

# Machine Learning

Part - 1





Machine Learning is all about teaching computers to learn from data — just like we learn from experience.

Akarsh Vyas



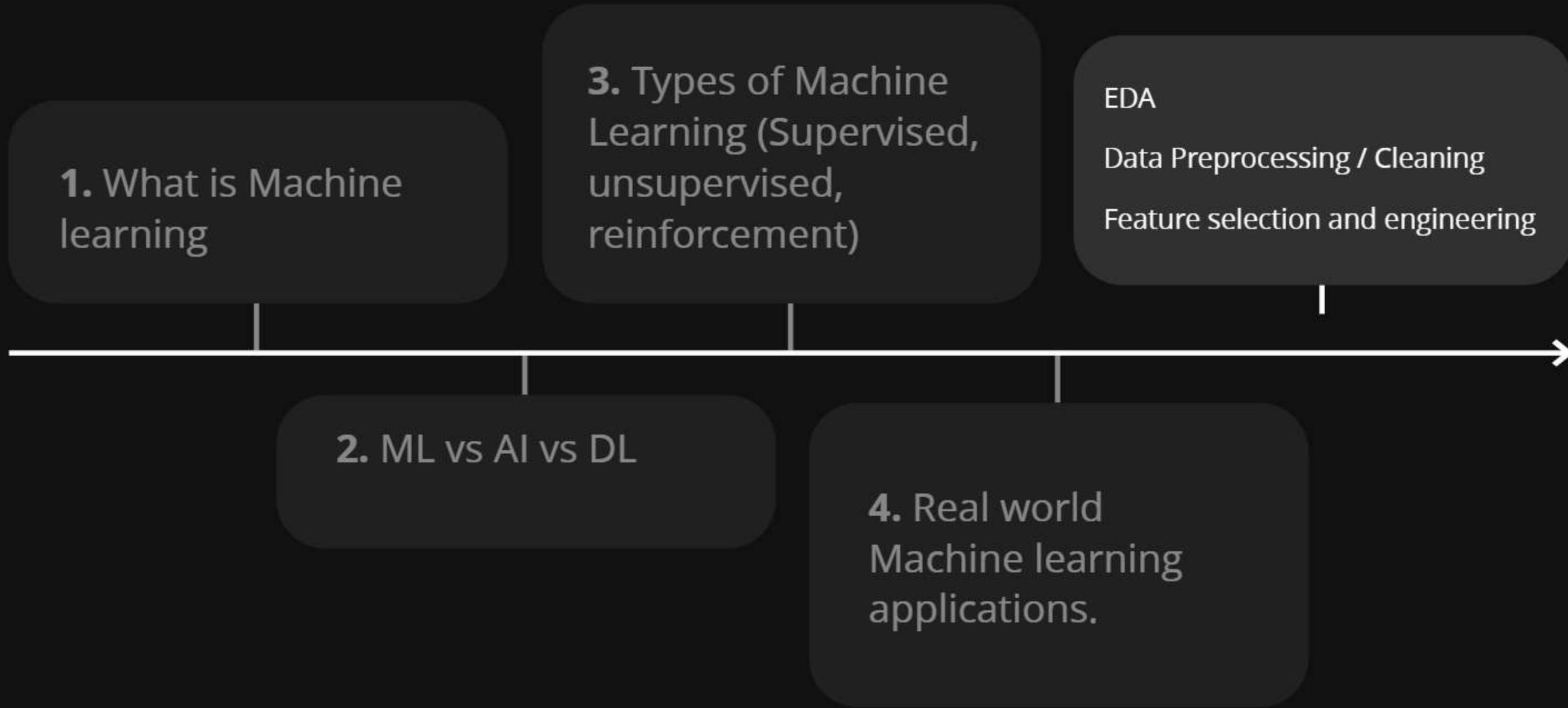
**1. What is Machine learning**

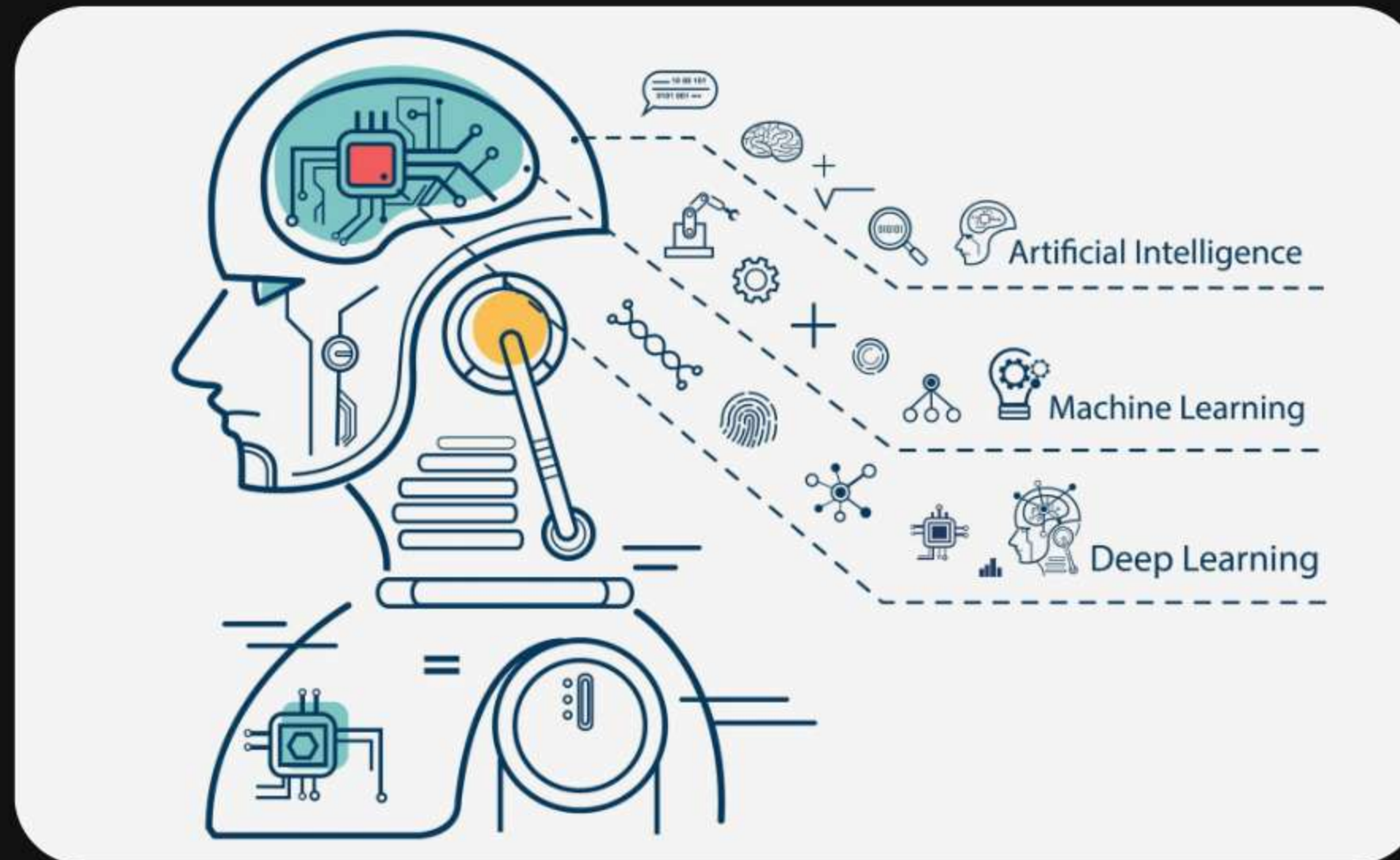
**3. Types of Machine Learning (Supervised, unsupervised, reinforcement)**

EDA  
Data Preprocessing / Cleaning  
Feature selection and engineering

**2. ML vs AI vs DL**

**4. Real world Machine learning applications.**





What is Machine Learning?



# Machine Learning

Have you ever wondered how Netflix knows what you want to watch next? Or how Google Translate works like magic? The secret behind it all... is **Machine Learning**.





# Machine Learning

Machine Learning is a way of teaching computers to learn from data — just like we humans learn from experience.



# Machine Learning

Instead of programming every step, we give machines lots of data and let them figure out patterns on their own.

