
1010101010
0101010101
1010101010
0101010101
Notes
DATA STRUCTURES
ALGORITHMS
COMPUTER NETWORKS
DATABASE SYSTEMS
Good Notes
Better Understanding
Better Results

Unit-I


1010101010
0101010101
1010101010
0101010101
10110
01001
10101
DATA STRUCTURES
ALGORITHMS
COMPUTER NETWORKS
DATABASE SYSTEMS
#include <iostream>
using namespace std;
int main()
{
cout << "Hello, World!";
return 0;
}

Introduction to Data Science

* **Definition :**
  Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

* **Purpose :**
  To analyze data and help in decision-making by discovering useful patterns, trends and relationships.

* **Key Components :**
  1. Data Collection
  2. Data Processing
  3. Data Analysis
  4. Data Visualization
  5. Machine Learning
  6. Communication of Results

* **Applications :**
  Healthcare, Finance, E-commerce, Social Media, Marketing, Education, Sports, and many more.

* **Importance :**
  In today's world, data is everywhere. Data Science helps organizations make better decisions, improve efficiency and solve complex problems.

Definition and Importance of Data Science

1. Definition of Data Science :

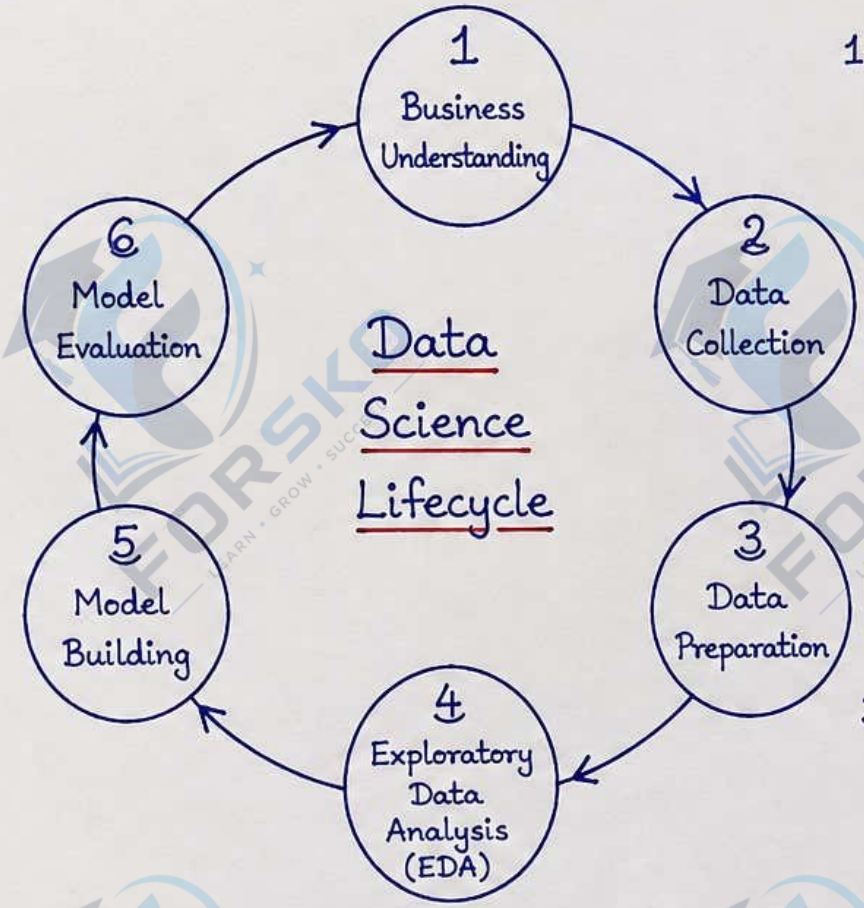
Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It combines statistics, programming, domain knowledge and data visualization to analyze data and support decision-making.

2. Importance of Data Science :

① Helps in making data-driven decisions based on facts, not guesses.
② Identifies hidden patterns, trends and correlations in large amounts of data.
③ Improves efficiency and reduces costs by optimizing processes.
④ Enables organizations to understand customers better and improve services.
⑤ Supports innovation by predicting future trends and outcomes.
⑥ Useful in many fields like healthcare, finance, marketing, education, e-commerce, sports and more.
⑦ Helps in solving complex problems and improving overall business performance.

Data Science Lifecycle

The Data Science Lifecycle is a series of iterative steps that helps in solving a problem using data and creating value from it.



1. **Business Understanding :**
   Define the problem, understand business objectives and identify the key questions to be answered.

2. **Data Collection :**
   Gather relevant data from various sources required to solve the problem.

3. **Data Preparation :**
   Clean, transform and prepare the data by handling missing values, removing duplicates, and formatting it for analysis.

4. **Exploratory Data Analysis (EDA) :**
   Explore and analyze the data to find patterns, trends, relationships and insights.

5. **Model Building :**
   Apply suitable machine learning algorithms to build the model using prepared data.

6. **Model Evaluation :**
   Evaluate the model performance using appropriate metrics and check if it meets the business objectives.

Applications of Data Science

Data Science is used in many real-world areas to solve problems, improve decision making and create value.

① **Healthcare** : Used to predict disease, analyze medical images, personalize treatment, and improve patient care.
Example – Predicting heart disease, drug discovery.

② **Finance** : Used for fraud detection, risk analysis, credit scoring, algorithmic trading, and customer segmentation.
Example – Credit card fraud detection, stock market prediction.

③ **Marketing & Sales** : Helps in understanding customer behavior, target marketing, recommendation systems, and sales forecasting.
Example – Product recommendation on e-commerce platforms.

④ **E-commerce** : Used for demand forecasting, dynamic pricing, customer segmentation, and inventory management.
Example – Predicting which products will be in high demand.

⑤ **Education** : Helps in student performance analysis, personalized learning, and dropout prediction.
Example – Identifying weak students and suggesting improvements.

⑥ **Transportation** : Used in route optimization, traffic prediction, and autonomous vehicles.
Example – Google Maps traffic prediction, Uber ride matching.

⑦ **Entertainment** : Used in movie/series recommendations, content personalization, and audience analysis.
Example – Netflix recommendation system, YouTube suggestions.

⑧ **Agriculture** : Helps in crop prediction, soil analysis, pest detection, and yield optimization.
Example – Predifying weather impact on crops, smart famhing.

Data Science is everywhere — from our daily lives to critical industries, making processes smarter, faster and more efficient.

Types of Data

In Data Science, data can be classified into different types based on their nature, structure and source.

1. Based on Structure

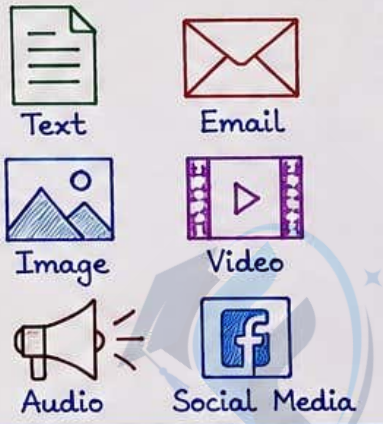
a) Structured Data :

- Highly organized and follows a fixed format.
- Easy to store, search and analyze.
- Example – Data in tables (Rows and Columns) in SQL databases, spreadsheets.

ID	Name	Age	City
1	Riya	25	Pune
2	Aman	30	Mumbai
3	Neha	28	Delhi
...	...	...	...

b) Unstructured Data :

- Does not follow any fixed structure.
- Difficult to store and analyze using traditional methods.
- Example – Text documents, emails, images, videos, social media posts, audio files.



2. Based on Source

a) Internal Data :

- Generated within the organization.
- Example – Sales records, customer details, employee data.

b) External Data :

- Collected from outside sources.
- Example – Government data, market reports, social media, weather data.

3. Based on Nature

a) Quantitative Data (Numerical) :

- Represents numbers and can be measured.
- Used for statistical analysis.
- Example – Age, salary, temperature, marks, sales figures.

b) Qualitative Data (Categorical) :

- Represents categories or labels.
- Describes qualities / attributes.
- Example – Gender, color, feedback, product type.

In short : Data can be Structured or Unstructured, Internal or External, Quantitative or Qualitative depending on its nature and source.

Data Science Tools and Platforms

Data Science makes use of various tools and platforms that help in collecting, processing, analyzing, visualizing and building machine learning models from data.

