly spherical in high-dimensional space.

- High-Dimensional Data Handling

1. Standardization: All features were scaled using `StandardScaler` to ensure equal weight across dimensions.
2. Dimensionality Reduction for Visualization: Principal Component Analysis (PCA) was used to project high-dimensional data into 2D for plotting, aiding visual evaluation of clustering quality.

- Preprocessing and Hyperparameters

1. Standardization：Used StandardScaler() to normalize features
2. Clustering：Used KMeans(n_clusters=4n-1, n_init=10)
3. Visualization：Used PCA to reduce dimensions to 2 for plotting

No missing values or categorical data were present, so no further data imputation or encoding was necessary.

Visualization (PCA 2D Projection)

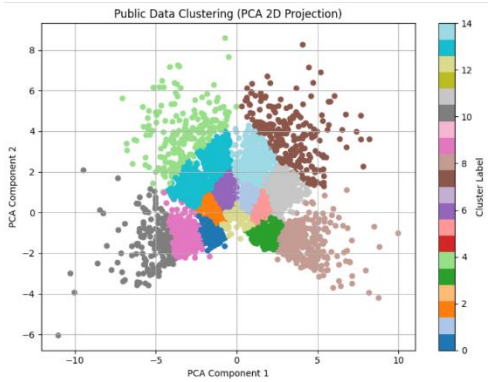

Figure 1. Public dataset clustered into 15 groups.

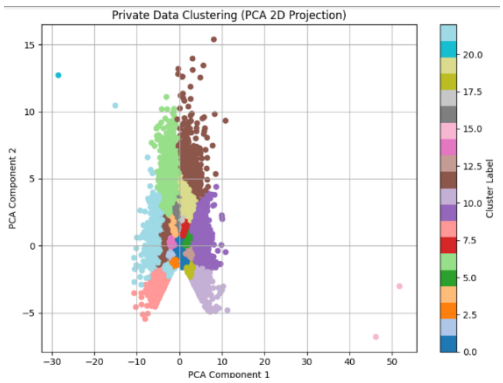

Figure 2. Private dataset clustered into 23 groups.

GitHub Repository

The complete source code, including data processing, clustering implementation, visualizations, and evaluation scripts, is available on GitHub:
> https://github.com/rita1015/bigdata_R13250028

Code link

https://colab.research.google.com/drive/1zizmjaaxjQP6t0uiNwwGxMcelJvVC7E3?usp=sharing