Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First / Second Semester 20**-20**

Mid-Semester Test
(EC-2 Regular/ EC-2 Make-up)

Course No.	: ****
Course Title	: ****
Nature of Exam	: Closed Book / Open Book (As per Course Handout)
Weightage	: 30% or 35% (As per Course Handout)
Duration	: 2 Hours
Date of Exam	: **** (FN/AN)

No. of Pages =
No. of Questions =

Note to Students:

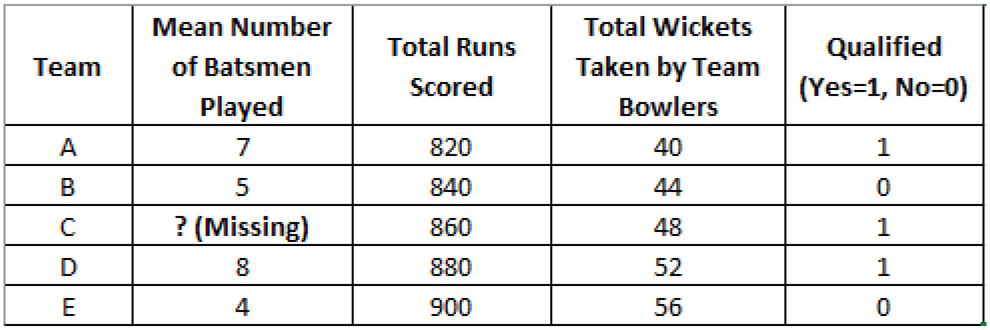
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Please Note:
A **maximum of 12 main questions** with Q.Nos. 1 to N**.**

Q.1.

[10 Marks]

An IPL franchise wants to predict whether a team will qualify for the playoffs based on its performance in the league stage (8 matches). The analyst has collected the following data for 5 teams:



Team	Mean Number of Batsmen Played	Total Runs Scored	Total Wickets Taken by Team Bowlers	Qualified (Yes=1, No=0)
A	7	820	40	1
B	5	840	44	0
C	? (Missing)	860	48	1
D	8	880	52	1
E	4	900	56	0

This dataset will be used as the training data for building a logistic regression model. One value is missing in the feature **"Mean Number of Batsmen Played."**

(a) Suggest appropriate methods to handle the missing value in the *Mean Number of Batsmen Played* feature before building the logistic regression model. **Justify your choice.**
(1 marks)

(b) Normalize **only** the *Total Runs* and *Total Wickets* features using **z-score normalization**.
Show the formula, compute the **mean** and **standard deviation**, and present the updated dataset.
(3 marks)

c) Perform **one iteration of Batch Gradient Descent** to update the weights for the logistic regression model, using:

- Initial weights:
  θ=[0.1, 0.1, 0.1, 0.2]
- Learning rate: α=0.1

(4 marks)

d) Calculate the value of the Cross-Entropy Loss (Log loss) function at the end of the first iteration (1 mark)

e) Calculate and interpret the **predicted probability** of playoff qualification for **Team D** after the first iteration.
(1 mark)

Solution:

(a) Handling the Missing Value (Team C)

Known x1 values: 7, 5, 8, 4

Mean = (7 + 5 + 8 + 4) / 4 = 6

Median = (5 + 7) / 2 = 6

Mean and median imputations are appropriate because the feature is numerical, continuous, and symmetric. Mode is inappropriate since values are all unique and not categorical.

Imputed value for Team C = 6

(b) Z-Score Normalization

Total Runs:

Mean μ = (820+840+860+880+900)/5 = 860

Variance = 800 → $\sigma = \sqrt{800} = 28.28$

Total Wickets:

Mean μ = 48

Variance = 32 → $\sigma = \sqrt{32} = 5.66$

Normalized Dataset:

Team	Mean(Batsmen)	Runs(z)	Wickets(z)
A	7	-1.41	-1.41
B	5	-0.71	-0.71
C	6	0	0
D	8	0.71	0.71
E	4	1.41	1.41

(c) Batch Gradient Descent

Given:

$\theta = [0.1, 0.1, 0.1, 0.2]$

$\alpha = 0.1, m = 5$

Computed gradients:

$\partial J/\partial \theta 0 = 0.065$

$\partial J/\partial \theta 1 = -0.1946$

$\partial J/\partial \theta 2 = 0.197$

$\partial J/\partial \theta 3 = 0.197$

Weight Updates:

$\theta 0 = 0.1 - 0.1(0.065) = 0.0935$

$\theta 1 = 0.1 - 0.1(-0.1946) = 0.1195$

$\theta 2 = 0.1 - 0.1(0.197) = 0.0803$

$\theta 3 = 0.2 - 0.1(0.197) = 0.1803$

(d) Cross Entropy Loss formula (0.5 mark)
Calculated value (0.5 mark)

(d) Predicted Probability for Team D

$z = 0.0935 + 0.1195(8) + 0.0803(0.71) + 0.1803(0.71) = 1.2342$

$h\theta(x) = 1 / (1 + e\text{^}{-1.2342}) = 0.7746$

Interpretation:

Team D has a 77.46% probability of qualifying for the playoffs.