1. Popular Data Science Tools :

Tool	Type	Key Features / Uses
Python	Programming Language	• Most popular tool for Data Science. • Rich libraries (NumPy, Pandas, Matplotlib, Scikit-learn).
R	Programming Language	• Widely used for statistical analysis and visualization.
Jupyter Notebook	Development Environment	• Interactive coding platform. • Easy to write and share data science projects.
Apache Spark	Big Data Processing	• Handles large scale data. • Fast and powerful in-memory data processing.
Tableau	Data Visualization	• Creates interactive and beautiful dashboards.
Microsoft Excel	Data Analysis	• Useful for basic data cleaning, analysis and visualization.

2. Data Science Platforms :

Google Colab
- Cloud based Jupyter Notebook.
- Free to use with GPU/TPU support.
- No setup required.

Kaggle
- Online platform for data science competitions.
- Provides free datasets and notebooks.
- Great for learning and practice.

Amazon Web Services (AWS)
- Cloud platform for building and deploying data science solutions.
- Services: SageMaker, Redshift etc.

Microsoft Azure
- Cloud platform by Microsoft.
- Provides tools for machine learning and big data analytics.

IBM Watson
- AI and data science platform.
- Offers machine learning and natural language processing tools.

In short : Data Science tools and platforms provide the necessary environment to work with data, build models and turn data into useful insights.

Introduction to Big Data and AI

1. What is Big Data?

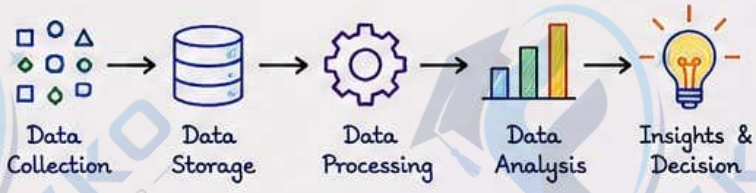
Big Data refers to extremely large and complex datasets that traditional data processing applications are not able to handle efficiently. It involves the collection, storage, management and analysis of massive amounts of data to extract valuable insights.

Importance of Big Data

- Helps organizations make data-driven decisions.
- Improves customer experience and services.
- Enables prediction of future trends and patterns.
- Supports innovation and new business opportunities.

Key Characteristics (5 Vs):

- Volume – Very large amount of data
- Velocity – High speed of data generation and processing
- Variety – Different types and formats of data
- Veracity – Data quality and reliability
- Value – Extracting meaningful insights from data



Sources of Big Data

- Social media
- Sensors and IoT devices
- Logs and clicks (web/app)
- Transactional data
- Images, videos, audio, etc.

5 Vs of Big Data

Volume	–	How much data
Velocity	–	How fast data is generated
Variety	–	Different types of data
Veracity	–	How accurate and reliable the data is
Value	–	How useful the data is

2. What is Artificial Intelligence (AI)?

Artificial Intelligence (AI) is the simulation of human intelligence in machines that are programmed to think, learn and act like humans. It enables machines to perform tasks such as learning, reasoning, problem solving, perception and decision making.

Importance of AI

- Automates repetitive and complex tasks.
- Improves accuracy and reduces human error.
- Enhances productivity and efficiency.
- Enables intelligent decision making.

Types of AI

- Narrow AI (Weak AI) – Designed to perform a specific task (e.g., voice assistant, recommendation system).
- General AI (Strong AI) – Machines with human-like intelligence that can perform any intellectual task. (Not yet achieved)
- Super AI – AI that is far smarter than humans. (Hypothetical)

Key Technologies in AI

- Machine Learning
- Deep Learning
- Natural Language Processing (NLP)
- Computer Vision
- Robotics
- Expert Systems

Applications of AI

- Virtual assistants (Siri, Alexa)
- Recommendation systems (Netflix, YouTube)
- Spam detection in email
- Self-driving cars
- Healthcare (disease prediction, medical imaging)
- Chatbots and customer support
- Fraud detection in banking
- Robotics in manufacturing

Big Data provides the raw material (data) and AI provides the intelligence to analyze that data and convert it into useful insights and actions. Together, they power the modern digital world.


1010101010
0101010101
1010101010
0101010101
Unit -II
10110
01001
10101
DATA STRUCTURES
ALGORITHMS
COMPUTER NETWORKS
DATABASE SYSTEMS
#include <iostream>
using namespace std;
int main()
{
cout << "Hello, World!";
return 0;
}

Data Collection and Preprocessing:

1. Data Collection

Data collection is the process of gathering raw data from various sources relevant to the problem or research.

Sources of Data:

- Databases (SQL, NoSQL)
- Web scraping
- APIs
- Surveys and questionnaires
- Sensors and IoT devices
- Social media
- Files (CSV, Excel, JSON, etc.)

2. Data Preprocessing

Data preprocessing is the process of cleaning and transforming raw data into a suitable format for analysis and modeling.

Key Objectives:

- Improve data quality
- Handle missing or inconsistent data
- Remove noise and outliers
- Transform data into a consistent format
- Prepare data for machine learning algorithms

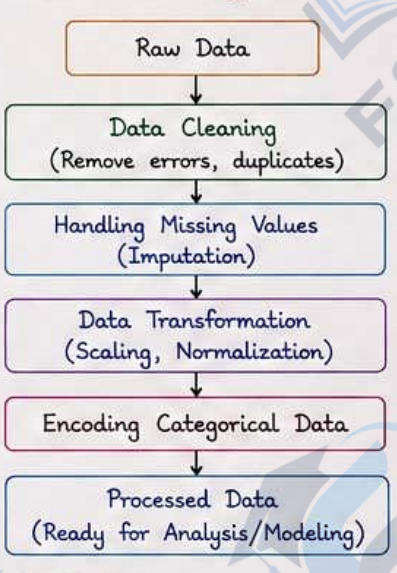
Common Preprocessing Tasks:

- Data Cleaning
- Handling Missing Values
- Removing Duplicates
- Outlier Detection and Removal
- Data Transformation
- Feature Scaling
- Encoding Categorical Data
- Data Integration

3. Steps in Data Preprocessing

Step	Description	Example	Before	After
① Data Cleaning	Handle errors, missing values, inconsistent data.	Fill missing age, remove duplicates.	ID | Age | Salary 1 | 25 | 50000 2 | — | 60000 2 | 30 | 60000	ID | Age | Salary 1 | 25 | 50000 2 | 30 | 60000 2 | 30 | 60000
② Handling Missing Values	Fill or impute missing data using mean, median, mode, or constant values.	Fill missing age with mean (28).	Age: 25, —, 30, 27	Age: 25, 28, 30, 27
③ Outlier Detection	Detect and remove or treat abnormal values.	Remove salary value 999999.	Salary: 45000, 50000, 999999, 52000	Salary: 45000, 50000, 52000
④ Data Transformation	Convert data into suitable format or scale.	Normalize salary using Min-Max scaling.	Salary: 45000, 50000, 52000	Salary: 0.00, 0.50, 1.00
⑤ Encoding Categorical Data	Convert categorical variables into numeric format.	Gender (M/F) → 0 (M), 1 (F)	Gender: M, F, M, F	Gender: 0, 1, 0, 1

4. Preprocessing Workflow



Note: Good quality data leads to accurate analysis and better results.

Data Collection Methods

Data collection methods are the techniques or approaches used to gather raw data from various sources for analysis and decision making.

✱ Common Data Collection Methods:

Method	Description	Examples / Sources
1. Surveys 	• Collect data by asking a set of questions to a group of people. • Can be conducted online, on paper, or by phone.	• Google Forms, Online Polls • Customer satisfaction surveys • Opinion polls
2. Interviews 	• Collect data by talking to individuals and asking questions. • Can be structured, semi-structured or unstructured.	• Job interviews • Expert interviews • User research interviews
3. Observations 	• Collect data by watching and recording behaviour or events as they happen. • Can be participant or non-participant.	• Observing customer behavior in a store • Traffic flow observation • Classroom observations
4. Experiments 	• Collect data by conducting controlled experiments and measuring results. • Helps in finding cause-and-effect relationships.	• A/B testing • Product testing • Scientific experiments
5. Documents and Records 	• Collect data from existing documents, reports, databases or records. • Cost-effective and saves time.	• Sales reports, annual reports • Government data • Academic research papers
6. Web Scraping 	• Collect data from websites using software tools or scripts. • Useful for large scale data.	• Price comparison websites • News websites • Social media data
7. Sensors and IoT Devices 	• Collect real-time data using sensors and IoT devices. • Useful for monitoring and tracking.	• Smartwatches, fitness bands • Weather stations • Smart home devices

Key Points:

- Choose the method based on the objective, data type, time, cost and accuracy required.
- Good data collection ensures reliable and meaningful analysis.

