Time: 3 hours **Total Marks: 80**

QP-10065528

Note: 1. Question no.1 is compulsory.
2. Attempt any three out of remaining five.
3. Assumptions made should be clearly indicated.
4. Figures to the right indicates full marks.
5. Assume suitable data whenever necessary.

Q. 1 **Solve any four.** **(05 marks each)**

A Every data structure in the data warehouse contains the time element. Why?

B Explain FP Growth Algorithm.

C Explain different types of attributes.

D Discuss different applications of Web Mining.

E Explain Holdout and Random subsampling method to evaluate the accuracy of classifier.

F Differentiate between Classification and Clustering.

Q.2 **(10 marks each)**

A For a supermarket chain, consider the following dimensions namely product, store, time and promotion. The schema contains a central fact table for sales with three measures unit_sales, dollars_sales and dollar_cost.

1. Draw a star schema.
2. Calculate the maximum number of base fact table records for warehouse with the following values given below:
   - Time period 5 years
   - Store-300 stores reporting daily sales
   - Product-40,000 products in each store (about 4000 sell in each store daily)
   - Promotion- a sold item may be in only one promotion in a store on a given day.

B Explain the different techniques to handle noisy data.

Suppose a group of sales price records has been sorted as follows:

3, 7, 8, 13, 22, 22, 22, 26, 26, 28, 30, 37.

Partition them into three bins by equal-frequency (Equi-depth) partitioning method. Perform data smoothing by bin mean and bin boundary.

Q.3 **(10 marks each)**

A Explain Updates to dimensional table in detail.

B Explain the following data pre-processing methods.

I) Dimensionality reduction II) Data transformation and Discretization

Q.4 **(10 marks each)**

A Given the training data for height classification, classify the tuple,

t=<Rohit, M, 1.95> using Naïve Bayes Classification.

Name	Gender	Height	Output
Kiran	F	1.6m	Short
Jatin	M	2m	Tall
Madhuri	F	1.09m	Medium
Manisha	F	1.88m	Medium
Shilpa	F	1.7m	Short
Bobby	M	1.85m	Medium
Kavita	F	1.6m	Short
Dinesh	M	1.7m	Short
Rahul	M	2.2m	Tall
Shree	M	2.1m	Tall
divya	F	1.8m	Medium
Tushar	M	1.95m	Medium
Kim	F	1.9m	Medium
Aarti	F	1.8m	Medium
Rajashree	F	1.75m	Medium

B Consider four objects with two attribute (X and Y). These four objects are to be grouped together into two clusters using k-means clustering algorithm. Following are the objects with their attribute values.

Object	X	Y
A	1	1
B	2	1
C	4	3
D	5	4

Q. 5 **(10 marks each)**

A Given the following data, apply the Apriori algorithm. Find frequent item set and strong association rules. Given Support threshold=50%, Confidence=60%

Transaction	Items
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

B What is Web Mining? Differentiate between Web Mining and Data Mining. Explain types of Web Mining.

Q. 6 **Write short note on.** **(5 marks each)**

A Decision Tree Induction Algorithm

B K-medoids clustering Algorithm

C Multilevel and multidimensional association rule mining

D Page Rank Algorithm

****

Time: 3 hours **Max. Marks: 80**

==========

Note: 1. Question no.1 is compulsory.
2. Attempt any three out of remaining five.
3. Assumptions made should be clearly indicated.
4. Figures to the right indicates full marks.
5.Assume suitable data whenever necessary.

Question 1 **Write a short note on the following. Solve any four.**

(5 marks each)

A Write a note on web usage mining. Also state its any two applications.

B Describe any five issues in data mining.

C Explain how Naive Bayes classification makes predictions and discuss the "naive" assumption in Naive Bayes. Provide an example to illustrate the application of Naive Bayes in a real-world scenario.

D Suppose the data for clustering is {6,14,18,22,1,40,50,11,25} consider k=2, cluster the given data using k means algorithm.

E Explain the concept of market basket analysis with example.

F Differentiate between ER modeling vs Dimensional modeling.

Question 2 **10 marks each**

A Describe in detail about how to evaluate accuracy of the classifier.

B Illustrate major steps in ETL process.

