Invigilator's Signature and Date

Set-1

KIIT DEEMED TO BE UNIVERSITY

Spring End Semester Examination-2022

Roll No.	
Registration No.	
Name	
Date of Exam	

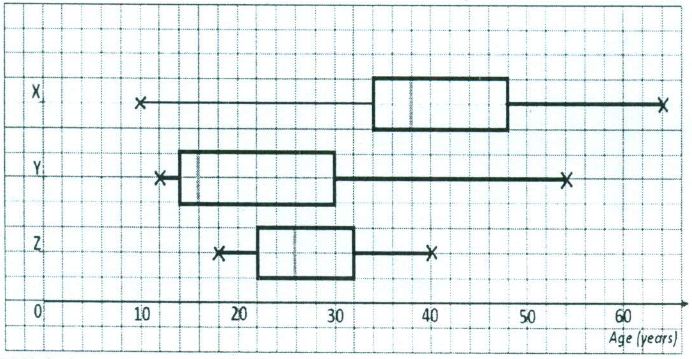
DATA MINING & DATA WAREHOUSING (IT 3031)

6^{th} Semester B.Tech

SECTION-A

(Answer All Questions)

Time: 30 Minutes **Full Marks = 2 × 7 = 14 Marks**

Question No	Question	Write the correct option here.
Q.No:1	Three different games are playing by three different 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	
Q.No:2	A production house has total 360 employee of four different departments. These are- manager: 36, Director: 54, hourly-paid staff: 90, contractual staff: 180. How many from each department will include in a stratified random sampling of size 20? A. 2, 3, 10, 5 B. 3, 2, 5, 10 C. 2, 3, 5, 10 D. 10, 5, 3, 2	

<table>
<tr><td>Q.No:3</td><td>How many frequent 3-itemsets are there for the following transaction data items with minimum_support = 50%?
<table>
<tr><td>TID</td><td>Items</td></tr>
<tr><td>T1</td><td>B, N, P</td></tr>
<tr><td>T2</td><td>B,C, P, N</td></tr>
<tr><td>T3</td><td>B, P, E</td></tr>
<tr><td>T4</td><td>B, N, E, M</td></tr>
<tr><td>T5</td><td>N, C, P, E, M</td></tr>
</table>
A. 1
B. 2
C. 0
D. 3</td><td></td></tr>
<tr><td>Q.No:4</td><td>Choose which data mining task is the most suitable for the following scenario: Customer behavior assessment for promotional offers
A. Prediction
B. Classification
C. Association Rules
D. Anomaly Detection</td><td></td></tr>
<tr><td>Q.No:5</td><td>Given a frequent itemset L, If |L| = k, then there are Select one:
A. 2k – 1 candidate association rules
B. 2k candidate association rules
C. 2k – 2 candidate association rules
D. 2^k -2 candidate association rules</td><td></td></tr>
<tr><td>Q.No:6</td><td>Considering the 2-median algorithm instead of 2-means algorithm. The data points (0, 3), (2, 1), and (-2, 2) are assigned to the first cluster now, what will the new centroid for this cluster?
A. (2, 2)
B. (0, 2)
C. (2, 0)
D. (2, 1)</td><td></td></tr>
<tr><td>Q.No:7</td><td>Which algorithm is expressed by the following list of keywords?
1. Forward Pass
2. Backword Pass
3. Chain Rule
4. Error Propagation

Options:
A. Back propagation
B. Feed-Forward neural network
C. Decision Tree
D. K-NN</td><td></td></tr>
</table>

****

Invigilator’s Signature and Date

Set-1

KIIT DEEMED TO BE UNIVERSITY

Spring End Semester Examination-2022

DATA MINING & DATA WAREHOUSING (IT 3031)

6th Semester B.Tech

SECTION-B

(Answer Any Three Questions.)

Time: 1 Hour and 30 Minutes **Full Marks = 12 × 3 = 36 Marks**

Q.No:8 “Data mining is the process of discovering interesting patterns from massive amounts of data.” Briefly explain the evolutionary process in your words. **(6 Marks)**

The following list of mid semester CSE students of one section of your institute and the median of the data is 24 are given. Find the value of x. **(6 Marks)**

Marks	Number of students
0-10	6
10-20	24
20-30	x
30-40	16
40-50	9

Q.No:9 Suppose that a data warehouse for KIIT University consists of the four dimensions, *student, course, semester, and instructor* and two measures *count* and *avg_grade*. At the lowest conceptual level(e.g., for a given *student, course, semester, and instructor combination*), the *avg_grade* measure stores the actual course grade of a student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

a) Draw a snowflake schema diagram for the data warehouse. **(9 marks)**

b) State the advantages of snowflake schema over star schema for the given problem. **(3 marks)**

Q.No:10 **A.** Given a survey data containing continuous performances in three different activity about the students who are like to opt end-semester examination in two operational modes, either in On-line or Off-line. Using manhattan distance, show how the KNN classifier (let K=3) with majority voting would classify the following student X with {Activity1 =5, Activity2 =7, Activity3 =8}

ID	Activity1	Activity2	Activity3	Opt-to
1	7	7	7	Off-line
2	4	4	5	On-line
3	6	8	8	Off-line
4	7	6	8	Off-line
5	9	5	5	On-line

B. Consider the same training data without the class labels(Opt-to). You are asked to labeling or grouping the data points into 2 groups(Off-line and On-line). Choose suitable approach for generating the clusters and explain the steps.

(6+6 Marks)

Q.No:11 **A.** Define medoid. Create two clusters for given data set using k-medoid clustering for the given data set. Let, the initial cluster medoids are C1 -(4, 5) and C2 -(8, 5) respectively. Execute the process upto 2 iterations or when the stopping condition is met, which one is earlier.

Id	X	Y
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5

B. What are the different methods used for spatial datamining? Explain in detail each of them. What are the major difficulties faced in it? **(6+6 Marks)**

****