# Machine Learning

- **YouTube Recommendations** → Learns from your watch history
- **Spam Detection in Gmail** → Learns patterns in spam emails
- **Voice Assistants (Siri, Alexa)** → Learn how you speak
- **Self-driving Cars** → Learn to identify stop signs, pedestrians, and roads
- **Face Unlock on Phones** → Learns to recognize your face





## Traditional programming

- Give rules + data → get result

Manual logic writing



## Machine Learning

- Give data + result → get rules (model)

Learns patterns automatically



# Machine Learning

Today, Machine Learning powers everything from the apps on your phone to the systems behind hospitals, banks, and even space research. Learning ML means you're learning the language of the future.

# Machine Learning

Now that you know what Machine Learning is, let's dive into the difference between AI ,ML , DL.



# AI vs ML vs DL

People often use the terms AI, Machine Learning, and Deep Learning like they mean the same thing. But guess what? They're **not** the same lets see the difference between them.

# AI vs ML vs DL

AI - Artificial Intelligence is the **big umbrella** — it's the science of making machines *smart*, just like humans.





# AI vs ML vs DL

AI - examples

- Playing chess like a human (AI)
- Talking to Alexa (AI)
- Driving a car on its own (AI)

AI = Any system that mimics human intelligence

# AI vs ML vs DL

ML - Machine Learning is a **subset of AI** — this is where machines learn from **data** and improve over time.

YouTube recommending videos

Netflix predicting your next binge

Gmail filtering spam


# AI vs ML vs DL

DL - Deep Learning is a **subset of Machine Learning** — it uses something called **neural networks**, which are inspired by the human brain.



# AI vs ML vs DL

## DL - examples

- Face recognition on phones
- ChatGPT 
- Self-driving car vision
- DL = ML using neural networks for big, complex data (like images or speech)

# Types of ML

Machine Learning has 3 major types. But most beginners mix them up!

Let me explain them in the simplest way possible — with real-life examples



# Supervised Learning

Supervised Learning is like a student being taught by a teacher.

We give the machine input and the correct answer — and it learns to predict

# Supervised Learning

Machine learning is a game of prediction we give the machine data and it predicts lets see an example with data.

# Supervised Learning

Here in this example Income and credit score are input variables and loan is the output variable.

And we have to predict the output variable.

Income(\$)	Credit Score	Loan
40,000	750	Yes
25,000	600	No
50,000	800	Yes
30,000	580	No

