AICERTs™

# AI+ Foundation™

Certification

# AI + Foundation

## Module-1

## Hands-on 2

**Title:** Identifying Supervised Learning (Labeled Datasets) vs. Unsupervised Learning (Unlabeled Datasets) in Machine Learning

### Problem Statement:

In machine learning, the type of dataset—labeled or unlabeled—determines which learning approach to use: supervised or unsupervised.

This hands-on lab activity helps learners visually identify the difference between the two types of datasets, understand their structure, and explore how machine learning models interact with each type. By analyzing examples and applying basic ML logic, learners will gain a foundational understanding of how data labeling influences model selection and training.

### Implementation in Detailed Manner

### Supervised Learning

The table below shows a labeled dataset that gives past lead data with the outcome (whether they are converted into clients).

It includes features like Lead Location, Gender, Company Size, Revenue, and Number of Emails. The 'Client' column is the label, showing whether the lead converted (YES) or not (NO). Green highlighting emphasizes the positive class ('YES'), ideal for visualizing classification tasks.

**Labeled Dataset: Lead Conversion**

| Lead Location | Gender | Lead Company Size | Lead Revenues | Number of Emails | Client |
|---|---|---|---|---|---|
| UK | F | 5 | 500000 | 3 | NO |
| Ireland | F | 6 | 600000 | 2 | NO |
| UK | M | 6 | 600000 | 5 | YES |
| USA | F | 10 | 1000000 | 12 | YES |
| USA | F | 3 | 300000 | 4 | YES |
| USA | M | 5 | 500000 | 7 | YES |
| Germany | M | 6 | 600000 | 1 | NO |

# ✅ Step-by-Step Walkthrough with Trainer Prompts

## ✅ Step 1: Recognize the Label

🧩 **Action:** We scan the dataset and notice the `Client` column contains values like **YES** and **NO**.

🎙️ **Trainer Prompt:**

"Take a look at this column – 'Client'. What kind of values does it hold?"

"Exactly—YES or NO. That tells us whether a lead converted into a client or not. Since we already know the outcome, is this a labeled or unlabeled dataset?"

✍️ **Learner Answer Expected:**
**Labeled dataset** → used in **supervised learning**

## ✅ Step 2: Understand the Goal

🧩 **Action:** Clarify what we want to predict using the data.

🎙️ **Trainer Prompt:**

"Now imagine you're working in a sales team. Wouldn't it be useful to know which leads are likely to convert before spending time on all of them?"

"Your goal is to use this dataset to train a model that predicts whether a new lead will become a client based on their profile and behavior."

💡 **Trainer Tip:**
Highlight real-world applications like **lead scoring**, **customer targeting**, and **conversion optimization**.

## ✅ Step 3: Choose the Algorithm

🧩 **Action:** Match the problem type (binary classification) to a suitable ML algorithm.

🎙️ **Trainer Prompt:**

"Given that our prediction is YES or NO, what type of ML task are we performing?"

"Correct – a **classification task**. Now, can anyone name some algorithms used for classification?"

🧠 **Common Responses You Can Guide Toward:**

- Logistic Regression (simple and interpretable)
- Decision Tree (good with categorical + numerical data)
- Random Forest (more accurate, but complex)
- Support Vector Machine (if mentioned by advanced learners)

🎙️ Follow-up:

"Let's say we start with **Logistic Regression** to keep things simple and interpretable."

## ✅ Step 4: Evaluate the Model

🧩 **Action:** Assess how well the model performs using evaluation metrics.

🎙️ **Trainer Prompt:**

"Once we build our model, we need to test it. Would it make sense to use the same data we trained on?"

"Right – we need a **test set**. That's why we split our data into **training and testing** parts."

📊 **Introduce Metrics:**

"Now, here are a few ways we measure model success:"

- **Accuracy:** Overall, how often is the model correct?
- **Precision:** Out of all predicted YES, how many were actually YES?
- **Recall:** Out of all real YES, how many did the model catch?

📊 **Trainer Add-on:**

"If we care about not missing high-potential clients, **recall** might be more important than precision."

# 🟢 Conclusion Prompt:

🎙️ **Wrap-Up Prompt:**

"Now that we've completed all four steps, what would you say are the key things you look for before building a supervised learning model?"

✅ **Learner Takeaways Expected:**

- Identify a label column
- Understand what outcome to predict
- Choose a classification algorithm
-

## 🎯 Key Takeaway for Learners:

"This is like a teacher showing students the answers during practice. The goal is to help the machine learn from known outcomes so it can predict future ones."

