Evaluating Model Performance with Metrics Like Accuracy, Precision, Recall, and F1 Score

Evaluating machine learning models is crucial to understand their **effectiveness and reliability**. Different metrics provide insights into how well a model performs, especially in classification tasks.

---

1. Accuracy

- Measures the proportion of **correct predictions** over total predictions.
- Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **TP:** True Positives, **TN:** True Negatives, **FP:** False Positives, **FN:** False Negatives.
- Best for **balanced datasets**, but can be misleading with imbalanced classes.

---

2. Precision

- Measures how many of the predicted positive instances are **actually positive**.
- Formula:

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

- High precision means **low false positive rate**.

---

3. Recall (Sensitivity or True Positive Rate)

- Measures how many of the **actual positive instances** are correctly predicted.
- Formula:

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

- High recall means **low false negative rate**.

---

4. F1 Score

- Harmonic mean of precision and recall, balancing both metrics.
- Formula:

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

- Useful for **imbalanced datasets** where accuracy alone is insufficient.

Python Example Using scikit-learn:

```python
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Sample true and predicted labels
y_true = [1, 0, 1, 1, 0, 1, 0, 0, 1, 0]
y_pred = [1, 0, 1, 0, 0, 1, 0, 0, 1, 1]

# Calculate metrics
accuracy = accuracy_score(y_true, y_pred)
precision = precision_score(y_true, y_pred)
recall = recall_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

Key Points

- **Accuracy** is simple but can be misleading with imbalanced classes.
- **Precision** is important when **false positives** are costly (e.g., spam detection).
- **Recall** is crucial when **false negatives** are costly (e.g., disease diagnosis).
- **F1 Score** provides a balanced metric when **both precision and recall matter**.

Using these metrics together provides a **comprehensive evaluation** of classification model performance, helping identify strengths and weaknesses.