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

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	SAT	What is descriptive and predictive data mining?	CO1	
	SAT	What are the factors that leads to mining the data?	CO1	
	SAT	Data mining is applicable for any kind of information repository. Justify.	CO1	
	SAT	Give examples of incomplete and inconsistent data.	CO1	
Q.No: 2	SAT	Given two objects represented by the tuples (23, 11, 42, 10) and (20, N, 36, 7). Compute the Manhattan distance between the two objects. N = right most significant digit of your Roll No. (Ex:- for Roll No. 180655, N=5)	CO2	
	SAT	Given two objects represented by the tuples (23, 11, 42, 10) and (20, N, 36, 7). Compute the Euclidean distance between the two objects. N = right most significant digit of your Roll No. (	CO2	

		Ex:- for Roll No. 180656, N=6)		
	SAT	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 26, 27, 30, N What is the mean of the data? What is the median? N = 20 + right most significant digit of your Roll No. (Ex:- for Roll No. 180652, N=20+2=22)	CO2	
	SAT	Obtain trimmed 20% mean for the following data: 40,36,48,50,52,52,56,60,63, 80.	CO2	
Q.No: 3	**SAT**	Explian the importance of cross validation.	CO3	
	SAT	How FP growth tree is better than Apriori?	CO3	
	SAT	Does SVM algorithm have high classification accuracy in high dimensional space? Justify your answer.	CO3	
	SAT	What is over fitting and what can you do to prevent it?	CO3	
Q.No: 4	**SAT**	State the advantages of tree pruning.	CO3	
	SAT	What is meant by spatial database? Mention its features.	CO5	
	SAT	List out two major strengths of decision tree method.	CO2	
	SAT	Data marts are becoming popular day by day. Comment.	CO3	
Q.No: 5	**SAT**	Differentiate between median and midrange of a data set.	CO4	
	SAT	List the disadvantages of the k-NN classification technique?	CO4	
	SAT	How to improve the performance of the Apriori method for association mining?	CO4	
	SAT	List out some social impacts of Data mining.	CO4	

Q.No: 6	**SAT**	Differentiate between supervised and unsupervised learning.	CO5	
	SAT	Differentiate between metadata and data mart.	CO5	
	SAT	What is the difference between discrimination and classification?	CO5	
	SAT	List the short comings of K-means algorithm.	CO5	
Q.No: 7	**SAT**	Explain web mining and write some applications of it.	CO6	
	SAT	Explain spatial mining and write some applications of it.	CO6	
	SAT	Explain text mining and write some applications of it.	CO6	
	SAT	Explain multimedia mining and write some applications of it.	CO6	

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

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

Question No	Question	CO Mapping (Each question should be from the same CO(s))
Q.No: 8	What is the need of preprocessing the data before mining it? Explain the various techniques use for data transformation and discretization.	CO2
	What are the different schemas to represent multidimensional data models? What is a data cube? Explain with an example. Describe	

<table>
<tr><td></td><td>the OLAP operations that can be performed on a data cube.</td><td></td></tr>
<tr><td></td><td>Explain different types of attributes with examples. What are the different measures to represent dispersion of data? Discuss the purpose of a boxplot.</td><td></td></tr>
<tr><td><u>Q.No: 9</u></td><td>What is frequent pattern analysis? Explain the apriori property and the general steps of the Apriori algorithm for association rule mining.

The database below has four transactions. Find all frequent itemsets and association rules, if the minimum support is 60% and the minimum confidence is 80%.

Trans_id Itemlist

T1 {K, A, D, B}

T2 {D, A C, E, B}

T3 {C, A, B, E}

T4 {B, A, D}</td><td>CO4</td></tr>
<tr><td></td><td>You are given the transaction data shown in the table below from a fast food restaurant. There are 9 distinct transactions and a total of 5 meal items that are involved in the transactions.

