

K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai-22
(Autonomous College Affiliated to University of Mumbai)

Subject Code: AIC502

Subject Name: Data Warehousing and Mining

Date: 05/12/2022

Nov – Dec 2022 (B.Tech) Program: <u>Artificial Intelligence and Data Science</u> Examination: TY Semester: V Course Code: <u>AIC502</u> and Course Name: <u>Data Warehousing and Mining</u> Duration: 2.5 Hours Max. Marks: 60				
Instructions: (1) All questions are compulsory. (2) Draw neat diagrams wherever applicable. (3) Assume suitable data, if necessary.				
		Max. Marks	CO	BT level
Q 1	Solve any six questions out of eight:	12		
i)	What is structured data? Give two drawbacks of structured data.	2	CO1	R
ii)	What are BASE properties of NoSQL databases?	2	CO1	U
iii)	List the different keys used in a fact table? Explain any two.	2	CO3	R
iv)	Write short notes on hierarchical method of clustering	2	CO4	R
v)	Justify that FP growth trees overcome a few drawbacks of Apriori algorithm.	2	CO6	U
vi)	List out any four methods to handle missing attribute values in a dataset.	2	CO2	R
vii)	Draw OLAP cube & perform rollup operation with suitable example	2	CO3	A
viii)	What is information gain? How is it used in a decision tree?	2	CO5	U
Q.2	Solve any four questions out of six.	16		
i)	Apply K means algorithm to divide the set of given values into 3 clusters {2,3,6,8,9,12,15,18,22}	4	CO5	A
ii)	What are the different types of NoSQL databases?	4	CO1	R
iii)	What is the purpose of an Information package	4	CO2	U

K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai-22
(Autonomous College Affiliated to University of Mumbai)

Subject Code: AIC502

Subject Name: Data Warehousing and Mining

Date:05/12/2022

iv)	What is slowly changing dimensions? How is this problem solved? Give an example.	4	CO3	U																					
v)	<p>Calculate Accuracy, Precision and Recall and specificity for the following Confusion Matrix</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th colspan="2" rowspan="2">Cancer Classes</th> <th colspan="3">Predicted</th> </tr> <tr> <th>Yes</th> <th>No</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th rowspan="3">Actual</th> <th>Yes</th> <td>100</td> <td>200</td> <td>300</td> </tr> <tr> <th>No</th> <td>150</td> <td>9550</td> <td>9700</td> </tr> <tr> <th>Total</th> <td>250</td> <td>9750</td> <td>10000</td> </tr> </tbody> </table>	Cancer Classes		Predicted			Yes	No	Total	Actual	Yes	100	200	300	No	150	9550	9700	Total	250	9750	10000	4	CO4	A
Cancer Classes				Predicted																					
		Yes	No	Total																					
Actual	Yes	100	200	300																					
	No	150	9550	9700																					
	Total	250	9750	10000																					
vi)	<p>The transaction details are given in the following table, what is the confidence and support of the association rule {Cake} ⇒ {Coffee, Cookies}?</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>T_id</th> <th>Items bought</th> </tr> </thead> <tbody> <tr> <td>101</td> <td>Butter,Cookies, Cake</td> </tr> <tr> <td>102</td> <td>Butter, Coffee, Cake, Cookies</td> </tr> <tr> <td>103</td> <td>Butter, Cake, Eggs</td> </tr> <tr> <td>104</td> <td>Butter, Cookies, Eggs, Milk</td> </tr> <tr> <td>105</td> <td>Cookies, Coffee, Cake, Eggs, Milk</td> </tr> </tbody> </table>	T_id	Items bought	101	Butter,Cookies, Cake	102	Butter, Coffee, Cake, Cookies	103	Butter, Cake, Eggs	104	Butter, Cookies, Eggs, Milk	105	Cookies, Coffee, Cake, Eggs, Milk	4	CO6	An									
T_id	Items bought																								
101	Butter,Cookies, Cake																								
102	Butter, Coffee, Cake, Cookies																								
103	Butter, Cake, Eggs																								
104	Butter, Cookies, Eggs, Milk																								
105	Cookies, Coffee, Cake, Eggs, Milk																								
Q.3	Solve any two questions out of three.	16																							
i)	Define a data warehouse and elaborate the architecture of a data warehouse	8	CO1	U																					
ii)	<p>Suppose that a data warehouse for sports organisation consists of four dimensions date,seat_category, location and game and two measures count and ticket_price, seat_category may be lower, middle, upper with each category having its own rate</p> <p>a) Draw a star scheme for the data warehouse.</p> <p>b) Starting with the basic cuboid [date,seat_category,location,game], what specific OLAP operation should be performed in order to list the total charge paid by upper level spectators at Mumbai in 2010.</p>	5+3	CO3	An																					