Q.2. [Marks]

A student claims:

"I doubled my dataset size, but my validation RMSE still stayed high while the training RMSE kept dropping. Therefore, increasing dataset size doesn't help."

As an instructor, critique this statement using the concepts of **model complexity**, **bias–variance trade-off**, and **training - validation behavior**.

a) Comment whether the model has (low / high) Bias and (low / high) Variance (0.5 marks for each).
b) Include one likely reason this observation occurred. (1 mark)
c) Suggest one corrective action for the above. (1 mark)

Solution

- The model is **too complex** (low bias, high variance).
- Adding more data *should* reduce variance, but if data distribution changed or noise increased, overfitting can persist.

- Corrective actions: reduce polynomial order, apply regularization, normalize inputs, or recheck data quality.

Q.3. [8 Marks]

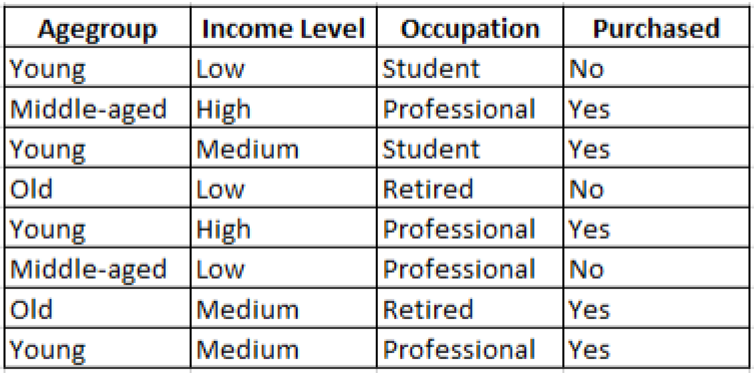
A. Consider the training dataset provided in the table below, which tracks the purchasing behaviour of customers based on Agegroup, Income Level, and Occupation.

i. Calculate the Entropy of the target class Purchased for the entire dataset. (**2 Marks**)

ii. Suppose we choose "Agegroup" as the root node. This creates three branches. Calculate the Entropy specifically for the "Young" branch. Is this branch pure? If not, which attribute would you use next to split this specific "Young" node to achieve pure leaf nodes? (**3 Marks**)

iii. Compare the attributes "Occupation" and "Income Level". Without performing full Information Gain calculations, argue which of these two attributes appears to separate the "Yes" and "No" classes more effectively. (**2 Marks**)

iv. A new customer walks in with the profile: {Age: Young, Income: Low, Occupation: Professional}. Based only on the majority voting of the "Young" subset in the training data, how would this customer be classified? (**1 Mark**)



Agegroup	Income Level	Occupation	Purchased
Young	Low	Student	No
Middle-aged	High	Professional	Yes
Young	Medium	Student	Yes
Old	Low	Retired	No
Young	High	Professional	Yes
Middle-aged	Low	Professional	No
Old	Medium	Retired	Yes
Young	Medium	Professional	Yes

Solution:

First, identify the instances where Agegroup = Young. Looking at the table, there are 4 such instances:

1. {Young, Low, Student} □ No
2. {Young, Medium, Student} □ Yes
3. {Young, High, Professional}□ Yes
4. {Young, Medium, Professional} □ Yes

- Total instances = 4
- Positive instances (Yes) = 3
- Negative instances (No) = 1

Calculate Entropy(Young)=0.811

- Is it pure? No. (Entropy is not 0; it contains both "Yes" and "No" classes).
- Next Attribute to Split: We look at the "Income Level" and "Occupation" for these 4 rows.

o Income Level:

□ Low □ No (1 instance)

□ Medium □ Yes (2 instances)
□ High □ Yes (1 instance)
□ Splitting on Income creates pure leaf nodes (Low=No, Med/High=Yes).
o Occupation:
□ Student □ 1 No, 1 Yes (Mixed/Impure)
□ Professional □ 2 Yes (Pure)
o You should use "Income Level" next, as it perfectly separates the remaining instances into pure classes.

B.iii.

