Evaluation Metrics in Machine Learning

Accuracy, Precision, Recall, and F1-Score

Introduction to Evaluation Metrics

Evaluation metrics are used to measure how well a machine learning model performs on unseen data. After training a model, it is essential to evaluate its predictions to understand its effectiveness, reliability, and real-world usability.

In classification problems, especially in NLP tasks such as sentiment analysis and text classification, evaluation metrics help us determine whether the model predictions are meaningful and correct.

Commonly used evaluation metrics include: - Accuracy - Precision - Recall - F1-Score

Confusion Matrix – The Foundation

Most classification metrics are derived from the confusion matrix. A confusion matrix summarizes prediction results by comparing actual labels with predicted labels.

For binary classification:

- True Positive (TP): Correctly predicted positive cases
- True Negative (TN): Correctly predicted negative cases
- False Positive (FP): Incorrectly predicted positive cases
- False Negative (FN): Incorrectly predicted negative cases

Understanding these four values is essential for calculating all evaluation metrics.

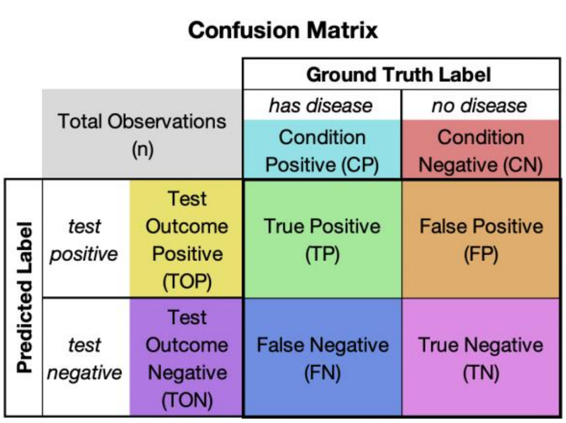
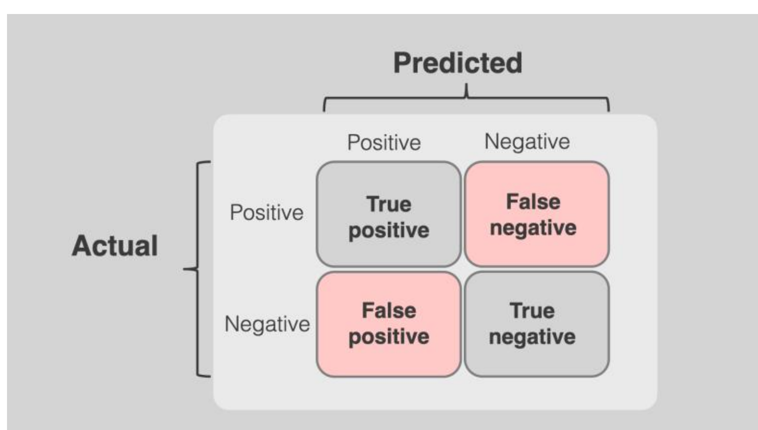


Figure 1: Basic colour coded confusion matrix with marginal sums



Accuracy

Definition

Accuracy measures the proportion of correct predictions out of all predictions made by the model.

Formula

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Interpretation

Accuracy indicates overall correctness of the model.

