Affiliated to CBSE, New Delhi (Affiliation No. 1030642)

GNPS

GOMTI NANDAN PUBLIC SCHOOL

Gandhi Ward, Bina (M.P.) 470113. | 9165 609 333. | info@mygnps.com | www.mygnps.com

CLASS X · ARTIFICIAL INTELLIGENCE (417) · SKILL SUBJECT · CBSE BOARD EXAMINATION 2027

PART B · UNIT 3

Evaluating Models

Unit Booklet · Short Notes · Board-Pattern Questions · Model Answers

Part B · 10 of 40 theory marks | 29 questions with model answers | Sonakshi Ma'am | Taught: Wed 25 Nov

INSIDE THIS BOOKLET

1	**Unit Notes** 8 note blocks, key terms, exam tips	2
A	**Objective Questions** 16 questions · 1 mark	5
B	**Short Answer Questions** 6 questions · 2 marks	6
C	**Long Answer Questions** 5 questions · 4 marks	6
D	**Case-Based Questions** 2 cases · 4 marks	7
E	**Answers** Key + model answers, marking-scheme style	8

STUDENT DETAILS

NAME _____

CLASS & SECTION _____

ROLL NO. _____

UNIT FINISHED ON _____

OBJECTIVE SELF-CHECK ____ / 16

1

READ THIS FIRST · SHORT NOTES

Unit Notes

Why models are evaluated, the train–test split, overfitting, accuracy and error, the confusion matrix, accuracy, precision, recall and F1 score with worked sums, choosing the right metric, and ethical concerns in evaluation.

THIS UNIT IN THE BOARD PAPER

10 / 40

Unit 3 carries **10 marks**. In the CBSE sample paper it gave **6 one-mark items** (including a precision sum), a **2-mark** question and a **4-mark confusion-matrix case** (fill TP, FP, TN, FN and find accuracy). Practise every formula with numbers.

SYLLABUS 2026–27 · WHAT CBSE LISTS FOR THIS UNIT

- What model evaluation is and why it is needed
- Splitting data into training and testing sets
- Accuracy and error
- Confusion matrix; accuracy, precision, recall, F1 score
- Choosing the right metric
- Ethical concerns: bias, transparency, accuracy

HOW THE CBSE SKILL-SUBJECT PAPER IS SET (THEORY 50 MARKS · 2 HOURS)

1 mark
Objective: MCQ, assertion–reason, fill in the blank, true/false, match. Q1 is Part A; Q2–Q5 (answer 5 of 6 in each) are Part B.

2 marks
Short answer in **20–30 words**: two clear points. Part B: answer any 4 of 6 (Q11–Q16).

4 marks
Long answer in **50–80 words**: four points, or 2 + 2. Answer any 3 of 5 (Q17–Q21); often case-based.

Part B = 40
Objective 20 + short 8 + long 12. Part A (Employability Skills) = 10. Practical = 50.

Write in points, start each point with the key word, and stay inside the word limit. Objective answers: write the option letter **and** the answer.

1 What is model evaluation?

> **Model evaluation:** checking how well a trained model performs on data it has **not seen**, using measures called **metrics**.

Why it is needed:

- To know whether the model can be **trusted** before it is deployed.
- To **compare** models and pick the best one.
- To find **weaknesses** (which cases it gets wrong) and improve it.
- To catch **overfitting** and **bias**.

2 Train–test split

- The dataset is divided into a **training set** (usually 70–80%) and a **testing set** (20–30%), e.g. an **80 : 20** split.
- The model learns from the training set only; the testing set is kept hidden and used at the end.
- Testing on training data is like giving students the exam paper in advance – the score looks high but shows nothing.

Problem	Meaning
Overfitting	Model does very well on training data but **badly on new data** – it has memorised, not learnt
Underfitting	Model does badly even on training data – it is too simple
Good fit	Does well on both training and testing data

3 Accuracy and error

For a model that predicts **numbers** (regression):

> Error = | Actual – Predicted | · Error rate = Error ÷ Actual · Accuracy = 1 – Error rate

WORKED EXAMPLE
Predicted sales 180, actual sales 200 → error 20 → error rate 20 ÷ 200 = 0.1 → accuracy 0.9 = **90%**.

