INTRO TO REG

By Rishabh Dubey

Quiz Time

useful links

https://colab.research.google.com/drive/1NsUk8quPn675V1sy66QZT_GaQcifld_l?usp=sharing

https://colab.research.google.com/drive/111yK-TF8mnANPE6YQMbulrHQDp0-IFcZ?usp=sharing

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 1: What does df.describe() do in Pandas?

- (A) Shows first 5 rows
- (B) Shows statistical summary
- (C) Deletes missing values
- (D) Merges two DataFrames

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 2: Which of the following is not a NumPy function?

- (A) np.mean()
- (B) np.array()
- (C) np.plot()
- (D) np.linspace()

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 3: What is the primary purpose of Matplotlib?

- (A) Data visualization
- (B) Machine Learning
- (C) Data cleaning
- (D) API development

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 4: Which method is used in Pandas to remove NaN values?

- (A) dropna()
- (B) fillna()
- (C) replace()
- (D) clear()

Introduction to Machine Learning

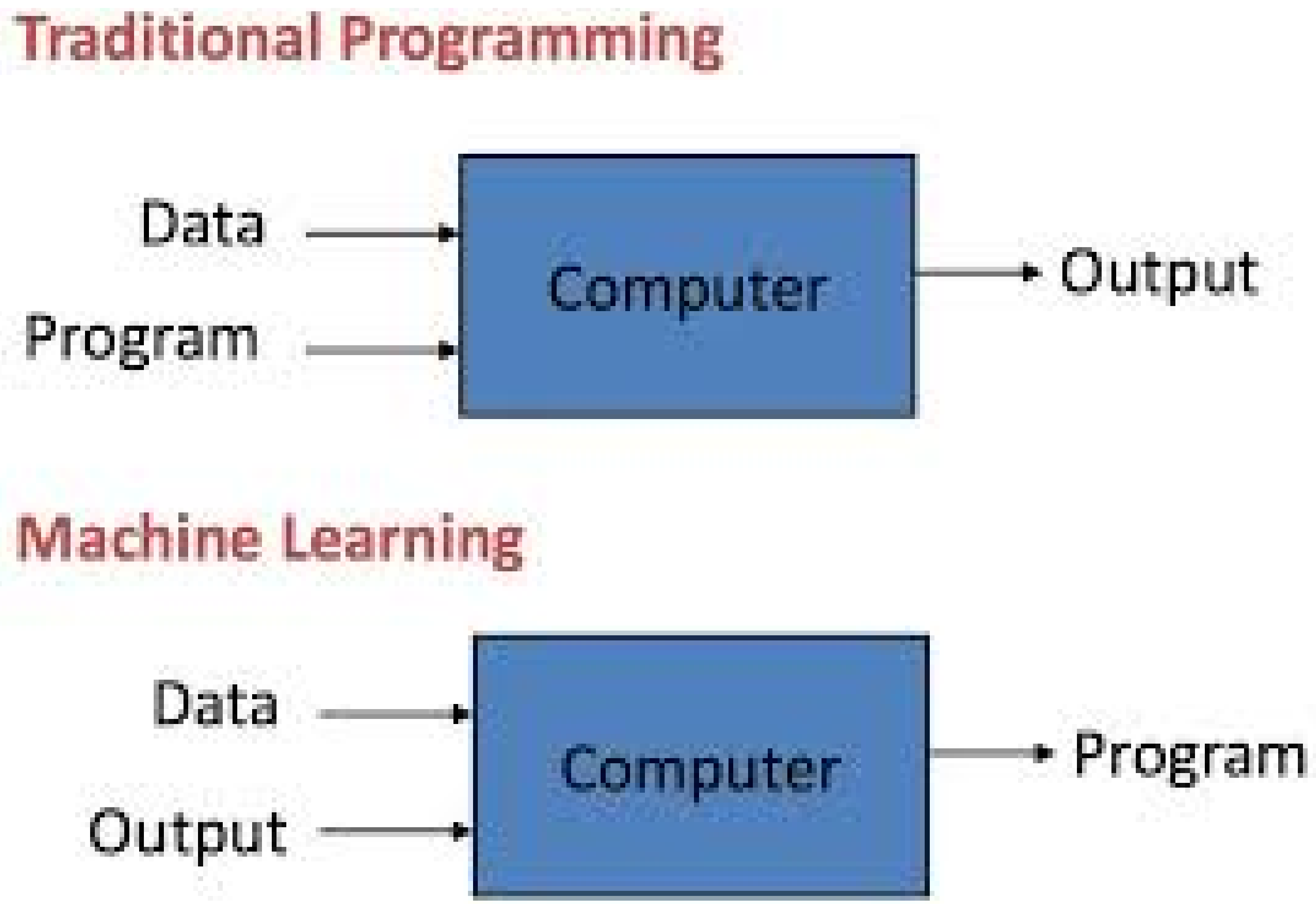
WHAT IS MACHINE LEARNING?

