Precision, Recall, and F1 Score | Classification Metrics

➢ When is Accuracy Misleading?

Accuracy simply measures the proportion of correct predictions. However, it can be misleading when data is **imbalanced**—for example, if 95% of cases belong to one class, predicting everything as that class will still yield 95% accuracy but be useless for detecting the minority class.

➢ Precision

Definition: Out of all the predictions the model made for a certain class, how many were actually correct?

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

Intuition: High precision means **few false alarms**.
Use case: When the cost of a **false positive** is high (e.g., diagnosing a healthy person as sick).

➢ Recall

Definition: Out of all actual cases of a certain class, how many did the model correctly identify?

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

Intuition: High recall means **catching most of the real cases**.
Use case: When missing a positive case is very costly (e.g., failing to detect cancer).

➢ F1 Score

Definition: The harmonic mean of Precision and Recall.

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

Why harmonic mean? It punishes extreme imbalance between precision and recall—both need to be high for a good F1 score.
Use case: When you need a balance between catching positives and avoiding false alarms.

➢ Multi-Class Precision and Recall

- **Macro averaging:** Treats all classes equally (good when class imbalance is not a concern).
- **Micro averaging:** Aggregates contributions of all classes (good for imbalanced data).
- **Weighted averaging:** Like macro but considers class size.

➢ MNIST Code Example

Applied these metrics to the **MNIST handwritten digit dataset**, demonstrating precision, recall, and F1 score in a real-world multi-class classification task.