## Unsupervised Learning

Below table shows an unlabeled dataset, which visually mirrors the structure of your labeled example — but with no 'Client' column (label).

You have lead data but no outcome labels.

**Unlabeled Dataset: Lead Information Without Outcome**

| Lead Location | Gender | Lead Company Size | Lead Revenues | Number of Emails |
|---|---|---|---|---|
| UK | F | 5 | 500000 | 3 |
| Ireland | F | 6 | 600000 | 2 |
| UK | M | 6 | 600000 | 5 |
| USA | F | 10 | 1000000 | 12 |
| USA | F | 3 | 300000 | 4 |
| USA | M | 5 | 500000 | 7 |
| Germany | M | 6 | 600000 | 1 |

## 👩‍🏫 Trainer Walkthrough with Prompts

## ✅ Step 1: Load the Unlabeled Dataset

**🧩 Action:**
Present a dataset that looks just like the labeled one, **but without the 'Client' column**.

**🎙️ Trainer Prompt:**

"Let's take a look at this dataset. What do you notice compared to the previous one?"

"Right—there's **no 'Client' column**. That means we don't know if these leads converted or not."

"Can we still use machine learning with this kind of data?"

**✅ Answer You're Guiding Toward:**
Yes, we can use **unsupervised learning** to explore **patterns**.

## ✅ Step 2: Understand the Features

**🧩 Action:**
Examine the input columns available.

**🎙️ Trainer Prompt:**

"Although we don't have labels, we do have several features:"

- Lead Location
- Gender
- Company Size

- Revenue
- Number of Emails

"What kind of information might we extract from these? Can we group similar leads?"

💡 **Trainer Insight:**
Explain that features provide **clues for similarity**, and clustering helps **group data points based on those clues**.

## ✅ Step 3: Apply Clustering Algorithm

🧩 **Action:**
Explain the purpose of clustering using K-Means or similar.

🎙️ **Trainer Prompt:**

"Let's say we want to divide these leads into 3 groups — without knowing anything about their outcomes."

"A clustering algorithm like **K-Means** can do this. It works by:"

- Measuring how similar data points are
- Grouping them into clusters
- Assigning each lead to one cluster ID (like Group 0, Group 1, Group 2)"

🧠 **Example Explanation:**

"So even if we don't know who converted, we might discover a group of high-revenue, highly engaged leads—useful for targeting!"

## ✅ Step 4: Visualize the Clusters

🧩 **Action:**
Plot a 2D graph of features like **Company Size vs. Number of Emails** to show grouping.

🎙️ **Trainer Prompt:**

"Once our algorithm groups the data, we can visualize it. Let's say we use two features — Company Size and Number of Emails."

"Each cluster will show up as a group on this plot. What might we expect to see?"

"Maybe one cluster has small companies with low email volume, while another has big companies that interact a lot."

📊 **Trainer Insight:**
Show how **patterns emerge visually**, even without labeled data — which can support strategies like segmentation or personalization.

# 🧠 Trainer Wrap-Up Prompt:

🎙️ **Final Reflection:**

"So even though we don't have an outcome column like 'Client', what did we still learn from this data?"

"Can we use these clusters to tailor campaigns or prioritize follow-ups?"

✅ **Expected Learner Insights:**

- You can find meaningful patterns without knowing the outcomes
- Clustering is about **discovery**, not prediction
- Useful for **grouping users, products, or behaviors** in business

## ☑️ Learning Outcome:

Students will:
- Understand the role of unlabeled data in unsupervised learning
- Learn how to apply clustering to reveal hidden patterns
- Differentiate between labeled and unlabeled data through hands-on analysis

## 📊 Comparison Table for Classroom Use:

| **Aspect** | **Supervised Learning** | **Unsupervised Learning** |
|---|---|---|
| Type of Data | Labeled (includes outcomes) | Unlabeled (no outcome column) |
| Goal | Predict future outcomes | Find hidden patterns or groupings |
| Output | Classification or Regression | Clusters or reduced features |
| Real-World Example | Predict customer conversion | Group similar customers by behavior |

| | | |
|---|---|---|
| Key Algorithm Types | Logistic Regression, Decision Tree, etc. | K-Means, DBSCAN, PCA |
| Use Case from Demo | "Will this lead become a client?" | "Which leads behave similarly?" |


**Contact**

252 West 37th St., Suite 1200W
New York, NY 10018