Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – May / June 2026

Date: 30 -05-2026 **Time:**01:00pm – 04:00pm

School: SOCSE	**Program:** B.Tech	
Course Code : CAI2500	**Course Name:** Machine Learning	
Semester: IV	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	CO1	CO2	CO3	CO4	CO5
Marks	24	14	14	24	24

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	Differentiate supervised and unsupervised ML with an example for each.	**2 Marks**	**L2**	**CO1**
2.	Find the Gini index and classification error of a population with a class distribution as {Male : 30, Female : 60}	**2 Marks**	**L3**	**CO1**
3.	Differentiate Gaussian and Bernoulli Naïve Bayes	**2 Marks**	**L2**	**CO2**
4.	For a multi-class data set with 6 class labels, how many binary classifiers are trained using the one-against-rest and one-against-one approach?	**2 Marks**	**L3**	**CO2**
5.	What is model underfitting? Write any two reasons for a model to underfit.	**2 Marks**	**L1**	**CO3**
6.	What is ensemble learning? Name any two such methods.	**2 Marks**	**L1**	**CO3**
7.	Consider two itemsets Y = {milk, honey}, Z = {milk}. Is the statement support(Z) < support(Y) true or false? Justify your answer.	**2 Marks**	**L3**	**CO4**
8.	For the association rule X → Y, define lift measure.	**2 Marks**	**L1**	**CO4**
9.	Define silhouette coefficient of an instance in clustering.	**2 Marks**	**L1**	**CO5**
10.	What are outliers? Give an application of outlier detection.	**2 Marks**	**L1**	**CO5**

Part B

Answer the Questions. **Total Marks 80M**

<table>
<tr><td>11.</td><td>a.</td><td>Using the following data set, predict the family of the given fruit with (length, width) as (4.3, 3.4) using KNN algorithm where k=3
<table>
<tr><th>S.No</th><th>Fruit length</th><th>Fruit Width</th><th>Fruit Family</th></tr>
<tr><td>1</td><td>7.2</td><td>4.2</td><td>Watermelon</td></tr>
<tr><td>2</td><td>3.2</td><td>2.6</td><td>Apple</td></tr>
<tr><td>3</td><td>2.7</td><td>1.9</td><td>Apple</td></tr>
<tr><td>4</td><td>2.6</td><td>2.7</td><td>Lemon</td></tr>
<tr><td>5</td><td>8.3</td><td>5.1</td><td>Watermelon</td></tr>
<tr><td>6</td><td>1.8</td><td>1.7</td><td>Lemon</td></tr>
<tr><td>7</td><td>4.3</td><td>3.1</td><td>Watermelon</td></tr>
<tr><td>8</td><td>3.2</td><td>2.1</td><td>Apple</td></tr>
<tr><td>9</td><td>2.1</td><td>2.3</td><td>Lemon</td></tr>
</table></td><td>10 Marks</td><td>L3</td><td>CO1</td></tr>
<tr><td colspan="6">Or</td></tr>
<tr><td>12.</td><td>a.</td><td>Given the following training set for classification problem into two classes “spam” and “email”. Find the best splitting feature at the root node of the decision tree using Gini index as the measure of node impurity.
<table>
<tr><th>A1</th><th>A2</th><th>Class</th></tr>
<tr><td>T</td><td>F</td><td>spam</td></tr>
<tr><td>T</td><td>T</td><td>spam</td></tr>
<tr><td>T</td><td>T</td><td>spam</td></tr>
<tr><td>T</td><td>F</td><td>email</td></tr>
<tr><td>T</td><td>T</td><td>spam</td></tr>
<tr><td>F</td><td>F</td><td>email</td></tr>
<tr><td>F</td><td>F</td><td>email</td></tr>
<tr><td>F</td><td>F</td><td>email</td></tr>
<tr><td>T</td><td>T</td><td>email</td></tr>
<tr><td>T</td><td>F</td><td>email</td></tr>
</table></td><td>10 Marks</td><td>L3</td><td>CO1</td></tr>
<tr><td>13.</td><td>a.</td><td>Consider a multi-class problem where the set of class labels $Y = \{y1, y2, y3\}$. A certain test instance is classified as (+, -, -) by all binary classifiers using the one-against-rest approach of multi-class classifiers. Answer the following:
i) How many binary classifiers are trained totally?
ii) Using a table clearly indicate the positive and negative examples of each binary classifier.
Calculate the final predicted class of the test instance.</td><td>10 Marks</td><td>L3</td><td>CO1</td></tr>
<tr><td colspan="6">Or</td></tr>
</table>

