Clustering Algorithms

Dr. Surya Prakash

Professor

Department of Computer Science & Engineering

Indian Institute of Technology Indore, Indore-453552, INDIA

E-mail: surya@iiti.ac.in

Introduction

- Machine Learning Techniques – 3 Types
  - Supervised
    - **Regression**: predict continuous values.
      - Linear regression
    - **Classification**: predict categories
      - Logistic regression, Decision tree, *k*-Nearest Neighbor
  - Unsupervised
    - Clustering problem: hierarchical clustering, *k*-means clustering
    - Dimensionality reduction problem
  - Reinforcement Learning
    - Reinforcement Learning is a type of machine learning where **an agent (for example, the algorithm)** learns to make decisions by receiving **rewards or penalties** from its environment.

Introduction

- A **common thread in all machine learning techniques** is that they **learn patterns or make decisions by being trained on data**.
  - **Supervised learning:** Trained on labeled data (input → correct output).
  - **Unsupervised learning:** Trained on unlabeled data to find patterns or structure.
  - **Reinforcement learning:** Learns by trial-and-error, receiving feedback (rewards/penalties) from interactions with the environment.

- So, whether it's labeled examples, raw data, or environment feedback, **data or experience drives the learning**.

Supervised Techniques

- Nature of Training data
  - **Labeled data**

- Labeled Data:
  - It refers to the data that has been tagged or annotated with **meaningful information (labels)** that describe what each piece of data represents.
    - input → correct output

Labeled Data - Examples

- In image recognition (cat *vs.* dog):
  - A photo of a cat 🐱 labeled as **"cat"**.
  - A photo of a dog 🐶 labeled as **"dog"**.

- In sentiment analysis:
  - Text: *"The movie was great!"* → Label: **Positive**.
  - Text: *"The food was terrible."* → Label: **Negative**.

- In medical diagnosis:
  - X-ray image → Label: **"Pneumonia"** or **"Healthy"**.

Labeled database: Example 1 - Play Tennis dataset

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Example: To play tennis or not

Inputs (Features)
→ **First four columns**

Output (Label)
→Last column **Play**

Labeled database: Example 2 - Loan dataset

Applicant Age	Annual Income	Credit Score	Existing Debts	Loan Approved (Label)
28	₹6,00,000	720	No	Yes
45	₹4,50,000	680	Yes	No
32	₹8,00,000	750	No	Yes
50	₹3,00,000	640	Yes	No
38	₹5,50,000	700	No	Yes
26	₹4,00,000	690	No	No

Example: Predicting if a loan application will be approved

Inputs (Features):

→ Age, Income, Credit Score, Debts — measurable and available in real banking systems.

Output (Label):

→ Whether the bank approved the loan or not.

Labeled database: Example 3 - Diabetes dataset

Age	BMI	Blood Pressure (mmHg)	Glucose Level (mg/dL)	Diabetes (Label)
45	28.5	130	180	Yes
33	24.0	120	95	No
50	31.2	140	210	Yes
29	22.1	118	88	No
60	34.8	150	170	Yes
42	26.3	125	102	No

Example: Predicting whether a patient has diabetes.

Inputs (Features):
→ Age, BMI, Blood Pressure, Glucose Level.

Output (Label):
→ Diabetes Test Outcomes: “Yes” or “No” for diabetes diagnosis (based on confirmed medical tests).

Labeled database: Example 4 - Fruit dataset

Color	Weight (grams)	Size (cm)	Sweetness (1-10)	Fruit Type (Label)
Red	180	8	8	Apple
Yellow	120	6	9	Banana
Green	150	7	6	Apple
Orange	160	7	8	Orange
Yellow	100	5	10	Banana
Orange	170	8	7	Orange
Red	200	9	9	Apple

Classification of different fruits

Inputs (Features):
→ Color, Weight, Size, Sweetness level.