<table><tr><th>Order</th><th>List of Item IDs</th></tr><tr><td>1</td><td>M1, M2, M5</td></tr><tr><td>2</td><td>M2, M4</td></tr><tr><td>3</td><td>M2, M3</td></tr><tr><td>4</td><td>M1, M2, M4</td></tr><tr><td>5</td><td>M1, M3</td></tr><tr><td>6</td><td>M2, M3</td></tr><tr><td>7</td><td>M1, M3</td></tr><tr><td>8</td><td>M1, M2, M3, M5</td></tr><tr><td>9</td><td>M1, M2, M3</td></tr></table>
Assuming the minimum support is 2/9 (.222) and the minimum confidence is 7/9 (.777),

i. Apply the Apriori algorithm to the transactions and identify all frequent itemsets</td><td>CO4</td></tr>
</table>

	ii. Find all strong association rules iii. Identify the interesting association rules out of the strong rules	
	Consider the following set of transactions, find all frequent itemsets and strong associations rules. Assume Support threshold=50%, Confidence threshold= 60%. TID Itemlist T1 I1, I2, I3 T2 I2, I3, I4 T3 I4, I5 T4 I1, I2, I4 T5 I1, I2, I3, I5 T6 I1, I2, I3, I4 List out the advantages and disadvantages of the Apriori algorithm and the different methods to improve its efficiency.	CO4
Q.No: 10	Explain K-means algorithm. Solve the following problem using K-means algorithm. Consider 10 points {X1, X2, X3, X4, X5, X6, X7, X8, X9, X10} with the following coordinates into two clusters: X1(1, 1), X2(1, 3), X3(2, 2), X4(3, 1), X5(4, 4), X6(8, 1), X7(8, 3), X8(9, 2), X9(10, 1), X10(10, 3) Suppose initial cluster center for each cluster are X1, X5 respectively. Use k-means algorithm to show the clusters. A. The two cluster centers after the first round of execution. B. The final two clusters.	CO3
	Define hierarchical methods of clustering. Explain each hierarchical clustering methods with a suitable example. Explain the detailed steps of the DBSCAN algorithm.	
	Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are 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	CO3

<table>
<tr><td></td><td>the k-means algorithm to show

(a) The three cluster centers after the first round of execution.

(b) The final three clusters.</td><td>CO3</td></tr>
<tr><td rowspan="3"><u>Q.No: 11</u></td><td>Suppose that a data warehouse for Big_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 the 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.
(b) Starting with the base cuboid (Student, Course, Semester, Instructor), what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big_University student.
(c) If each dimension has five levels (including all), such as “Student < Major < Status < University < all”, how many cuboids will this cube contain (including the base)?</td><td>CO5</td></tr>
<tr><td>Compare and contrast the characteristics of the four major clustering methods. Give examples of each category.

Discuss the two types of hierarchical clustering methods. In your opinion, which method is more preferable and why? Explain a dendrogram and its use in representing hierarchical clustering.</td><td>CO4</td></tr>
<tr><td>State the Bayes theorem and explain the working of a Naive bayes classifier? What is the major assumption that the Naive Bayes classifier follows?

<table>
<tr><th>T_id</th><th>Refund</th><th>Marital Status</th><th>Taxable Income</th><th>Evade</th></tr>
<tr><td>1</td><td>Yes</td><td>Single</td><td>125K</td><td>No</td></tr>
<tr><td>2</td><td>No</td><td>Married</td><td>100K</td><td>No</td></tr>
<tr><td>3</td><td>No</td><td>Single</td><td>70K</td><td>No</td></tr>
<tr><td>4</td><td>Yes</td><td>Married</td><td>120K</td><td>No</td></tr>
<tr><td>5</td><td>No</td><td>Divorced</td><td>95K</td><td>Yes</td></tr>
<tr><td>6</td><td>No</td><td>Married</td><td>60K</td><td>No</td></tr>
<tr><td>7</td><td>Yes</td><td>Divorced</td><td>220K</td><td>No</td></tr>
<tr><td>8</td><td>No</td><td>Single</td><td>85K</td><td>Yes</td></tr>
<tr><td>9</td><td>No</td><td>Married</td><td>75K</td><td>No</td></tr>
<tr><td>10</td><td>No</td><td>Single</td><td>90K</td><td>Yes</td></tr>
</table>

Given the above dataset for tax evaders, use a Naive bayes classifier to predict whether the following customer will evade tax.

Y = (Refund = No, Marital Status=Married, Income = 120K)</td><td>CO5</td></tr>
</table>