Data in Machine Learning (Detailed & Diagram Based)

Introduction

Data is the foundation of Machine Learning. The quality, quantity, and type of data directly influence how well a machine learning model learns and performs. Without data, machine learning systems cannot function.

1. What is Data?

Data refers to raw facts, figures, observations, or measurements collected from various sources. In machine learning, data is used to train models to recognize patterns and make predictions.

Examples: Numbers, text, images, audio, video, sensor readings.

2. Types of Data Based on Structure

1. **Structured Data:** Organized data in tables (rows & columns). Example: Excel sheets.
2. **Semi-Structured Data:** Partially organized data. Example: JSON, XML.
3. **Unstructured Data:** No fixed format. Example: Images, videos, text.

Diagram:

```
Data
 ■■ Structured (Tables)
 ■■ Semi-Structured (JSON, XML)
 ■■ Unstructured (Text, Images, Video)
```

3. Types of Data Based on Nature

1. **Numerical Data:** Integer and floating values. Example: Age, Salary.
2. **Categorical Data:** Category-based values. Example: Gender, City.

Diagram:

```
Data
 ■■ Numerical → Discrete / Continuous
 ■■ Categorical → Nominal / Ordinal
```

4. Training Data, Validation Data & Test Data

Datasets are usually divided into three parts to build and evaluate models correctly.

1. **Training Data:** Used to train the model.
2. **Validation Data:** Used for tuning parameters.
3. **Test Data:** Used for final evaluation.

Diagram:

```
Dataset
```

```
■■ Training Set (70%)
■■ Validation Set (15%)
■■ Test Set (15%)
```

5. Labeled vs Unlabeled Data

1. **Labeled Data:** Data with correct output labels.
2. **Unlabeled Data:** Data without output labels.

Diagram:

```
Labeled Data   → Input + Output
Unlabeled Data → Only Input
```

6. Data Quality in Machine Learning

Good quality data improves model accuracy and reliability. Poor data quality leads to incorrect predictions.

1. Accuracy
2. Completeness
3. Consistency
4. Timeliness

7. Real-Life Example

In house price prediction, data may include features such as location, size, number of rooms, and age of the house. Clean and relevant data helps the model predict prices accurately.

Summary

Data plays a crucial role in machine learning. Understanding data types, structure, and quality is essential for building effective machine learning models.