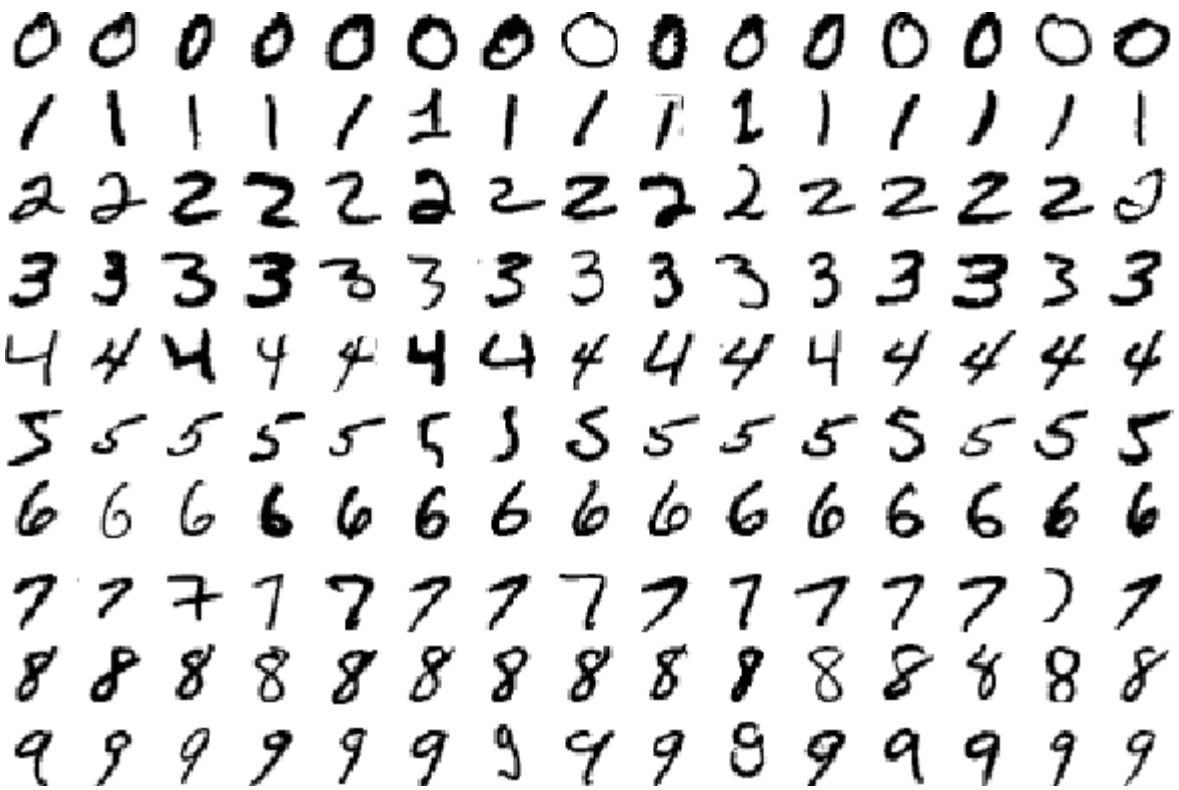
Output (Label):
→ Type of fruit (Apple, Banana, Orange).

Image-based Labeled Public Datasets

Dataset Name	Type	Description	Example Labels
MNIST	Supervised	70,000 grayscale images of handwritten digits (28×28 pixels)	0–9
CIFAR-10	Supervised	60,000 color images (32×32 pixels) of 10 object categories	Airplane, Dog, Truck…
ImageNet	Supervised	14M+ high-resolution images, 20,000+ categories	Cat, Car, Chair…
COCO	Supervised	330,000 images with object detection & segmentation	Person, Bicycle, Pizza…
LFW	Supervised	13,000 face images	Person's name
ChestX-ray14	Supervised	100,000+ chest X-ray images for medical diagnosis	Pneumonia, Healthy…

