INDIAN SCHOOL AL WADI AL KABIR

Class: X	Department: Computer Science
WORKSHEET1	ARTIFICIAL INTELLIGENCE (417) Part B UNIT 3: **Evaluating Models**

Section A: Fill in the Blanks

Fill in the blanks with the most appropriate words or mathematical formulas based on the Unit 3 curriculum.

1. Evaluating a machine learning model using the same dataset that was used to build it can cause the model to simply memorize the data points, which is a critical problem known as _____
2. Recall, which measures how well the model avoids missing true positive cases, is also commonly referred to as Sensitivity or _____
3. The cell in a confusion matrix representing the outcome where a model wrongly predicts the negative class as a positive class is designated as a _____
4. In classification scenarios where the dataset is unbalanced and we are unable to easily decide whether False Positives (FPs) or False Negatives (FNs) are more important to minimize, the _____ is the most suitable combined metric to evaluate
5. The complete mathematical formula used to calculate Classification Accuracy from the four cells of a confusion matrix is _____

Section B: Multiple Choice Questions

Choose the correct option for each of the following questions.

1. Relying solely on standard Classification Accuracy can produce highly misleading results under which of the following conditions?

A) When the dataset size is less than 1,000 samples.
B) When the dataset is unbalanced (unequal observations in each class).
C) When evaluating a regression model rather than a classification model.
D) When the model experiences no training errors.

2. In a COVID-19 screening system, predicting an infected individual (Positive class) as healthy (Negative class) is highly dangerous. To minimize these False Negatives (FNs), which metric must be maximized?2

A) Precision B) Accuracy C) Recall D) Error Rate

3. Which ethical concern in model evaluation focuses on providing an honest explanation of how the chosen evaluation metrics function, ensuring no critical performance details are hidden?

A) Bias B) Transparency C) Accountability D) Autonomy

4. Assertion and Reasoning Question:

Assertion (A): The train-test split technique divides a dataset into two subsets: a training dataset and a testing dataset.

Reasoning (R): Evaluating a model with testing data that was not used during training provides an unbiased estimate of how well the model generalizes to new, unseen data

A) Both A and R are true and R is the correct explanation for A.
B) Both A and R are true and R is not the correct explanation for A.
C) A is True but R is False.
D) A is False but R is True.

5. Assertion and Reasoning Question:
Assertion (A): For a satellite launching station, predicting a bad weather day (Negative class) as a good weather day (Positive class) can lead to a disastrous launch.
Reasoning (R): Therefore, the system must prioritize Recall to ensure all potential bad weather days are correctly flagged
A) Both A and R are true and R is the correct explanation for A.
B) Both A and R are true and R is not the correct explanation for A.
C) A is True but R is False.
D) A is False but R is True.

Section C: Short Answer Questions (2 Marks)
Answer the following questions in complete sentences.

1. **What is a Confusion Matrix, and how are the actual and predicted classes typically arranged within it?**
2. **Define "Overfitting" and explain how a proper train-test split addresses this issue.**
3. **Can classification evaluation metrics (such as Precision, Recall, and Confusion Matrices) be used for a Regression model? Justify your answer.**
4. **State the difference between "Bias" and "Accountability" as ethical concerns in model evaluation.**

Section D: Long Answer / Case-Study Questions (4 Marks)
Read the scenarios carefully and provide complete, step-by-step answers.

1. Scenario Analysis: Precision vs. Recall
Identify whether **Precision** or **Recall** is the more critical evaluation metric for each of the following scenarios, and provide a brief explanation justifying your choice:

- **a) Email Spam Detection:** Identifying whether an incoming email is "spam" or "not spam".
- **b) Heart Disease Diagnosis:** A diagnostic model used to detect patients with critical heart conditions.
- **c) Legal Trial System:** An AI predicting whether a defendant is "guilty" (under the principle of "innocent until proven guilty").
- **d) Safe Content Filtering for Kids:** An AI flagging videos as "safe for children".

2. Case Study: Diagnostic System Evaluation
A medical diagnosis system is designed to classify patients as either having a specific respiratory disease (1) or not having it (0). Out of a total test group of 1,000 patients:

- **True Positives (TP):** 120 patients were correctly diagnosed with the disease.
- **False Positives (FP):** 20 healthy patients were incorrectly diagnosed with the disease.
- **True Negatives (TN):** 800 patients were correctly diagnosed as not having the disease.
- **False Negatives (FN):** 60 infected patients were incorrectly diagnosed as healthy.

