BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: B.Tech. **SEMESTER : V**
BRANCH: Food Engineering & Technology **SESSION : MO/2025**

SUBJECT: FE325 STATISTICAL MACHINE LEARNING II

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	The following table consists of training data from an employee database. The data has been generalized. For example, "31 . . . 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.	[5]	CO1	2,3,4

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let status be the class label attribute.
i. How would you modify the basic decision tree algorithm to take into consideration the count of each generalised data tuple (i.e., of each row entry)?
ii. Use your algorithm to construct a decision tree from the given data.

Q.1(b)	Given two clusters: Cluster 1: (1,1), (1.2,1.1), (0.9,1.0), (1.1,0.9) Cluster 2: (10,10), (10.2,10.1), (9.9,10.0) Identify the DBSCAN clusters, with ε = 0.5 and MinPts = 3.	[5]	CO1	1,2,3

Q.2(a)	Evaluate Ant Colony Optimization (ACO) algorithm for the Travelling Salesman Problem (TSP) consisting of four cities, A, B, C, and D. The distances between the cities are given in the table below (the matrix is symmetric):	[5]	**CO2**	2,3,4

From	To	Distance
A	B	4
A	C	6
A	D	8
B	C	3
B	D	7
C	D	5

Initially, the pheromone level on every edge is τ=1.0. The algorithm uses parameters α=1, β=2, evaporation rate ρ=0.1, and pheromone constant Q=100.
An ant starts at city A and constructs the tour: $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$

Q.2(b)	Describe the concept of Reproducing Kernel Hilbert Spaces (RKHS) and the significance of Mercer's theorem in RKHS.	[5]	CO2	2,3,5

PTO

Q.3(a) Describe the AdaBoost Algorithm. Evaluate the algorithm using decision stumps. on the following binary classification dataset, which contains 6 training instances: [5] CO3 3,4

Instance	x	Label (y)
1	1	+1
2	2	+1
3	3	−1
4	4	+1
5	5	−1
6	6	−1

Q.3(b) You are given one-**dimensional data** $X = \{2,4,5,9\}$. Assume the data is generated from a mixture of **two Gaussian components** with **known variances** $\sigma^2 = 1$, where mixing weights are given as $\pi_1 = 0.5, \pi_2 = 0.5$ and means are given as $\mu_1 = 3, \mu_2 = 8$. Perform **one complete EM iteration**. [5] CO3 3,4

Q.4(a) A Random Forest classifier is being trained on the following dataset with binary class labels: [5] CO4 3,4,5

Instance	Feature X1	Feature X2	Class
1	2	1	A
2	3	2	A
3	4	1	B
4	5	3	B
5	6	2	B
6	7	3	A

A Random Forest of size **2 trees** are constructed. For each tree:

- A **bootstrap sample** of size 6 is drawn (sampling with replacement).
- At each split, the algorithm randomly selects **one feature** (m = 1) out of the two features and performs a split that **minimizes Gini impurity**.

You are given the following bootstrap samples. Compute the Gini impurity for each split point on feature X1 and identify the best split.

Tree 1 Bootstrap Sample:	Tree 2 Bootstrap Sample:
Instances: {1, 2, 2, 3, 5, 6}	Instances: {2, 3, 4, 4, 6, 6}
Feature chosen at root: **X1**	Feature chosen at root: **X2**
Candidate split points: 2.5, 3.5, 4.5, 5.5	Candidate split points: 1.5, 2.5, 3.5

Q.4(b) Describe Convolutional Neural Network. Evaluate the feature map for the given input space [5] CO4 4,5

1	0	1	1	0	1
1	1	1	1	1	0
0	0	0	1	1	1
0	0	1	1	1	0
1	1	1	1	0	0
0	1	0	1	0	0

and the feature detector $W=\begin{bmatrix}1 & 0 & 1\\ 1 & 0 & 1\\ 1 & 0 & 1\end{bmatrix}$

Q.5(a) You are given a **minority class dataset** containing the following 2-dimensional points: $P_1 = (2, 3), P_2 = (4, 5), P_3 = (3, 6)$. Assume SMOTE is applied with **Oversampling rate of** 200% and **Number of nearest neighbors** $k = 2$. For a selected point $P_1 = (2,3)$, the two nearest minority neighbors are P_2 and P_3. The SMOTE formula for generating a synthetic point is $x_{\text{new}} = x_i + \lambda(x_{nn} - x_i)$, where λ is a random number in $[0,1]$. Generate two synthetic samples for P1. [5] CO5 3,4,5

Q.5(b) Perform Principal Component Analysis (PCA) on the following dataset consisting of 3 observations in 2 dimensions. [5] CO5 5,6

Observation	(x_1)	(x_2)
1	2	0
2	0	2
3	3	1

:::::25/11/2025:::::M