MNIST Dataset

- Link: http://yann.lecun.com/exdb/mnist/



0
1
2
3
4
5
6
7
8
9

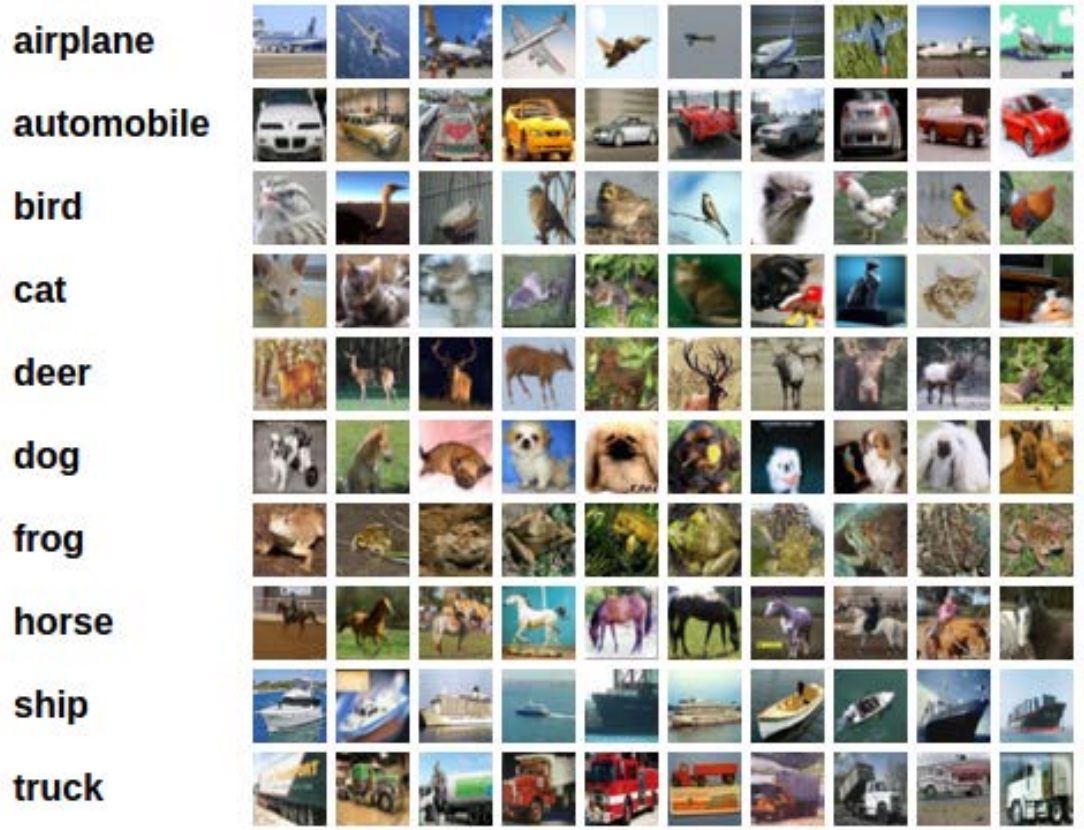
→ Training samples: 60,000
→ Test samples: 10,000

→ Each image is represented by 28x28 pixels,
→ Each containing a value 0 - 255 with its grayscale value.

CIFAR10 Dataset

- It has 10 classes:
  - ‘airplane’, ‘automobile’, ‘bird’, ‘cat’, ‘deer’, ‘dog’, ‘frog’, ‘horse’, ‘ship’, ‘truck’.

- The images in CIFAR-10 are of size 3x32x32
  - 3-channel color images of 32x32 pixels in size.



Unsupervised Techniques

- Nature of Training data
  - **No target labels** are provided – the algorithm only has input features.
  - The algorithm tries to **discover hidden patterns** in the data by exploring the similarities in features of different input samples.

Example: Fruit Dataset (Unlabeled)

Discover hidden patterns in the data by exploring the similarities in features of different input samples.

Example: Customer Shopping Data (Unlabeled)

Customer ID	Age	Annual Income (₹)	Spending Score
1	22	25,000	39
2	35	60,000	81
3	26	45,000	6
4	29	30,000	77
5	45	80,000	40
6	34	25,000	5

Problem:

→ To group customers with similar spending patterns.

Why is it unsupervised?

- There is **no "target" column** (like "Will Buy Again: Yes/No").
- The model can use clustering (e.g., **K-Means**) to group customers with similar spending patterns.

Applications of Clustering

- What is the use of grouping the customers with similar spending patterns.

- **1. Targeted Marketing**
  - Instead of sending the same offers to everyone, the company can send **customized promotions** to each group.
  - **Example:**
    - High spenders → Premium product recommendations
    - Low spenders → Discount coupons to encourage more purchases

Applications of Clustering

- **2. Product Recommendations**
  - Suggest products that similar customers bought.
  - **Example:**
    - "Customers like you also purchased…" on Amazon.

- **3. Improve Customer Retention**
  - Identify groups at risk of leaving and give them **special incentives**.
  - **Example:**
    - People with low spending score but high income might need better engagement.

Applications of Clustering

- **4. Strategic Decision-Making**
  - Helps decide **where to open new stores**, **which products to stock more**, or **which regions to focus on**.

- **5. Personalized Customer Experience**
  - Make customers feel valued by understanding their **preferences**.

