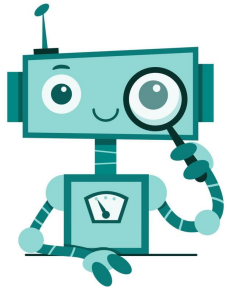
Artificial Intelligence (AI)



Week-10

Agenda

- What is a Dataset
- Types of data
- Data Preprocessing

What is a Dataset

A dataset is a collection of data that has been organized into a structured form, usually as a table of rows and columns. Each row represents an individual record, and each column represents a particular variable or attribute of the record. Datasets are the foundation of data analysis, statistics, and machine learning models.

Characteristics of a Good Dataset:

- **Relevant:** Contains data pertinent to the problem or analysis at hand.
- **Comprehensive:** Includes all necessary variables and observations.
- **Accurate:** Free from errors and inconsistencies.
- **Current:** Up-to-date with recent information.
- **Accessible:** Available in a format that can be easily used and processed.

Sources for Downloading Datasets:

UCI Machine Learning Repository: A collection of databases, domain theories, and data generators widely used by the machine learning community.

- **Kaggle:** Offers a variety of datasets along with competitions and notebooks.
- **Google Dataset Search:** A tool that enables the discovery of datasets stored across the web.
- **Government Databases:** Many governments provide open access to a range of datasets (e.g., data.gov, data.gov.uk).
- **AWS Public Data Sets:** Large datasets available on Amazon's cloud services.
- **Academic Databases:** Repositories hosted by universities and

Types of Data

- Structured Data
- Unstructured Data
- Semi Structured Data

Structured Data

- Structured data is organized and follows a **predefined format**, usually stored in **databases** or **spreadsheets**.
- **Example:** A customer database with columns for name, age, email, and purchase history.

  - Ordinal data
  - Nominal data
  - Numerical data

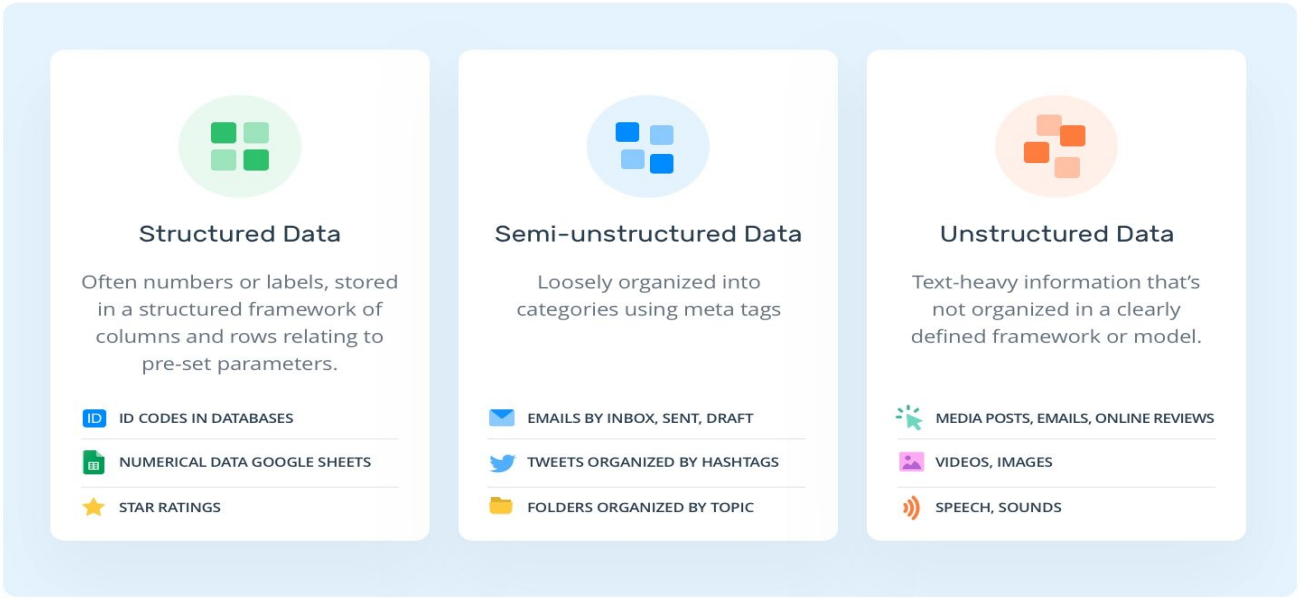
Unstructured Data

- Unstructured data is **not organized** and lacks a predefined format, often in the form of **text, images, audio, or video**.
- **Example:** Social media posts, customer reviews, or images from a surveillance camera.

Semi-Structured Data

- Semi-structured data has some organization but does not adhere to a **strict schema**, often **containing tags** or **labels.**
- **Example:** Emails, JSON files that contain data with tags or key-value pairs.

Unstructured vs Structured Data



What is Data Preprocessing?

Data preprocessing is the process of **cleaning**, **transforming**, and **organizing raw** data to make it suitable for analysis and machine learning models.

Importance: High-quality data preprocessing is crucial for accurate and meaningful insights.

Common Steps in Data Preprocessing:

- Importing the Data
- Handling Missing Data
- Handling Duplicate Data
- Handling Outliers
- Encoding Categorical Variables
- Scaling and Normalization