# Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It provides a detailed breakdown of correct and incorrect predictions for each class.

<table>
<tr><td colspan="2"></td><td colspan="2">Predicted class</td><td></td></tr>
<tr><td colspan="2"></td><td>Classified positive</td><td>Classified negative</td><td></td></tr>
<tr><td rowspan="2">Actual class</td><td>Actual positive</td><td>TP</td><td>FN</td><td>TPR: $\frac{TP}{TP + FN}$</td></tr>
<tr><td>Actual negative</td><td>FP</td><td>TN</td><td>FPR: $\frac{TN}{TN + FP}$</td></tr>
<tr><td colspan="2"></td><td>Precision: $\frac{TP}{TP + FP}$</td><td colspan="2">Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$</td></tr>
</table>

Where:

- **TP (True Positives)**: Correctly predicted positive cases
- **TN (True Negatives)**: Correctly predicted negative cases
- **FP (False Positives)**: Incorrectly predicted as positive (Type I error)
- **FN (False Negatives)**: Incorrectly predicted as negative (Type II error)

**Key Benefits:**

- Shows exactly where the model is making mistakes
- Reveals class-specific performance patterns
- Foundation for calculating other metrics
- Helps identify if model has bias toward certain classes

# Precision, Recall, F1-Score

## Precision

**Formula:** TP / (TP + FP)

Precision answers: "Of all the positive predictions made, how many were actually correct?"

- High precision = Low false positive rate
- Critical when the cost of false positives is high
- Example: Email spam detection (you don't want important emails marked as spam)

### Recall (Sensitivity)

**Formula:** TP / (TP + FN)

Recall answers: "Of all the actual positive cases, how many did we correctly identify?"

- High recall = Low false negative rate
- Critical when the cost of missing positives is high
- Example: Medical diagnosis (you don't want to miss actual diseases)

### F1-Score

**Formula:** 2 × (Precision × Recall) / (Precision + Recall)

The F1-score is the harmonic mean of precision and recall:

- Balances both precision and recall
- Useful when you need a single metric that considers both
- Particularly valuable with imbalanced datasets
- Ranges from 0 to 1, where 1 is perfect

**Precision-Recall Tradeoff:** There's typically an inverse relationship between precision and recall. Adjusting the classification threshold can shift this balance:

- Lower threshold → Higher recall, lower precision
- Higher threshold → Higher precision, lower recall

# When Accuracy is Misleading

Accuracy = (TP + TN) / (TP + TN + FP + FN) = Correct Predictions / Total Predictions

**Problem Scenarios:**

### 1. Biased Datasets

**Example:** Disease detection where 95% of patients are healthy

- A model that always predicts "healthy" achieves 95% accuracy
- But it has 0% recall for actually detecting the disease
- Accuracy masks the model's complete failure at its primary task

### 2. Unequal Misclassification Costs

**Example:** Fraud detection

- Missing fraud (false negative) costs much more than flagging legitimate transaction (false positive)
- High accuracy might hide poor performance on the critical minority class

## 3. Multi-class Imbalance

**Example:** Text classification with 10 categories where one category represents 80% of data

- Model might perform well on the dominant class but poorly on others
- Overall accuracy appears good but specific class performance is poor

## 4. Context-Dependent Performance Requirements

**Example:** Medical screening vs. final diagnosis

- Screening: High recall crucial (don't miss cases)
- Final diagnosis: High precision crucial (don't overtreat)
- Same accuracy could represent very different utility

**Better Alternatives:**

- Use precision, recall, and F1-score for imbalanced datasets
- Consider area under ROC curve (AUC-ROC) for threshold-independent evaluation
- Examine per-class metrics separately
- Use domain-specific cost functions when misclassification costs vary
- Consider balanced accuracy: (Sensitivity + Specificity) / 2