Question 3 **10 marks each**

A Explain KDD process with neat diagram. Also state any five applications of data mining.

B For the table given perform Apriori algorithm and show frequent item set and strong association rules. Assume Minimum Support of 30% and Minimum confidence of 70%.

TID	Items
1	1,4,6,8
2	2,5,3
3	7,1,3,8
4	9,10
5	1,5

Question 4 **10 marks each**

A A social media platform wants to analyze user engagement data to improve content recommendations and user experience. The INTERACTIONS fact table contains information about user interactions, including interaction details, user information, content details, and time periods. The dimension tables provide additional context about users, content, categories, and time periods. Design a star schema and snowflake schema for the same.

B Explain Multilevel Association Rules Mining and Multidimensional Association Rules Mining with examples.

Question 5 **10 marks each**

A A company wants to predict whether a customer will subscribe to a premium membership based on their demographic and browsing behavior data. The dataset contains information about customers, including age, gender, income, browsing time, and subscription status.

Age	Gender	Income	Browsing Time	Subscription
20-30	Male	High	10am-12pm	Yes
20-30	Female	Medium	2pm-4pm	Yes
30-40	Male	Low	8am-10am	No
30-40	Female	High	4pm-6pm	Yes
>40	Male	Medium	6pm-8pm	Yes
>40	Female	Medium	8am-10am	No
>40	Male	High	12pm-2pm	Yes
20-30	Female	Low	10am-12pm	No
20-30	Male	Medium	2pm-4pm	Yes
30-40	Female	High	8am-10am	Yes

Use ID3 to build the decision tree and predict the following example:

Age	Gender	Income	Browsing Time
20-30	Male	Medium	10am-12pm

B Illustrate page rank algorithm with example.

Question 6 **10 marks each**

A Following table gives fat and proteins content of items. Apply single linkage clustering and construct dendrogram.

Food Item	*Protein*	*Fat*
1	*1.1*	*60*
2	*8.2*	*20*
3	*4.2*	*35*
4	*1.5*	*21*
5	*7.6*	*15*
6	*2.0*	*55*
7	*3.9*	*39*

B Explain in brief what is data discretization and concept hierarchy generation.

Duration:(3 Hours) **[80 Marks]**

N.B. 1) Question No. 1 is compulsory.
2) Attempt any Three questions out of the remaining.
3) Assume suitable data wherever necessary and state them clearly.

Q.1 Solve any four of the following (20)

A. Compare OLTP vs OLAP systems.
B. Explain the KDD process of data mining.
C. Explain any two methods of evaluating the accuracy of a Classifier.
D. Explain K-means clustering algorithm and draw flowchart.
E. Explain multilevel association rule mining with example.
F. Write a short note on web usage mining.

Q.2 A. Consider the following transaction database with minimum support 50% and minimum confidence 66%. Find the frequent patterns and strong association rules. (10)

Tid	Items
10	A,C,D
20	B,C,E
30	A,B,C,E
40	B,E

Q.2 B. Explain different steps involved in data preprocessing. (10)

Q.3 A. Find the clusters for the following dataset using a single link technique. Use Euclidean distance and draw the dendrogram. (10)

Sample No	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Q.3.B. The college wants to record the Marks for the courses completed by students using the dimensions: I) Course, II)Student, III) Time & a measure Aggregate marks .
Create a cube and describe following OLAP operations :
I) Slice II) Dice III) Roll up IV) Drill Down V)Pivot (10)

Q.4.A. What is dimensional modeling? Design the data warehouse dimensional model for a wholesale furniture Company. The data warehouse has to analyze the company's situation at least with respect to the Furniture, Customer and Time. Moreover, the company needs to analyze: The furniture with respect to its type, category and material. The customer with respect to their spatial location, by considering at least cities, regions and states. The company is interested in learning the quantity, income and discount of its sales.. (10)

Q.4 B. A data sample is given below. Find whether Patient X has flu or not using Naïve Bayes classifier.

