Exercise 1:

Part 1: Confusion Matrix and Accuracy

We are building models to detect a **rare disease** (**1 = sick (positive)** and **0 = healthy (negative)**). To analyze classification performance in detail, we use the confusion matrix. Below is an empty confusion matrix for a binary classification problem:

	Actual Positive	**Actual Negative**
Predicted Positive	(A)	(B)
Predicted Negative	(C)	(D)

(a) Assign the correct terms (TP, FP, FN, TN) to fields (A), (B), (C), (D) in the confusion matrix above and briefly explain what each term represents.

(b) Look at the row sums and column sums of the confusion matrix. What do these sums represent?

(c) Express the accuracy metric using TP, FP, FN, and TN.

(d) For each of the 8 metrics in the table below, categorize which question (Question 1 or Question 2) it answers, and write a one-line interpretation of what the metric measures. Many metrics naturally group into two categories based on what question they answer:

- **Question 1:** "Among all actual positives/negatives, how many did we correctly/incorrectly classify?" (Metrics condition on real class: denominators positives (P) or negatives (N))
- **Question 2:** "Among all predicted positives/negatives, how many were actually correct/incorrect?" (Metrics condition on predictions: denominators predicted positives (PP) or predicted negatives (PN))

Metric	**Formula**
Precision (Positive Predictive Value, PPV)	$\frac{TP}{TP+FP}$
False Positive Rate (FPR, Fallout)	$\frac{FP}{FP+TN}$
True Negative Rate (TNR, Specificity)	$\frac{TN}{TN+FP}$
False Omission Rate (FOR)	$\frac{FN}{TN+FN}$
True Positive Rate (TPR, Recall, Sensitivity)	$\frac{TP}{TP+FN}$
False Discovery Rate (FDR)	$\frac{FP}{TP+FP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$
False Negative Rate (FNR, Miss Rate)	$\frac{FN}{TP+FN}$

Part 2: Metrics Derived from the Confusion Matrix

Now we apply this to actual data. We have 10 patients with true labels and predictions from two models:

Patient ID	1	2	3	4	5	6	7	8	9	10
True label y	0	0	0	1	0	0	0	0	0	0
Model A	1	0	0	0	0	0	0	0	0	0
Model B	0	1	0	1	0	0	1	0	1	0

(a) For both models, compute the accuracy. Based only on accuracy, which model seems better?

(b) For both models, compute: TPR, TNR, Precision, and NPV. For each of these four metrics, identify which model performs better and discuss what this means in practice (in particular for which aspects of the confusion matrix each model performs well or poorly, and which model you would choose in this medical setting).

Part 3: F1 Score (Harmonic Mean of Precision and Recall)

Sometimes we want a single metric that balances precision and recall. We define the F1 score as the harmonic mean of precision and recall:

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

(a) Using your values for Precision and Recall from Part 2, compute the F1 score for both models. What result do you obtain for Model A, and how does the F1 formula behave in this case? Explain your answer.

(b) Why might we prefer the *harmonic* mean instead of the arithmetic mean for combining precision and recall?

Part 4: Thresholds and the Choice of 0.5

So far we have worked with hard predictions (0 or 1). But many models output probability scores instead. The choice of threshold (the cutoff above which we predict class 1) affects the confusion matrix.

Consider Model C with probabilistic outputs for 6 patients (see below). Compute and compare TPR and FPR at thresholds 0.5 and 0.3. Describe how lowering the threshold from 0.5 to 0.3 changes the confusion matrix and the derived metrics, and explain why the choice of threshold is important.

Patient ID	1	2	3	4	5	6
True y	0	1	0	0	1	0
Model C score	0.20	0.40	0.10	0.30	0.35	0.05

Solution 1:

Part 1: Confusion Matrix and Accuracy

(a) Assigning terms to the confusion matrix:

	Actual Positive	**Actual Negative**
Predicted Positive	$(A)TP$	$(B)FP$
Predicted Negative	$(C)FN$	$(D)TN$

- **TP (True Positives)**: Cases where model correctly predicted positive and actual label is positive.
- **FP (False Positives)**: Cases where model incorrectly predicted positive but actual label is negative. Also called false alarms.
- **FN (False Negatives)**: Cases where model incorrectly predicted negative but actual label is positive. Also called misses.
- **TN (True Negatives)**: Cases where model correctly predicted negative and actual label is negative.

