Classification Evaluation Metrics

When training a classification model, we need to *measure its performance*. We can not rely only on accuracy because real-world problems often involve imbalanced datasets. Instead, we use *multiple evaluation metrics* to *analyze different aspects* of the model's performance.

Confusion Matrix

A **confusion matrix** is a table that summarizes the performance of a classification model by comparing actual vs. predicted values.

Actual / Predict	Predicted Negative (0)	Predicted Positive (1)
Actual Negative (0)	True Negative (TN)	False Positive (FP)
Actual Positive (1)	False Negative (FN)	True Positive (TP)

Definitions

- *True Positive (TP)* - The model correctly predicts a positive case
- *True Negative (TN)* - The model correctly predicts a negative case.
- *False Positive (FP) (Type I Error)* - The model incorrectly predicts a positive case when it's actually negative (a false alarm).
- *False Negative (FN) (Type II Error)* - The model incorrectly predicts a negative case when it's actually positive (a missed detection)

Example

Imagine a medical test for detecting COVID-19:

- **TP** - The test correctly detects an infected person.
- **TN** - The test correctly identifies a healthy person.
- **FP** - The test says a healthy person has COVID-19 (false alarm).
- **FN** - The test fails to detect an infected person (missed case).

Why is the confusion matrix useful?

- It helps us see different types of errors (FP & FN).
- It is the foundation for all other evaluation metrics.

Accuracy

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

Accuracy measures the proportion of correctly predicted cases out of all cases.

Example

If we have 100 patients and our model correctly classifies 90 (both positive and negative), the accuracy is:

$$\frac{90}{100} = 90\%$$

When to use Accuracy?

- *Good for balanced datasets* (equal positives and negatives).
- *Not good for imbalanced datasets* (e.g., fraud detection, where 99% are non-fraud).

Example of a problem with accuracy

If 99% of transactions are non-fraudulent, a model that always predicts "no fraud" will have 99% accuracy but it’s useless because it never detects fraud!

Precision (Positive Predictive Value)

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

Precision measures how many predicted positive cases are actually positive.

Example

- Suppose we predict **100 people as having COVID-19**.
- Out of those, **80 actually have** COVID-19, and **20 are false positives**.

$\Rightarrow \text{Precision} = \frac{80}{80+20} = 80\%$

When to use Precision?

- When **false positives are costly**.
- Examples:
  - Spam detection because we don't want many good emails marked as spam.
  - Fraud detection because a false positive means blocking a real customer's card.

Recall (Sensitivity / True Positive Rate)

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

Recall measures how many actual positives were correctly identified.

Example

- Suppose **100 people have COVID-19**.
- The model detects **80**, but **misses 20 cases**.

$\Rightarrow \text{Recall} = \frac{80}{80+20} = 80\%$

When to use Recall?

- When **false negatives are costly**.
- Examples:
  - Legal systems because falsely convicting an innocent person.
  - Drug testing because falsely flagging athletes for doping.

Specificity (True Negative Rate)

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Specificity measures how many actual negatives were correctly identified.

Example

- Suppose **100 people don’t have COVID-19**.
- The model correctly identifies **90 as negative** but **wrongly classifies 10 as positive**.

$\Rightarrow \text{Specificity} = \frac{90}{90+10} = 90\%$

F1 Score (Balance between Precision and Recall)

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

The **F1 Score** is the **harmonic mean** of Precision and Recall, balancing both.

Example

If *Precision* = 80% and *Recall* = 60%

$$\text{F1 Score} = 2 \times \frac{0.8 \times 0.6}{0.8 + 0.6} = 68.5\%$$

When to use F1 Score?

- When both **false positives** and **false negatives matter**.
- Examples:
  - Chatbots as detecting user intent without too many wrong classifications.
  - Medical AI balancing false positives & false negatives.

ROC Curve & AUC Score

The **ROC (Receiver Operating Characteristic) Curve** plots *True Positive Rate (Recall)* vs. *False Positive Rate (1 - Specificity)* at different classification thresholds.

The **AUC (Area Under Curve) Score** measures the overall model performance.

Interpretation of AUC Score

AUC = 1 $\rightarrow$ Perfect model

AUC = 0.5 $\rightarrow$ Random guessing (useless model)

AUC < 0.5 → Worse than random (bad model)

When to use ROC-AUC?

- When **choosing the best classification threshold**.
- Examples:
  - Medical tests. Should we label a patient as "positive" at 90% confidence or 70% confidence?
  - Credit scoring as adjusting fraud detection sensitivity.

Final Summary

Metric	Measures	Best Use Case
Accuracy	Overall correctness	Balanced datasets
Precision	Correctly predicted positives	Avoiding false positives
Recall	Correctly detected actual positives	Avoiding false negatives
Specificity	Correctly detected actual negatives	Avoiding false positives
F1 Score	Balance between Precision & Recall	When both errors matter and for imbalanced data
ROC-AUC	Model performance at different thresholds	Comparing different models