Sample Question Format
(For all courses having end semester Full Mark=50)

KIIT Deemed to be University
Online End Semester Examination(Autumn Semester-2021)

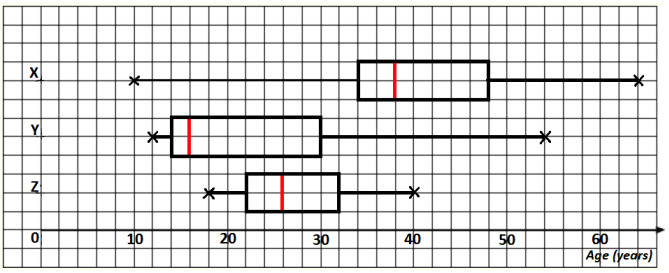
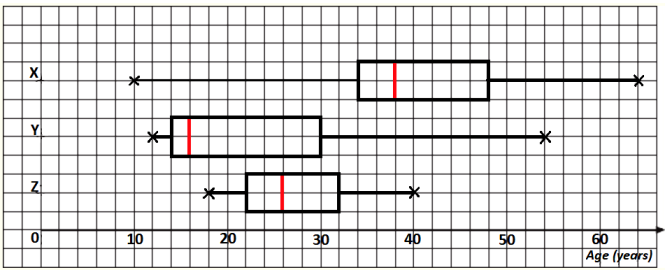
Subject Name & Code: Data Mining and Data Warehousing (IT-3031)
Applicable to Courses:
B.Tech

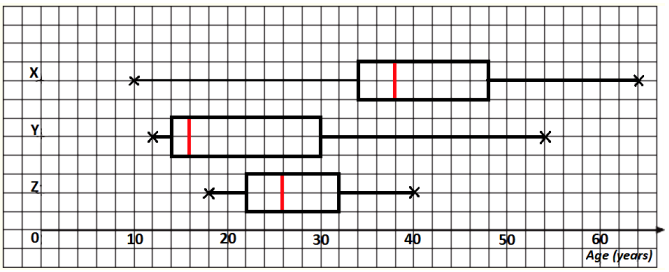
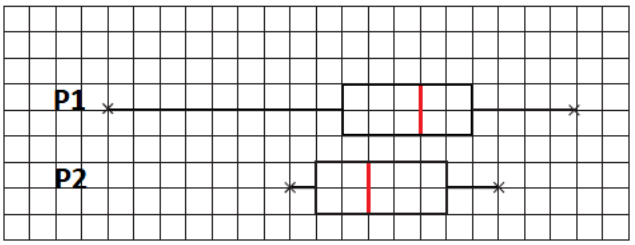
Full Marks=50
Time:2 Hours

SECTION-A(Answer All Questions. Each question carries 2 Marks)

Time:30 Minutes
(7×2=14 Marks)

Question No	Question Type (MCQ/ SAT)	Question	CO Mapping	Answer Key (For MCQ Questions only)
Q.No: 1	MCQ	Major data mining activities include the following general operations except, A. Exploratory data analysis B. Predictive modeling C. Discovering patterns and rules D. Data interpretation	CO1	D
	MCQ	Assumption is satisfied when the probability of missing values in one variable is unrelated to the value of the variable itself or to values of any other variable, A. Assumption of Missing at Random B. Assumption of Missing Completely at Random C. Assumption of Missing Not at Random D. Assumption of Missing Completely not at Random	CO1	B
	MCQ	What are the diffificulties in implementing a data warehouse? (I). Construction, (II). Administration, (III). Quality control, (IV). Building blocks A. I, II, III B. I, II, IV	CO2	A

		C. II, III, IV D. I, III, IV		
	MCQ	Which one of the following is helps to store the data closer to users to enhance the performance. A. Data warehouse B. Data mart C. Metadata D. None	CO2	B
Q.No: 2	**MCQ**	Three different games are playing by three differnt age group (X, Y, and Z) of the people. Players behaviours are visualizing in the following box ploat with whisker, answer the question;  Which game do you think ***you*** (according to your age) would not be allowed to play? A. Game X B. Game Y C. Game Z D. None	CO1/ CO2	C
	MCQ	Three different games are playing by three differnt age group (X, Y, and Z) of the people. Players behaviours are visualizing in the following box ploat with whisker, answer the question;  Which game would ***you*** probably enjoy most? A. Game X B. Game Y C. Game Z D. None	CO1/ CO2	B
	MCQ	Three different games are playing by three differnt age group (X, Y, and Z) of the people. Players behaviours are visualizing in the following box ploat with whisker, answer the question;	CO1/ CO2	A