(b) Looking at the confusion matrix structure:

- **Column sums:** The first column sum $(TP + FN)$ represents all actual positives (P). The second column sum $(FP + TN)$ represents all actual negatives (N).
- **Row sums:** The first row sum $(TP + FP)$ represents all predicted positives (PP). The second row sum $(FN + TN)$ represents all predicted negatives (PN).

(c) Accuracy represents the proportion of all cases that were correctly classified:
Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{P+N}$

(d) The row and column sums are often used as denominators in metrics and can be understood as the different "universes" we condition on when calculating these metrics. Simplified forms of each metric using the universes P, N, PP, and PN:

Metric	Formula	Simplified
Precision (Positive Predictive Value, PPV)	$\frac{TP}{TP+FP}$	$\frac{TP}{PP}$
False Positive Rate (FPR, Fallout)	$\frac{FP}{FP+TN}$	$\frac{FP}{N}$
True Negative Rate (TNR, Specificity)	$\frac{TN}{TN+FP}$	$\frac{TN}{N}$
False Omission Rate (FOR)	$\frac{FN}{TN+FN}$	$\frac{FN}{PN}$
True Positive Rate (TPR, Recall, Sensitivity)	$\frac{TP}{TP+FN}$	$\frac{TP}{P}$
False Discovery Rate (FDR)	$\frac{FP}{TP+FP}$	$\frac{FP}{PP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$	$\frac{TN}{PN}$
False Negative Rate (FNR, Miss Rate)	$\frac{FN}{TP+FN}$	$\frac{FN}{P}$

Categorization and interpretations:

Question 1 (conditioned on real class): "Among all actual positives/negatives, how many did we correctly/incorrectly classify?" (metrics condition on real class: denominators positives (P) or negatives (N))

- **TPR (Recall, Sensitivity):**
  - Proportion of actual positives correctly identified.
  - Interprets how many sick patients we detect.
- **TNR (Specificity):**
  - Proportion of actual negatives correctly identified.
  - Interprets how many healthy patients we correctly classify as healthy.
- **FPR (Fallout):**
  - Proportion of actual negatives incorrectly classified as positive.
  - Interprets the false alarm rate among healthy patients (healthy people wrongly called sick).
- **FNR (Miss Rate):**
  - Proportion of actual positives incorrectly classified as negative.
  - Interprets how many sick patients we miss (sick people wrongly told they are healthy).

Question 2 (conditioned on predictions): "Among all predicted positives/negatives, how many were actually correct/incorrect?" (metrics condition on predictions: denominators predicted positives (PP) or predicted negatives (PN))

- **Precision (PPV):**
  - Proportion of positive predictions that are correct.
  - Interprets how trustworthy our positive predictions are (if we say a patient is sick, what is the chance they really are sick?).
- **NPV:**
  - Proportion of negative predictions that are correct.
  - Interprets how trustworthy our negative predictions are (if we say a patient is healthy, what is the chance they are really healthy?).
- **FDR:**
  - Proportion of positive predictions that are incorrect.
  - Interprets the false discovery rate among our positive predictions (if we say a patient is sick, how often are we wrong?).
- **FOR:**
  - Proportion of negative predictions that are incorrect.
  - Interprets the false omission rate among our negative predictions (if we say a patient is healthy, how often are we wrong and miss a sick patient?).

Part 2: Metrics Derived from the Confusion Matrix

(a) Computing accuracy:

Model A: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+8}{0+8+1+1} = \frac{8}{10} = 0.8$

Model B: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{1+6}{1+6+3+0} = \frac{7}{10} = 0.7$

Based on accuracy alone, Model A appears better (0.8 vs 0.7).