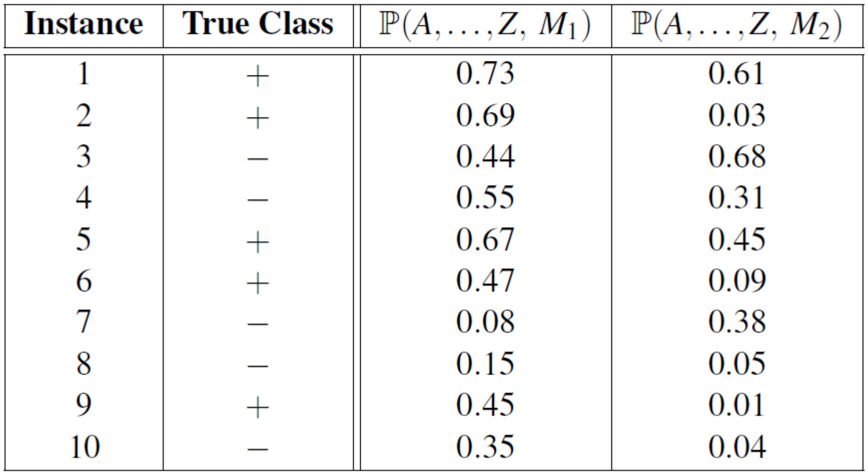
• Occupation:
o Student: Contains both "Yes" and "No". (Impure)
o Professional: Contains both "Yes" and "No". (Impure)
o Retired: Contains both "Yes" and "No". (Impure)
o Conclusion: Knowing someone's occupation gives very little certainty about whether they purchased the item. The entropy is high.

• Income Level :
o Low Income: Rows 1, 4, 6 are all No. (Pure)
o Medium Income: Rows 3, 7, 8 are all Yes. (Pure)
o High Income: Rows 2, 5 are all Yes. (Pure)
o Conclusion: Income Level provides a perfect classification. Knowing the Income Level tells you the target class with 100% certainty.

• IncomeLevel is the better attribute. Without doing calculation, we can see that it perfectly separates the data: all "Low" income customers said No, while all "Medium" and "High" income customers said Yes. Occupation, conversely, has mixed results (yes/no) for every category, implying it has a much lower Information Gain.

.iv Yes

Q.4. [5 Marks]



Instance	True Class	$\mathbb{P}(A,\ldots,Z, M_1)$	$\mathbb{P}(A,\ldots,Z, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

You are asked to evaluate the performance of two classification models, M_1 and M_2. The test set you have chosen contains 26 binary attributes, labeled as *A* through *Z*.

The above table shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(−) = 1−P(+) and $P(- \mid A,\ldots,Z) = 1-P(+ \mid A,\ldots,Z)$. Assume that, we are mostly interested in detecting instances from the positive class.

For both models, M_1 and M_2, suppose you choose the cutoff threshold to be $t = 0.5$. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example.

a) Construct the Confusion Matrix for both models (2 marks)

b) Compute the precision, recall, and F-measure for both models at this threshold value. (3 marks)

Solution:

When $t = 0.5$, the confusion matrix for M_1 and M_2 are shown in the tables.

M_1		Prediction	
		+	−
Actual	+	3	2
	−	1	4

M_2		Prediction	
		+	−
Actual	+	1	4
	−	1	4

For M_1:

Precision =$\frac{3}{4} = 0.75$. Recall = $\frac{3}{5} = 0.6$. F-score = $\frac{2\times0.75\times0.6}{0.75+0.6} = \frac{2}{3} = 0.67$.

For M_2:

Precision =$\frac{1}{2} = 0.5$. Recall = $\frac{1}{5} = 0.2$. F-score = $\frac{2\times0.5\times0.2}{0.5+0.2} = \frac{2}{7} = 0.29$.

Q.5. [4 Marks]

A. In a regression model predicting *house price = 50 + 200×(area) + 5×(age)*,
   a) Interpret both coefficients. (1 mark)
   b) What happens if we change the units of "area" from m^2 to ft^2? (1 mark)

B. In gradient descent, $\partial J/\partial\theta_1$ is large while $\partial J/\partial\theta_2$ is near zero. From this, what can you infer about the model parameters? (1 mark)

C. Which feature has a stronger influence on house price—area or age? Justify. (1 mark)

Solution:

A.

a) `200` means each additional m^2 adds ₹200 to the price (keeping age constant). (1 mark)
b) Changing units rescales the coefficient but not model behavior. The relationship stays identical — only θ_1 changes numerically. (1 mark)

B. θ_1 contributes more to reducing error; θ_2 has reached near-optimal value or irrelevant feature. (1 mark)

C. The **area** has a much stronger influence because its coefficient (200) is far larger than the coefficient of age (5), meaning each unit increase in area raises the price far more than a unit increase in age. (1 mark)

Q.6. [Marks]

Q.7. [Marks]

Q.8. [Marks]

Q.9. [Marks]

Q.10. [Marks]

Q.11. [Marks]

Q.12. [Marks]

****