Problem Statement

Title: *Predicting House Sale Status Based on Size and Number of Bedrooms Using Logistic Regression*

Objective

To develop a **logistic regression model** that predicts whether a house **sells within 30 days (1) or not (0)** based on:

- **Size in square feet**
- **Number of bedrooms**

Background

In real estate, a key metric for sellers is how quickly a house sells after being listed. Various features (like size and bedroom count) influence this. This project uses logistic regression to model the **probability** that a house will sell within 30 days.

Dataset Description

- **X_1:** Size (in square feet)
- **X_2:** Number of bedrooms
- **Y:** Sold within 30 days (binary: 1 = Yes, 0 = No)

Sample Data

House	Size (sqft)	Bedrooms	Sold in 30 Days (Y)
1	1400	3	0
2	1600	3	1
3	1700	4	1
4	1875	3	1
5	1100	2	0
6	1550	4	0
7	2350	4	1
8	2450	5	1
9	1425	3	0
10	1700	3	1

Goals

- Fit a **logistic regression** model: $ P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2)}} $
- Interpret (b_1) and (b_2) to understand how size and bedrooms influence likelihood of a fast sale.
- Make binary predictions using a probability threshold (e.g., 0.5)
- Evaluate performance using:
  - **Confusion Matrix**
  - **Accuracy**
  - **Precision, Recall, F1 Score**
  - **Log Loss**
  - **ROC-AUC (if desired)**

Assumptions

- The relationship between predictors and the log-odds of the target is linear.
- Observations are independent.
- The output is a **binary classification** (house sells within 30 days or not).

Solution

Problem

Predict whether a house **sells within 30 days (1)** or **not (0)** based on:

- Size (X_1: in square feet)
- Bedrooms (X_2: count)

Step 1: Dataset

House	X_1 (Size sqft)	X_2 (Bedrooms)	Y (Sold in 30 Days)
1	1400	3	0
2	1600	3	1
3	1700	4	1
4	1875	3	1
5	1100	2	0
6	1550	4	0
7	2350	4	1
8	2450	5	1
9	1425	3	0

House	X_1 (Size sqft)	X_2 (Bedrooms)	Y (Sold in 30 Days)
10	1700	3	1

Step 2: Logistic Regression Model

We model the **probability** that a house sells in 30 days as:

$P(Y=1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2)}}$

Let’s assume the model has been trained (using maximum likelihood estimation), and yields the following coefficients:

$b_0 = -7.5,\quad b_1 = 0.003,\quad b_2 = 0.9$

Step 3: Final Equation

$P(\text{Sold}) = \frac{1}{1 + e^{-(-7.5 + 0.003 \cdot \text{Size} + 0.9 \cdot \text{Bedrooms})}}$

Step 4: Make Predictions

Example: Predict for House 3 (1700 sqft, 4 bedrooms)

$z = -7.5 + 0.003 \cdot 1700 + 0.9 \cdot 4 = -7.5 + 5.1 + 3.6 = 1.2$

$P = \frac{1}{1 + e^{-1.2}} \approx 0.768 \Rightarrow \text{Predicted Class} = 1$

Step 5: Predict All and Compare with Actual

House	Size	Beds	Actual Y	(z)	Predicted P	Pred Y
1	1400	3	0	0.7	0.668	1 ❌
2	1600	3	1	1.3	0.786	1 ✅
3	1700	4	1	1.2	0.768	1 ✅
4	1875	3	1	1.575	0.828	1 ✅
5	1100	2	0	-0.4	0.401	0 ✅
6	1550	4	0	1.05	0.741	1 ❌
7	2350	4	1	2.55	0.927	1 ✅
8	2450	5	1	3.45	0.969	1 ✅
9	1425	3	0	0.775	0.685	1 ❌
10	1700	3	1	0.3	0.574	1 ✅

Step 6: Confusion Matrix

	Actual 1	**Actual 0**
Predicted 1	6	3
Predicted 0	1	0

- **TP** = 6, **FP** = 3
- **FN** = 1, **TN** = 0

Step 7: Evaluation Metrics

✅ Accuracy

$\text{Accuracy} = \frac{TP + TN}{Total} = \frac{6 + 0}{10} = 0.6$

✅ Precision

$\text{Precision} = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667$

✅ Recall

$\text{Recall} = \frac{TP}{TP + FN} = \frac{6}{6 + 1} = 0.857$

✅ F1 Score

$F1 = 2 \cdot \frac{0.667 \cdot 0.857}{0.667 + 0.857} \approx 0.75$

Final Model Summary

- **Logistic Regression Equation**:
  $P(\text{Sold}) = \frac{1}{1 + e^{-(-7.5 + 0.003X_1 + 0.9X_2)}}$

- **Evaluation Metrics**:

  - Accuracy: **60%**
  - Precision: **66.7%**
  - Recall: **85.7%**
  - F1 Score: **75%**

Example Prediction

- **2000 sqft, 4 bedrooms:**

$z = -7.5 + 0.003 \cdot 2000 + 0.9 \cdot 4 = -7.5 + 6 + 3.6 = 2.1$ $P = \frac{1}{1 + e^{-2.1}} \approx 0.891 \Rightarrow \text{Likely to Sell in 30 Days (✓)}$