Discipline : Computer Science and Engineering

Paper Code: Paper Name: **DATA MINING & KNOWLEDGE DISCOVERY**

Time Allotted : 3 hrs Full Marks : 70

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – B

2 a) Define the mathematical model for Naïve Bayes Classifier. 4 + 8 = 12

b) The following table provides the data of a set of officers. Use Naïve Bayes classifier to classify an officer's gender who is blue-eyed, over 170cm tall and has long hair.

Sl No	Over 170CM	Eye	Hair length	Gender
1	No	Blue	Short	Male
2	Yes	Brown	Long	Female
3	No	Blue	Long	Female
4	No	Blue	Long	Female
5	Yes	Brown	Short	Male
6	No	Blue	Long	Female
7	Yes	Brown	Short	Female
8	Yes	Blue	Long	Male

3 a) Write the k-NN (k nearest neighbour) algorithm to classify a set of data points having *n* features belonging to *m* classes. 4 + 8 = 12

b) Consider the following training set in the 2-dimensional Euclidean space:

x	y	Class
-1	1	1
0	1	2
0	2	1
1	-1	1
1	0	2
1	2	2
2	2	1
2	3	2

(i) What is the prediction of the 3-nearest-neighbor classifier at the point (1,1)?
(ii) What is the prediction of the 5-nearest-neighbor classifier at the point (1,1)?
(iii) What is the prediction of the 7-nearest-neighbor classifier at the point (1,1)?
Note: Show the computations.

Please do not write questions below this line

Paper setter: ……

Moderator : ……

Discipline : Computer Science and Engineering

Paper Code: Paper Name: **DATA MINING & KNOWLEDGE DISCOVERY**

Group - C

4 Define Information gain, Gain Ratio and Gini Index. 5 + 7 = 12

Consider the following set of training examples:

Instance	Classification	A1	A2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

What are the information gains of a1 and a2 relative to these training examples? Provide the equation for calculating the information gain as well as the intermediate results.

5 a) Write short notes on any two of the followings: 4 X 2 + 4 = 12

I. Data space and feature space
II. Support vector
III. Margin in Support Vector Machine

b) Suppose a support vector machine for separating pluses from minuses finds a plus support vector at the point x_1= (1, 0), a minus support vector at x_2= (0, 1). You are to determine values for the classification vector w and the threshold value b.

Group – D

6 (a) Define support and confidence in mining frequent pattern mining. 3+ 4 + 5 = 12

A market basket dataset is given in the following table

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

(b) Compute the support for item-sets {e}, {b, d} and {b, d, e} by treating each transaction ID as a market basket.

Please do not write questions below this line

Paper setter: ……

Moderator : ……

<table>
<tr><td></td><td>(c) Use the results in part (b) to compute the confidence for the association rules {b, d} ⟶ {e} and {e} ⟶ {b, d}. Is confidence a symmetric measure?</td><td></td></tr>
<tr><td>7</td><td>Construct the FP-tree for the transaction database provided in question 1 and find all frequent item-sets using FP-growth approach.</td><td>5 + 7 = 12</td></tr>
<tr><td></td><td>Group – E</td><td></td></tr>
<tr><td>8</td><td>a) Perform K-means clustering on all the points in the following table, where K=2. Randomly select the initial seeds and perform the algorithm for two iterations.
<table>
<tr><th>Points</th><th>X co-ordinate</th><th>Y co-ordinate</th></tr>
<tr><td>p1</td><td>1</td><td>9</td></tr>
<tr><td>p2</td><td>2</td><td>10</td></tr>
<tr><td>p3</td><td>7</td><td>4</td></tr>
<tr><td>p4</td><td>10</td><td>3</td></tr>
<tr><td>p5</td><td>5</td><td>9</td></tr>
<tr><td>p6</td><td>7</td><td>2</td></tr>
<tr><td>p7</td><td>3</td><td>8</td></tr>
<tr><td>p8</td><td>4</td><td>10</td></tr>
<tr><td>p9</td><td>8</td><td>1</td></tr>
<tr><td>p10</td><td>9</td><td>3</td></tr>
</table>
b) Describe the major drawbacks of K-means algorithm for clustering.</td><td>8 + 4= 12</td></tr>
<tr><td>9</td><td>a) Define Core point, Border Point and Noise point in the perspective of DBSCAN clustering algorithm.
b) Describe the DBSCAN Algorithm
c) Explain why DBSCAN does not work well for the data having varying density.</td><td>3 + 6 + 3 = 12</td></tr>
</table>

Discipline : Computer Science and Engineering

Paper Code: Paper Name: **DATA MINING & KNOWLEDGE DISCOVERY**

Please do not write questions below this line

Paper setter: ……

Moderator : ……