Problem Statement

Title: *Predicting Student Exam Outcome Using Logistic Regression*

Objective

To develop a **logistic regression model** that predicts whether a student will **pass or fail** an exam based on the **number of hours studied**.

Background

Educational researchers and instructors often seek to understand how study habits impact student success. This project uses logistic regression to model the relationship between the number of hours studied and the probability of passing an exam. Unlike linear regression, logistic regression is suitable here because the target variable is **binary**: pass (1) or fail (0).

Dataset

The dataset contains two variables:

- **X (Hours Studied)** — a continuous variable representing the number of hours a student studied before the exam.
- **Y (Exam Result)** — a binary variable:
  - `1` = Pass
  - `0` = Fail

Sample Data:

Hours Studied (X)	Result (Y)
1.0	0
2.0	0
3.0	0
4.0	0
5.0	1
6.0	1
7.0	1
8.0	1
9.0	1
10.0	1

Goals

- Fit a **logistic regression model** to estimate the probability of passing based on study time.
- Interpret the model parameters to understand the effect of study hours on passing likelihood.
- Classify students as "pass" or "fail" using a decision threshold (e.g., 0.5).
- Evaluate the model using:
  - **Accuracy**
  - **Confusion Matrix**
  - **Precision, Recall, F1 Score**
  - **Log Loss**

Assumptions

- There is a **logistic (sigmoidal)** relationship between hours studied and the probability of passing.
- Observations are independent.
- The response variable is **binary**.

Solution

Problem

Predict whether a student will **pass or fail** based on how many hours they studied, using **logistic regression**.

Step 1: Dataset

Hours Studied (X)	Result (Y)
1.0	0
2.0	0
3.0	0
4.0	0
5.0	1
6.0	1
7.0	1
8.0	1
9.0	1
10.0	1

- **X**: Independent variable (hours studied)
- **Y**: Dependent binary variable (0 = Fail, 1 = Pass)

Step 2: Logistic Regression Model

We fit a model of the form:

$$ P(Y = 1 \mid X) = \frac{1}{1 + e^{-(b_0 + b_1 X)}} $$

Let's assume the model was trained using maximum likelihood estimation and yielded:

$$ b_0 = -7, \quad b_1 = 1.2 $$

So, the final model:

$$ P(\text{Pass}) = \frac{1}{1 + e^{-(-7 + 1.2X)}} = \frac{1}{1 + e^{7 - 1.2X}} $$

Step 3: Make Predictions

Calculate the predicted probabilities and binary classifications for each student:

X	Y (Actual)	$(P(\text{Pass}))$	Predicted Y (Threshold = 0.5)
1	0	0.00091	0
2	0	0.00247	0
3	0	0.00669	0
4	0	0.018	0
5	1	0.047	0 ❌
6	1	0.119	0 ❌
7	1	0.269	0 ❌
8	1	0.537	1 ☑
9	1	0.802	1 ☑
10	1	0.943	1 ☑

Step 4: Confusion Matrix

	Actual 1	Actual 0
Predicted 1	3	0
Predicted 0	3	4

- **TP** = 3
- **FP** = 0
- **FN** = 3
- **TN** = 4

Step 5: Evaluation Metrics

Accuracy

$ \text{Accuracy} = \frac{TP + TN}{Total} = \frac{3 + 4}{10} = 0.7 $

Precision

$ \text{Precision} = \frac{TP}{TP + FP} = \frac{3}{3 + 0} = 1.0 $

Recall (Sensitivity)

$ \text{Recall} = \frac{TP}{TP + FN} = \frac{3}{3 + 3} = 0.5 $

F1 Score

$ F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}} = 2 \cdot \frac{1.0 \cdot 0.5}{1.5} = 0.667 $

Log Loss (Simplified Calculation Example)

$ \text{Log Loss} = -\frac{1}{n} \sum \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] $

For accurate computation, plug in all (y_i) and (p_i). Example for first value:

$ y = 0,\ p = 0.00091 \Rightarrow -\log(1 - 0.00091) \approx 0.00091 $

Summing all and averaging gives approximate log loss:
$ \text{Log Loss} \approx 0.31 $

Final Summary

- **Model Equation:**
  $ P(\text{Pass}) = \frac{1}{1 + e^{-(-7 + 1.2X)}} $
- **Accuracy:** 70%
- **Precision:** 100%
- **Recall:** 50%
- **F1 Score:** 0.667
- **Log Loss (approx.):** 0.31
- **Prediction Example:**
  A student who studies **8 hours** has a **53.7%** chance of passing.