Introduction to Machine Learning - Clustering & Dimensionality Reduction

Dataset

- Synthetic Circle - This dataset comprises 10000 two-dimensional points arranged into 100 circles, each containing 100 points. It was designed to evaluate clustering algorithms by providing a clear and structured clustering challenge.
  - Dataset Characteristics - Multivariate
  - Feature Type - Real Numbers
  - # Instances - 10000
  - # Features - 2
  - Target Variable - Integer

- Bank Marketing - The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).
  - Dataset Characteristics - Multivariate
  - Feature Type - Categorical, Integers
  - # Instances - 45211
  - # Features - 17
  - Target Variable - Boolean

Objectives

- Get hands-on with clustering methods: understand, implement, and compare.

- Explore how dimensionality reduction affects clustering results.

- Practice data exploration, preprocessing, clean code, and clear reporting.

- Improve ability to interpret results: strengths, weaknesses, and use cases.

Deliverable

Submit a **single Jupyter Notebook** that–for both datasets–is:

- Well-structured with sections, headings, and subheadings
- Clean and readable, with comments and meaningful variable names
- Written like a report: includes explanations, visualizations, and discussion (not just code)
- Fully cited: any references (papers, blogs, documentation) must be acknowledged

---

Tasks / Outline

1. Introduction

- Briefly describe the datasets: its features and what “conflicting” means.
- State the objective of this notebook.

2. Exploratory Data Analysis (EDA)

- Summarize the datasets: number of samples, features, data types, missing values.
- Visualize important feature distributions and correlations.
- Identify any potential data quality issues.

3. Preprocessing

- Handle missing values (imputation or removal).
- Encode categorical variables (One-Hot Encoding or similar).
- Normalize or standardize numerical features.
- (Optional) Perform feature selection or engineering, undersampling and other techniques.

4. Clustering Methods

Use **at least two clustering algorithms** (choose from K-Means, ROCK, Hierarchical, DBSCAN, Gaussian Mixture Models, etc.) across the 2 datasets.

For each method:

- Choose and explain hyperparameters (e.g., number of clusters, distance metric).
- Visualize results (cluster assignments, cluster sizes, 2D plots).
- Evaluate clusters with metrics (e.g., silhouette score, Davies–Bouldin index).
- Discuss pros and cons and use cases of this method.

5. Dimensionality Reduction

Apply **at least one dimensionality reduction** techniques (e.g., PCA, LDA etc) to each dataset.

- Visualize the data after reduction.
- Explain why each method was chosen.

6. Clustering After Dimensionality Reduction

- Re-run your clustering methods on the reduced data.
- Compare results before vs. after reduction:
  - Are clusters more compact, more separated, or less meaningful?
  - Compare quantitative metrics and provide qualitative observations.

7. Comparison & Discussion

- Which clustering methods worked best?
- How did dimensionality reduction affect clustering performance?
- What are the limitations of your approach?

8. Conclusion

- Summarize key findings and insights from the exercise.

9. References

- Cite all external sources used (papers, documentation, tutorials).

Important Notes

- Both teammates are required to submit the same notebook
- The emphasis will be more on the process than the results.
- AI-generated solutions are not allowed. All work must be your own.
- Use markdown cells and comments to explain your decisions.
- Use clear, labeled visualizations to support discussion.
- Improper referencing will result in grade penalties.

Bonus

In addition to the main assignment, there will be a **Bonus Clustering Challenge** held **during a regular class session**.

- A **new dataset** will be released **at the start of class**.
- You will have to **run your implemented clustering algorithms** on it in real-time.
- Your task will be to:
    - Plot the **cluster centers** and cluster visualization
    - Write a short paragraph explaining what **insights** you can infer from the clusters