(b) Computing metrics and comparing models:

Model A:

		Predicted	
		1	0
Actual	1	0 (TP)	1 (FN)
	0	1 (FP)	8 (TN)

$$\text{TPR} = \frac{TP}{TP+FN} = \frac{0}{0+1} = \frac{0}{1} = 0$$
$$\text{TNR} = \frac{TN}{TN+FP} = \frac{8}{8+1} = \frac{8}{9} \approx 0.889$$
$$\text{Precision} = \frac{TP}{TP+FP} = \frac{0}{0+1} = \frac{0}{1} = 0$$
$$\text{NPV} = \frac{TN}{TN+FN} = \frac{8}{8+1} = \frac{8}{9} \approx 0.889$$

Model B:

		Predicted	
		1	0
Actual	1	1 (TP)	0 (FN)
	0	3 (FP)	6 (TN)

$$\text{TPR} = \frac{TP}{TP+FN} = \frac{1}{1+0} = \frac{1}{1} = 1.0$$
$$\text{TNR} = \frac{TN}{TN+FP} = \frac{6}{6+3} = \frac{6}{9} \approx 0.667$$
$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+3} = \frac{1}{4} = 0.25$$
$$\text{NPV} = \frac{TN}{TN+FN} = \frac{6}{6+0} = \frac{6}{6} = 1.0$$

Model comparison and practical implications:

- **TPR (Recall) – catching sick patients:**
  - Model B is better (1.0 vs 0): it catches all sick patients, while Model A misses all.
  - For medical diagnosis, Model B is safer because no sick patients are overlooked.
- **TNR (Specificity) – avoiding false alarms in healthy patients:**
  - Model A is better (0.889 vs 0.667): it correctly identifies more healthy patients.
  - This leads to fewer false alarms and fewer unnecessary medical interventions and costs.
- **Precision – reliability of positive predictions:**
  - Model B is better (0.25 vs 0), though both are low.
  - A positive prediction from Model B is correct 25% of the time, but never for Model A.
  - If Model B predicts someone is sick, there is a 1 in 4 chance they actually are sick.
- **NPV – reliability of negative predictions:**
  - Model B is better (1.0 vs 0.889).
  - Its negative predictions are always correct, while Model A's have a small error rate.
  - Model B's "healthy" predictions are completely reliable; Model A's are slightly less trustworthy.

Overall: Model B is much preferred in medicine, since it never misses a sick patient, even though it makes more false positive errors. Model A is less useful as it completely fails to identify any sick patient.

Part 3: F1 Score (Harmonic Mean of Precision and Recall)

(a) Using your values for Model B, compute F1 for Model B:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot 0.25 \cdot 1.0}{0.25 + 1.0} = \frac{0.5}{1.25} = 0.4$$

(b) Compute F1 Score for Model A. What happens with the formula?

Since both Precision and Recall are 0 for Model A, the F1 formula would result in $\frac{0}{0}$ (undefined). By convention, we define F1 = 0 in this case, since the model has no predictive value for the positive class. (In practice, most implementations set F1 = 0 whenever either precision or recall is 0.)

(c) Why might we prefer the *harmonic* mean instead of the arithmetic mean for combining precision and recall?

The **harmonic mean** is more sensitive to low values than the arithmetic mean: if either precision or recall is low, the F1 score will also be low, which discourages a model from having poor performance in either metric.

- For example, with precision = 0.25 and recall = 1.0:
  - Arithmetic mean: $(0.25 + 1.0)/2 = 0.625$ (may seem acceptable)
  - Harmonic mean (F1): 0.4 (more accurately reflects the imbalance)

F1 is only high when *both* precision and recall are high:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Each false positive (FP) or false negative (FN) directly lowers F1, so the F1 score will sharply penalize poor precision or recall, unlike the arithmetic mean.