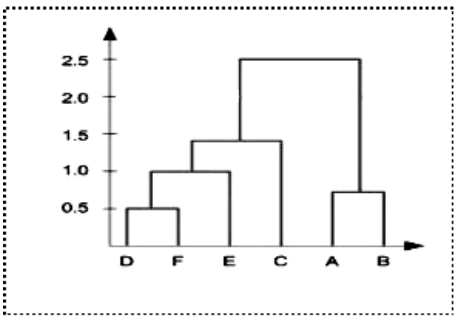
		 Which game would ***your parents*** probably enjoy most? A. Game X B. Game Y C. Game Z D. None		
	MCQ	In an industry, the quality inspector will check the two different types of product that are marked in P1 & P2. The box and whisker plots below shows the results of quality tests on the quality of the respective products.  "The inspector might prefer to use the product (P2) because that P2 has the smaller _____, and the larger _____, which means that industry is less likely to produce a poor product. A. range, minimum B. range, maximum C. median, range D. lower quartile, median	CO1/ CO2	A
<u>Q.No:</u> 3	**MCQ**	In a shoping mall the products are numbered from 01 to 30 are Sampoos and numbered from 31 to 50 are Conditioners. Which product numbers would you include in a systematic sample of size 10 ? A. 10, 20, 30, 40, 50 B. 01, 06, 11, 16, 21, 26, 31, 36, 41, 46 C. 01, 11, 21, 31, 41 D. 05, 10, 15, 20, 25, 30, 35, 40, 45, 50	CO2/CO3	D
	MCQ	Our KIIT University employs the following numbers of faculty members in 3 different positions as; Professor: 10, Associate Prof.: 20, Asst. Prof: 20. How many from each positions should be included in a quota sample of size: 25? A. 3, 11, 11 B. 5, 10, 10	CO1/ CO2	B

<table>
<tr><td></td><td></td><td>C. 10, 5, 10
D. 10, 10, 5</td><td></td><td></td></tr>
<tr><td></td><td>MCQ</td><td>In a picnic trip consists of 40 members of whom 15 are gents. A quota of size 8 is to be selected for site visit. How many ladies and how many gents should be included in the sample?
A. 6, 2
B. 5, 3
C. 2, 6
D. 3, 5</td><td>CO1/ CO2/ CO3</td><td>B</td></tr>
<tr><td></td><td>MCQ</td><td>In a DMDW class 75 students are present of of whom 15 are girls. A quota of size 15 is to be selected for site visit. How many boys and how many girls should be included in the sample?
A. 5, 10
B. 10, 5
C. 3, 12
D. 12, 3</td><td>CO1</td><td>D</td></tr>
<tr><td><u>Q.No: 4</u></td><td>MCQ</td><td>The ring sizes for the customers of a jewellery shop are shown in the table below. What will be the mean value?
<table>
<tr><td>Waist size</td><td>Frequency</td></tr>
<tr><td>4</td><td>2</td></tr>
<tr><td>5</td><td>4</td></tr>
<tr><td>6</td><td>7</td></tr>
<tr><td>7</td><td>5</td></tr>
<tr><td>8</td><td>6</td></tr>
<tr><td>9</td><td>3</td></tr>
<tr><td>10</td><td>3</td></tr>
</table>
A. 7
B. 4.28
C. 30
D. 1.63</td><td>CO1/ CO2</td><td>A</td></tr>
<tr><td></td><td>MCQ</td><td>The number of minutes in a bus stop for a particular bus service was late have been shown in the table as;
<table>
<tr><td>Minutes Late</td><td>Frequency</td></tr>
<tr><td>on time</td><td>19</td></tr>
<tr><td>1-5</td><td>12</td></tr>
<tr><td>6-10</td><td>9</td></tr>
<tr><td>11-20</td><td>4</td></tr>
<tr><td>21-40</td><td>4</td></tr>
<tr><td>41-60</td><td>2</td></tr>
<tr><td>over 60</td><td>0</td></tr>
</table>
Estimate the probability of a bus being more than 20 minutes late.
A. 8%</td><td>CO1/ CO2</td><td>B</td></tr>
</table>