Factors Affecting Data Collection:

- Objective of the study
- Type of data required
- Time and budget
- Availability of resources
- Accuracy and reliability

Structured and Unstructured Data

Data can be broadly classified into two types based on its organization and format.

1. Structured Data

Structured data is well-organized, formatted and stored in a predefined manner, usually in rows and columns. It is easy to search, analyze and process.

Characteristics:

- Organized in tables (rows & columns)
- Follows a fixed schema
- Easy to store and retrieve
- Suitable for quantitative analysis
- **Examples:**
  - Customer details in a bank database
  - Sales records
  - Student information (roll no., name, marks)

Example (Table):

ID	Name	Age	City	Marks
101	Rahul	20	Pune	85
102	Aisha	21	Mumbai	90
103	Karan	19	Nashik	78

2. Unstructured Data

Unstructured data does not have a predefined format or organization. It is more complex and difficult to store, search and analyze using traditional methods.

Charructeristics:

- No fixed format or structure
- Does not follow a schema
- Difficult to analyze
- Suitable for qualitative analysis
- **Examples:**
  - Social media posts
  - Images and videos
  - Emails, text documents
  - Audio recordings
  - PDF files, presentations, etc.

Examples (Unstructured Data):

In short: Structured data is organized and easy to handle. | Unstructured data is not organized and requires advanced techniques to extract useful information.

Data Cleaning and Handling Missing Values

Data cleaning is the process of detecting and correcting errors or inconsistencies in data to improve its quality and reliability.

1. Data Cleaning

It involves identifying and fixing problems such as :

- Missing values
- Duplicate records
- Inconsistent data
- Incorrect data types
- Outliers and noise

Example :

Consider the following dataset.

ID	Name	Age	City	Salary
1	Asha	25	Pune	25000
2	Rohit	—	Mumbai	30000
3	Neha	30	—	28000
4	Vijay	28	Nashik	—
5	—	27	Pune	26000

After Cleaning and Handling Missing Values :

ID	Name	Age	City	Salary
1	Asha	25.0	Pune	25000
2	Rohit	27.5	Mumbai	30000
3	Neha	30.0	Pune	28000
4	Vijay	28.0	Nashik	28000
5	Unknown	27.0	Pune	26000

2. Handling Missing Values

Missing values occur when no data is recorded for a variable.

They can be handled in the following ways :

① **Delete records** : Remove rows that contain missing values. (Use when missing data is less)

② **Delete columns** : Remove columns with too many missing values.

③ **Mean / Median imputation** : Replace missing values with the mean or median of the column.

④ **Mode imputation** : Replace with the most frequent value (for categorical data).

⑤ **Forward fill / Backward fill** : Use the previous or next valid value to fill the missing data (useful in time series data).

⑥ **Interpolation** : Estimate missing values using other values (used in numerical/time series data).

Key Points :

- Always understand the nature of missing data.
- Choose the appropriate method based on the data and problem.
- Proper handling improves data quality and the performance of models.

In short : Data cleaning removes errors and inconsistencies, and handling missing values ensures complete, accurate and reliable data for analysis.

Data Transformation and Normalization

Data transformation is the process of converting data from one form or scale to another to make it suitable for analysis and modeling. Normalization is a type of transformation that scales numerical data to a standard range.

1. Data Transformation

Data transformation includes techniques that change the structure, scale or distribution of data.

Common Transformation Techniques :

- Aggregation : Summarizing data (e.g., daily → monthly sales)
- Generalization : Replacing specific values with general categories (e.g., age 25 → 20-30)
- Encoding : Converting categorical data to numerical form (e.g., Male → 1, Female → 0)
- Log Transformation : Reduces skewness in data ($x' = \log(x)$)
- Feature Construction : Creating new meaningful features from existing ones (e.g., BMI = $\frac{\text{weight}}{\text{height}^2}$)

Example – Transformation (Log Transformation)

Original Data (x)		Log Transformation $x' = \log_{10}(x)$
10	→	1.0000
100	→	2.0000
1000	→	3.0000
10000	→	4.0000

2. Normalization

Normalization scales numerical data to a standard range so that all features contribute equally to the analysis.

Common Normalization Techniques :

(a) Min-Max Normalization (Rescaling)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Scales data to range [0,1]

Example :

Original Data (x)	Calculation	Normalized Data (x')
10	(10-10)/(50-10) = 0/40	0.00
20	(20-10)/(50-10) = 10/40	0.25
30	(30-10)/(50-10) = 20/40	0.50
40	(40-10)/(50-10) = 30/40	0.75
50	(50-10)/(50-10) = 40/40	1.00

(b) Z-Score Normalization (Standardization)

$$z = \frac{x - \mu}{\sigma}$$

Transforms data to have mean = 0 and standard deviation = 1

Where, μ = mean of data
σ = standard deviation

In short : Data transformation changes the form or structure of data to make it useful. Normalization scales numerical data to a standard range for fair comparison and better model performance.

Data Integration and Reduction

Data integration combines data from multiple sources into a consistent format, while data reduction reduces the volume of data while retaining important information.

1. Data Integration

Data integration is the process of combining data from multiple sources the provide a unified view.

Why Data Integration?

- Data comes from different sources and formats.
- Integration removes redundancy and inconsistency.
- Provides accurate and comprehensive information for analysis and decision making.

Methods of Data Integration

1. Schema Integration: Resolving differences in naming conventions and data formats.
2. Entity Identification: Identifying the same real-world entity across different sources.
3. Redundancy Detection: Eliminating duplicate data.
4. Data Value Conflict Resolution: Handling inconsistent data values.

Example: Consider two data sources.

Source 1 (Sales_Online)

ID	Name	Product	Amount
1	Aisha	Mobile	15000
2	Rohit	Laptop	35000
3	Neha	Tablet	12000

→

Source 2 (Sales_Store)

Cust_ID	Customer	Item	Amt
2	Rohit	Laptop	36000
3	Neha	Tablet	12000
4	Vijay	Mobile	18000

Integrated Data

Customer_ID	Customer_Name	Product	Amount
1	Aisha	Mobile	15000
2	Rohit	Laptop	35500*
3	Neha	Tablet	12000
4	Vijay	Mobile	18000

* Conflict resolved by taking average of 35000 and 36000.

2. Data Reduction

Data reduction reduces the size of data while preserving important information.

Why Data Reduction?

- Reduces storage space and computational cost.
- Improves efficiency of data processing.
- Helps in focusing on relevant information.

Techniques of Data Reduction

1. Data Cube Aggregation: Summarizing data by aggregating values (e.g., sum, avg, count).
2. Dimensionality Reduction: Reducing number of attributes (e.g., PCA – Principal Component Analysis).
3. Data Compression: Encoding data to reduce size without losing information.
4. Discretization: Converting continuous data into fewer intervals or categories.
5. Sampling: Using a subset of data to represent the whole.

Example:

Original Sales Data

Date	Product	Region	Sales
01-05-24	Mobile	North	12000
02-05-24	Mobile	North	13000
03-05-24	Laptop	South	35000
04-05-24	Laptop	South	34000
05-05-24	Tablet	East	11000
06-05-24	Tablet	East	10000

→

Reduced Data (Aggregated)

Product	Region	Total_Sales
Mobile	North	25000
Laptop	South	69000
Tablet	East	21000

In short: Data Integration combines data from multiple sources into a consistent and unified format. Data Reduction reduces the volume of data while keeping the most useful information.

Introduction to Python for Data Science

Python is a high-level, interpreted, general-purpose programming language. It is simple, readable and has a rich ecosystem of libraries that make it ideal for Data Science.