# Supervised Learning

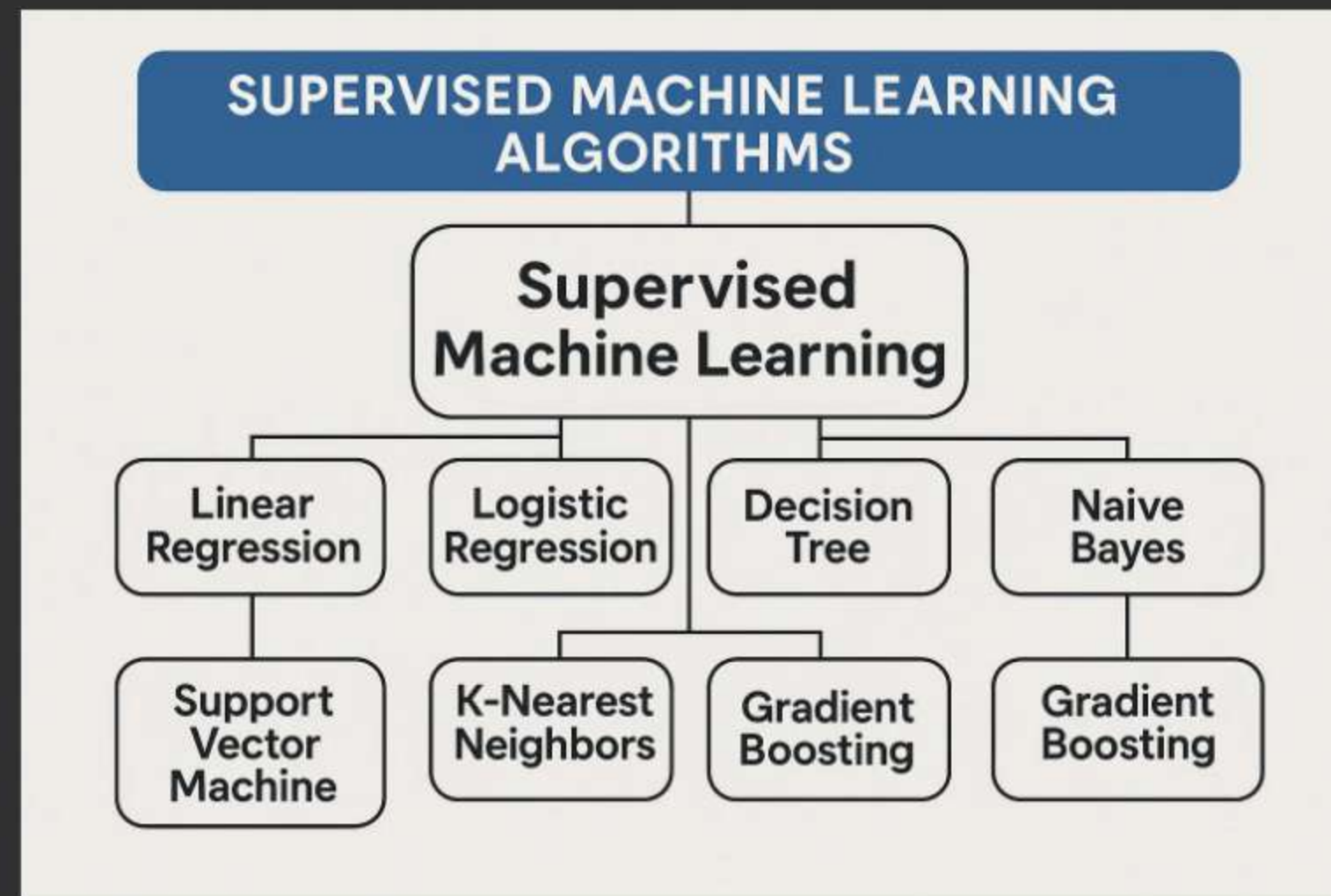
So here we start training the model with input and output variables both.

So in short we can say we have both input and output variables.

Income(\$)	Credit Score	Loan
40,000	750	Yes
25,000	600	No
50,000	800	Yes
30,000	580	No

# Supervised Learning

But the training of a model are done using various methods and these, methods are known as algorithms.





# unsupervised Learning

Here also we will get the data but we are not going to predict anything we will just find patterns, Group similar items, Reduce complexity, spot outliers.

# unsupervised Learning

Some of the examples are Google search see  
the implementation...



# unsupervised Learning

so unsupervised learning are used with supervised learning to make models

# Reinforcement learning

Reinforcement Learning is like training a dog. You don't teach it everything directly — instead, you reward good behavior and ignore or punish bad actions.

Over time, the dog learns what actions give it treats. That's the core idea of RL — learning by **trial and error**.

# Our Story

Now without wasting much time lets move towards the main part of video what actually is EDA how it works and what is the use case of EDA .





# EDA

EDA stands for **Exploratory Data Analysis** — it's the **process of analyzing, visualizing, and understanding your data** before you build any machine learning model.



# EDA

Lets see the steps involved in making a machine learning model.

- |                                     |                                   |
|-------------------------------------|-----------------------------------|
| 1 - Problem definition              | 2 - Data collection               |
| 3 - Exploratory Data Analysis (EDA) | 4 - Data Preprocessing / Cleaning |
| 5 - Feature Selection & Engineering | 6 - Split the Dataset             |
| 7 - Model Selection.                | 8 - Model Training                |
| 9 - Model Evaluation                | 10 - Hyperparameter Tuning        |
| 11 - Model Testing / Validation     |                                   |

# EDA

EDA stands for Exploratory Data Analysis — it's the step where you explore the data to:

Understand it

Discover patterns

Spot anomalies

Generate insights

And decide what to do next

# EDA

Think of EDA as detective work — you're looking at your data with curiosity before building any models or doing fancy transformations.

# EDA Steps

## 1) Viewing the Data

- `head()`, `tail()`, `shape`, `info()`
- What columns do I have? What types of data?