Two-leg journey analogy (harmonic vs. arithmetic mean)

This behavior is analogous to averaging speeds on a two-leg journey with speeds 30 km/h and 60 km/h:

- **Harmonic mean (equal distances):** With equal distances (30 km each):

$$t_1 = \frac{30}{30} = 1 \text{ hour}, \quad t_2 = \frac{30}{60} = 0.5 \text{ hours},$$

$$d_{\text{total}} = 30 + 30 = 60 \text{ km}, \quad t_{\text{total}} = 1 + 0.5 = 1.5 \text{ hours},$$

$$\text{Average speed} = \frac{d_{\text{total}}}{t_{\text{total}}} = \frac{60}{1.5} = 40 \text{ km/h} = \frac{2ab}{a+b} = \frac{2 \cdot 30 \cdot 60}{30+60} = 40 \text{ km/h}.$$

  The slower speed dominates total time.

- **Arithmetic mean (equal times):** With equal times (1 hour each):

$$d_1 = 30 \cdot 1 = 30 \text{ km}, \quad d_2 = 60 \cdot 1 = 60 \text{ km},$$

$$d_{\text{total}} = 30 + 60 = 90 \text{ km}, \quad t_{\text{total}} = 1 + 1 = 2 \text{ hours},$$

$$\text{Average speed} = \frac{d_{\text{total}}}{t_{\text{total}}} = \frac{90}{2} = 45 \text{ km/h} = \frac{a+b}{2} = \frac{30+60}{2} = 45 \text{ km/h}.$$

  Each speed is weighted equally.

- **Sensitivity to low values:** Reducing the slower leg to 15 km/h (equal distances):

$$H = \frac{2ab}{a+b} = \frac{2 \cdot 15 \cdot 60}{15+60} = \frac{1800}{75} = 24 \text{ km/h},$$

$$A = \frac{a+b}{2} = \frac{15+60}{2} = 37.5 \text{ km/h}.$$

  - Harmonic mean drops from 40 to 24 km/h (closer to 15).
  - Arithmetic mean drops from 45 to 37.5 km/h.
  - Harmonic mean reacts more strongly to low values.

Part 4: Thresholds and the Choice of 0.5

(a) For Model C, compute TPR and FPR at threshold 0.5 and at threshold 0.3. Describe qualitatively how the confusion matrix and the metrics change when moving from threshold 0.5 to 0.3, and explain why this demonstrates the importance of choosing an appropriate threshold.

TPR and FPR for Model C at two thresholds:

Threshold 0.5

- **Predictions:** All scores (0.40, 0.35, 0.30, 0.20, 0.10, 0.05) are below 0.5 ⇒ all cases predicted as 0 (negative).
- **Confusion matrix:**

		Predicted	
		1	0
Actual	1	0	2
	0	0	4

- **Metrics:**
  - TPR $= \dfrac{TP}{P} = \dfrac{0}{2} = 0$
  - FPR $= \dfrac{FP}{N} = \dfrac{0}{4} = 0$
- **Interpretation:** Extremely conservative: no false positives, but misses all truly sick patients (no true positives).

Threshold 0.3

- **Predictions:** IDs 2 (0.40) and 5 (0.35) predicted positive (the two truly sick patients); IDs 1, 3, 4, 6 predicted negative.
- **Confusion matrix:**

		Predicted	
		1	0
Actual	1	2	0
	0	0	4

- **Metrics:**
  - TPR $= \dfrac{2}{2} = 1.0$ (perfect recall)
  - FPR $= \dfrac{0}{4} = 0$ (no false alarms)
  - Precision $= \dfrac{2}{2} = 1.0$
  - Accuracy $= \dfrac{6}{6} = 1.0$
- **Interpretation:** With a lower threshold, the model becomes perfect on this dataset: all sick patients are detected and no healthy patients are flagged.

Lowering the threshold from 0.5 to 0.3 changes the confusion matrix and metrics dramatically. This illustrates how crucial the threshold choice is and motivates tools like ROC curves that evaluate performance across thresholds.