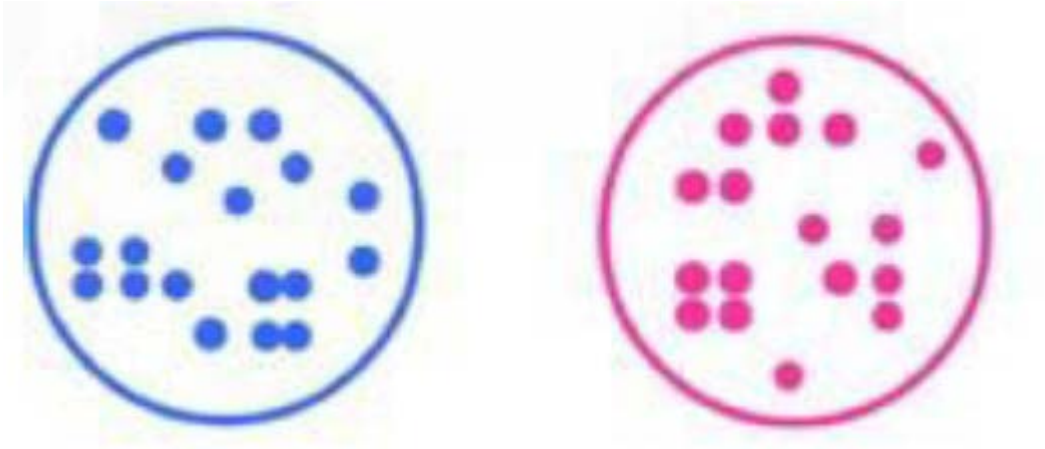
What is Clustering?

- **Clustering** is an **unsupervised machine learning technique** used to group similar items together based on their features, **without using predefined labels**.
  - The idea is that items in the **same cluster** are more similar to each other, while items in **different clusters** are less similar.

- **Example: Shopping Mall Customers**
  - A mall collects data about customers' **age** and **annual income**.
  - A clustering algorithm might group them into:
    - **Cluster 1:** Young with low income (students, interns)
    - **Cluster 2:** Middle-aged with high income (professionals)
    - **Cluster 3:** Elderly with medium income (retirees)

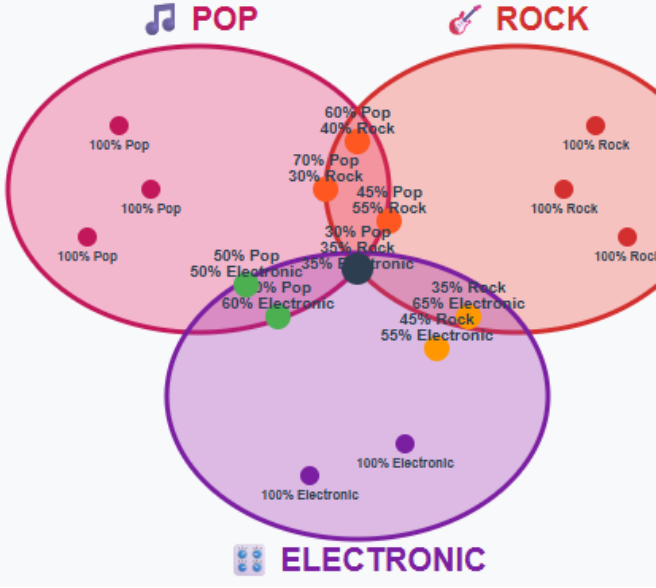
Different Clustering Techniques

- **Exclusive Clustering (Hard Clustering)**
  - Each data point belongs to **exactly one cluster**.
  - **Key Idea:** No overlaps — once assigned, a point can't be part of another cluster.
  - **Example Algorithm:** K-Means, K-Medoids.
  - **Example Use:**
    - Dividing students into groups based on grades (A-group, B-group, etc.) where each student fits only one group.



Different Clustering Techniques

- **Overlapping Clustering (Soft Clustering / Fuzzy Clustering)**
  - A data point can belong to **multiple clusters** with varying degrees of membership (probabilities or weights).
  - **Key Idea:** Real-world objects can share traits with more than one group.
  - **Example Algorithm:** Fuzzy C-Means, Gaussian Mixture Models (GMM).
  - **Example Use:**
    - Music classification → A song can be both "Pop" (60% membership) and "Rock" (40% membership).



Different Clustering Techniques

- **Hierarchical Clustering**
  - Creates a **tree-like structure** (dendrogram) of clusters by merging or splitting them step-by-step.
  - **Types:**
    - **Agglomerative (Bottom-Up):** Start with each point as a cluster → merge.
    - **Divisive (Top-Down):** Start with one cluster → split.

