

Data Preprocessing

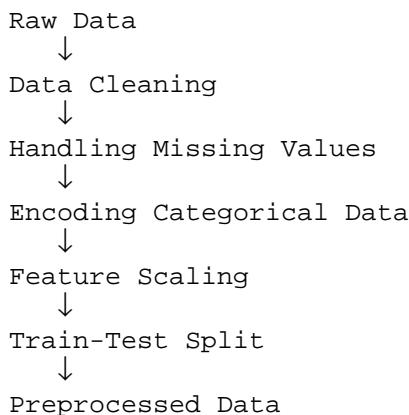
Introduction

Data preprocessing is one of the most important steps in machine learning. Real-world data is often incomplete, noisy, and inconsistent. This step converts raw data into a clean and usable format so that machine learning models can perform accurately.

Why Data Preprocessing is Required

- 1 Improves model accuracy
- 2 Removes noise and inconsistencies
- 3 Handles missing and incorrect data
- 4 Makes data suitable for ML algorithms

Data Preprocessing Workflow



Data Cleaning

- 1 Removing duplicate data
- 2 Fixing incorrect values
- 3 Removing irrelevant information

Handling Missing Values

- 1 Remove rows or columns
- 2 Replace with mean, median, or mode
- 3 Predict missing values

Missing Data
■■■ Remove
■■■ Replace
■■■ Predict

Encoding Categorical Data

- 1 Label Encoding
- 2 One-Hot Encoding

3 Ordinal Encoding

Color → Red:0, Blue:1, Green:2

Feature Scaling

1 Normalization

2 Standardization

Before: [1, 100, 1000]

After: [0.01, 0.5, 1.0]

Handling Outliers

1 Detect using IQR method

2 Detect using Z-score

3 Remove or cap outliers

Train-Test Split

Dataset

■■■ Training Set (80%)

■■■ Test Set (20%)

Summary

Data preprocessing improves data quality and ensures better model performance. It is a critical step in every machine learning project.