KIIT Deemed to be University
Online End Semester Examination(Autumn Semester-2020)

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

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	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8). Compute the Euclidean distance between the two objects. a) 6.7082 b) 6.6209 c) 5.6034 **d)** 5.7062	CO1	A
	MCQ	What is the interquartile range for an even sample size (81,62, 77,63,72,64,70,76,81,64) a) Interquartile range (79-64)=15 b) Interquartile range (77-64)=13 c) Interquartile range (76-64)=12 **d)** Interquartile range (81-64)=17	CO2	B
	MCQ	Which one is correct normalized data after performing min-max normalization by setting min = 0 and max = 1 for the data 200, 300, 400, 600, 1000. a) 0, 0.5, 0.25, 0.125, 1 b) 0, 0.125, 0.25, 0.5, 1 c) 1, 0.5, 0.125, 0.25, 0 **d)** 1, 0.125, 0.25, 0.5, 0	CO1	B
	MCQ	Suppose that the data for analysis includes the attribute age. 13, 15, 16, 46, 19, 20, 20, 21, 22, 22, 25, 25, 25, 52, 30, 33, 35, 35, 33, 35, 35, 36,40, 45, 16, 25, 70. What is the mode, first quartile and third quartile of the data? a) mode=25, first quartile=22, third quartile=35 b) mode=22,25, first quartile=22, third quartile=35	CO1	D

<table>
<tr><td></td><td></td><td>c) mode=25, first quartile=20, third quartile=35
d) mode=22,25, first quartile=20, third quartile=35</td><td></td><td></td></tr>
<tr><td><u>Q.No:2</u></td><td><u>SAT</u></td><td>Define association and co-relation.</td><td>CO3</td><td></td></tr>
<tr><td></td><td><u>SAT</u></td><td>What is Jaccard co-efficient? Explain with example.</td><td>CO1</td><td></td></tr>
<tr><td></td><td><u>SAT</u></td><td>List out the functionality of metadata.</td><td>CO2</td><td></td></tr>
<tr><td></td><td><u>SAT</u></td><td>What is correlation analysis.</td><td>CO3</td><td></td></tr>
<tr><td><u>Q.No:3</u></td><td><u>SAT</u></td><td>What is the need of Information Gain in decision tree classifier?</td><td>CO4</td><td></td></tr>
<tr><td></td><td><u>SAT</u></td><td>What is discrete and continuous data in data mining?</td><td>CO1</td><td></td></tr>
<tr><td></td><td><u>SAT</u></td><td>How to handle tuples with missing values for some attributes?</td><td>CO1</td><td></td></tr>
<tr><td></td><td><u>SAT</u></td><td>Define anti-monotone property.</td><td>CO3</td><td></td></tr>
<tr><td><u>Q.No:4</u></td><td><u>MCQ</u></td><td>For d items, there are _____ possible candidate item set and _____ rules can be created.
a) pow(3,d), pow(2,d)-pow(3,d+1)+1
b) pow(2,d), pow(3,d)-pow(2,d+1)+1
c) pow(2,d), pow(2,d)-pow(3,d+1)+1
d) pow(3,d), pow(3,d)-pow(2,d+1)+1</td><td>CO3</td><td>B</td></tr>
<tr><td></td><td><u>MCQ</u></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</td><td>D</td></tr>
<tr><td></td><td><u>MCQ</u></td><td>Perform KNN for "K=3" on the following dataset and 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</td><td>CO4</td><td>A</td></tr>
</table>

		a) none		
	MCQ	Bayesian classifiers is a) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory. b) Any mechanism employed by a learning system to constrain the search space of a hypothesis c) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation. **d)** None	CO4	A
Q.No:5	**SAT**	What the condition two item sets A and B have no co-relation between them?	CO3	
	SAT	What is tree pruning explain with example.	CO4	
	SAT	Explain ETL with respect to data warehouse.	CO2	
	SAT	What is Bassel's correction?	CO3	
Q.No:6	**MCQ**	Knowledge is referred to a) Non-trivial extraction of implicit previously unknown and potentially useful information from data b) Set of columns in a database table that can be used to identify each record within this table uniquely c) collection of interesting and useful patterns in a database **d)** none of these	CO1	C
	MCQ	A star schema has what type of relationship between a dimension and fact table? a) Many-to-many b) One-to-one c) One-to-many **d)** All of the above	CO6	C
	MCQ	A Snowflake schema is which of the following types of table? a) Fact b) Dimension c) Helper **d)** All of the above	CO6	D
	MCQ	Bias is a) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory	CO6	B

		b) Any mechanism employed by a learning system to constrain the search space of a hypothesis c) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation. **d)** None		
Q.No:7	**SAT**	Differentiate Between classification and clustering?	CO4	
	SAT	What is the difference between discrimination and classification?	CO4	
	SAT	Differentiate Between classification and regression?	CO4	
	SAT	Differentiate between metadata and data mart.	CO2	