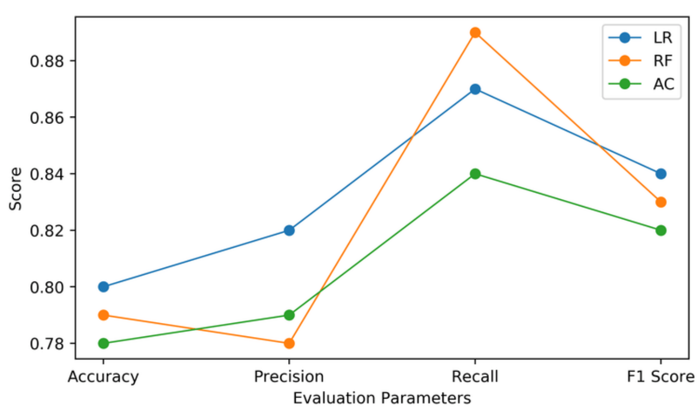
Advantages

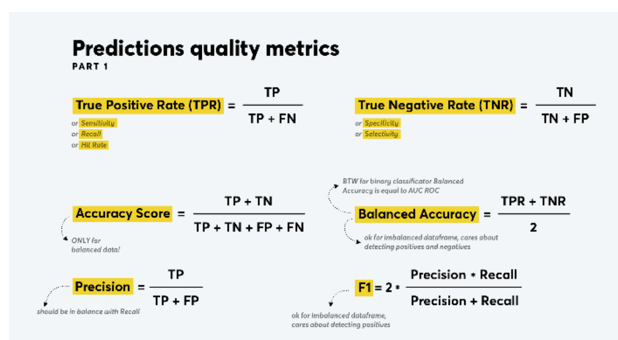
- Easy to understand
- Useful when classes are balanced

Limitations

- Misleading for imbalanced datasets
- Does not distinguish between types of errors

Example: If a model correctly predicts 90 out of 100 samples, accuracy = 90%.





Precision

Definition

Precision measures how many of the predicted positive cases are actually positive.

Formula

Precision = TP / (TP + FP)

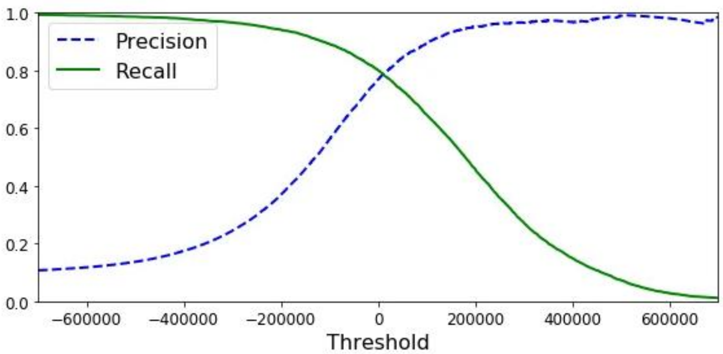
Interpretation

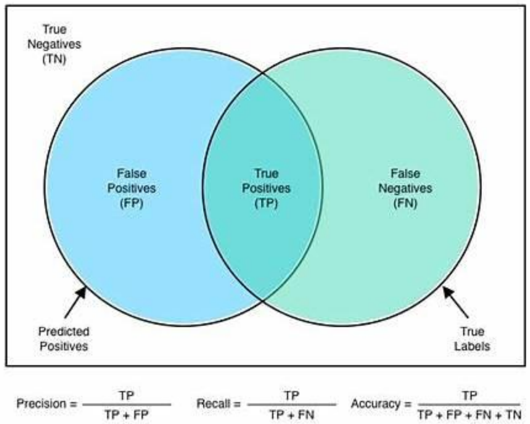
High precision means fewer false positives.

When to Use

- Spam detection
- Medical diagnosis where false positives are costly

Precision focuses on prediction quality rather than completeness.





Recall (Sensitivity)

Definition

Recall measures how many actual positive cases are correctly identified by the model.

Formula

Recall = TP / (TP + FN)

Interpretation

High recall means fewer false negatives.

When to Use

- Disease detection
- Fraud detection

Recall focuses on capturing all relevant instances.

Precision vs Recall Trade-Off

Precision and recall often have an inverse relationship.

- Increasing precision may decrease recall
- Increasing recall may decrease precision

This trade-off depends on the classification threshold and problem requirements.

Example: - Email spam filtering prefers high precision - Cancer detection prefers high recall

Choosing the right balance is critical.

F1-Score

Definition

F1-score is the harmonic mean of precision and recall. It balances both metrics into a single score.

Formula

F1 = 2 × (Precision × Recall) / (Precision + Recall)

Interpretation

- High F1-score indicates good balance
- Useful for imbalanced datasets

F1-score is widely used in NLP tasks.

Accuracy vs F1-Score

Metric	Accuracy	F1-Score
Dataset Type	Balanced	Imbalanced
Focus	Overall correctness	Error balance
Sensitivity to imbalance	High	Low

Accuracy may look high even when minority class performance is poor, whereas F1-score provides a more realistic evaluation.

Best Practices

Best Practices

- Always analyze confusion matrix
- Use F1-score for imbalanced datasets
- Do not rely on accuracy alone
- Choose metrics based on business impact