

Types of Data in Machine Learning

In Machine Learning, data comes in various **formats** and **types**, and the way we process it depends on the format.

1 CSV (Comma-Separated Values)

- **What is it?**

CSV is one of the **most common and simple formats** for storing tabular data, where values are separated by commas.

- **Why is it useful for ML?**

CSV files are **easy to load into ML frameworks** and **ideal for structured data** (rows and columns).

It is usually used for **datasets** like customer info, sales data, and more.

A	B	C	D
Gender	Grade in 12	is_bachlour_pass	Job
Male	3.6	Yes	Yes
Male	3.54	No	Yes
Male	3	Yes	No
Female	2.57	No	Yes
Female	3.65	Yes	No
Female	3.23	No	Yes

JSON (JavaScript Object Notation)

- **What is it?**

JSON is a lightweight format often used for **storing data** in a **nested structure**, making it great for **hierarchical data** like **user profiles** or **API responses**.

- **Why is it useful for ML?**

It's perfect for **storing unstructured data**, especially when the data is more **complex** or **has nested relationships**.

```
json

{
  "name": "John",
  "age": 30,
  "address": {
    "city": "New York",
    "state": "NY"
  },
  "purchases": ["Laptop", "Phone"]
}
```

SQL (Structured Query Language) 📦

- **What is it?**
SQL is used to query relational databases. It's ideal for retrieving large, structured datasets from a database.
- **Why is it useful for ML?**
SQL allows you to query and retrieve data directly from databases like MySQL or PostgreSQL, which is helpful for large-scale projects.

```
sql

SELECT name, age, purchase_history FROM customers WHERE age > 25;
```

Images 📷 (Pixel Data) 🖼️

- **What is it?**
Images are a form of unstructured data where each image is made of pixels. Each pixel contains RGB values (Red, Green, Blue) for color representation.
- **Why is it useful for ML?**
Used in Computer Vision tasks like image classification, object detection, etc.

Text 📄

- **What is it?**
Text data is made up of words and sentences. It is unstructured data and requires processing before using it in ML models.
- **Why is it useful for ML?**
Text is used in Natural Language Processing (NLP) tasks like sentiment analysis, language translation, and more.

Time Series Data ⌚

- **What is it?**
Time series data refers to data that is recorded in sequences at specific time intervals. This type of data is ordered and often used to forecast future events.
- **Why is it useful for ML?**
Used in tasks like stock price prediction, weather forecasting, and sales forecasting.

Date	Stock Price (\$)
2023-01-01	100
2023-01-02	105

Audio (Sound Data) 🎧

- **What is it?**
Audio data is made up of sound waves, typically represented as waveforms or spectrograms.
- **Why is it useful for ML?**
Used in speech recognition, audio classification, and music recommendation systems.