UPES
End Semester Examination, May 2024

Course: Data Analytics and Machine Learning   **Semester: IV**
Program: B. Tech (Chemical)   **Time   : 03 hrs.**
Course Code: CSBA2013   **Max. Marks: 100**

Instructions: (a) This is a closed book exam. Possession of mobile phone or any other electronic communicating devices is strictly prohibited during the exam.
(b) Use of scientific calculators is allowed (not mobile phone).

SECTION A (5Q x 4M = 20Marks)

S. No.	Statement of the question	Marks	CO
Q 1	For the data set: **4, 5, 6, 6, 6, 7, 7, 7, 7, 8** Show the necessary calculation, analysis and reasons to identify the nature of skewness of the data set.	4	**CO3[2]** **CO4[2]**
Q 2	A random experiment consist of flipping three fair coins. **(a)** What is the probability of getting THH ? Here, THH means tails after tossing 1st coin and then Head after tossing the second coin and then Head after tossing the third coin, **(b)** What is the probability of getting at-least one head?	2 + 2	**CO1[2]** **CO3[2]**
Q3	What do you understand by bias and variance? (explain with figures) Give an example where we utilize this concept in choosing one algorithm over another	4	**CO1**
Q4	Discuss the situations when should we choose unsupervised KNN clustering techniques and when should we choose unsupervised Hierarchical clustering technique? Discuss about the parameters used in unsupervised DB scan clustering algorithm.	4	**CO2[2]** **CO3[2]**
Q5	Discuss about any two activation functions in neural network with their advantages and disadvantages.	4	**CO1[2]** **CO3[2]**

SECTION B (4Q x 10M = 40 Marks)

| Q 6 | Because a new medical procedure has been shown to be effective in the early detection of an illness, a medical screening of the population is proposed. The probability that the test correctly identifies someone with the illness as positive is 0.99, and the probability that the test correctly identifies someone without the illness as negative is 0.95. The incidence of the illness in the general population is 0.0001. You take the test, and the result is positive. What is the probability that you have the illness?
OR
Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the | 10 | **CO1[3]**
CO2[2]
CO3[3]
CO4[2] |

	presence of the pollutant. (a) Find the probability that in the next 18 samples, exactly 2 contain the pollutant. (b) Determine the probability that at least four samples contain the pollutant, (c) Determine the probability that $3 \le X < 7$. Here, X is the random variable.	**10**

Q7	In a chemical plant, you are tasked with predicting whether a batch of chemical product will be of high quality or low quality based on two predictors: temperature and pH level. You have historical data of several batches with their temperature, pH levels, and corresponding quality. Use the Naive Bayes Theorem to classify a new batch as high or low quality.	**10**	**CO1[3]** **CO2[2]** **CO3[3]** **CO4[2]**

- Temperature: High (H), Moderate (M), Low (L)
- pH level: Basic (B), Neutral (N), Acidic (A)

Temp	H	M	L	H	M	L	H	M	L
Ph Level	B	N	A	N	B	N	A	A	B
Quality	HIGH	LOW	LOW	HIGH	HIGH	LOW	LOW	LOW	HIGH

Predict the quality of a new batch with Moderate Temperature (M) and Basic pH level (B) using Naive Bayes classification

Q8	A chemical engineer is monitoring a process where three chemical compounds (A, B, and C) are produced. The engineer collects data on four different variables that are believed to influence the quality and yield of the products: temperature (X1), pressure (X2), feed rate (X3), and catalyst activity (X4). The engineer wants to reduce the dimensionality of the data to identify which variables explain the most variance in the production process. Discuss the steps of performing PCA using the data of X1, X2, X3, and X4.	**10**	**CO1[3]** **CO2[2]** **CO3[3]** **CO4[2]**
Q9	Discuss the chain rule used in the back propagation training of typical neural network to tune the values of weights and biases. Consider a very simplified network with one node in every layer to explain the chain rule.	**10**	**CO1[4]** **CO2[4]** **CO3[2]**

SECTION-C (2Q x 20M = 40 Marks)

Q10	Derive the expression of (a) probability mass function, $f(x)$ and (b) mean for a discrete Poisson distribution. Include all the mathematical steps in detail, diagram, and all necessary assumptions/ reasons. **OR** With the help of a suitable example, derive the expression of (a) probability mass function, $f(x)$ and (b) variance for a discrete Poisson distribution. Include all detailed steps of the mathematical formulation, diagram, and all necessary assumptions/ reasons.	**15 + 5**	**CO1[5]** **CO2[5]** **CO3[10]**
Q11	Design a strategy of choosing between the "Weather" and "Temperature" features in decision tree classification algorithm.	**20**	**CO1[4]** **CO2[4]** **CO3[7]** **CO4[5]**

ID	Weather	Temperature	Play
1	Sunny	Hot	No
2	Sunny	Hot	No

	3	Not Sunny	Hot	Yes			
	4	Not Sunny	Not Hot	Yes			
	5	Not Sunny	Not Hot	Yes			
	6	Not Sunny	Not Hot	Yes			
	7	Not Sunny	Not Hot	Yes			
	8	Sunny	Not Hot	No			
	9	Sunny	Not Hot	Yes			
	10	Not Sunny	Not Hot	Yes			
	11	Sunny	Not Hot	Yes			
	12	Not Sunny	Not Hot	Yes			
	13	Not Sunny	Hot	Yes			
	14	Not Sunny	Not Hot	No			