1. Why Python for Data Science?

- Easy to learn and use
- Large collection of libraries
- Supports data analysis, visualization, machine learning and more
- Open source and community support
- Works on multiple platforms

2. Features of Python

- Simple and readable syntax
- Interpreted language
- Dynamically typed
- Object-oriented programming support
- Extensive standard library

3. Basic Syntax

```
# This is a comment
# Printing output
print ("Hello, Data Science!")
# Variables
x = 10
y = 3.5
name = "Python"
# Basic operations
sum = x + y
diff = x - y
prod = x * y
div = x / y
```

4. Data Structures in Python

- List : `[1, 2, 3, 4]`
- Tuple : `(1, 2, 3, 4)`
- Dictionary : `{'name': 'Aisha', 'age': 21}`
- Set : `{1, 2, 3, 4}`
- String : `"Data Science"`

5. Useful Built-in Functions

- `print()` → Display output
- `type()` → Check data type
- `len()` → Get length/size
- `range()` → Generate sequence of numbers
- `sum()`, `min()`, `max()` → Common operations

6. Popular Libraries for Data Science

Library	Purpose	Example Use
NumPy	Numerical computing	Arrays, matrices, mathematical functions
pandas	Data manipulation & analysis	DataFrames, reading CSV, data cleaning
Matplotlib	Data visualization	Plots, charts, graphs
Seaborn	Statistical visualization	Advanced visualizations
Scikit-learn	Machine learning	Model building, training, evaluation

7. Simple Example

```
import pandas as pd
data = {'Name': ['A', 'B', 'C'],
        'Score': [85, 90, 78]}
df = pd.DataFrame(data)
print(df)
```

In short: Python provides a powerful and easy environment for data analysis, visualization and building machine learning models.

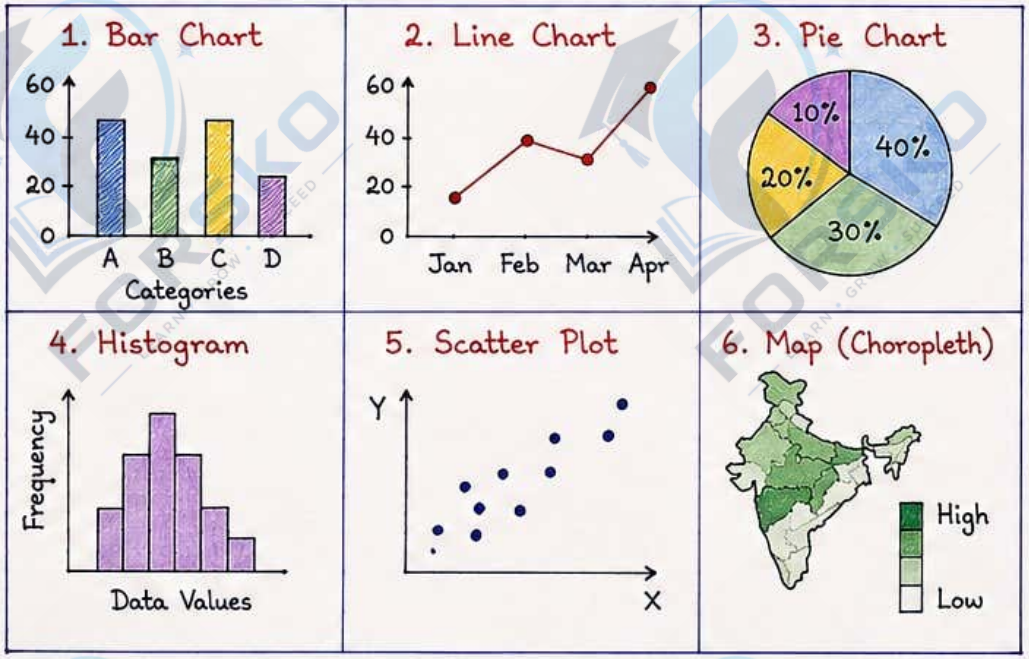

1010101010
0101010101
1010101010
0101010101
Unit - III
10110
01001
10101
DATA STRUCTURES
ALGORITHMS
COMPUTER NETWORKS
DATABASE SYSTEMS
#include <iostream>
using namespace std;
int main()
{
cout << "Hello, World!";
return 0;
}

Data Visualization and Importance of Data Visualization

1. Data Visualization

Data Visualization is the graphical representation of data and information using charts, graphs, plots, maps and dashboards. It helps to communicate complex data in a visual format that is easy to understand.

Common Types of Data Visualization



2. Importance of Data Visualization

Data visualization plays a crucial role in data analysis and decision making. It helps to:

1. **Simplifies Complex Data :** It converts large and complex datasets into simple visual formats that are easy to understand.
2. **Improves Understanding :** Visuals help to quickly grasp patterns, trends and relationships in data that may be difficult to notice in raw data.
3. **Enhances Communication :** Charts and graphs communicate insights clearly and effectively to different audiences.
4. **Supports Better Decision Making :** It provides clear and actionable insights, helping in making informed and data-driven decisions.
5. **Identifies Patterns and Trends :** It helps in discovering hidden patterns, trends, outliers and correlations.
6. **Saves Time :** Visuals allow us to analyze and interpret data faster than reading through tables of numbers.
7. **Increases Impact :** Well-designed visualizations make reports, presentations and dashboards more engaging and impactful.

In short : Data visualization turns data into meaningful pictures. It helps us understand the past, monitor the present and plan for the future. → Better Visuals, Better Insights, Better Decisions.

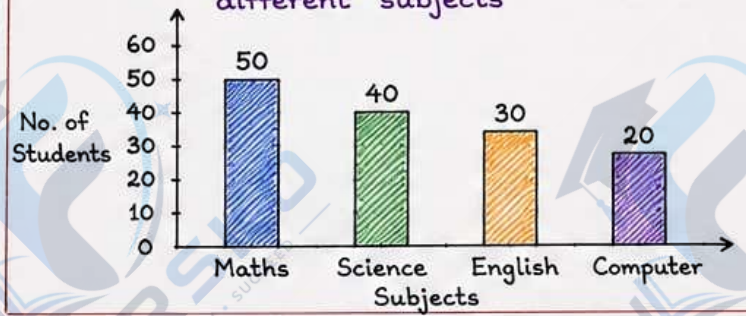
Charts and Graphs: Bar Chart, Pie Chart, Histogram, Scatter Plot

Charts and graphs are visual tools used to represent data in a clear and meaningful way. Different types of charts help to understand data from different perspectives.

1. Bar Chart

Bar chart is used to compare data across different categories.

Example: Number of students in different subjects

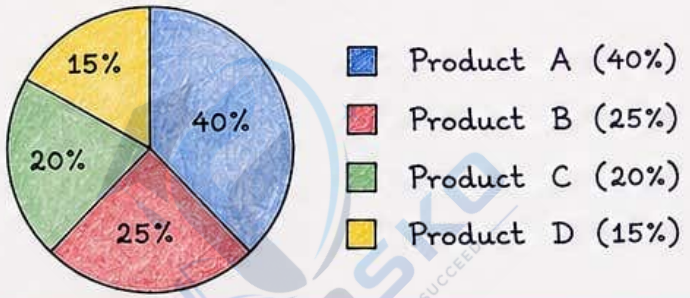


Use:
To compare quantities among different categories.

2. Pie Chart

Pie chart is used to show parts of a whole in percentage.

Example: Percentage of sales by product

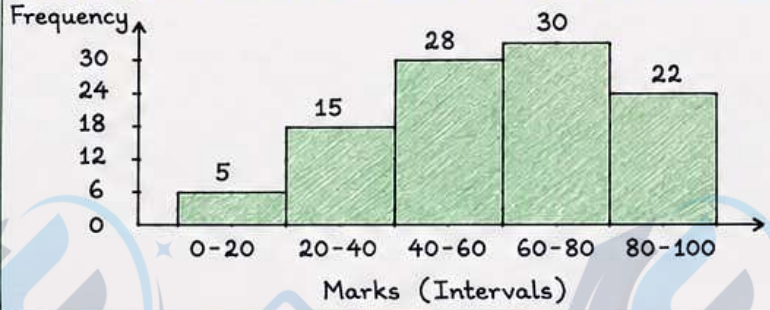


Use:
To show how a whole is divided into parts.

3. Histogram

Histogram is used to show the distribution of continuous data in intervals (bins).

