# Skewed Datasets

**Accuracy không còn đáng tin cậy khi dữ liệu bị lệch**, đặc biệt với bài toán **phát hiện sự kiện hiếm (rare class)**.

### **Confusion Matrix** – Ma trận nhầm lẫn (2x2)

| | **Actual: Positive (1)** | **Actual: Negative (0)** |
|---|---|---|
| **Predicted: 1** | ✅ **True Positive (TP)** | ❌ **False Positive (FP)** |
| **Predicted: 0** | ❌ **False Negative (FN)** | ✅ **True Negative (TN)** |

### **Precision** (Độ chính xác): Trong số các trường hợp mô hình dự đoán là dương tính, thì bao nhiêu phần trăm là đúng?

$$\text{Precision} = \frac{TP}{TP + FP}$$

### **Recall** (Độ bao phủ / Tỷ lệ phát hiện): Trong số các ca thực sự là dương tính, thì mô hình phát hiện được bao nhiêu phần trăm?

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

# Trade off giữa Precision và Recall

Chúng ta luôn muốn mô hình có:

- **Precision cao**: Khi mô hình dự đoán có bệnh (y=1), thì **thật sự** có khả năng cao là đúng.
- **Recall cao**: Trong tất cả các trường hợp thật sự có bệnh, mô hình phát hiện được **nhiều nhất có thể**.

Tuy nhiên, **trong thực tế**, bạn phải **chọn điểm cân bằng** giữa precision và recall tùy theo từng tình huống cụ thể.

## F1 Score — Chỉ số kết hợp Precision và Recall

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Hoặc tương đương:

$$F_1 = \frac{1}{\frac{1}{P} + \frac{1}{R}} \cdot 2$$

- Đây là **harmonic mean** (trung bình điều hòa), chú trọng vào giá trị thấp hơn.
- Nếu **Precision hoặc Recall rất thấp**, thì F1 cũng sẽ rất thấp → giúp tránh chọn mô hình "lệch".