For a model that predicts **classes**: Accuracy = correct predictions ÷ total predictions; **error rate = 1 – accuracy**.

4 Confusion matrix

A table that compares the model's **predictions** with **reality** for a yes/no problem.

	Reality: Yes	Reality: No
Predicted: Yes	**TP** – True Positive	**FP** – False Positive
Predicted: No	**FN** – False Negative	**TN** – True Negative

- **TP:** predicted yes, really yes (fire predicted, fire happened).
- **TN:** predicted no, really no.
- **FP:** predicted yes, really no – a false alarm (Type I error).
- **FN:** predicted no, really yes – a miss (Type II error).

Tip: the **second word** (Positive / Negative) is what the model **said**; the first word (True / False) says if it was **right**.

5 The four metrics

Metric	Formula	Question it answers
Accuracy	(TP + TN) ÷ (TP + TN + FP + FN)	Out of all cases, how many were right?
Precision	TP ÷ (TP + FP)	Of all the 'yes' predictions, how many were really yes?
Recall	TP ÷ (TP + FN)	Of all the real 'yes' cases, how many did the model catch?
F1 score	2 × Precision × Recall ÷ (Precision + Recall)	One balanced score of precision and recall

All four lie between 0 and 1 (or 0–100%). Higher is better.

6 Worked example

A model screens 200 people for a disease: TP = 60, FP = 10, FN = 20, TN = 110.

- Accuracy = (60 + 110) ÷ 200 = 170 ÷ 200 = **0.85 (85%)**
- Precision = 60 ÷ (60 + 10) = 60 ÷ 70 = **0.857**
- Recall = 60 ÷ (60 + 20) = 60 ÷ 80 = **0.75**
- F1 = 2 × 0.857 × 0.75 ÷ (0.857 + 0.75) = 1.286 ÷ 1.607 = **0.80**

The model missed 20 sick people (FN) – serious for a disease test, even though accuracy looks high.

7 Which metric to use?

Situation	Costly mistake	Focus on
Disease test, forest-fire alarm, fraud alert	**FN** – missing a real case	**Recall**
Spam filter (important mail sent to spam), punishing a suspect	**FP** – false alarm	**Precision**
Both mistakes matter	FP and FN	**F1 score**

WHY ACCURACY IS NOT ENOUGH
100 bank transactions: 98 genuine, 2 fraud. A model that calls every transaction "genuine" has **98% accuracy** but catches **no** fraud (recall = 0). So accuracy can mislead when one class is rare.

8 Ethical concerns in evaluation

Concern	What to do
Bias	Test data must include **all groups** (age, gender, region, skin tone). Check the model's results for each group separately.
Transparency	Say clearly **how** the model was tested, on what data, with which metrics, and what its limits are.
Accuracy (honesty)	Report true results; choose the metric that fits the problem; do not hide mistakes or test on training data.

Key terms

15 terms

- **Model evaluation:** checking a model's performance on unseen data
- **Metric:** measure of performance
- **Train–test split:** dividing data to train and to test
- **Overfitting:** great on training data, poor on new data
- **Underfitting:** poor even on training data
- **Error:** | actual – predicted |
- **Confusion matrix:** table of TP, FP, FN, TN
- **True Positive:** predicted yes, really yes
- **False Positive:** false alarm (Type I)
- **False Negative:** a miss (Type II)
- **Accuracy:** (TP+TN) ÷ total
- **Precision:** TP ÷ (TP+FP)
- **Recall:** TP ÷ (TP+FN)
- **F1 score:** balance of precision and recall
- **Error rate:** 1 – accuracy

Common mistakes and exam tips

AVOID THESE

- Precision divides by **predicted** positives (TP + FP); recall divides by **actual** positives (TP + FN).
- In a confusion-matrix case, first write the four values with labels – the marking scheme gives ½ mark each.
- Write the **formula** before putting numbers in: ½ mark for the formula, ½ for the answer.
- Never test a model on the data it was trained on.
- High accuracy does not always mean a good model when one class is rare.