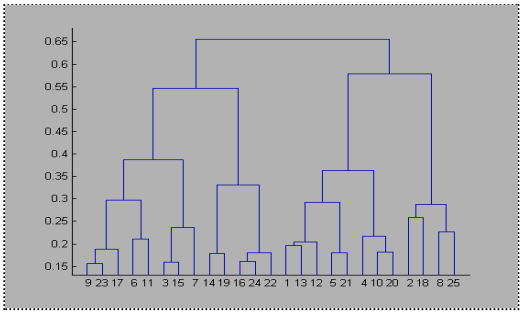
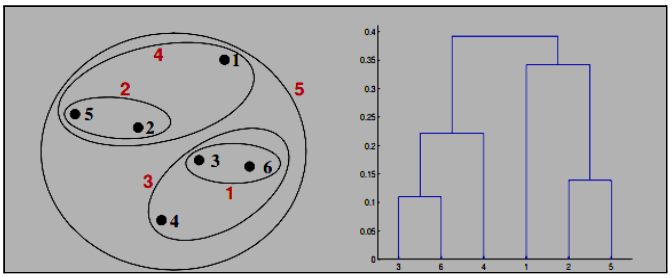
<table>
<tr><td></td><td></td><td>B. 12%
C. 80%
D. 88%</td><td></td><td></td></tr>
<tr><td></td><td>MCQ</td><td>The different items of a customer's basket in a shoping mall are shown in the table below. What will be the mean value?
<table>
<tr><th>Item Codes</th><th>Frequency</th></tr>
<tr><td>3</td><td>2</td></tr>
<tr><td>4</td><td>4</td></tr>
<tr><td>5</td><td>7</td></tr>
<tr><td>6</td><td>5</td></tr>
<tr><td>7</td><td>6</td></tr>
<tr><td>8</td><td>3</td></tr>
<tr><td>9</td><td>3</td></tr>
</table>
A. 4
B. 6
C. 5
D. 1</td><td>CO1/ CO2</td><td>B</td></tr>
<tr><td></td><td>MCQ</td><td>The number of minutes in a bus stop for a particular bus service was late have been shown in the table as;
<table>
<tr><th>Minutes Late</th><th>Frequency</th></tr>
<tr><td>on time</td><td>15</td></tr>
<tr><td>1-5</td><td>10</td></tr>
<tr><td>6-10</td><td>11</td></tr>
<tr><td>11-20</td><td>8</td></tr>
<tr><td>21-40</td><td>4</td></tr>
<tr><td>41-60</td><td>2</td></tr>
<tr><td>over 60</td><td>0</td></tr>
</table>
Estimate the probability of a bus being late of 10 minutes or less.
A. 42%
B. 72%
C. 28%
D. 88%</td><td>CO1/ CO2</td><td>A</td></tr>
<tr><td><u>Q.No:</u>
5</td><td>MCQ</td><td>An itemset {Bread, Milk, Butter} whose support value is 10 ≥ a minimum support threshold is considered as,
A. Itemset
B. Frequent Itemset
C. Infrequent items
D. Threshold values</td><td>CO3</td><td>B</td></tr>
<tr><td></td><td>MCQ</td><td>How do you calculate Confidence (Shoes→Socks), if support of Shoes is 5, support of Socks is 12, and support of togeather purchased is 60?
A. 5
B. 12</td><td>CO3</td><td>B</td></tr>
</table>

<table>
<tr><td></td><td></td><td>C. 1
D. 25</td><td></td><td></td></tr>
<tr><td></td><td>MCQ</td><td>In Association Rule Mining, which combination is correct?
I.Support is never be equal to its confidence
II.Support is always equal to its confidence
III.Support is always greater than its confidence.
IV. Support is always less than its confidence
A. I, II, & III
B. II, III, & IV
C. ALL
D. None</td><td>CO3</td><td>D</td></tr>
<tr><td></td><td>MCQ</td><td>Which of the following assumptions will satisfy, if the same minimum value of support is maintained at each level of an Association Rule Mining?
I. An itemset is not frequent, then none of its supersets can be frequent.
II. An itemset is frequent, then all its supersets are also frequent.
III. An itemset is not frequent, then none of its subsets can be frequent.
IV. An itemset is frequent, then all its subsets are also frequent.
A. I & II
B. II & III
C. I & IV
D. III & IV</td><td>CO3/ CO4</td><td>C</td></tr>
<tr><td><u>Q.No: 6</u></td><td>MCQ</td><td>In a classification model if the model is predicted true for a class value whose actual value was false. Then this is a___.
A. False positive
B. False negative
C. True positive
D. True negative</td><td>CO4</td><td>A</td></tr>
<tr><td></td><td>MCQ</td><td>Which one is incorrect option for support and confidence value for the following transaction data ?
<table><tr><th>TID</th><th>ITEMS</th></tr><tr><td>1</td><td>Bread, milk</td></tr><tr><td>2</td><td>Bread, Diaper, Beer, Eggs</td></tr><tr><td>3</td><td>Milk, diaper, beer, coke</td></tr><tr><td>4</td><td>Bread, milk, diaper, beer</td></tr><tr><td>5</td><td>Bread, milk, diaper, coke</td></tr></table>A. {Diaper,Beer} → {Milk} (s=0.4, c=0.67)
B. {Milk,Diaper} → {Beer} (s=0.4, c=0.67)
C. {Milk} → {Diaper,Beer} (s=0.4, c=0.5)
D. {Milk,Beer} → {Diaper} (s=0.4, c=0.6)</td><td>CO3/ CO4</td><td>D</td></tr>
<tr><td></td><td>MCQ</td><td>Perform KNN for "K=3" on the following dataset and</td><td>CO4</td><td>A</td></tr>
</table>

