BENNETT
UNIVERSITY
THE TIMES GROUP

Enrolment No: E22CSEU0827 Name of Student: MADHAV GUPTA

Department/ School: SCSET

MID TERM EXAMINATION, ODD SEMESTER JULY 2023

COURSE CODE: CSET211 MAX. DURATION 1 HR
COURSE NAME: STATISTICAL MACHINE LEARNING
PROGRAM: ***B. Tech. - Computer Science & Engineering*** TOTAL MARKS 15

Mapping of Questions to Course and Program Outcomes										
Q.No.	1	2	3	4	5	6	7	8	9	10
CO	CO1	CO1	CO2	CO1	CO3	CO2	CO1	-	-	-
PO	PO1	PO1	PO2	PO2	PO3	PO2	PO4	-	-	-

GENERAL INSTRUCTIONS: -

1. Do not write anything on the question paper except name, enrolment number and department/school.
2. Carrying mobile phone, smart watch and any other non-permissible materials in the examination hall is an act of UFM.

COURSE INSTRUCTIONS:

a) All workings must be shown.

b) You can use calculator.

c) No clarifications can be provided during the exam. Make reasonable assumptions if necessary, and state any assumptions made.

SECTION A

(Answer all questions, Section A carries maximum 6 marks.) *Marks*

1) Differentiate between overfitting and underfitting. How can it affect model generalization? **(1+1=2)**

2) Describe ensemble methods. Differentiate between bagging and boosting. **(1+1=2)**

3) Given the set of values $X = (3, 9, 11, 5, 2)^T$ and $Y = (1, 8, 11, 4, 3)^T$.
Evaluate the regression coefficients. What is the value of variable Y when X = 7? **(1+1=2)**

SECTION B

(Answer any three full questions, Section B carries maximum 9 marks.) *Marks*

4) What are the benefits of pruning in decision tree induction? Explain different approaches to tree pruning? (1.5+1.5=3)

5) Look at the table below, which provides data related to Hired Professionals Listings. We aim to construct a decision tree classifier to anticipate the likelihood of hiring a professional, classifying them into 'YES' or 'NO' categories. (2+1=3)

a) Calculate the ***GINI Index*** and **Information Gain** for features (Major, Experience, Tie) of root node of decision tree of following training **data.**

Major	Experience	Tie	Hired?
CS	programming	pretty	NO
CS	programming	pretty	NO
CS	management	pretty	YES
CS	management	ugly	YES
business	programming	pretty	YES
business	programming	ugly	YES
business	management	pretty	NO
business	management	pretty	NO

b) Information gain has a disadvantage that it prefers splits having large number of small but pure partitions. How do we overcome this?

6) Explain two reasons why Linear regression is not ideal for use in classification.

Why is ***mean squared error cost*** function not used with logistic regression? Write the cost function that is used instead. (1+1+1=3)

7) Calculate ***Accuracy, Precision, Recall*** & ***F1 Score*** for following confusion matrix.

Give two different examples when we need ***High precision, Low recall*** and ***High recall, Low precision*** for a specific positive class. (0.5+0.5+0.5+0.5+1=3)

		Predicted Class	
		Spam	Non-Spam
Actual Class	Spam	TP=45	FN=20
	Non-Spam	FP=5	TN=30

-ALL THE BEST-