Binary Classification Metrics Exercises

Exercise 1 — Basic Metrics

A classifier produces the following confusion matrix for a test set of 100 samples:

	Predicted Positive	**Predicted Negative**
Actual Positive	40	10
Actual Negative	20	30

Tasks:

1. Compute **Accuracy**, **Precision**, **Recall (Sensitivity)**, **Specificity**, and **F1-score**.
2. Interpret what each metric tells you about the model's performance.
3. Discuss how increasing the classification threshold might affect precision and recall.

Exercise 2 — Comparing Two Models

Two classifiers (A and B) yield the following confusion matrices on the same dataset of 200 samples:

Model A	**TP=70**	**FP=30**	**FN=20**	**TN=80**
Model B	TP=90	FP=50	FN=10	TN=50

Tasks:

1. Compute the precision, recall, F1-score, and accuracy for both models.
2. Which model would you prefer if **false negatives** are more costly (e.g., disease detection)?
3. Which would you prefer if **false positives** are more costly (e.g., fraud detection)?
4. Which has better overall balance (based on F1)?

Exercise 3 — ROC and AUC (Manual Points)

A binary classifier outputs the following **predicted probabilities** and true labels for 8 samples:

Sample	True Label	Predicted Probability
1	1	0.95
2	0	0.90
3	1	0.85
4	0	0.70
5	1	0.60
6	0	0.40
7	1	0.20
8	0	0.10

Tasks:

1. Compute the **TPR (Recall)** and **FPR** at thresholds = {0.9, 0.8, 0.6, 0.4, 0.2, 0.1}.
2. Plot the **ROC curve** (TPR vs FPR).
3. Estimate the **AUC (Area Under Curve)** using the trapezoidal rule.
4. Interpret what the AUC value means in terms of model performance.

Exercise 4 — Threshold Sensitivity

A medical test gives the following results for 10 patients:

Patient	True	Score
P1	1	0.95
P2	1	0.92
P3	0	0.89
P4	1	0.75
P5	0	0.72
P6	1	0.65
P7	0	0.55
P8	0	0.40
P9	1	0.35
P10	0	0.10

Tasks:

1. For thresholds **0.5** and **0.8**, compute confusion matrices.
2. For each threshold, compute **precision, recall, and F1-score**.

3. Plot how precision and recall change as the threshold increases (Precision–Recall trade-off).
4. Explain why AUC-PR may be more informative than AUC-ROC for imbalanced datasets.

Exercise 5 — Imbalanced Dataset Scenario

Out of 10,000 emails:

- 200 are spam (positive class)
- 9,800 are not spam (negative class)

A model flags 300 emails as spam, correctly identifying 150 of the 200 spams.

Tasks:

1. Compute precision, recall, and F1-score.
2. Compute accuracy.
3. Explain why **accuracy** can be misleading in this scenario.
4. Discuss which metric would be more informative (and why).