ML allows computers to learn from data instead of being explicitly programmed.

Examples: Spam email detection

- Netflix recommendations
- Self-driving cars

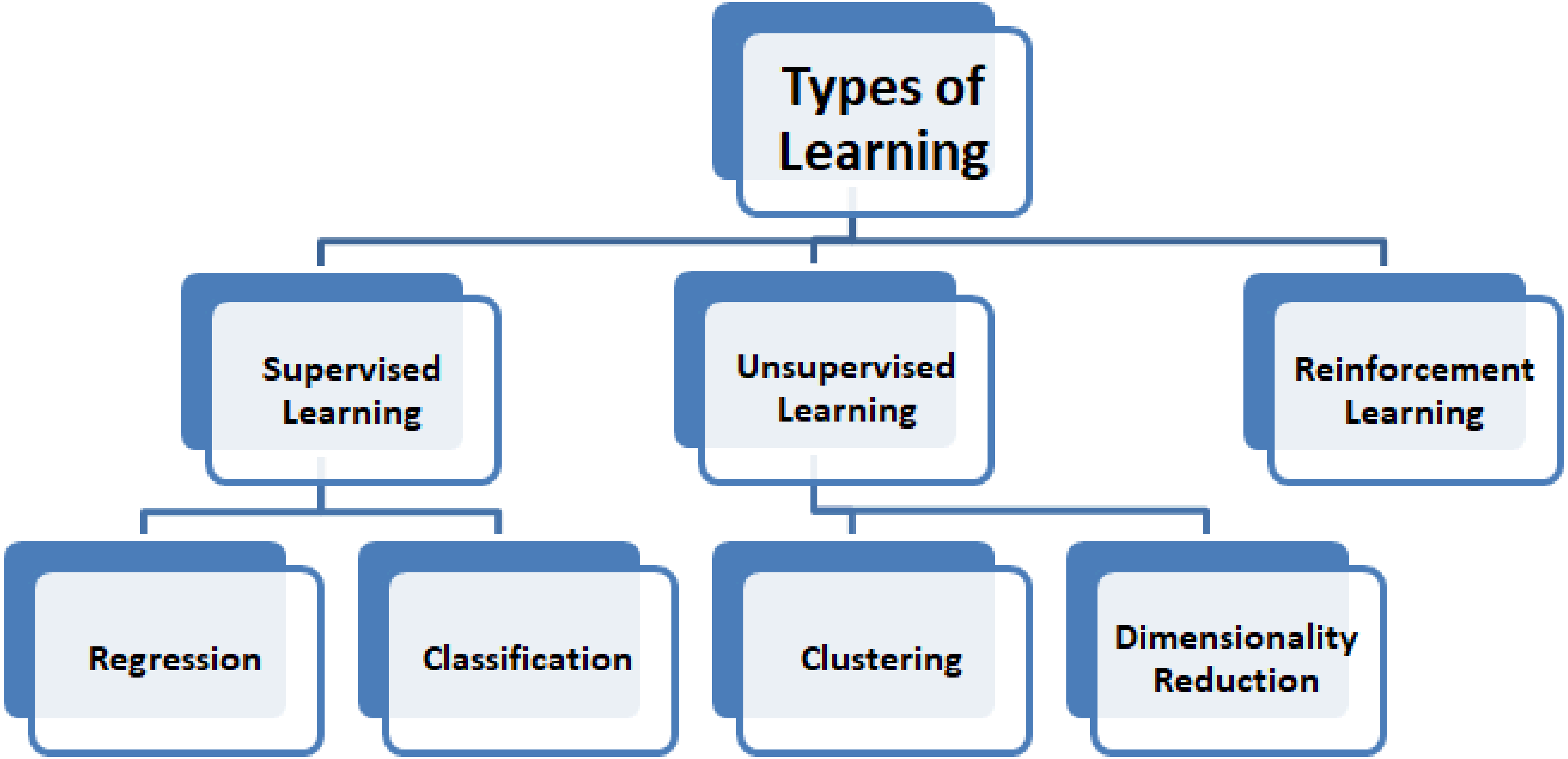
HOW ML DIFFERS FROM TRADITIONAL PROGRAMMING