A

PART A · 1 MARK EACH

Objective Questions

16 questions: MCQs, assertion–reason, fill in the blanks and true/false, like Q2–Q5 of the board paper. Answers in Part E.

ASSERTION–REASON: HOW TO ANSWER

(a) Both A and R are true, and R is the correct explanation of A.
(b) Both A and R are true, but R is not the correct explanation of A.
(c) A is true, but R is false.
(d) A is false, but R is true.

A1. Evaluating a model on the same data it was trained on usually gives: [1]

(a) a true picture
(b) a misleadingly high score
(c) a zero score
(d) a lower score than on new data

A2. A model scores 99% on training data but 60% on new data. This is: [1]

(a) Underfitting **(b)** Overfitting **(c)** Good fit **(d)** Clustering

A3. In a confusion matrix, the model predicted "rain" but it did not rain. This is a: [1]

(a) True Positive
(b) False Positive
(c) False Negative
(d) True Negative

A4. A model predicted "no fire" and a fire broke out. This is a: [1]

(a) True Negative
(b) False Positive
(c) False Negative
(d) True Positive

A5. TP = 30, FP = 10. Precision is: [1]

(a) 0.25 **(b)** 0.75 **(c)** 0.33 **(d)** 3

A6. TP = 40, FN = 10. Recall is: [1]

(a) 0.8 **(b)** 0.25 **(c)** 4 **(d)** 0.5

A7. TP = 50, TN = 30, FP = 15, FN = 5. Accuracy is: [1]

(a) 50% **(b)** 65% **(c)** 80% **(d)** 85%

A8. For a cancer-screening model, which metric matters most? [1]

(a) Precision
(b) Recall
(c) Error
(d) Training accuracy

A9. For a spam filter that must not send important emails to spam, which metric matters most? [1]

(a) Recall **(b)** Precision **(c)** Error rate **(d)** Training size

A10. A model predicted 450 units of sales; actual sales were 500. Its accuracy for this prediction is: [1]

(a) 90% **(b)** 10% **(c)** 95% **(d)** 50%

A11. ASSERTION–REASON [1]

Assertion (A): Accuracy alone can be misleading when one class is very rare.
Reason (R): A model can label every case as the common class and still get high accuracy.

(a) Both A and R are true, and R is the correct explanation of A.
(b) Both A and R are true, but R is not the correct explanation of A.
(c) A is true, but R is false.
(d) A is false, but R is true.

A12. ASSERTION–REASON [1]

Assertion (A): The F1 score is the average of accuracy and error.

Reason (R): The F1 score balances precision and recall.

(a) Both A and R are true, and R is the correct explanation of A.
(b) Both A and R are true, but R is not the correct explanation of A.
(c) A is true, but R is false.
(d) A is false, but R is true.

A13. FILL IN THE BLANK [1]

Recall = TP ÷ (TP + _____).

A14. FILL IN THE BLANK [1]

If a model's accuracy is 0.92, its error rate is _____.

A15. TRUE OR FALSE [1]

Test data should include examples from all groups of people who will use the model.

A16. TRUE OR FALSE [1]

Precision tells us how many of the actual positive cases the model found.

B — Short Answer Questions

PART B · 2 MARKS EACH

Answer in 20–30 words: two clear points. Model answers in Part E.

B1. Why is model evaluation important? Give two reasons. MUST DO [2]

B2. What is a train–test split? Why is it needed? [2]

B3. Define precision and recall with formulas. MUST DO [2]

B4. Differentiate between a false positive and a false negative with an example. [2]

B5. Differentiate between overfitting and underfitting. [2]

B6. Precision is 0.6 and recall is 0.9. Find the F1 score. [2]

C — Long Answer Questions

PART C · 4 MARKS EACH

Answer in 50–80 words: four points, or two points with an example each. Model answers in Part E.