<table>
<tr><td></td><td></td><td>generate the class level for the input (Acid durability =3 , strength=7, class=?).

<table><tr><th>Name</th><th>Acid durablity</th><th>strength</th><th>class</th></tr><tr><td>Type 1</td><td>7</td><td>7</td><td>bad</td></tr><tr><td>Type 2</td><td>7</td><td>4</td><td>bad</td></tr><tr><td>Type 3</td><td>3</td><td>4</td><td>good</td></tr><tr><td>Type 4</td><td>1</td><td>4</td><td>good</td></tr></table>A. Good
B. Bad
C. Invalid
D. None</td><td></td><td></td></tr>
<tr><td></td><td>MCQ</td><td>In a brute-force approach for mining association rules, the total number of possible rules extracted from a data set that contains 4 items is,
A. 40
B. 50
C. 51
D. 41</td><td>CO5</td><td>B</td></tr>
<tr><td><u>Q.No:</u>
7</td><td>MCQ</td><td>What will be the no of clusters present in the below dendrogram?
 

A. 4
B. 5
C. 2
D. 3</td><td>CO6</td><td>C</td></tr>
<tr><td></td><td>MCQ</td><td>At first iteration the 3 cluster observations using K-means algorithm are: C1: {(3,3), (5,5), (7,7)}; C2: {(0,4), (4,0), (8,8)}; C3: {(0,6), (6,0)}, What will be the cluster centroids if you want to proceed for second iteration?
A. C1: (5,5), C2: (4,0), C3: (6,0)
B. C1: (5,5), C2: (4,4), C3: (3,3)
C. C1: (5,5), C2: (0,4), C3: (0,6)
D. C1: (5,5), C2: (4,0), C3: (6,6)</td><td>CO4</td><td>B</td></tr>
<tr><td></td><td>MCQ</td><td>What will be the possible no of clusters present in the below dendrogram?</td><td>CO5/
CO6</td><td>B</td></tr>
</table>

		 A. 2 B. 4 C. 7 D. 10		
	MCQ	From the following clustering representations and dendrogram, identify the type of link proximity function in hierarchical clustering.  A. MAX or Complete link B. MIN or Single link C. Average link D. None	CO5/ CO6	A

<u>SECTION-B(Answer Any Three Questions. Each Question carries 12 Marks)</u>

<u>Time: 1 Hour and 30 Minutes</u>
<u>(3×12=36 Marks)</u>

Question No	**Question**	**CO Mapping (Each question should be from the same CO(s))**

Q.No:8 CO1/CO2

Find the group mean, median, & mode of the following data scored by students.

Scores	Frequency
1-10	6
11-20	9
21 - 30	11
31 - 40	32
41 - 50	17
51 - 60	22
61 - 70	27
71 - 80	15
81-90	2
91 - 100	3

As per the data list given below, prepare the partition into 4 bins using Equi-depth binning method and perform the smoothing by bin mean, bin median, and bin boundaries.

13,15,15,17,17,18,21,22,22,22,23,23,24,25,26,32,42,47,47,47,73,74, 75,77

The ages of the 112 people who admitted in a hospital are grouped as follows:

Age	Number
0 - 9	20
10 - 19	21
20 - 29	23
30 - 39	16
40 - 49	11
50 - 59	10
60 - 69	7
70 - 79	3
80 - 89	1

Calculate the mean, median & mode of these grouped data according to the above data.

