(1)

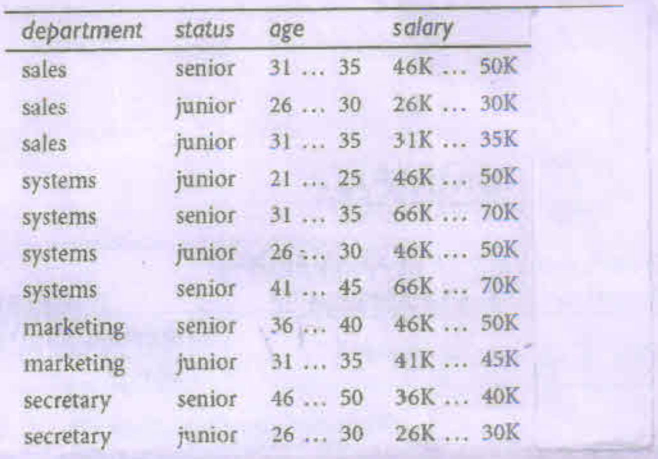
25.05.2023 (E)

SOMAIYA
VIDYAVIHAR UNIVERSITY

Maximum Marks: 100	Semester: January 2023 –May 2023 Examination: ESE Examination	Duration:3 Hrs.
Programme code: 01 Programme: B.Tech Computer Engineering	Class: TY	Semester:_ VI _(SVU 2020)
Name of the Constituent College: K. J. Somaiya College of Engineering		Name of the department: Computer
Course Code: 116U01E622	Name of the Course: Data Mining and Business Intelligence	
Instructions: 1)Draw neat diagrams 2) All questions are compulsory 3) Assume suitable data wherever necessary		

<table>
<tr><th>Que. No.</th><th>Question</th><th>Max. Marks</th></tr>
<tr><td>Q1</td><td>Solve any <b>TWO</b></td><td><b>20</b></td></tr>
<tr><td>i)</td><td>Explain Multilevel Association Rules and Multidimensional Association Rules with relevant examples.</td><td>10</td></tr>
<tr><td>ii)</td><td>Differentiate between simple linkage, average linkage and complete linkage algorithms. Use complete linkage algorithm to find the clusters from the following dataset.
<table>
<tr><td>X</td><td>4</td><td>8</td><td>15</td><td>24</td><td>24</td></tr>
<tr><td>Y</td><td>4</td><td>4</td><td>8</td><td>4</td><td>12</td></tr>
</table></td><td>10</td></tr>
<tr><td>iii)</td><td>Describe various methods for handling missing values in real-world data.</td><td>10</td></tr>
<tr><td>Q2 A</td><td>Perform k-mediods for the following data points with k=2
(8,7), (3,7), (4,9), (9, 6), (8, 5), (5,8)</td><td>10</td></tr>
<tr><td></td><td>OR</td><td></td></tr>
<tr><td>Q2 A</td><td>A node in the decision tree represents a splitting attribute and the branch of the node represents the different outcomes i.e. different values the node can take. 1. Explain the different types of partitioning the tuples when the splitting attribute is
a) discrete-valued
b) discrete-valued with only 2 outcomes
c) continuous valued.
2. Exemplify your answers</td><td>10</td></tr>
<tr><td>Q 2 B</td><td>Solve any <b>One</b></td><td><b>10</b></td></tr>
<tr><td>i)</td><td>Explain in detail Click stream mining.</td><td>10</td></tr>
<tr><td>ii)</td><td>The following table consists of training data from an employee database. The data have been generalized. For example, "31 … 35" for <i>age</i> represents the age range of 31 to 35. For a given row entry, [illegible] represents the number of data tuples</td><td>10</td></tr>
</table>

having the values for *department, status, age*, and *salary* given in that row.



department	*status*	*age*	*salary*
sales	senior	31 … 35	46K … 50K
sales	junior	26 … 30	26K … 30K
sales	junior	31 … 35	31K … 35K
systems	junior	21 … 25	46K … 50K
systems	senior	31 … 35	66K … 70K
systems	junior	26 … 30	46K … 50K
systems	senior	41 … 45	66K … 70K
marketing	senior	36 … 40	46K … 50K
marketing	junior	31 … 35	41K … 45K
secretary	senior	46 … 50	36K … 40K
secretary	junior	26 … 30	26K … 30K

Let *status* be the class label attribute.

(i) Use your algorithm to construct a decision tree from the given data.

(ii) Given a data tuple having the values *"systems," "26 . . . 30,"* and *"46–50K"* for the attributes *department, age*, and *salary*, respectively, what would a naive Bayesian classification of the *status* for the tuple be?

Q3	Solve any **Two**	**20**
i)	Briefly outline with example, how to compute the dissimilarity between objects described by the following: i. Nominal Attributes ii. Asymmetric binary attributes	10
ii)	Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters. A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only (i) The three cluster centers after the first round of execution and (ii) The final three clusters	10
iii)	Explain why data integration is required as a part of pre-processing in data mining? List all the methods of data reduction. Explain in detail the stratified sampling technique with an example applying it for any real time data.	10
Q4	Solve any **Two**	**20**
i)	Discuss examples of any 5 data mining tasks using real world database.	10
ii)	Explain KDD Process with a relevant diagram. Discuss the applications and issues in data mining.	10
iii)	A database has five transactions. Let min sup = 60% and min conf = 80%. (a) Find all frequent itemsets using Apriori (b) Generate association rules	10

Transaction – ID	Items
t100	M, O, N, K, E, Y
t200	D, O, N, K, E, Y
t300	M, A, K, E
t400	M, U, C, K, Y
t500	C, O, O, K, I, E

Q5	Write short notes on any **four**	**20**
i)	Business Intelligence Systems	5
ii)	Lift	5
iii)	Correlation Analysis	5
iv)	Logistic Regression.	5
v)	Decision Support System	5