Department of Computer Engineering
BE Computer-B (2025-26 Sem I)
LP-III Machine Learning
Practical Assignment 5: K-Means clustering/ hierarchical clustering

[CO1, CO3, BT: L3 (Apply)] [Max Marks: 10]

Problem Definition: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset.

Determine the number of clusters using the elbow method.

Dataset link: https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

Learning Outcomes:

By completing this practical, students will be able to:

Learning Outcome	**Bloom's Taxonomy (BT) Level**
Understand clustering as an unsupervised learning technique	Remember (L1)
Apply K-Means and Hierarchical clustering on real-world data	Apply (L3)
Determine optimal number of clusters using the Elbow method	Analyze (L4)
Interpret cluster results for business insights	Evaluate (L5)
Visualize clustering results effectively	Create (L6)

Software / Hardware Requirements

Software:

- Python 3.x
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, scipy
- Jupyter Notebook / Google Colab

Hardware:

- Minimum 4 GB RAM
- Intel i3 or higher
- Stable internet connection (for data access)
- Recommended: GPU support for faster training

Theory:

The aim of this experiment is to implement the K-Nearest Neighbors algorithm on the PIMA Indians Diabetes Dataset to classify whether a patient is diabetic or not based on given medical parameters. The dataset contains diagnostic measurements such as glucose level, blood pressure, insulin level, body mass index (BMI), and others. The model will be trained and tested to evaluate its classification performance using metrics like confusion matrix, accuracy, error rate, precision, and recall.

1. Dataset Description:

The dataset used is **sales_data_sample.csv** (from Kaggle), which contains **sales transactions data** of a company. It includes details about **customers, products, sales, orders, and geographical information**.

Attributes in diabetes.csv:

Attribute	**Description**
ORDERNUMBER	Unique identifier for each sales order.
QUANTITYORDERED	Quantity of items ordered in the transaction.
PRICEEACH	Price of a single item.
ORDERLINENUMBER	Order line sequence number within the order.
SALES	Total sales amount for the order line (Quantity × Price).
ORDERDATE	Date when the order was placed.
STATUS	Status of the order (e.g., Shipped, Cancelled, On Hold).
QTR_ID	Quarter of the year in which the order was placed.
MONTH_ID	Month (numeric) of the order.
YEAR_ID	Year in which the order was placed.
PRODUCTLINE	Category of the product (e.g., Classic Cars, Motorcycles, Trucks, Ships, Trains).
MSRP	Manufacturer's Suggested Retail Price.
PRODUCTCODE	Unique code for the product.
CUSTOMERNAME	Name of the customer.
PHONE	Contact number of the customer.
ADDRESSLINE1/2	Address details of the customer.
CITY	City of the customer.
STATE	State/Province of the customer.
POSTALCODE	Postal/ZIP code.
COUNTRY	Country of the customer.
TERRITORY	Sales territory.
CONTACTLASTNAME	Contact person's last name.
CONTACTFIRSTNAME	Contact person's first name.
DEALSIZE	Size of the deal (Small, Medium, Large).

Acknowledgement:

The dataset is referred from Kaggle: https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

2. Clustering

- Unsupervised learning technique that groups similar data points.
- Helps in customer segmentation, anomaly detection, sales analysis.

2.1 K-Means Clustering

- Partitioning method where dataset is divided into K clusters.

- Steps: Initialize centroids → Assign points → Update centroids → Repeat until convergence.

Objective Function (WCSS):

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

where μi is centroid of cluster Ci.

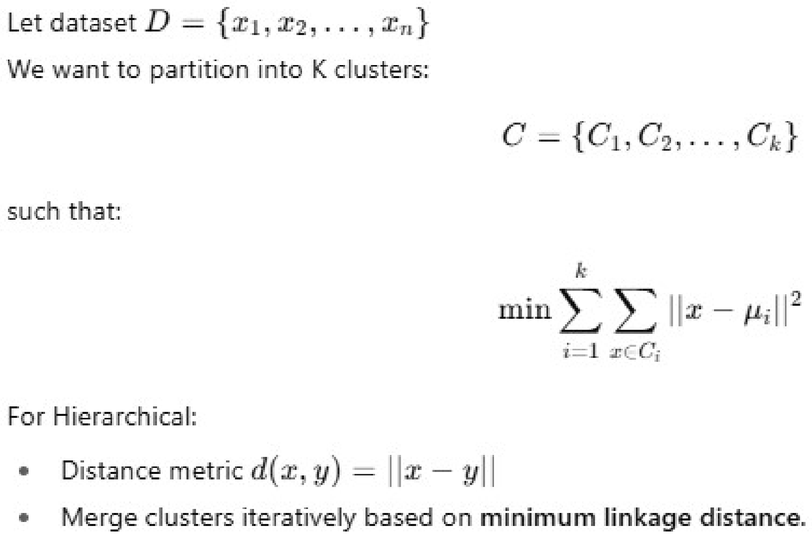
2.2 Hierarchical Clustering

- Builds nested clusters.
- Agglomerative (bottom-up) and Divisive (top-down) approaches.
- Represented using dendrogram.