Q.No:9 CO2/CO3

1. The given contingency table summarizes supermarket transaction data, where *Bread* refers to the transactions containing *Bread*, and $\overline{Bread}$ refers to the transactions that do not contain *Bread*. Similarly, *Butter* refers to the transactions containing *Butter*, and $\overline{Butter}$ refers to the transactions that do not contain *Butter*.

	$Bread$	$\overline{Bread}$	$\sum_{Row}$
$Butter$	2000	500	2500
$\overline{Butter}$	1000	1500	2500

$\sum_{Col}$	3000	2000	5000

(a) Suppose that the association rule "***Bread→Butter***" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

(b) Based on the given data, is the purchase of ***Bread*** independent of the purchase of ***Butter***? If not, what kind of *correlation* relationship exists between the two?

Explain what is meant by association rule mining? For the table perform apriori algorithm,

i. Determine the k-itemsets (frequent) obtained.
ii. Justify the strong association rule that has been determined.

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3, 5, 8
04	2, 5, 9, 10
05	1, 4

Assume min_sup=30% and min_conf=75%

Given the following transactional database:

Transactions			
1	C	B	H
2	B	F	S
3	A	F	G
4	C	B	H
5	B	F	G
6	B	E	O

(a) We want to mine all the frequent itemsets in the data using the Apriori algorithm. Assume the minimum support level is 30%.

(b) Find all the association rules that involve only B, C, H (in either left or right hand side of the rule). The minimum confifi dence is 70%.

Q.No:10

Consider the following data table where "Sale" is a class attribute. The training data is shown below.

Price	**Feedback**	**Warranty**	**Sale**
LOW	NO	NO	YES
LOW	NO	YES	YES
HIGH	YES	NO	YES
HIGH	YES	YES	NO
LOW	YES	NO	NO
LOW	YES	YES	YES
HIGH	NO	NO	NO

Build a conditional probability table. Show how Naïve Bayesian method is used to classify the following test data

Price	**Feedback**	**Warranty**
HIGH	YES	NO

CO1/ CO4/ CO5

A simple example from the stock market involving only discrete ranges has Profit as categorical attributes, with values (up, down) and the training data is,

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Apply the decision tree algorithm and show the generated rules.

The following data table shows the details of a second hand car sale company.Apply the ID3 decision tree algorithm and show the suitable generated rules.

Color	Type	Doors	Tires	Class
Red	SUV	2	Whitewall	+
Blue	Minivan	4	Whitewall	-
Green	Car	4	Whitewall	-
Red	Minivan	4	Blackwall	-
Green	Car	2	Blackwall	+
Green	SUV	4	Blackwall	-
Blue	SUV	2	Blackwall	-
Blue	Car	2	Whitewall	+
Red	SUV	2	Blackwall	-
Blue	Car	4	Blackwall	-
Green	SUV	4	Whitewall	+
Red	Car	2	Blackwall	+
Green	SUV	2	Blackwall	-
Green	Minivan	4	Whitewall	-

Q.No:11

CO5/CO6

1. To cluster the following 8 data examples representing locations with (x, y) coordinates & distributed into 3 clusters: A1=(3,11), A2=(3,6), A3=(9,5), A4=(6,9), A5=(8,6), A6=(7,5), A7=(2,3), A8=(5,10).
Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7.

Run the k-means algorithm for 3 epoch and show the clusters.

2. Discuss how Spatial data mining is societally important.

1. The following eight points representing locations with (x, y) coordinates. Initial cluster centers are: A1(4, 12), A4(7, 10) and A7(3, 4). The distance function between two points a = (x_1, y_1) and b = (x_2, y_2) is defined as $D(a, b) = |x_2 - x_1| + |y_2 - y_1|$

Points	x	y
A1	4	12
A2	4	7
A3	10	6
A4	7	10
A5	9	7
A6	8	6
A7	3	4
A8	6	11

Use K-Means Algorithm to find the three cluster centers after the second iteration.

2. Discuss how text data mining is societally important.

1. What is hierarchical clustering? Apply Agglomerative Hierarchical Clustering and draw single link and average link dendogram for the following distance matrix.

	A	B	C	D	E
A	0	2	6	10	9
B	2	0	3	9	8
C	6	3	0	7	5
D	10	9	7	0	4
E	9	8	5	4	0

2. Discuss how multimedia data mining is societally important.