Example: Marks distribution of 100 students

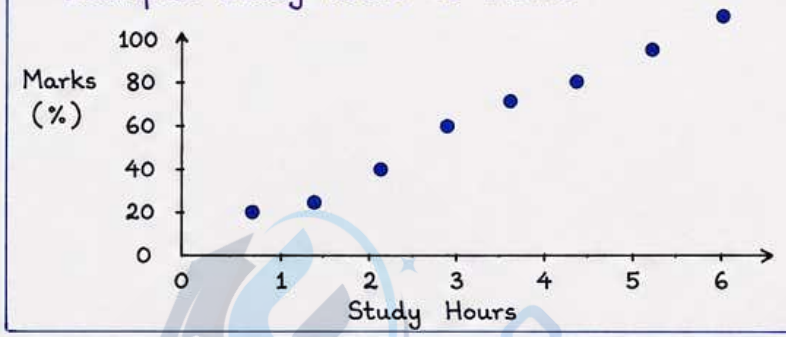


Use:
To show the distribution of data in frequency across ranges.

4. Scatter Plot

Scatter plot is used to show the relationship between two numerical variables.

Example: Study hours vs Marks



Use:
To find the relationship or correlation between two variables.

In short: Bar chart compares categories, Pie chart shows parts of a whole, Histogram shows distribution of continuous data, and Scatter plot shows relationship between two variables.

Statistics : Descriptive Statistics, Mean, Median, Mode

Statistics is the branch of mathematics that deals with collecting, organizing, analyzing, interpreting and presenting data.

1. Descriptive Statistics

Descriptive statistics involves organizing, summarizing and presenting data in a meaningful way. It helps to describe the main features of a dataset.

It includes :

- Measures of central tendency (Mean, Median, Mode)
- Measures of dispersion (Range, Variance, Std. Deviation)
- Graphical representation (Charts, Graphs, Tables)
- Frequency distribution

2. Mean (Arithmetic Mean)

Mean is the average of all observations.

$$\text{Mean } (\bar{x}) = \frac{\text{Sum of all observations}}{\text{Number of observations}}$$

Example : Find the mean of 5, 7, 9, 11, 13.

Solution : Sum = 5 + 7 + 9 + 11 + 13 = 45

Number of observations (n) = 5

$$\text{Mean } (\bar{x}) = \frac{45}{5} = 9$$

3. Median

Median is the middle value of the data when it is arranged in ascending or descending order.

Steps :

1. Arrange the data in ascending order.
2. If n is odd → Median is the middle value.
3. If n is even → Median is the average of the two middle values.

Example : Find the median of 4, 1, 7, 3, 9.

Solution :

Original Data	4	1	7	3	9

↓

Ascending Order	1	3	4	7	9

Here, n = 5 (odd)

Median = 4 (middle value)

4. Mode

Mode is the value that appears most frequently in the data.

Example : Find the mode of 2, 3, 3, 5, 7, 3, 8, 5, 3.

Solution :

Value	2	3	5	7	8
Frequency	1	4	2	1	1

↓

The value 3 occurs most frequently (4 times).

Mode = 3

Example (All Together)

Data : 2, 4, 4, 6, 8, 10

$$\text{Mean} = \frac{2+4+4+6+8+10}{6} = \frac{34}{6} = 5.67$$

$$\text{Median} = \frac{4+6}{2} = 5$$

Mode = 4 (appears most frequently)

Summary

* Descriptive statistics helps in summarizing and describing data.
* Mean gives the average value.
* Median gives the middle value.
* Mode gives the most frequent value.

Correlation and Regression Basics

1. Correlation

Correlation measures the strength and direction of a linear relationship between two numerical variables.

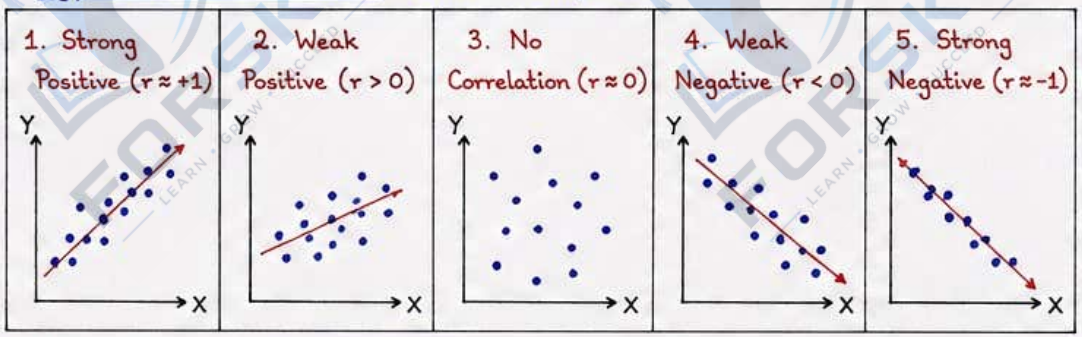
Key Points :

- Value ranges from -1 to $+1$.
- $+1$ → Perfect positive correlation
- -1 → Perfect negative correlation
- 0 → No linear correlation
- It does NOT imply causation.

Correlation Coefficient (r):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\ \sqrt{\sum (y_i - \bar{y})^2}}$$

Types of Correlation (with examples)



Example :

Study hours and exam marks → Positive correlation
Price of a product and demand → Negative correlation
Height and weight → Usually positive correlation

2. Regression

Regression is used to model the relationship between a dependent variable (Y) and one or more independent variable(s) (X) and to predict the value of Y.

Key Points :

- Finds the best fit line (line of regression).
- Minimizes the difference between actual and predicted values.
- Used for prediction and forecasting.

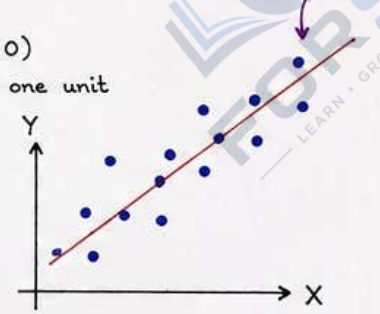
Simple Linear Regression

Relationship between one independent variable (X) and one dependent variable (Y).

Regression Equation (Line of Best Fit):

$$Y = a + bX$$

Where,
Y = Predicted value of dependent variable
X = Independent variable
a = Intercept (Y when $X = 0$)
b = Slope (change in Y for one unit change in X)



How are a and b found?

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad a = \bar{y} - b\,\bar{x}$$

Example :

Predicting salary (Y) from experience (X).
The regression line helps estimate salary for a given experience.

In Short :

- Correlation tells us how two variables are related (strength & direction).
- Regression helps us find the equation of the line and predict Y from X.
- High correlation does not always mean good prediction (use regression for prediction).

Remember :
Correlation → Association
Regression → Prediction

Probability Concepts in Data Science

Probability is the branch of mathematics that deals with the likelihood of an event occurring. It forms the foundation for many data science techniques such as statistics, machine learning and predictive modeling.

1. What is Probability?

- Probability measures how likely an event is to occur.
- The value of probability lies between 0 and 1.
- 0 → Impossible event 1 → Certain event

2. Basic Terminology

- **Experiment** : An action or process with a set of possible outcomes.
- **Sample Space (S)** : The set of all possible outcomes.
- **Event (E)** : A subset of the sample space.
- **Outcome** : A single result of an experiment.

3. Classical (Theoretical) Probability

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Properties :

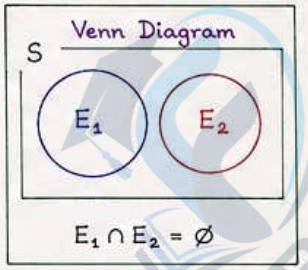
- $0 \le P(E) \le 1$
- $P(S) = 1$
- $P(\emptyset) = 0$
- If $E_1, E_2, \ldots$ are mutually exclusive, $P(E_1 \cup E_2 \cup \ldots) = P(E_1) + P(E_2) + \ldots$

Example :
A fair dice is rolled.
Sample Space $S = \{1, 2, 3, 4, 5, 6\}$
$P(\text{rolling a } 3) = \frac{1}{6}$
$P(\text{rolling an even number}) = \{2, 4, 6\} = \frac{3}{6} = \frac{1}{2}$

4. Types of Events

- **Mutually Exclusive Events** : Cannot occur at the same time. $(E_1 \cap E_2 = \emptyset)$
- **Exhaustive Events** : At least one of them must occur.
- **Independent Events** : Occurrence of one does not affect the other. $P(E_2 \mid E_1) = P(E_2)$
- **Dependent Events** : Occurrence of one affects the other.