Elbow Method

- Plots WCSS vs K.
- The "elbow point" indicates the optimal number of clusters.

Mathematical Model



Let dataset $D = \{x_1, x_2, \ldots, x_n\}$

We want to partition into K clusters:

$$C = \{C_1, C_2, \ldots, C_k\}$$

such that:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

For Hierarchical:

- Distance metric $d(x, y) = ||x - y||$
- Merge clusters iteratively based on **minimum linkage distance.**

Model Evaluation Metrics:

1. **Within-Cluster-Sum of Squares (WCSS) – Compactness of clusters.**
2. **Silhouette Score – Quality of clustering:**

s= {b-a}/max(a,b)

where a = intra-cluster distance, b = nearest cluster distance.

Algorithm

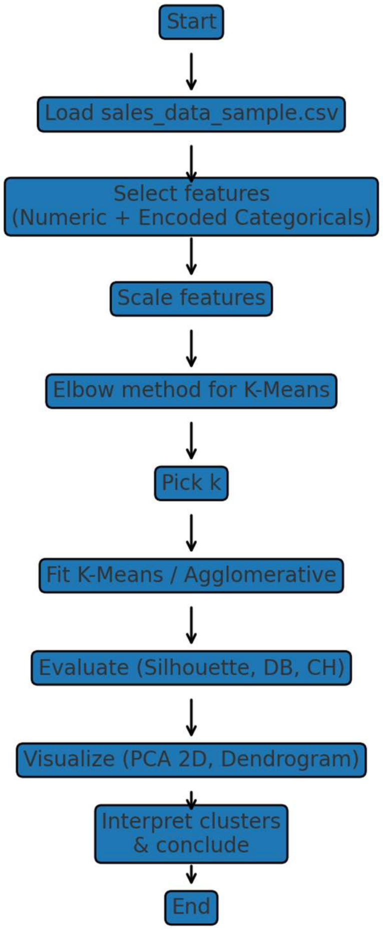
K-Means Algorithm

1. Choose number of clusters K.
2. Randomly initialize centroids.
3. Assign each data point to nearest centroid.
4. Update centroids as mean of assigned points.
5. Repeat until centroids stabilize.

Hierarchical Algorithm (Agglomerative)

1. Start with each point as a single cluster.
2. Merge two closest clusters based on linkage distance.
3. Repeat until only one cluster remains.
4. Cut dendrogram at chosen level to form clusters.

Flow Chart


Start
Load sales_data_sample.csv
Select features
(Numeric + Encoded Categoricals)
Scale features
Elbow method for K-Means
Pick k
Fit K-Means / Agglomerative
Evaluate (Silhouette, DB, CH)
Visualize (PCA 2D, Dendrogram)
Interpret clusters
& conclude
End

Expected Results (Screen Shots of Output)

- Elbow Method graph shows optimal clusters (likely 3–5).
- K-Means groups sales data into clusters (e.g., high-value, medium-value, low-value customers).
- Hierarchical Dendrogram shows hierarchical grouping.
- Segmentation provides useful business insights.

Conclusion

- K-Means provides **efficient partition-based clustering** for sales data.
- Hierarchical clustering provides **tree-based structure** useful for visualization.
- Elbow method determines **optimal number of clusters** for meaningful segmentation.
- Clustering results can help in **customer profiling, product strategy, and revenue analysis**.

Viva Questions

Question	BT Level
Differentiate between supervised and unsupervised learning.	Understand (L2)
Why is standardization important before clustering?	Apply (L3)
How do you decide the number of clusters in K-Means?	Apply (L3)
Explain the concept of Within-Cluster-Sum of Squares.	Analyze (L4)
What is the difference between K-Means and Hierarchical clustering?	Analyze (L4)
How does Silhouette Score help in evaluating clusters?	Evaluate (L5)
Suggest a business scenario where clustering can be applied effectively.	Create (L6)

Rubrics for Evaluation (Example)

Criteria	**Best**	**Good**	**Average**	**Poor**
Correctness of Program / Demonstration of Program [4]	Perfect [4]	Can be better [3]	Satisfied [2]	Poor [0]
File Submission/Documentation [2]	Perfect [2]	Can be better [1.5]	Satisfied [1]	Poor [0]
Timely Submission [2]	On Time [2]	After 1 week [1.5]	After 1.5 week [1]	At the end of semester [0]
Viva [2]	Perfect [2]	Can be better [1.5]	Satisfied [1]	Poor [0]

Dr. Nikita Singhal

(Subject Incharge)

DR SR Dhore

(HoD)