Abdelhamid Ibn Badis University – Mostaganem
Faculty of Exact Sciences and Computer Science
3rd Year Engineering – Data Science
Advanced Programming

Project 3: K-Means Clustering (5 points)

This project will deepen your understanding of clustering and unsupervised learning by implementing the K-Means algorithm from scratch.

By the end of this project, students will be able to:

- Implement K-Means manually with Python functions
- Compute distances, update centroids, and assign clusters
- Evaluate clustering results using error metrics
- Use the elbow method to identify the optimal number of clusters

Each project may be completed individually or in pairs

Part A: Preparing the Data

Dataset

You will use the dataset **College_Data.csv**, which contains 777 observations and 19 variables.

- **Name**: The name of the college or university
- **Private**: A factor with levels *No* and *Yes* indicating whether the university is private or public
- **Apps**: Number of applications received
- **Accept**: Number of accepted applications
- **Enroll**: Number of newly enrolled students
- **Top10perc**: Percentage of new students from the top 10% of their high school class
- **Top25perc**: Percentage of new students from the top 25% of their high school class
- **F.Undergrad**: Number of full-time undergraduate students
- **P.Undergrad**: Number of part-time undergraduate students
- **Outstate**: Tuition fees for out-of-state students
- **Room.Board**: Housing and meal plan costs

- **Books**: Estimated book costs
- **Personal**: Estimated personal expenses
- **PhD**: Percentage of faculty members with a PhD
- **Terminal**: Percentage of faculty members with a terminal degree
- **S.F.Ratio**: Student-to-faculty ratio
- **perc.alumni**: Percentage of alumni who donate
- **Expend**: Instructional expenditure per student
- **Grad.Rate**: Graduation rate

Tasks

Each group will:

1. Load and Explore the Dataset

- Load **College_Data.csv** using Pandas
- Display the first 10 rows
- Print summary statistics

2. Select Variables for Clustering

Because K-Means requires numerical features:

- Remove “Name” (string)
- Convert Private to numeric (Yes=1, No=0)
- Keep all other numeric columns
- Store the final numerical dataframe as df_numeric

Part B: Implementing K-Means

You must write all K-means core functions manually, without using sklearn.cluster.KMeans or any similar packages

Function	Parameters	Description
initialize_centroids(data, k)	data, k	Randomly choose k points as initial centroids
compute_distance(point, centroids)	point, centroids	Compute the Euclidean distance from one point to all centroids
assign_clusters(data, centroids)	data, centroids	Assign each data point to the nearest centroid
update_centroids(data, labels, k)	data, labels, k	Compute the new centroid of each cluster
compute_inertia(data, labels, centroids)	data, labels, centroids	Compute total SSE within clusters
kmeans(data, k, max_iter=100, patience=2)	Data, k, max_iter, patience	Runs the K-Means clustering algorithm using the data. The maximum number of iterations is max_iter. patience is the number of consecutive iterations with unchanged cluster assignments before stopping early

Part C: Apply the Implemented K-Means on the Dataset

Test your implementation on the given dataset and make sure that it is fully functional

Part D: Apply K-Means with Several K Values

- Implement the elbow_method(data, k_range) function, where data is your dataset, and k_range a set k values to use
- Run your K-Means implementation for k = 2, 3, 4, ..., 10
- Store each model's inertia (SSE)
- Generate the elbow plot to visualize the best K

Technical Requirements

You can use the following libraries:

- pandas to manipulate the data
- matplotlib or seaborn for optional visualization

You **cannot:**

- Use sklearn.cluster.Kmeans or any similar packages
- Use AI to implement the functions

Your code must:

- Contain clear comments explaining each major step
- Include a header comment with:
  - Student name(s)
  - Project description

Deliverables

Submit a single .zip file containing:

1. Python script(s)
2. Short report (2 pages max) in .pdf or .docx, including:
   - Description of any additional function
   - Output of the clusters
   - Challenges encountered

Evaluation Criteria

Criterion	Points
Correct dataset loading & preprocessing (df_numeric)	1.0
Correct implementation of core K-Means functions	3.0
Correct elbow method implementation and interpretation	1.0
Total	5.0

Project Timeline

- Start Date: November 30, 2025
- Due Date: December 12, 2025