26/05/2022(E)

K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)

End Semester Examinations
May-June 2022

Max. Marks: 100 **Duration: 3 hrs.**
Class: T Y B.Tech Semester: VI
Name of the Course: **Data Mining and Business Intelligence**
Branch: Computer
Course Code: 2UCE602

Instructions:
(1) **All Questions are Compulsory**
(2) **Draw neat diagrams**
(3) **Assume suitable data if necessary**

Question No.		Marks
Q 1 (a)	i) Suppose that you are employed as a data mining consultant for an Somaiya Vidhyavihar University. Describe how data mining can help the university by giving specific examples of any 2 data mining techniques that can be applied. ii) With diagram, explain the architecture of typical data mining system you would choose for the given above case study? What is purpose of each component?	10
Q 1 (b)	i) Suppose the dataset for analysis includes the attributes - frequency of stop words in documents. The values given are: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52,70. Use the smoothing by bin means with a depth of 3. Give the five-box summary of the data. ii) Consider the sample dataset of loan dispersed by Peer to peer lending firm. Discuss any 2 better strategies to handle missing values. The description of columns in the dataset is as follows. Total Loan Repayments- Total number of loans previously paid by the borrower on the platform, Credit Score- Credit score of the borrower of the loan, State- State in which the loan was issued, Monthly EMI- Monthly EMI paid by the borrower, Monthly Income- Monthly income of the borrower, Liquidity ratio- Monthly EMI/ Monthly Income, Defaulted- If the borrower defaulted on the loan (dependent variable)	10

Total loan repayments	Credit Score	State	Monthly EMI	Monthly Income	Liquidity ratio	Defaulted
0	680	AP	15000	32000	High	Y
1	600	GUJ	16000	20000	Very High	N
0	700	GUJ	9000	45000	High	N
2	720	MAH	4000	34000	Low	N
0	650		23000	33000	High	Y

Q2 (a) The following table shows the data collected by a state highway patrol safety division on stopping distances. 10

Speed(mph)	Stopping Distance (ft)
12	17.5
15	28
20	41.5
25	56
32	77.5
37	104
42	122
47	158
52	177.5
32	80

Use the method of least squares to formulate an equation for the prediction of stopping distance of a vehicle according the speed traveled.
Predict the stopping distance for a vehicle travelling at 48 mph.

OR

i. Suppose a bank would like to develop a classifier that guards against fraudulent credit card transactions. Discuss different strategies as to how to induce a quality classifier based on a large set of non-fraudulent examples and a very small set of fraudulent cases.

ii. You are given 10 training samples. They are divided into 4 classes: A,B,C,D. One sample belongs to A, two to B, three belong to C, and four belong to D. Use the following table to answer the following questions:

p	$Log_2(p)$
0.1	-3.32
0.2	-2.32
0.3	-1.74
0.4	-1.32
0.5	-1.0
0.6	-0.74

What is the total information contained in the samples?
What is the total Gini index?

<table>
<tr><td>Q2 (b)</td><td>Explain the steps of the ID3 algorithm for generating Decision trees.

OR

Imagine that you created a large and complex decision tree from a training dataset that contain many attributes.Your tree has a high accuracy on the training set, but the accuracy on the test set is very low. Explain why the accuracy on the test data could be so much lower than the training data.
Discuss briefly methods to improve your decision tree accuracy on test data set</td><td>10</td></tr>
<tr><td>Q3 (a)</td><td>Consider the 2-D points and Euclidean distance matrix given below.
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2) , A8=(4,9)

<table><tr><th></th><th>A1</th><th>A2</th><th>A3</th><th>A4</th><th>A5</th><th>A6</th><th>A7</th><th>A8</th></tr><tr><td>A1</td><td>0</td><td>$\sqrt{25}$</td><td>$\sqrt{36}$</td><td>$\sqrt{13}$</td><td>$\sqrt{50}$</td><td>$\sqrt{52}$</td><td>$\sqrt{65}$</td><td>$\sqrt{5}$</td></tr><tr><td>A2</td><td></td><td>0</td><td>$\sqrt{37}$</td><td>$\sqrt{18}$</td><td>$\sqrt{25}$</td><td>$\sqrt{17}$</td><td>$\sqrt{10}$</td><td>$\sqrt{20}$</td></tr><tr><td>A3</td><td></td><td></td><td>0</td><td>$\sqrt{25}$</td><td>$\sqrt{2}$</td><td>$\sqrt{2}$</td><td>$\sqrt{53}$</td><td>$\sqrt{41}$</td></tr><tr><td>A4</td><td></td><td></td><td></td><td>0</td><td>$\sqrt{13}$</td><td>$\sqrt{17}$</td><td>$\sqrt{52}$</td><td>$\sqrt{2}$</td></tr><tr><td>A5</td><td></td><td></td><td></td><td></td><td>0</td><td>$\sqrt{2}$</td><td>$\sqrt{45}$</td><td>$\sqrt{25}$</td></tr><tr><td>A6</td><td></td><td></td><td></td><td></td><td></td><td>0</td><td>$\sqrt{29}$</td><td>$\sqrt{29}$</td></tr><tr><td>A7</td><td></td><td></td><td></td><td></td><td></td><td></td><td>0</td><td>$\sqrt{58}$</td></tr><tr><td>A8</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>0</td></tr></table>
Use the hierarchical agglomerative clustering on the above 8 points and draw dendrograms, clearly showing the relevant distance information using Single link, Complete link.</td><td>10</td></tr>
<tr><td>Q3 (b)</td><td>Suppose that we have the following data:
A(2, 0), B(1,2), C(2,2), D(3,2), E(2,3), F(3,3), G(2,4), H(3,4), I(4,4), J(3,5)
Identify the cluster by applying the k-means algorithm, with k=2. Try using initial cluster centers as far apart as possible. Calculate the quality of the clusters

OR

i) K-means and K-mediods are similar techniques. In what way using mediod instead of mean change clustering. Which of these methods are preferred?

ii) Discuss the strength and weakness of BIRCH clustering algorithm.</td><td>10</td></tr>
<tr><td>Q4 (a)</td><td>How apriori property is used in apriori algorithm ?What are limitation of apriori algorithm. Discuss any 2 approaches to improve the efficiency of apriori algorithm

OR

How do we use different variations of apriori algorithm for mining multilevel association rules.</td><td>10</td></tr>
</table>

<table>
<tr><td>Q4 (b)</td><td>Given the minimum support is 2 and minimum confidence is 70%. Find frequent item set and strong association rules using mining with vertical data format algorithm.
<table><tr><th>TID</th><th>Items in transaction</th></tr><tr><td>1</td><td>A,B,C</td></tr><tr><td>2</td><td>B,C,D,E</td></tr><tr><td>3</td><td>C,D</td></tr><tr><td>4</td><td>A,B,D</td></tr><tr><td>5</td><td>A,B,C</td></tr></table></td><td>10</td></tr>
<tr><td>Q5 (a)</td><td>What is need of Business Intelligence? With suitable diagram, discuss the three main components of Business Intelligence architecture</td><td>10</td></tr>
<tr><td>Q5 (b)</td><td>State the areas where Business intelligence and data mining can be used for Banking and Finance.

OR

State the areas where Business intelligence and data mining can be used for Click stream mining, Telecommunication industry</td><td>10</td></tr>
</table>