Example : Toss a coin
$S = \{H, T\}$
Let $E_1 = \{H\}$, $E_2 = \{T\}$
$P(E_1) = \frac{1}{2}$, $P(E_2) = \frac{1}{2}$
E_1 and E_2 are mutually exclusive and exhaustive.



5. Addition Rule

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

If E_1 and E_2 are mutually exclusive,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Example : A bag contains 5 red, 3 blue and 2 green balls. One ball is drawn at random.

$P(\text{Red}) = \frac{5}{10}$, $P(\text{Blue}) = \frac{3}{10}$, $P(\text{Green}) = \frac{2}{10}$

$P(\text{Red or Blue}) = \frac{5}{10} + \frac{3}{10} = \frac{8}{10} = 0.8$

6. Conditional Probability

Probability of event E_2 given that E_1 has already occurred.

$$P(E_2 \mid E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} , \quad P(E_1) > 0$$

Example : Drawing cards from a deck
$P(\text{Ace}) = \frac{4}{52}$
After drawing an Ace, remaining cards = 51
$P(\text{King} \mid \text{Ace drawn}) = \frac{4}{51}$

7. Independent Events

If E_1 and E_2 are independent,

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2)$$

Example : Toss two coins

Outcomes	HH	HT	TH	TT
Probability	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$P(\text{First coin} = H) = \frac{1}{2}$
$P(\text{Second coin} = T) = \frac{1}{2}$
$P(H \text{ and } T) = \frac{1}{4}$
$= \frac{1}{2} \times \frac{1}{2}$ (Independent)

8. Random Variables (Basics)

A random variable is a numerical value assigned to the outcomes of an experiment.

- **Discrete Random Variable** : Takes countable values (e.g., number of heads in coin tosses)
- **Continuous Random Variable** : Takes any value in an interval (e.g., height, weight, time)

Why Probability is Important in Data Science?

- Helps in understanding uncertainty and variability in data.
- Used in statistical inference, hypothesis testing and model evaluation.
- Forms the basis for algorithms in machine learning and AI.

In Short :
Probability helps us quantify uncertainty and make better decisions from data.


1010101010
0101010101
1010101010
0101010101
Unit - IV
10110
01001
10101
DATA STRUCTURES
ALGORITHMS
COMPUTER NETWORKS
DATABASE SYSTEMS
#include <iostream>
using namespace std;
int main()
{
cout << "Hello, World!";
return 0;
}

Introduction to Machine Learning

What is Machine Learning?

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that enables computers to learn from data and improve their performance on a task without being explicitly programmed.

How does it work?

1. Collect data
2. Use algorithms to learn patterns
3. Make predictions or decisions
4. Improve automatically with more data

Why Machine Learning?

- Handles large amounts of data
- Finds hidden patterns and insights
- Makes accurate predictions
- Automates tasks and saves time
- Used in many real-world applications

Types of Machine Learning

1. Supervised Learning

Learn from labeled data.

Example: Email spam detection

2. Unsupervised Learning

Find patterns in unlabeled data.

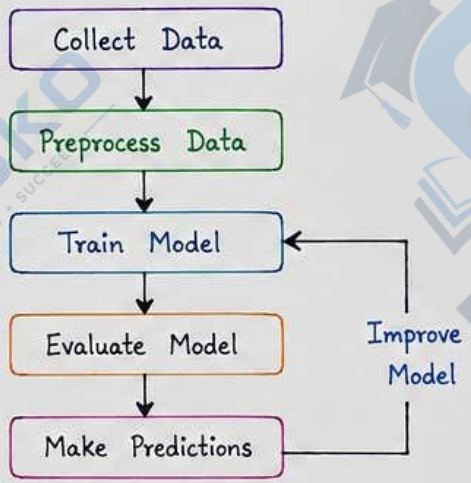
Example: Customer segmentation

3. Reinforcement Learning

Learn by interacting with environment and getting rewards.

Example: Training a robot to walk

Common ML Workflow



Real-World Applications

- Email Spam Detection
- Recommendation Systems (YouTube, Netflix)
- Voice Assistants (Siri, Alexa)
- Self-Driving Cars
- Medical Diagnosis
- E-commerce (Product Suggestion)

Key Terms

- ★ Data: Raw facts and figures collected from various sources.
- ★ Algorithm: Step-by-step procedure or formula to solve a problem.
- ★ Model: A system that learns from data and makes predictions.
- ★ Training: The process of teaching the model using data.
- ★ Prediction: Output or result predicted by the trained model.

In Simple Words

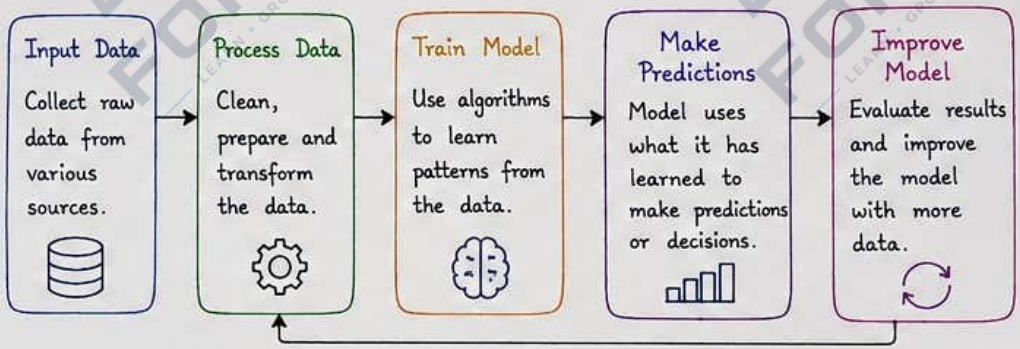
Machine Learning is all about giving computers the ability to learn from experience, just like humans do, and make smart decisions. ☺

Machine Learning Overview

What is Machine Learning?

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that enables computers to learn from data and improve their performance on a task without being explicitly programmed.

How Does Machine Learning Work?



Why Machine Learning?

- Handles large amounts of data
- Finds hidden patterns and insights
- Makes accurate predictions
- Automates tasks and saves time
- Improves with experience
- Used in many real-world applications

Types of Machine Learning

1. Supervised Learning

Learns from labeled data.

Goal: Predict output for new data.

Example:
Email spam detection, House price prediction

2. Unsupervised Learning

Finds patterns in unlabeled data.

Goal: Discover hidden structure or groups.

Example:
Customer segmentation, Market basket analysis

3. Reinforcement Learning

Learns by interacting with an environment and getting rewards.

Goal: Maximize cumulative reward.

Example:
Game playing, Self-driving cars

Common Machine Learning Workflow

1. Problem Definition – Define the problem and objectives.
2. Data Collection – Gather relevant data.
3. Data Preparation – Clean and preprocess data.
4. Feature Engineering – Select or create useful features.
5. Model Training – Train the model using appropriate algorithm.
6. Evaluation – Evaluate model performance.
7. Deployment – Use the model in real-world applications.
8. Monitoring & Improvement – Monitor results and update the model.

Examples of Real-World Applications

- Email Spam Detection
- Recommendation Systems (YouTube, Netflix)
- Fraud Detection in Banking
- Voice Assistants (Siri, Alexa, Google Assistant)
- Self-Driving Cars
- Medical Diagnosis
- E-commerce (Product Suggestions)
- Stock Market Prediction

Key Terms

- ★ Data: Raw facts and figures collected from various sources.
- ★ Feature: An individual measurable property or characteristic of data.
- ★ Model: A mathematical representatioal learned from data.
- ★ Training: The process of teaching the model using training data.
- ★ Validation: Checking model performance during training.
- ★ Test Data: Unseen data used to evaluate final model performance.
- ★ Overfitting: When a model learns too much from training data, including noise, and performs poorly on new data.
- ★ Underfitting: When a model is too simple to capture the pattern in data, resulting in poor performance.

In Simple Words

Machine Learning is all about giving computers the ability to learn from data, just like humans learn from experience, and make better decisions without being told every time.