Traditional Programming

Machine Learning

TYPES OF ML


Types of Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Regression
Classification
Clustering
Dimensionality Reduction

TYPES OF ML

Supervised Learning

- Uses labeled data
    - Examples:Spam detection
    - House Price Prediction

Unsupervised Learning

- Uses unlabeled data
    - Examples:Customer Segmentation
    - Anomaly Detection

TYPES OF ML

Reinforcement Learning

Learning via rewards and penalties

Examples:

- Game AI
- Robotics
- Self-driving cars

💡 Example: Surveys, Elections, Market Research

TYPES OF DATASETS

What are Datasets?

- Collection of structured or unstructured data used for training ML models.

A label is additional information (a classification, category, or meaningful value) mapped to raw data.

Labeled vs Unlabeled Data

Type	Definition	Example
Labeled Data	Contains input-output pairs	(X-ray, Disease Yes/No)
Unlabeled Data	No predefined labels	(Customer purchase history)

Supervised Learning Overview

SUPERVISED LEARNING

📌WHAT IS SUPERVISED LEARNING?

Model learns from labeled data (X → Y).

Example: Email Spam Detection → (Email content → Spam/Not Spam).

Used for both classification and regression tasks.

SUPERVISED LEARNING

📌 HOW SUPERVISED LEARNING WORKS?

1. Input Data (X) → Features.
2. Model Trains on Data.
3. Model Learns Patterns.
4. Output Prediction (Y).
5. Evaluates Performance & Adjusts.

REGRESSION VS. CLASSIFICATION

Regression: Predicts continuous values (e.g., house price).

Classification: Predicts categories (e.g., dog vs. cat).

Key Difference:

Regression → Numeric Output,

Classification → Labels.

Introduction to Regression

WHAT IS REGRESSION?

Regression is used when the target variable is continuous.
Example: Predicting a person’s salary based on years of experience.
Applied in finance, sales forecasting, healthcare.

Model Selection

- Choosing the right algorithm (Linear Regression, Decision Trees, Neural Networks).

WHAT IS REGRESSION?

- Predicts continuous values.
- Examples: House prices, Sales prediction.

Types of Regression

Type	**Use Case**
Simple Linear Regression	One input, one output
Multiple Regression	Multiple inputs
Polynomial Regression	Non-linear relationships

SIMPLE LINEAR REGRESSION

Equation: **y=mx+c** (Line of Best Fit).

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

Example: Predicting student scores based on study hours.

VISUALIZING SIMPLE LINEAR REGRESSION

Equation: **$y=mx+c$** (Line of Best Fit).

- A scatter plot with a best-fit line.
- The line minimizes the distance between actual points and predictions.
- Helps identify trends in data.

HOW DOES LINEAR REGRESSION WORK?

Equation: **$y=mx+c$** (Line of Best Fit).

- Finds the line that best fits the data using Least Squares Method.
- Minimizes error between actual and predicted values.
- Uses Mean Squared Error (MSE) as a loss function.

WHAT IS A COST FUNCTION?

- A cost function measures how well the model predicts values.
- Mean Squared Error (MSE): Average squared difference between predicted & actual values.
- Lower MSE = Better model fit.