14.	**a.**	**In the following table discretize satisfaction level into 2 bins using 3.5 and 5.5 as candidate splits using entropy based discretization. Find the best of these two splits.**	**10 Marks**	**L3**	**CO1**

Satisfaction Level	**1.0**	**6.0**	**5.0**	**4**	**7**	**3**	**8**	**7**	**5**
Class Label	**+**	**+**	**-**	**+**	**-**	**-**	**-**	**+**	**-**

15.	**a.**	**For the training set given below, using Gaussian Naïve Bayes estimate the class of the given test data** **X = (Refund = Yes, Divorced, Taxable_Income = 120K).** **The sample mean and variance of Evade = 'Yes' for the given taxable income are {90,25} respectively and for Evade = 'No' the mean and variance are {110, 2975} respectively.**	**10 Marks**	**L3**	**CO2**

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Or

16.	**a.**	**Explain SVM classifiers with respect to soft margin, hard margin and kernel tricks.**	**10 Marks**	**L1**	**CO2**

17.	**a.**	**Explain bagging and pasting approaches using its algorithm/step by step procedure.**	**10 Marks**	**L1**	**CO3**

Or

18.	**a.**	**Discuss each of the following with the corresponding formula used in AdaBoost algorithm.** **a) Error rate ε_i and importance α_i of the base classifiers.** **b) Weight update at the end of each iteration** **c) Final ensemble C^***	**10 Marks**	**L1**	**CO3**

<table>
<tr><td>19.</td><td>a.</td><td>With a neat flow chart, explain the Apriori algorithm with its property used for candidate generation.</td><td>10 Marks</td><td>L1</td><td>CO4</td></tr>
<tr><td></td><td>b.</td><td>Using Apriori, generate the candidate 4-itemsets C_4 from the frequent 3-itemset $L_3 =$ $\{\{1,2,3\},\{1,2,4\},\{1,3,4,\},\{1,3,5\},\{2,3,4\}\}$</td><td>10 Marks</td><td>L3</td><td>CO4</td></tr>
<tr><td colspan="6">Or</td></tr>
<tr><td>20.</td><td>a.</td><td>Explain the support, confidence and lift measures used to evaluate association rules with an example for each.</td><td>10 Marks</td><td>L1</td><td>CO4</td></tr>
<tr><td></td><td>b.</td><td>Assume min_conf = 75% and Support(X, Y) = 36% , Support(X) = 48%, Support(Y) = 60%, find the strong rules that can be generated from the frequent itemset L_2 = {X, Y}.</td><td>10 Marks</td><td>L3</td><td>CO4</td></tr>
</table>

<table>
<tr><td>21.</td><td>a.</td><td>Explain the types of hierarchical clustering namely, agglomerative and divisive clustering.</td><td>10 Marks</td><td>L1</td><td>CO5</td></tr>
<tr><td></td><td>b.</td><td>Using k-means clustering with K = 2 and with A and C as initial seeds, cluster the following points. Show the detailed working and the cluster memberships in each iteration.
<table>
<tr><th>Objects</th><th>X1</th><th>X2</th></tr>
<tr><td>A</td><td>1</td><td>1</td></tr>
<tr><td>B</td><td>1</td><td>0</td></tr>
<tr><td>C</td><td>0</td><td>2</td></tr>
<tr><td>D</td><td>2</td><td>4</td></tr>
<tr><td>E</td><td>3</td><td>5</td></tr>
</table></td><td>10 Marks</td><td>L3</td><td>CO5</td></tr>
<tr><td colspan="6">Or</td></tr>
<tr><td>22.</td><td>a.</td><td>Discuss the cluster validity measure WCSS and the elbow method used to find the optimal value of k for k-means clustering.</td><td>10 Marks</td><td>L1</td><td>CO5</td></tr>
<tr><td></td><td>b.</td><td>Given the distance matrix of 5 objects below, cluster them using hierarchical agglomerative complete linkage clustering. Show the updated distance matrix in each iteration and draw the final dendrogram.
<table>
<tr><th>Object</th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th></tr>
<tr><td>A</td><td>0</td><td></td><td></td><td></td><td></td></tr>
<tr><td>B</td><td>2</td><td>0</td><td></td><td></td><td></td></tr>
<tr><td>C</td><td>6</td><td>3</td><td>0</td><td></td><td></td></tr>
<tr><td>D</td><td>11</td><td>9</td><td>7</td><td>0</td><td></td></tr>
<tr><td>E</td><td>9</td><td>8</td><td>5</td><td>4</td><td>0</td></tr>
</table></td><td>10 Marks</td><td>L3</td><td>CO5</td></tr>
</table>