Name:

Enrolment No:

UPES

End Semester Examination, December 2024

Course: Data analytics and machine learning **Semester: III**
Program: B. Tech (Chemical) **Time : 03 hrs**
Course Code: CSBA2013 **Max. Marks: 100**

Instructions: (a) This is a closed book exam. Possessing a mobile phone and any other communication devices during the exam is strictly prohibited.
(b) Use of calculator is allowed after prior approval from the invigilators.

<table>
<tr><th colspan="4">SECTION A (5Q x 4M = 20 Marks)</th></tr>
<tr><th>S. No.</th><th>Statement (s) of the question (s)</th><th>Marks</th><th>CO</th></tr>
<tr><td>Q 1</td><td>Discuss how the number of parameters affects the bias and variance of an algorithmic model</td><td>4</td><td>CO2</td></tr>
<tr><td>Q 2</td><td>Discuss the advantages and limitations of commonly used clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering. In what scenarios would you prefer one method over the others?</td><td>4</td><td>CO2</td></tr>
<tr><td>Q 3</td><td>What is the need of dimensionality reduction? Discuss the two different ways of dimensionality reductions in supervised and unsupervised machine learning?</td><td>4</td><td>CO1</td></tr>
<tr><td>Q 4</td><td>For the sample dataset below, estimate the root mean squared error. (show the steps)
<b>Predicted values: 2.5 3 4.8 8.6</b>
<b>Actual values: 2.1 3 4.2 9</b></td><td>4</td><td>CO2</td></tr>
<tr><td>Q 5</td><td>Provide examples of numeric data and categorical data (four examples each).</td><td>4</td><td>CO1</td></tr>
<tr><th colspan="4">SECTION B (4Q x 10M = 40 Marks)</th></tr>
<tr><td>Q 6</td><td>A chemical engineer is monitoring a production process and collects the following data on Temperature (X1), measured in °C and Pressure (X2), measured in kPa
<table>
<tr><td><b>Temperature (X1)</b></td><td>70</td><td>75</td><td>80</td><td>85</td><td>90</td></tr>
<tr><td><b>Pressure (X2)</b></td><td>200</td><td>215</td><td>220</td><td>235</td><td>238</td></tr>
</table>
To understand the underlying patterns and reduce the dimensionality of the data, the engineer aims to perform <b>Principal Component Analysis (PCA)</b>. You need to:
1. <b>Standardize</b> the data (both variables).
2. Calculate the <b>covariance matrix</b> of the standardized data.</td><td>10</td><td>CO3</td></tr>
</table>

	3. Find the **eigenvalues** and **eigenvectors** of the covariance matrix. 4. Determine the **principal components** and the proportion of variance explained by each. Project the original data onto the first principal component.		
Q 7	Explain the structure of a typical neural network. How do layers in such a network interact? Discuss training of neural network with the back propagation method using the cost function. Use the chain rule to show how the weights and bias values are trained.	**10**	**CO2**
Q 8	Analyze the dataset shown in **Table 1** , and construct a decision tree for performing binary classification. Show all the necessary calculations and detailed reasonings involved in the process.	**10**	**CO3**
Q 9	(a) For 100 test patients, 60 patients actually have disease, and 40 patients actually do not have disease. After running a classification model, the following test (predicted by the model) results were obtain *i.e.* 50 patients were predicted to have disease and 50 patients were predicted not to have the disease. However, out of 50 patients predicted to have disease, only 45 patients actually have the disease. And, out of the other 50 patients predicted not to have disease, only 15 patients actually have the disease. Construct the confusion matrix for present test result. (b) Calculate the Euclidean distance between the two points, *i.e.* x_1 (0, 1, 3, 5) and x_2 (2, -1, 2, 1). The features values are mentioned inside the brackets.	**6 + 4**	**CO3**
SECTION C (2Q x 20M = 40 Marks)			
Q 10	(i) The training dataset for a support vector machine (SVM) is shown in **Table 2**. Use the trained model and determine (or predict) the target of the unknown data, $\hat{y}_i$ (i.e. shown in data for prediction). Use the Lagrange multiplier = (last digit of roll number + 1.2) × 0.01 (compulsory) (ii) Use appropriate evaluation metric to determine the performance of the SVM model. (The actual target is provided for evaluating the performance)	**16 + 4**	**CO3**
Q 11	(i) Applying gradient descent algorithm (compulsory), estimate the coefficients to be used in a linear regression model. The dataset is provided in **Table 3**. Show the detailed calculations (for at-least two iterations) and report the value of the cost function in each iteration. Here, learning rate = (last digit of roll number + 1.2) × 0.01 (compulsory) (ii) Use the model for predict of target value for the data point, $x_1 = 4$, $x_2 = 5$, $x_3 = 0.4$	**18 + 2**	**CO4**

Table 1: Dataset for reaction yield at various conditions.

Features			Target
Reactor type	**Reactant concentration**	**Temperature (°C)**	**Reaction yield is High**
Batch	Low	450	Yes
Continuous	Medium	550	No
Batch	High	400	No
Continuous	Low	500	Yes
Batch	Medium	475	Yes
Continuous	High	600	No

Table 2: Dataset for training and prediction using support vector machine.

Training dataset				*Data for prediction*			
x_1	x_2	Target, y_i		x_1	x_2	Target predicted by model, $\hat{y}_i$	Actual Target, y_i
1	2	2		7	7		49
3	3	9		3	4		12
2	4	8		2	3		6
5	5	25		--	--	--	--
4	6	24		--	--	--	--

Table 3: Dataset for linear regression.

x_1	x_2	x_3	y_i
1	2	0.1	5.01
2	3	0.2	8.04
3	4	0.3	11.09
5	6	0.5	17.25
6	7	0.6	20.36