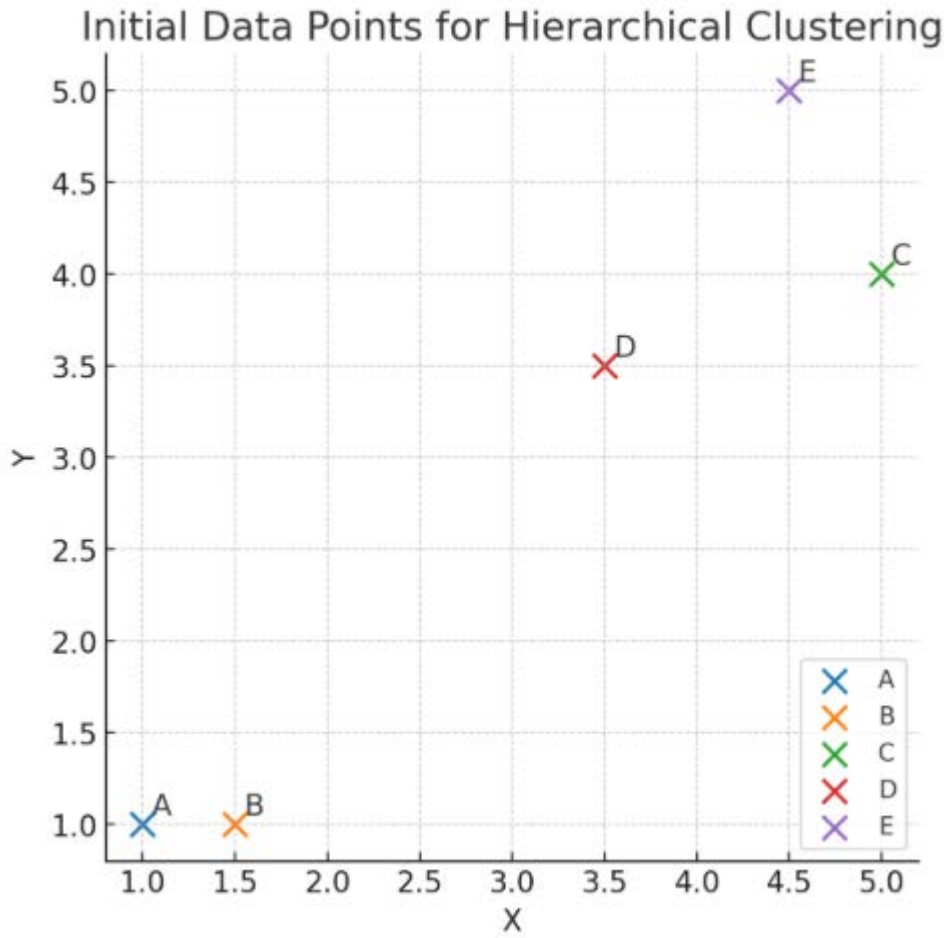
Quick Comparison Table

Feature	Exclusive Clustering	Overlapping Clustering	Hierarchical Clustering
Membership	One cluster only	Multiple clusters	Multiple levels
Output	Flat groups	Membership weights	Dendrogram
Example Algorithm	K-Means	Fuzzy C-Means, GMM	Agglomerative, Divisive
Best For	Clear separation	Shared characteristics	Data with hierarchy

Hierarchical Clustering - Example

Dataset

Point	X	Y
A	1.0	1.0
B	1.5	1.0
C	5.0	4.0
D	3.5	3.5
E	4.5	5.0



Hierarchical Clustering - Example

Step 0 — Initial distances

	A	B	C	D	E
A	0	**0.500**	5.000	3.905	5.315
B	0.500	0	4.301	3.202	4.716
C	5.000	4.301	0	1.581	**1.118**
D	3.905	3.202	1.581	0	1.803
E	5.315	4.716	1.118	1.803	0

Merge: A and B (0.500) → Cluster **AB**

Hierarchical Clustering - Example

Step 1 — After merging A & B

Clusters: **AB**, C, D, E

	AB	C	D	E
AB	0	**4.301**	3.202	4.716
C	4.301	0	1.581	**1.118**
D	3.202	1.581	0	1.803
E	4.716	1.118	1.803	0

Merge: C and E (1.118) → Cluster **CE**

In **single-linkage hierarchical clustering**, when we merge two points into a new cluster (**AB**), the distance from that cluster to any other point/cluster is the **minimum distance** between *any* member of the new cluster and the other point.