SECTION-B
(Answer Any Three Questions. Each Question carries 12 Marks)

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

Q No	Question	co
Q.No:8	Following data given for the attribute age: 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. (a) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0]. (b) Use z-score normalization to transform the value 35 for age, where the standard deviation of ageis 12.94 years. (c) Use normalization by decimal scaling to transform the value 35 for age.	CO2
	Discuss on Normalization? Explain min-max normalization, z-score normalization and decimal scalingmethods using appropriate example.	
	What is Normalization? Use min-max normalization, z-score normalization and decimal scaling methods to normalize the following group of data: 200, 300, 400, 600, 1000	
Q.No:9	Consider the following dataset represented by 5 training example. The target attribute is acceptable which can have values yes and no. Construct a decision tree from the given table. Show the value of information gain for each candidate attribute at each step in the construction of the tree.	CO4

	<table><tr><th>House</th><th>Furniture</th><th>Nos. rooms</th><th>New kitchen</th><th>Acceptable</th></tr><tr><td>1</td><td>No</td><td>3</td><td>Yes</td><td>Yes</td></tr><tr><td>2</td><td>Yes</td><td>3</td><td>No</td><td>No</td></tr><tr><td>3</td><td>No</td><td>4</td><td>No</td><td>Yes</td></tr><tr><td>4</td><td>No</td><td>3</td><td>No</td><td>No</td></tr><tr><td>5</td><td>Yes</td><td>4</td><td>No</td><td>Yes</td></tr></table>	
	Consider the following transactional database T. Let min sup = 60% and min conf = 80%. <table><tr><td>TID</td><td>Items bought</td></tr><tr><td>T100</td><td>{M, O, N, K, E, Y}</td></tr><tr><td>T200</td><td>{D, O, N, K, E, Y }</td></tr><tr><td>T300</td><td>{M, A, K, E}</td></tr><tr><td>T400</td><td>{M, U, C, K, Y}</td></tr><tr><td>T500</td><td>{C, O, O, K, I ,E}</td></tr></table>a) Find all frequent itemsets using Apriorialgorithms. b) Which of the itemsets from a) are closed? Which of the itemsets from a) are maximal? c) Determine strong association rules.	
	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, <table><tr><th>Age</th><th>Competitio</th><th>Type</th><th>Profit</th></tr><tr><td>Old</td><td>Yes</td><td>Software</td><td>Down</td></tr><tr><td>Old</td><td>No</td><td>Software</td><td>Down</td></tr><tr><td>Old</td><td>No</td><td>Hardware</td><td>Down</td></tr><tr><td>Mid</td><td>Yes</td><td>Software</td><td>Down</td></tr><tr><td>Mid</td><td>Yes</td><td>Hardware</td><td>Down</td></tr><tr><td>Mid</td><td>No</td><td>Hardware</td><td>Up</td></tr><tr><td>Mid</td><td>No</td><td>Software</td><td>Up</td></tr><tr><td>New</td><td>Yes</td><td>Software</td><td>Up</td></tr><tr><td>New</td><td>No</td><td>Hardware</td><td>Up</td></tr><tr><td>New</td><td>No</td><td>Software</td><td>Up</td></tr></table>Apply the decision tree algorithm and show the generated rules.	
Q.No:10	What is data mining? Describe the steps involved in data mining when viewed as a process of knowledge discovery.	CO1
	What is data pre-processing. Explain various data pre-processing methods used in data mining.	
	Describe data mining from " Business Intelligence" perspective. What are the various application in data mining.	
Q.No:11	Explain Hierarchical method clustering of classification with example?Construct the single link agglomerative hierarchical clustering for the given distance matrix:	CO5

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 3 & 0 & & \\ 10 & 9 & 7 & 0 & \\ 9 & 8 & 5 & 4 & 0 \end{bmatrix} \end{array}$$

Consider the following data set consisting of the scores of two variables on each of seven individuals.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Apply k-Means algorithm to this data set and grouped into two clusters. As a first step in finding a sensible initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving;

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Consider five points {X1, X2, X3, X4, X5} with the following coordinates as a 2D sample for clustering : X1 =(0,0.25), X2=(0, 0), X3=(1.5, 0), X4=(5, 0), X5=(5, 2). Illustrate K-Means partitioning algorithm using the given dataset for two cluster.