Popular Machine Learning Algorithms

Algorithm	Used For
Linear Regression	Prediction of continuous values
Logistic Regression	Classification (Yes/No, 0/1)
Decision Tree	Classification and Regression
Random Forest	Improved accuracy for classification/regression
K-Means Clustering	Clustering / Grouping data
Support Vector Machine (SVM)	Classification problems

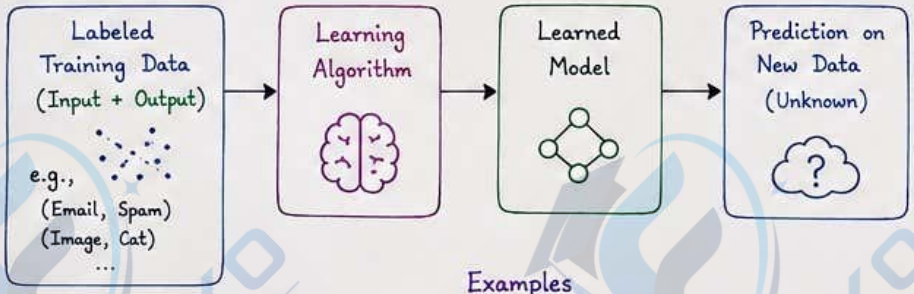
Supervised and Unsupervised Learning

Machine Learning can be broadly divided into two main types based on the kind of data used for training.

1. SUPERVISED LEARNING

In Supervised Learning, the model is trained on labeled data.
Each training example has an input and the correct output.
The model learns the mapping from input to output and makes predictions on new, unseen data.

How it Works?



Characteristics

- Uses labeled data.
- Goal is to predict the output.
- Higher accuracy with more quality data.
- Works well for classification and regression problems.

Examples

- Email spam detection (Spam / Not Spam)
- House price prediction (Regression)
- Disease prediction (Yes / No)
- Sentiment analysis (Positive / Negative)
- Sales prediction

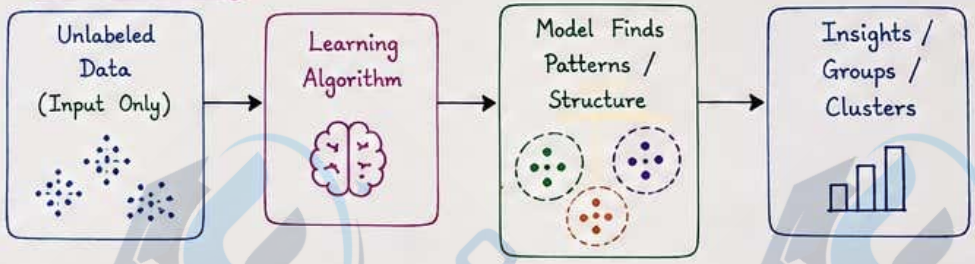
Common Algorithms

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

2. UNSUPERVISED LEARNING

In Unsupervised Learning, the model is trained on unlabeled data.
The data has only inputs and no correct output.
The model finds hidden patterns, structure or relationships in the data on its own.

How it Works?



Characteristics

- Uses unlabeled data.
- Goal is to discover patterns or structure.
- No correct output is provided.
- Useful for exploratory data analysis.

Examples

- Customer segmentation
- Market basket analysis
- Image clustering
- Anomaly / outlier detection
- Document clustering

Common Algorithms

- K-Means Clustering
- Hierarchical Clustering
- DBSCAN
- Principal Component Analysis (PCA)
- Apriori Algorithm
- Autoencoders

In Simple Words

Supervised Learning is like teaching a child with answers.

Unsupervised Learning is like giving a child some data and asking him to find patterns by himself.

Key Difference

Aspect	Supervised Learning	Unsupervised Learning
Data	Labeled data (Input + Output)	Unlabeled data (Input only)
Goal	Predict output for new data	Find hidden patterns / structure
Output	Specific value / class	Groups, clusters or associations
Examples	Classification, Regression	Clustering, Association, Dimensionality reduction
Evaluation	Easy (Accuracy, Precision, etc.)	Relatively difficult (Silhouette score, etc.)

Remember!

Supervised Learning tells the computer "what to do"

Unsupervised Learning helps the computer "discover what's there"

Classification and Regression

Two important types of Supervised Learning used for prediction problems in Machine Learning.

1. Classification

What is Classification?

Classification is a Supervised Learning technique used to predict a discrete category or class label for a given input.

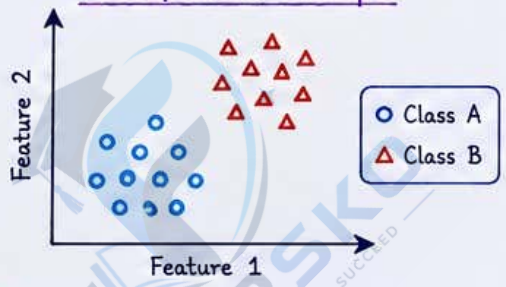
In Simple Words

Classification answers the question "Which class does this data belong to?"

How it Works?

1. Train the model using labeled data.
2. Model learns the patterns of each class.
3. For new data, it predicts the most probable class.

Classification Example



Real-World Applications

- Email Spam Detection (Spam / Not Spam)
- Disease Prediction (Positive / Negative)
- Loan Approval (Approve / Reject)
- Customer Churn (Will leave / Not leave)
- Image Recognition (Cat / Dog / Car)

Popular Classification Algorithms

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Naive Bayes

2. Regression

What is Regression?

Regression is a Supervised Learning technique used to predict a continuous numerical value for a given input.

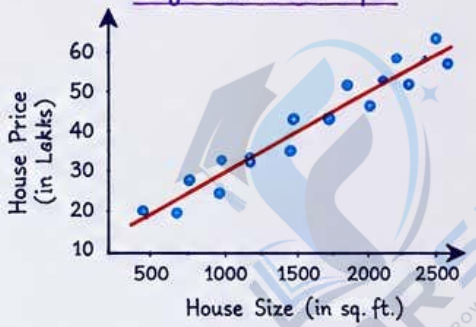
In Simple Words

Regression answers the question "What will be the exact value?"

How it Works?

1. Train the model using labeled data.
2. Model learns the relationship between input features and the target value.
3. For new data, it predicts a numerical output.

Regression Example



Real-World Applications

- House Price Prediction
- Stock Market Forecasting
- Weather Temperature Prediction
- Fuel Efficiency Prediction
- Sales / Revenue Forecasting

Popular Regression Algorithms

- Linear Regression
- Ridge & Lasso Regression
- Decision Tree Regression
- Random Forest Regression
- Support Vector Regression (SVR)

Key Differences

Aspect	Classification	Regression
Output Type	Discrete class label (Category)	Continuous numerical value
Goal	Predict which class / category	Predict exact value
Example Output	Spam or Not Spam, Cat or Dog	Price = 45.6, Temperature = 32.5°C
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score	MAE, MSE, RMSE, R^2 Score
Type of Problem	Classification Problem	Regression Problem

Conclusion

Classification and Regression are the two most common types of Supervised Learning. Use Classification when you need to predict a category, and Regression when you need to predict a numerical value.

CLUSTERING BASICS

Clustering is an unsupervised learning technique that groups similar data points together based on their patterns or characteristics.

1. What is Clustering?

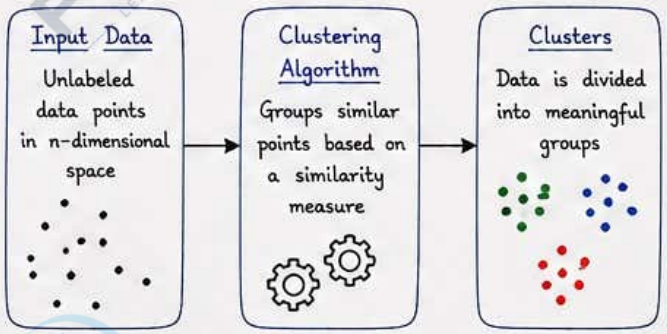
Clustering aims to divide data into groups (clusters) such that:

- Points in the same cluster are more similar
- Points in different clusters are less similar

Real-life Examples

- Customer segmentation
- Grouping documents by topic
- Image segmentation
- Social network analysis
- Anomaly / Outlier detection