iii)	Divya is a student who loves listening to music while she works. Find below a typical schedule of her day. Identify and apply the appropriate algorithm to find the kind of music divya listens for the tuple {morning,yes, no}.	8	CO4	A																																				
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Time of Day</th> <th>Homework</th> <th>Maths</th> <th>Type of music</th> </tr> </thead> <tbody> <tr> <td>Morning</td> <td>yes</td> <td>no</td> <td>piano</td> </tr> <tr> <td>Morning</td> <td>no</td> <td>no</td> <td>violin</td> </tr> <tr> <td>morning</td> <td>no</td> <td>yes</td> <td>piano</td> </tr> <tr> <td>afternoon</td> <td>yes</td> <td>no</td> <td>piano</td> </tr> <tr> <td>afternoon</td> <td>yes</td> <td>yes</td> <td>violin</td> </tr> <tr> <td>afternoon</td> <td>no</td> <td>no</td> <td>violin</td> </tr> <tr> <td>evening</td> <td>no</td> <td>yes</td> <td>violin</td> </tr> <tr> <td>evening</td> <td>yes</td> <td>yes</td> <td>piano</td> </tr> </tbody> </table>					Time of Day	Homework	Maths	Type of music	Morning	yes	no	piano	Morning	no	no	violin	morning	no	yes	piano	afternoon	yes	no	piano	afternoon	yes	yes	violin	afternoon	no	no	violin	evening	no	yes	violin	evening	yes	yes	piano
Time of Day	Homework	Maths	Type of music																																					
Morning	yes	no	piano																																					
Morning	no	no	violin																																					
morning	no	yes	piano																																					
afternoon	yes	no	piano																																					
afternoon	yes	yes	violin																																					
afternoon	no	no	violin																																					
evening	no	yes	violin																																					
evening	yes	yes	piano																																					
Q.4	Solve any two questions out of three.	16																																						
i)	a)Data for length of an item is given as 9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7 i) Calculate the mean, median, and standard deviation ii) Draw the boxplots b) Calculate Euclidean distance and Manhattan distance between ((30, 70), (40, 54))	8 (3+2+2+1)	CO2	A																																				
ii)	Consider the transaction database given in the table below. Apply Apriori Algorithm with minimum support of 40% and confidence of 50%. Find all frequent itemsets and all the association rules.	8	CO6	An																																				
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>TID</th> <th>I1</th> <th>I2</th> <th>I3</th> <th>I4</th> <th>I5</th> <th>I6</th> <th>I7</th> <th>I8</th> <th>I9</th> </tr> </thead> <tbody> <tr> <td>Items</td> <td>1,3,4</td> <td>2,3,5</td> <td>1,2,3,5</td> <td>2,5</td> <td>1,2,3</td> <td>3,5</td> <td>1,2,3,5</td> <td>1,5</td> <td>1,3</td> </tr> </tbody> </table>					TID	I1	I2	I3	I4	I5	I6	I7	I8	I9	Items	1,3,4	2,3,5	1,2,3,5	2,5	1,2,3	3,5	1,2,3,5	1,5	1,3																
TID	I1	I2	I3	I4	I5	I6	I7	I8	I9																															
Items	1,3,4	2,3,5	1,2,3,5	2,5	1,2,3	3,5	1,2,3,5	1,5	1,3																															
iii)	Show the dendrogram created by single link clustering. Use Manhattan method and agglomerative clustering	8	CO5	A																																				
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>I1</th> <th>I2</th> <th>I3</th> <th>I4</th> <th>I5</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>2</td> <td>11</td> <td>8</td> <td>1</td> <td>10</td> </tr> <tr> <td>B</td> <td>4</td> <td>2</td> <td>3</td> <td>5</td> <td>2</td> </tr> </tbody> </table>						I1	I2	I3	I4	I5	A	2	11	8	1	10	B	4	2	3	5	2																		
	I1	I2	I3	I4	I5																																			
A	2	11	8	1	10																																			
B	4	2	3	5	2																																			