

**Q1.** You will work with the synthetic 2D **Q1 dataset** containing two features and three underlying clusters. You are required to implement, tune, compare, and interpret three clustering algorithms on the generated dataset. Your analysis should demonstrate both technical correctness and conceptual understanding.

**Part A** – Select and Implement the three clustering algorithms from this list: K-Means, GMM, DBSCAN, Hierarchical. Clearly state any assumptions each algorithm makes about cluster shape, size, or density.

**Part B:** For each algorithm, identify and tune the key hyperparameters. Examples include (but are not limited to):

- K-Means: number of clusters (k), initialization method
- GMM: number of components, covariance type
- DBSCAN:  $\epsilon$  (epsilon), min\_samples
- Hierarchical: linkage criterion, distance metric

Requirements:

- Describe your method
- Report the best configuration and justify why it is optimal.

**Part C:** Evaluate and compare the three algorithms using three clustering metrics like:

- Silhouette Score
- Davies–Bouldin Index
- Calinski–Harabasz Index

Tasks:

1. Present results in a comparison table.
2. Rank the algorithms based on overall performance.
3. Explain cases where different metrics disagree.

**Part D:**

1. Plot the clustered data for each algorithm.
2. Overlay decision boundaries or cluster regions where applicable.
3. Comment on:
  - Cluster shape and separation
  - Misclustered or ambiguous regions
  - Differences in how algorithms partition the same space

**Part E:** Answer the following questions:

1. Which algorithm best matches the true data-generating process, and why?

## Final Exam

2. How would the results change if:
  - The number of clusters were unknown?
  - Significant noise or outliers were added?
  - The clusters had very different densities?