How to check classifiers… the measures taken.

1. Matthews Correlation Coefficient (MCC)

- Measures the **quality of binary classifications**
- Takes into account **all four** confusion matrix categories (TP, TN, FP, FN)
- Range: **-1 (worst)** to **+1 (perfect)**; 0 = random
- Good for **imbalanced datasets**

2. Cohen's Kappa

- Measures **agreement** between predicted and actual labels, adjusted for **chance agreement**
- Especially useful when evaluating **human vs model agreement**

3. Balanced Accuracy

- Average of **sensitivity (recall)** and **specificity**
- Good when classes are **imbalanced**

4. Log Loss (Cross-Entropy Loss)

- Used when model predicts **probabilities** (not just classes)
- Penalizes **confident but wrong** predictions heavily
- Lower = better

5. Brier Score

- Measures **probability accuracy** (how close your predicted probabilities are to true labels)
- Lower = better
- Often used in **calibration** of classifiers

6. G-Mean (Geometric Mean)

- Geometric mean of **sensitivity** and **specificity**
- Used for **imbalanced datasets** to ensure both classes are treated fairly

7. Top-K Accuracy (Top-1, Top-5, etc.)

- Common in **image classification** (e.g., ImageNet)
- "Was the correct label in the top K predictions?"

8. Average Precision (AP) / Mean Average Precision (mAP)

- Used in **object detection, ranking**, and **retrieval**

- Average of precisions at different thresholds

9. Precision@K, Recall@K

- Popular in **recommender systems** or **search engines**
- How many of the top-K recommendations were actually relevant?

10. Hamming Loss

- Used in **multi-label classification**
- Fraction of wrong labels to total labels

Use Case	Recommended Metrics
Imbalanced datasets	MCC, F1, Balanced Accuracy, G-Mean
Probabilistic predictions	Log Loss, Brier Score, ROC-AUC
Multi-class classification	Macro/Micro F1, Top-K Accuracy, Cohen’s Kappa
Ranking or recommendations	Precision@K, mAP, Recall@K
Multi-label classification	Hamming Loss, Subset Accuracy, Jaccard Score

Metric	Formula	Think of it as...	Easy Mnemonic 💡
Accuracy	(TP + TN) / (TP + FP + FN + TN)	How often your prediction is **overall correct**	“How many total correct”
Precision	TP / (TP + FP)	Of all predicted positives, how many were **actually correct**	“Trust your **positive**s”
Recall (Sensitivity)	TP / (TP + FN)	Of all real positives, how many did you **catch**?	“Catch all real cases”
Specificity	TN / (TN + FP)	Of all real negatives, how many did you **correctly ignore**?	“Ignore negatives well”
F1 Score	2 × (Precision × Recall) / (Precision + Recall)	Balances **precision and recall** (harmonic mean)	“Trade-off score”
ROC Curve	Plot of TPR vs. FPR	Visual curve showing model performance at all thresholds	“Curve for classifier quality”
AUC	Area under ROC	One number summarizing the ROC curve — 1 is best	“Bigger area = better”