Final Project–Big Data

This is Big data final project for two courses. Please do your best to uncover the hidden secrets of this Big Data. While exploring the boundaries of cosmic knowledge, Professor Liao discovered a series of particle accelerator datasets. These datasets may help him identify the so-called “God Particle” — or perhaps the “Devil Particle” — a potential breakthrough that could lead to the next Nobel Prize.

Task:

Your task is to analyze the relationships within this dataset and classify the data into several distinct categories. It is known that if the data has n dimensions, you should be able to clearly observe 4n – 1 clusters. Please attempt to group the data into these 4n – 1 clusters. The actual numerical labels you assign to the clusters are not important — what matters is whether the clustering itself is accurate.

Your results will be evaluated based on the **Fowlkes–Mallows Index (FMI)**, which measures the similarity between your clustering results and a hidden ground truth. **Note that the FMI is a ratio ranging from 0 to 1**, rather than an absolute score.

There are two types of datasets:

- A public dataset with 4 dimensions, for which we will provide the grading script so you can check your performance.
- A private dataset with 6 dimensions.

Please write a short report explaining why your algorithm is effective at clustering the data.

Your report should briefly describe:

- The algorithm or method you used
- Why it is suitable for this dataset
- How it handles high-dimensional data
- Any preprocessing, hyperparameters, or assumptions involved

Rules:

1. Individual project – each student must work independently.
2. You may use any clustering methods or algorithms that you find suitable.
3. No plagiarism or cheating. Any violations will result in a zero for the final project and may lead to academic dismissal.

Grading Criteria:

- 60% Public dataset score (FMI_pub)
- 30% Private dataset score (FMI_priv)
- 20% Report quality (R)

= Total: 110%

$$\text{Final Score} = 60 \times FMI_{\text{pub}} + 30 \times FMI_{\text{priv}} + 0.2 \times R$$

Report scoring criteria:

Item	Description	Score
Task fulfillment	Use an unsupervised learning method to cluster the data into 4n-1 groups, as specified in the project guidelines.	50%
Technical execution and creativity	Beyond applying standard clustering methods, additional effort to improve accuracy (e.g., through preprocessing or method innovation) will be rewarded.	30%
Report clarity	Technical writing must be structured, readable, and clearly describe the analysis pipeline and clustering algorithms. Good formatting and visual aids are encouraged. **Students are encouraged to include visualizations of their clustering results in the report.**	20%

GitHub (for reference/testing): https://github.com/Jackbear8868/Final-Project-Big-Data

Please create your own repository and commit your code to GitHub.

Please include the GitHub link with your algorithm implementation in the report.

Submission Structure

R12XXXXXX.zip

```
|
├── public_submission.csv        ← Results on public dataset
├── private_submission.csv       ← Results on private dataset
└── Report.pdf (with GitHub link)  ← Method + analysis + code link
```

Hint:

In the public dataset, there appear to be some unique relationships among the four dimensions. Students are encouraged to visualize the data by plotting different pairs of dimensions to observe whether alternative methods can enhance the data representation. For example, in the plot of the second and third dimensions, it is visually apparent that the data can be separated into five distinct clusters.

Since both the public and private datasets originate from the same type of physical event, their characteristics should be similar. This implies that a similar cluster structure should also be visible in the second and third dimensions of the private dataset.

Inter-dimensional Relationships in the Public Dataset

S1 vs S2	S2 vs S3	S3 vs S4