C1. A model predicts whether students will need extra classes. Out of 100 students: correctly predicted 'needs help' 35, wrongly predicted 'needs help' 15, wrongly predicted 'no help' 5, correctly predicted 'no help' 45. Draw the confusion matrix and find the accuracy. MUST DO [4]

C2. Using TP = 45, FP = 15, FN = 5, TN = 35, calculate precision, recall and F1 score. MUST DO [4]

C3. Explain with examples when to focus on recall and when on precision. [4]

C4. Explain the ethical concerns in evaluating AI models. [4]

C5. Why can a model with 98% accuracy still be useless? Explain with an example. [4]

D

PART D · CASE AND COMPETENCY BASED

Case-Based Questions

Read the situation, then answer from it and from the unit. These are the 4-mark board questions that test application.

D1. Case-Based · Forest-fire alert system [4]

An AI system predicts forest fires from temperature, humidity and wind data. Over 100 days its results were: it predicted fire on 25 days and fire actually happened on 20 of them; it predicted no fire on 75 days, but fire broke out on 5 of those days.

(i) Write TP, FP, FN and TN. [2]

(ii) Find the accuracy and the recall. [1]

(iii) Why is recall the key metric here? [1]

D2. Competency-Based · The face-unlock complaint [4]

A company says its face-unlock feature is 97% accurate. Users from one region complain that it often fails for them. It turns out the testing photos came mostly from people of another region.

(i) Which ethical concern does this show? [1]

(ii) Why was the 97% figure misleading? [1]

(iii) Suggest two ways to evaluate the model fairly. [2]

E

PART E · ANSWERS IN CBSE MARKING-SCHEME STYLE

Answers

Objective key with a reason, then model answers: numbered value points, key word first, the mark for each point in orange, and the word count under each answer. Any other correct, relevant point also earns the mark.

Objective questions – answer key

Part A

Write the option letter **and** the answer, as the marking scheme does.

Q	Answer	Why (one line)
A1	**(b) a misleadingly high score**	The model has already seen the answers.
A2	**(b) Overfitting**	It memorised the training data.
A3	**(b) False Positive**	Predicted yes, really no – false alarm.
A4	**(c) False Negative**	Predicted no, really yes – a miss.
A5	**(b) 0.75**	30 ÷ (30 + 10) = 0.75.
A6	**(a) 0.8**	40 ÷ (40 + 10) = 0.8.
A7	**(c) 80%**	(50 + 30) ÷ 100 = 80%.
A8	**(b) Recall**	Missing a patient (FN) is the costliest mistake.
A9	**(b) Precision**	False positives must be few.
A10	**(a) 90%**	Error 50; error rate 0.1; accuracy 0.9.
A11	**(a) Both true; R explains A**	R explains A.
A12	**(d) A false, R true**	A is false; F1 combines precision and recall.
A13	**FN**	Actual positives = TP + FN.
A14	**0.08**	1 – 0.92.
A15	**True**	This checks for bias.
A16	**False**	Correct: that is recall; precision = correct share of predicted positives.

Short answers

Part B · 2 marks

B1. Why is model evaluation important? Give two reasons. [2]

1. **Trust:** shows whether the model works well on new data before deployment. 1
2. **Improve:** finds weaknesses and helps choose the best model. 1

21 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | Any two, 1 mark each | MUST DO

B2. What is a train–test split? Why is it needed? [2]

1. **Meaning:** dividing data into a training set and a testing set, e.g. 80 : 20. 1
2. **Need:** testing on unseen data gives a true measure of performance. 1

26 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

B3. Define precision and recall with formulas. [2]

1. **Precision:** TP ÷ (TP + FP): correct share of predicted positives. 1
2. **Recall:** TP ÷ (TP + FN): share of actual positives caught. 1

22 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1 | MUST DO

B4. Differentiate between a false positive and a false negative with an example. [2]

1. **False positive:** predicted yes, really no – fire alarm with no fire. 1
2. **False negative:** predicted no, really yes – sick person called healthy. 1