Task:

1. Draw the corresponding Confusion Matrix layout.
2. Calculate the following metrics (show your formulas and calculations step-by-step):
    - **i) Classification Accuracy (in %)**
    - **ii) Precision**
    - **iii) Recall**
    - **iv) F1-Score**

ANSWER KEY

Section A: Fill in the Blanks

1. **Overfitting**
2. **True Positive Rate**
3. **False Positive (FP)**
4. **F1-Score**
5. **Accuracy=(TP+TN) / (TP+TN+FP+FN)**

Section B: Multiple Choice Questions

1. **B) When the dataset is unbalanced** *Reasoning:* If a dataset is highly unbalanced, a faulty model that always predicts the majority class can yield a deceptively high accuracy score5more_horiz.
2. **C) Recall** *Reasoning:* Recall measures how well the model avoids False Negatives (FNs). In medical situations, we must minimize missing infected cases2.
3. **B) Transparency** *Reasoning:* Transparency requires an honest, unhidden explanation of how evaluation metrics calculate performance6.
4. **A) Both A and R are true and R is the correct explanation for A.** *Reasoning:* Splitting the dataset prevents the model from evaluating itself on data it has memorized, giving an accurate assessment of generalization18.
5. **C) A is True but R is False.** *Reasoning:* The Assertion is true (launching in bad weather is disastrous)9. However, the Reasoning is false because to reduce False Positives (FPs), we must prioritize **Precision**, not Recall9.

Section C: Short Answer Questions (2 Marks)

1. **Answer:** A Confusion Matrix is a tabular presentation of the prediction accuracy of a classification model10. The table plots actual values along the vertical y-axis and predicted values along the horizontal x-axis, with each cell showing the count of predictions that fall into that category10.
2. **Answer:** Overfitting occurs when an AI model simply memorizes the training dataset rather than learning its underlying patterns, causing it to fail on new data1. A train-test split prevents this by holding back a testing subset that the model has never seen, ensuring we evaluate how well the model can generalize to future predictions78.
3. **Answer:** No, classification metrics cannot be used for Regression models1113. Classification models predict discrete categorical labels (e.g., "Yes/No")24, whereas Regression models predict continuous, changing numerical values (e.g., temperature or house prices), which require error-distance evaluation rather than confusion matrices11more_horiz.
4. **Answer: Bias** refers to ensuring that the chosen evaluation metrics do not result in systematically unfair or unequal performance across groups6. **Accountability** means taking active responsibility for the real-world consequences of chosen metrics and methodologies if users are disadvantaged by them6.

Section D: Long Answer / Case-Study Questions (4 Marks)
1. Scenario Analysis: Precision vs. Recall

- **a) Email Spam Detection: Precision.** It is crucial to minimize False Positives because marking a legitimate, important email as spam is highly disruptive for the user.
- **b) Heart Disease Diagnosis: Recall.** False Negatives must be minimized because missing a patient's critical heart condition is life-threatening, while a False Positive can be corrected with further tests..
- **c) Legal Trial System: Precision.** In a system where a suspect is innocent until proven guilty, committing an innocent person to prison (False Positive) is far worse than failing to convict a guilty suspect (False Negative), so FPs must be minimized

- **d) Safe Kids YouTube Filter: Precision.** Letting an inappropriate video slide through to kids is unacceptable; we must be absolutely certain that any video labeled as "safe" (Positive class) is genuinely safe, requiring FPs to be minimized.

2. Case Study: Diagnostic System Evaluation

1. Confusion Matrix Layout:

	Predicted Positive (1)	**Predicted Negative (0)**
Actual Positive (1)	**TP = 120**	**FN = 60**
Actual Negative (0)	**FP = 20**	**TN = 800**

2. Step-by-Step Calculations:

- **i) Classification Accuracy:** 5

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{120 + 800}{120 + 800 + 20 + 60} = \frac{920}{1000} = 0.9$$

- **ii) Precision:** 5 9

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{120}{120 + 20} = \frac{120}{140} \approx 0.857 \text{ or } \mathbf{85.7\%}$$

- **iii) Recall:** 4

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{120}{120 + 60} = \frac{120}{180} \approx 0.667 \text{ or } \mathbf{66.7\%}$$

- **iv) F1-Score:** 21

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0.857 \times 0.667}{0.857 + 0.667} = \frac{1.143}{1.524} \approx \mathbf{0.75}$$

****