Hierarchical Clustering - Example

Step 2 — After merging C & E
Clusters: **AB**, **CE**, D

	AB	CE	D
AB	0	**4.301**	3.202
CE	4.301	0	**1.581**
D	3.202	1.581	0

Merge: CE and D (1.581) → Cluster **CDE**

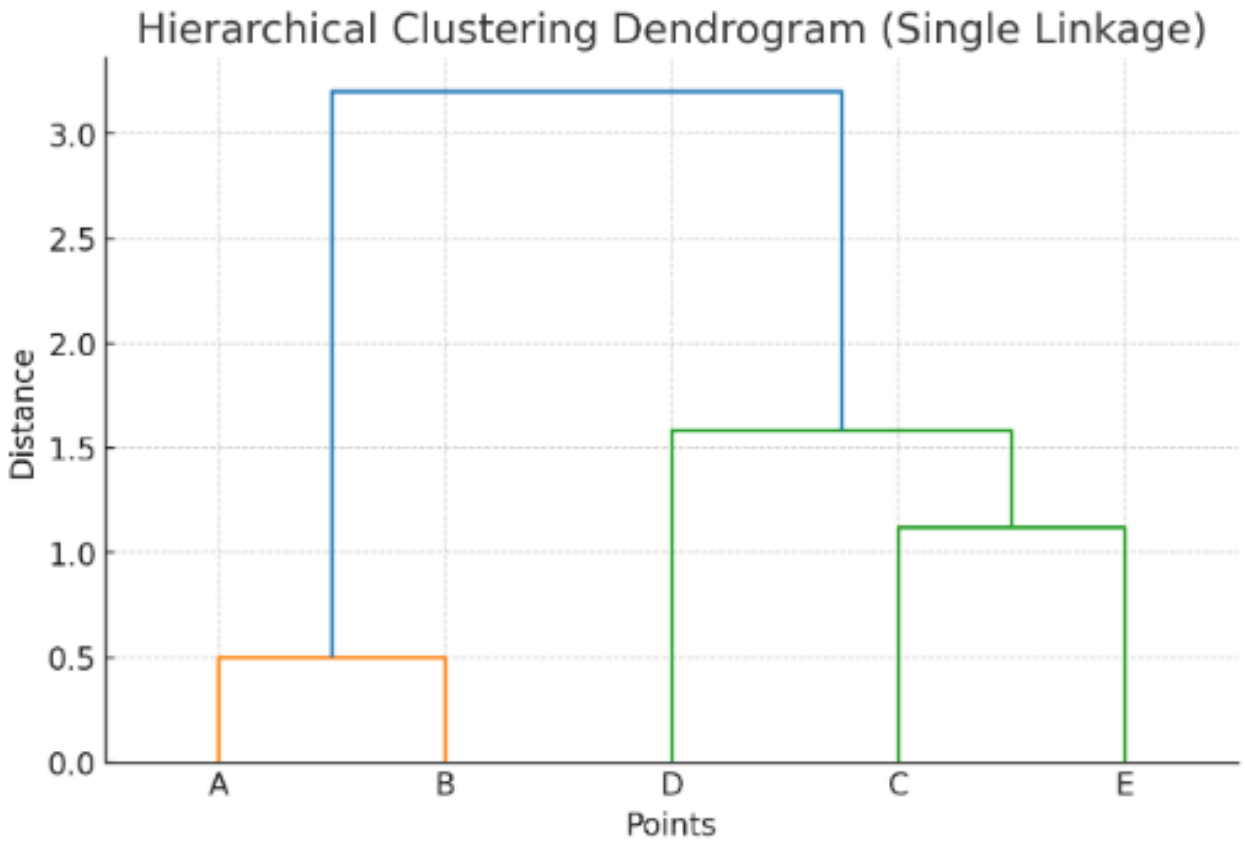
Hierarchical Clustering - Example

Step 3 — After merging CDE & AB
Clusters: **AB**, **CDE**

	AB	CDE
AB	0	**3.202**
CDE	3.202	0

Merge: AB and CDE (3.202) → Final cluster **ABCDE**

Dendrogram



To get clusters from a dendrogram, you basically **choose a cut height** and group everything that merges **below** that height together.

Here,
Cut at height = 2
{A, B} is one cluster
{D, C, E} is another

Cut at height = 1.5
{A, B} becomes one cluster
{D} becomes one cluster
{C, E} is another

Thank You