2. How Clustering Works?



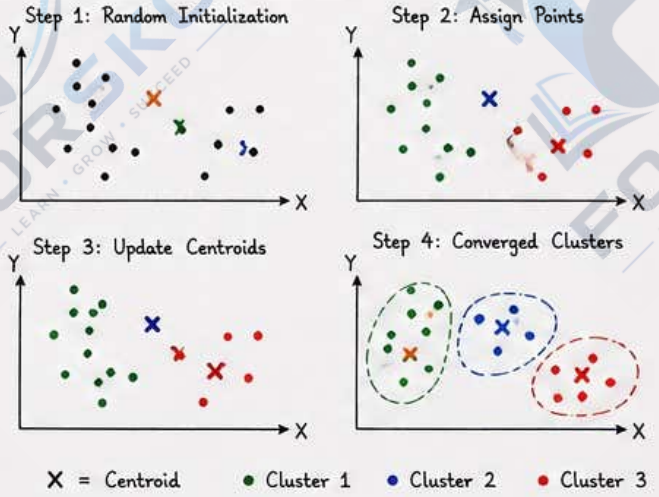
3. Types of Clustering

1. **Partitioning Clustering** : Divides data into k partitions.
   Example: K-Means
2. **Hierarchical Clustering** : Creates a hierarchy of clusters.
   Example: Agglomerative, Divisive
3. **Density-Based Clustering** : Forms clusters based on dense regions.
   Example: DBSCAN
4. **Grid-Based Clustering** : Divides the data space into a grid structure.
   Example: ST-DBSCAN

4. Popular Clustering Algorithms

Algorithm	Type	Key Idea	Best Used For
K-Means	Partitioning	Groups data into K clusters by minimizing within-cluster variance	Spherical clusters, medium to large datasets
Hierarchical Clustering	Hierarchical	Builds a tree (dendrogram) by merging or splitting clusters	Small to medium datasets, nested clusters
DBSCAN	Density-Based	Forms clusters in dense regions and marks outliers	Arbitrary shape clusters, noise handling
K-Medoids	Partitioning	Similar to K-Means but uses medoids (actual points)	Robust to outliers

5. Example: K-Means Clustering (K = 3)



6. Similarity / Distance Measures

- Euclidean Distance : $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- Manhattan Distance : $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
- Cosine Similarity : $\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|}$

7. Evaluation of Clustering

- Silhouette Score : Measures how similar a point is to its own cluster compared to other clusters.
- Davies-Bouldin Index : Lower value indicates better clustering.
- Calinski-Harabasz Index : Higher value indicates better clustering.
- Elbow Method (for K) : Choose K at the "elbow" point. where improvement slows.

8. Advantages and Limitations

Advantages

- No need for labeled data
- Discovers hidden patterns
- Useful for exploratory data analysis
- Helps in decision making & segmentation

Limitations

- Choosing the right number of clusters (K) is hard
- Sensitive to noise and outliers
- Results may vary with different algorithms
- Hard to evaluate without ground truth

In Simple Words

Clustering is like grouping your books by topics without being told the topics. The algorithm finds the natural groups on its own.

Remember!

Good clustering = High similarity within cluster + Low similarity between clusters.

APPLICATIONS OF MACHINE LEARNING IN REAL LIFE

Machine Learning is used in many real-life applications around us. It helps machines learn from data and make intelligent decisions without being explicitly programmed.

1. Where is ML Used?

From the apps we use daily to the services we rely on, Machine Learning is everywhere!

1. Virtual Assistants

- Siri, Google Assistant, Alexa use ML to understand your voice and respond.
- Tasks: set reminders, play music, answer questions, etc.

2. Recommendation Systems

- YouTube, Netflix, Amazon use ML to recommend movies, shows and products you may like.
- Based on your past behavior and preferences.

3. Email Spam Detection

SPAM

- ML models learn from previous emails and detect spam.
- Keeps your inbox clean and safe.

4. Image Recognition

- Used in Google Photos, Facebook to tag friends automatically.
- Helps in medical imaging, security, and object detection.

5. Speech Recognition

- Converts speech into text.
- Used in dictation tools, voice typing, call transcription, etc.
- Improves accessibility for everyone.

6. Fraud Detection

- Banks and credit card companies use ML to detect unusual transarptions.
- Helps prevent fraud and protects users.

7. Healthcare

- ML helps in disease prediction, drug discovery, and personalized treastment.
- Analyzes medical reports, scans, and patient data.

8. Self-Driving Cars

- ML enables cars to detect objects, recognize lanes, and make driving decisions.
- Improves safety and reduces accidents.

9. Online Shopping

- ML predicts what products you may like.
- Helps in dynamic pricing, search suggestions and customer support (chatbots).

10. Social Media

- ML powers news feed ranking, friend suggestions, and content moderation.
- Detects fake accounts, hate speech and spam.

11. Finance & Stock Market

- ML analyzes market trends, predicts stock prices and manages risk.
- Used in algorithmic trading and portfolio management.

12. Entertainment & Gaming

- ML creates realistic game environments and smart NPCs.
- Used in game playing agents and personalized gaming experience.

Examples Around You (Think about it!)

- Spotify suggests songs you may like.
- Google Maps predicts traffic and best routes.
- Google Translate translates text in real-time.
- Face Unlock in phones uses ML to recognize you.
- You type "teh" keyboard corrects it to "the". teh → the
- Fitness watches predict your health and activity.

2. Benefits of ML in Real Life

- ✓ Automates tasks and saves time
- ✓ Improves accuracy and efficiency
- ✓ Handles large amounts of data
- ✓ Learns and improves from experience
- ✓ Helps in better decision making
- ✓ Solves complex problems

3. Impact on Our Daily Life

- ★ Makes technology smarter and more useful
- ★ Enhances user experience
- ★ Helps businesses grow and reduce costs
- ★ Supports safety, health and convenience
- ★ Improves future innovations

4. In Simple Words

Machine Learning is behind many things we use every day. It learns from data, finds patterns and helps machines make smart decisions like humans! ☺

Remember!

ML is not just about robots or computers. It is about making our lives easier, smarter and better every day. ☆

ETHICAL ISSUES IN DATA SCIENCE

Data Science and AI have the power to improve lives, but they also bring ethical challenges.
Using data responsibly and fairly is just as important as building accurate models.

1. What is Ethics in Data Science?

It refers to doing the right thing while collecting, analyzing and using data.

Goal: Make sure data-driven decisions are fair, safe, transparent and respectful to people.

2. Why Ethics Matters?

- Protects people's rights and privacy
- Builds trust in data-driven systems
- Prevents harm and discrimination
- Ensures fairness and accountability
- Follows laws and avoids legal issues

3. Key Ethical Principles

Fairness
Treat everyone fairly, avoid bias.

Privacy
Protect personal data and respect privacy.

Transparency
Be open about how data is used and how decisions are made.

Accountability
Take responsibility for decisions and their impact.

Beneficence
Use data to do good and benefit people.

Non-maleficence
Do no harm – avoid causing negative impact.

4. Major Ethical Issues

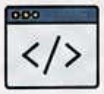
1. Bias and Discrimination

- Biased data or models can lead to unfair treatment.
- Example: Biased hiring system against certain groups.

2. Privacy and Data Security

- Collecting too much personal data.
- Risk of data leaks and misuse.

3. Lack of Transparency



- Complex models (like Black Box models) are hard to explain.
- Users don't know how decisions are made.

4. Data Misuse

- Using data for purposes people didn't agree to.
- Example: Selling user data without permission.

5. Lack of Accountability

- No one takes responsibility when things go wrong.
- Hard to find who is accountable.

6. Over-reliance on Automation

- Blindly trusting models without human oversight.
- Can cause wrong or harmful decisions.

5. Real-world Examples

- ★ **Social Media:** Personal data used for targeted ads without clear consent.
- ★ **Facial Recognition:** Misidentification and privacy invasion.
- ★ **Loan Approval Systems:** Biased models may reject qualified applicants.
- ★ **Recommendation Systems:** Can create filter bubbles and manipulate users.

6. How to Address Ethical Issues?

- ✓ Use diverse and representative data.
- ✓ Check and reduce bias in data and models.
- ✓ Be transparent and explainable.
- ✓ Get informed consent from users.
- ✓ Protect data with strong security.
- ✓ Involve humans in the loop.
- ✓ Regularly audit models and data.
- ✓ Follow laws and ethical guidelines.
- ✓ Promote a culture of ethics and responsibility.

In Simple Words

Data Science is powerful, but with great power comes great responsibility.
Use data fairly, respect privacy, be transparent and always think about people first.

Remember!

Ethics in Data Science is not just about following rules, it's about doing the right thing for a better and fairer world.