Let’s break it down further with a **detailed example** to understand **precision**, **recall**, and **F1-score**, especially when dealing with imbalanced datasets.

Scenario:

Imagine you're working with a **medical diagnostic model** to detect a rare disease (Class 1) in a population of **1000 people**. Out of these 1000 people, only **50 people** have the disease (Class 1), and the remaining **950 people** are healthy (Class 0).

You train the model and it gives the following **confusion matrix**:

	Predicted Class 0 (Healthy)	Predicted Class 1 (Disease)
Actual Class 0 (Healthy)	940	10
Actual Class 1 (Disease)	30	20

This matrix means:

- **True Negatives (TN)**: The model correctly predicted **940 healthy** people as healthy (Class 0).
- **False Positives (FP)**: The model incorrectly predicted **10 healthy people** as having the disease (Class 1).
- **False Negatives (FN)**: The model missed **30 people with the disease**, predicting them as healthy (Class 0).
- **True Positives (TP)**: The model correctly predicted **20 people with the disease**.

Step-by-Step Calculation of Precision, Recall, and F1-Score:

1. Precision

- **Precision** tells us how many of the instances predicted as positive (disease) are actually positive.
- Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- In this case:

$$\text{Precision} = \frac{20}{20 + 10} = \frac{20}{30} = 0.67$$

This means **67%** of the instances the model predicted as having the disease (Class 1) are actually correct.

2. Recall (Sensitivity or True Positive Rate)

- **Recall** tells us how many of the actual positives (people who really have the disease) were correctly predicted as positive.
- Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- In this case:

$$\text{Recall} = \frac{20}{20 + 30} = \frac{20}{50} = 0.40$$

This means **40%** of the people who actually have the disease were correctly identified by the model. The model **missed 60%** of the actual disease cases, which is a significant problem in a medical context.

3. F1-Score

- The **F1-Score** is the **harmonic mean** of precision and recall. It is useful when you need a balance between precision and recall. It's particularly helpful when both false positives and false negatives are important.
- Formula:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- In this case:

$$F1 = 2 \times \frac{0.67 \times 0.40}{0.67 + 0.40} = 2 \times \frac{0.268}{1.07} = 0.50$$

So, the **F1-score** is **0.50**, which gives a single metric balancing both precision and recall.

Why Accuracy Is Not Enough:

- **Accuracy** is the proportion of correct predictions to total predictions.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

- In this case:

$$\text{Accuracy} = \frac{20 + 940}{1000} = \frac{960}{1000} = 0.96$$

So, the accuracy is **96%**, which sounds great, but here's the issue: The model is heavily biased toward predicting **Class 0 (Healthy)** because the dataset is imbalanced. In fact, the model could simply predict all instances as healthy (Class 0), and it would still achieve **94% accuracy** (because 940 out of 1000 are healthy).

- The model **misses 30% of people who actually have the disease** (high false negatives) but still has a high **accuracy**. This demonstrates why **accuracy** isn't a reliable metric for **imbalanced**

datasets.

When to Use Precision, Recall, and F1-Score:

1. **Precision**:
    - Use **precision** when **false positives** are more costly or undesirable.
    - For example, in spam email detection, you don't want to wrongly classify important emails as spam (false positives), so you care more about precision.
2. **Recall**:
    - Use **recall** when **false negatives** are more important.
    - For example, in **medical diagnosis**, you don’t want to miss any actual cases of disease (false negatives), even if that means predicting some healthy people incorrectly (false positives).
3. **F1-Score**:
    - Use **F1-score** when you want to balance precision and recall. This is especially useful when both false positives and false negatives are important, as in our medical example.

Summary with Example:

In an imbalanced dataset:

- **Accuracy** might give a false impression of model performance.
- **Precision** tells you how good the model is at predicting the positive class.
- **Recall** tells you how good the model is at identifying all the positive cases.
- **F1-Score** provides a balance between precision and recall.

In the medical example, even though the model has **96% accuracy**, the **F1-score** is only **0.50**, which indicates the model is not performing well in detecting the disease.