GRADIENT DESCENT – OPTIMIZATION ALGORITHM

- Goal: Find the best values for mmm and ccc (slope & intercept).
- Starts with random values & iteratively adjusts using learning rate.
- Moves in the direction of the steepest decline in cost function.

VISUALIZING GRADIENT DESCENT

- Cost function plotted as a curve.
- Model updates parameters to reach the lowest point (optimal values).
- Smaller learning rate = Slower but stable learning.

LIVE CODING - SIMPLE LINEAR REGRESSION

- Use sklearn.linear_model.LinearRegression to fit a model.
- Dataset: Study Hours vs. Exam Score.
- Code: (Already provided in your original request)

STUDENT HANDS-ON ACTIVITY

- Implement Linear Regression on a simple dataset.
- Task: Predict Exam Scores based on Study Hours.
- Train & evaluate the model using sklearn.

MULTIPLE LINEAR REGRESSION

Equation: **y=b0+b1x1+b2x2+...+bnxn**

- Predicting using multiple independent variables.

Example: Predicting house prices using area, number of rooms, location..

VISUALIZING MULTIPLE LINEAR REGRESSION

Equation: **Equation: y=b0+b1x1+b2x2+...+bnxn**

- Instead of a line, multiple regression fits a plane/hyperplane.
- More features → More dimensions in the model.
- Helps in complex decision-making.

LIVE CODING - MULTIPLE LINEAR REGRESSION

- Example using sklearn.
- Dataset: Predict house prices using multiple features.
- Train, test, and evaluate the model.

MODEL EVALUATION

- How Do We Measure Performance?

- Mean Absolute Error (MAE): Average absolute error.
- Root Mean Squared Error (RMSE): Square root of MSE.
- R^2 Score: Measures how well the model explains variance.

UNDERSTANDING R^2 SCORE

- How Do We Measure Performance?

- $R^2 = 1$: Perfect model.
- $R^2 = 0$: Model explains no variance.
- $R^2 < 0$: Model performs worse than random guessing.

OVERFITTING VS. UNDERFITTING

- Overfitting: Model memorizes data, fails on new data.
- Underfitting: Model too simple, fails to learn patterns.
- Solution: Train-test split, regularization techniques.

CONFIDENCE INTERVALS IN REGRESSION

- Represents the uncertainty in predictions.
- Example: 95% confidence interval → 95% chance actual value is within range.
- Helps assess model reliability.

HYPOTHESIS TESTING IN REGRESSION

- Null Hypothesis (H_0): No relationship between X and Y.
- Alternative Hypothesis (H_1): X affects Y.
- P-value: If $p<0.05$, we reject H_0

INTRODUCTION TO THE CONFUSION MATRIX

- Used for evaluating classification models.
- TP, FP, FN, TN: Measures model's accuracy.
- Helps in fraud detection, medical diagnoses.

KEY TAKEAWAYS

- Regression helps predict continuous values.
- Simple Linear Regression fits one feature; Multiple Regression handles multiple.
- Evaluation metrics like RMSE & R^2 help assess model performance.
- Avoid overfitting and underfitting with proper training techniques.

VISUALIZING GRADIENT DESCENT

- Cost function plotted as a curve.
- Model updates parameters to reach the lowest point (optimal values).
- Smaller learning rate = Slower but stable learning.

ERROR METRICS FOR REGRESSION

MSE (Mean Squared Error)

RMSE (Root Mean Squared Error)

MAE (Mean Absolute Error)

Introduction to Scikit-Learn

INTRODUCTION TO SCIKIT-LEARN

What is Scikit-Learn?

- Popular Python ML Library

Key Functions in Scikit-Learn

Function	Purpose
train_test_split()	Splits data into train-test
LinearRegression()	Creates a regression model
fit()	Trains the model
predict()	Makes predictions

TASK - IMPLEMENT SIMPLE LINEAR REGRESSION

```
import numpy as np
from sklearn.linear_model import LinearRegression
X = np.array([1, 2, 3, 4, 5]).reshape(-1,1)
y = np.array([2, 4, 5, 4, 5])

model = LinearRegression()
model.fit(X, y)
print("Predictions:", model.predict(X))
```

THANK YOU