21 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

B5. Differentiate between overfitting and underfitting. [2]

1. **Overfitting:** good on training data, poor on new data. 1
2. **Underfitting:** poor even on training data; model too simple. 1

18 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

B6. Precision is 0.6 and recall is 0.9. Find the F1 score. [2]

1. **Formula:** F1 = 2 × P × R ÷ (P + R). 1
2. **Answer:** 2 × 0.6 × 0.9 ÷ 1.5 = 1.08 ÷ 1.5 = 0.72. 1

26 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | ½ formula + 1½ working and answer

Long answers

Part C · 4 marks

C1. A model predicts whether students will need extra classes. Out of 100 students: correctly predicted 'needs help' 35, wrongly predicted 'needs help' 15, wrongly predicted 'no help' 5, correctly predicted 'no help' 45. Draw the confusion matrix and find the accuracy. [4]

1. **Matrix:** TP = 35, FP = 15, FN = 5, TN = 45. 2
2. **Formula:** Accuracy = (TP + TN) ÷ total. 1
3. **Answer:** (35 + 45) ÷ 100 = 80%. 1

29 words · CBSE asks 50–80 words | Marking: 2 + 1 + 1 = 4 | ½ each value; 1 formula; 1 answer | MUST DO

C2. Using TP = 45, FP = 15, FN = 5, TN = 35, calculate precision, recall and F1 score. [4]

1. **Precision:** 45 ÷ 60 = 0.75. 1
2. **Recall:** 45 ÷ 50 = 0.9. 1
3. **F1:** 2 × 0.75 × 0.9 ÷ 1.65 = 1.35 ÷ 1.65 ≈ 0.82. 2

26 words · CBSE asks 50–80 words | Marking: 1 + 1 + 2 = 4 | 1 + 1 + 2 (formula ½ each) | MUST DO

C3. Explain with examples when to focus on recall and when on precision. [4]

1. **Recall:** when missing a real case is costly. 1
2. **Example:** disease screening, forest-fire alarm. 1
3. **Precision:** when false alarms are costly. 1
4. **Example:** spam filter; flagging innocent people. 1

25 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each

C4. Explain the ethical concerns in evaluating AI models. [4]

1. **Bias:** test data must cover all groups; check results for each. 1
2. **Transparency:** explain how and on what data it was tested. 1
3. **Honest accuracy:** report true results; never test on training data. 1
4. **Right metric:** choose the metric that fits the harm. 1

40 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each

C5. Why can a model with 98% accuracy still be useless? Explain with an example. [4]

1. **Rare class:** when one class is very rare, accuracy hides failures. 1
2. **Example:** 98 genuine, 2 fraud transactions. 1
3. **Lazy model:** calling all 'genuine' gives 98% accuracy. 1
4. **Result:** recall is 0; no fraud is caught, so use recall or F1. 1

38 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each

Case-based answers

Part D

D1. Case-Based · Forest-fire alert system (passage printed in Part D) [4]

(i) Write TP, FP, FN and TN. [2]

TP = **20**, FP = **5**, FN = **5**, TN = **70**.

(ii) Find the accuracy and the recall. [1]

Accuracy = (20 + 70) ÷ 100 = **90%**; recall = 20 ÷ (20 + 5) = **0.8**.

(iii) Why is recall the key metric here? [1]

A **missed fire (FN)** can destroy forest and lives, so catching every real fire matters most.

Marking: 2 + 1 + 1 = 4 | ½ × 4 + (½ + ½) + 1

D2. Competency-Based · The face-unlock complaint (passage printed in Part D) [4]

(i) Which ethical concern does this show? [1]

Bias in the evaluation data.

(ii) Why was the 97% figure misleading? [1]

It was measured on test data that **did not represent all users**.

(iii) Suggest two ways to evaluate the model fairly. [2]

(1) Test on data from **every region, age and skin tone** and report results for each group. (2) Be **transparent**: publish how it was tested and its known limits.

Marking: 1 + 1 + 2 = 4 | 1 + 1 + 2