# EDA Steps

## 2) Summary Statistics

- mean, median, mode, std, min, max, quartiles
- Helps understand spread and central tendency

# EDA Steps

## 3) Value Counts

- How many unique values in a column?
- Great for categorical columns



# EDA Steps

## 4) Missing Value Analysis

- Where are the gaps? What percent of the data is missing?

# EDA Steps

## 5) Visualizations

- Histograms → distribution of values
- Boxplots → outliers and spread
- Bar plots → comparisons of categories
- Correlation heatmaps → linear relationships between numerical features
- Scatter plots → bivariate relationships

# EDA Steps

## 6) Target Variable Exploration

- How does your output (like 'charges' in your dataset) relate to other variables?



# EDA Importance

- You can't clean or preprocess what you don't understand.
- It helps you identify mistakes, biases, or limitations in the data.
- It guides the direction of your data cleaning and feature engineering.
- It gives your audience or stakeholders a "story" or overview of what the data is telling you.

# Data cleaning

## 1. Handle Missing Values

- Check which columns have missing values (nulls)



# Data cleaning

- Strategies to handle:
  - Drop missing rows/columns (only if very few)
  - Impute with:
    - Mean/Median → for numerical data
    - Mode → for categorical data
    - Advanced: Linear regression, KNN, or interpolation (for future learning)

# Data cleaning

## 2. Remove Duplicates

- Detect and drop exact duplicate rows

# Data cleaning

## 3. Fix Data Types

- Convert wrong types (e.g., numbers stored as strings, dates as text)



# Data cleaning

## 4. Handle Inconsistent Categories

- Clean up categorical values like:
  - "Male", "male", "MALE" → should all become "male"
  - "Yes", "yes", "Y" → unify to one format

# Data cleaning

## 5. Detect and Handle Outliers

- Use boxplots, IQR, or Z-score
- Handle by:
  - Removing (if clearly wrong)
  - Capping (e.g., to 95th percentile)

# Data cleaning

## 7. Fix Logic or Domain Errors

- E.g., age = -5 is invalid, or BMI = 150 likely an error
- Can replace with mean, median, or remove

# Data cleaning

You can say:

EDA tells you what's wrong. Data Cleaning fixes it.

Cleaning is not glamorous, but it's 80% of the work in real-world projects.



# Data Preprocessing

To prepare clean data so it can be analyzed or used in a machine learning model."

If Data Cleaning is about fixing mistakes,  
Data Preprocessing is about transforming valid data into a usable format.



# Data Preprocessing

## 1. Encoding Categorical Variables

Convert text labels (like "male", "yes", "southeast") into numbers.

# Data Preprocessing

Two common methods:

- Label Encoding (Ordinal):  
Good for ordered categories like "Low", "Medium", "High"
- One-Hot Encoding (Nominal):  
For non-ordered categories like region

# Data Preprocessing

2) Feature Transformation (Log, Square root, etc.)  
Used to handle skewed data, like right-skewed or left skewed data.

# Data Preprocessing

## 3. Feature Scaling (Normalization or Standardization)

Bring numerical values to the same scale — especially useful for distance-based algorithms.

# Data Preprocessing

Normalization (Min-Max Scaling):  
Scales values between 0 and 1

# Data Preprocessing

Standardization (Z-score Scaling):  
Transforms data to have mean 0 and std 1



# Feature Engineering

Creating new features or transforming existing ones to expose useful patterns that ML models can learn from.

# Feature Engineering

Why do we need it?

Because ML models don't know domain logic — we have to give them the right signals.

# Feature Engineering

Common Feature Engineering Techniques in ML:

Mathematical Combinations

Target-Based Flags

Binning (when it helps)

Time-Based Features (if time exists)

# Feature Selection (for ML)

Selecting the most useful features and removing the rest.

Why is it important?

- Reduces noise and overfitting
- Speeds up training
- Improves model accuracy
- Makes model interpretation easier

# Feature Selection (for ML)

## 1. Filter Methods (Pure Statistics)

- Correlation Matrix → Remove highly correlated features
- Chi-square test (categorical vs categorical)
- ANOVA F-test (numerical vs categorical target)



# Feature Selection (for ML)

## 3. Embedded Methods (Selection built into the model)

- Lasso Regression → Shrinks coefficients to 0
- Tree-based models (Random Forest, XGBoost) → Feature importance scores