If X= (chills=Y, runny nose=N, headache=Mild, fever=Y, flu=?) (10)

chills	Runny nose	headache	fever	Flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Q.5 A.Explain Page Rank algorithm with example. (10)
B. Explain different data visualization techniques. (10)

Q.6. Write short notes on following: (20)
A. Applications of Data Mining.
B. FP Tree
C. Web content Mining
D. Techniques of data Loading

****

TE comp| sem V| R-19| ATKT |FH 2023| 31|05|2023

QP code: 27279

Time: 3 hours **Max. Marks: 80**

Note: 1. Question no.1 is compulsory.
2. Attempt any three out of remaining five.
3. Assumptions made should be clearly indicated.
4. Figures to the right indicates full marks.
5. Assume suitable data whenever necessary.

Question 1 **Solve any four.** **5 marks each**

A What are the basic building blocks of Data warehouse?

B Explain Page Rank technique in detail.

C Compare OLTP and OLAP.

D Differentiate between Agglomerative and Divisive clustering method.

E Discuss data visualization Technique.

F Explain issues in Data mining.

Question 2 **10 marks each**

A Explain Decision Tree based Classification Approach with example. Discuss Metrics for evaluating Classifier Performance.

B Describe the steps involved in Data Mining when viewed as a process of Knowledge Discovery.

Question 3 **10 marks each**

A Differentiate between Star schema and Snowflake schema. Design Star schema for company sales with three dimensions such as Location, Item and Time.

B Explain Data Pre-processing.

Question 4 **10 marks each**

A Differentiate between top-down and bottom-up approaches for building data warehouse. Discuss the merits and limitations of each approach. Also explain the practical approach for designing a data warehouse.

B What is Web mining? Explain Web structure Mining and Web Usage Mining in detail.

20

Question 5 **10 marks each**

A Explain multilevel and multidimensional association rule mining in detail.

B A database has five transactions. Let minimum support count = 2 and minimum confidence =60 %. Find all frequent item sets using Apriori Algorithm. List strong association rules.

TID	Items
100	1,3,4
200	2,3,5
300	1,2,3,5
400	2,5
500	1,3,5

Question 6 **10 marks each**

A Explain K-Means clustering algorithm. Discuss its advantages and limitations. Apply K-Means algorithm for the following data set with 3 clusters.

Data Set={2,3,6,8,9,12,15,18,22}

B Consider the data given below. Create adjacency matrix. Apply complete link algorithm to cluster the given data set and draw the dendogram.

	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

TE(COMP) / SEM V / R-19 / FH-22 / DWM / 10.06.22

QP CODE: 91926

University of Mumbai

Examinations Summer 2022

Examinations Commencing from 17th May 2021 to _____

Program: **_Computer Engineering_____**

Curriculum Scheme: Rev2019

Examination: TE Semester: V

Course Code: **CSC504** and Course Name: **Data Warehousing and Mining**

Time: 2 hour 30 minutes Max. Marks: 80

Q1.	**Choose the correct option for following questions. All the Questions are compulsory and carry equal marks**
1.	For the given attribute marks values: 35,45,50,55,60,65,75. Identify the first quartile and third quartile of data.
Option A:	35, 50
Option B:	35,75
Option C:	45,65
Option D:	45,60
2.	OLTP is not
Option A:	Operational database
Option B:	Subject oriented
Option C:	For continuous updates from many sources
Option D:	For high read, write update delete activity
3.	Correcting the customer flat number is
Option A:	Type1 change
Option B:	Type 2 change
Option C:	Type 3 change
Option D:	Type 4 change
4.	In Banking scenario following dimension tables are identified. Point out the inappropriate one.
Option A:	Account
Option B:	Branch
Option C:	Time
Option D:	Transaction
5.	A company would like to improve its sales by analyzing its past data. Which of the following tasks will occupy maximum time required for the return on investment?
Option A:	identifying sources of data
Option B:	ETL process
Option C:	data analysis
Option D:	preparing reports
6.	Suppose we have the following values for salary (in thousands of dollars),

	in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. The mid range of data would be:
Option A:	50,000
Option B:	70,000
Option C:	77,000
Option D:	52,000
7.	Consider the data transaction given below:T1:{F,A,D,B} T2:{D,A,C,E,B} T3:{C,A,B,E} T4:{B,A,D} With minimum support = 60% and the minimum confidence = 80% , which of the following is not valid association rule?
Option A:	A -----> B
Option B:	B ----->A
Option C:	D -----> A
Option D:	A -----> D
8.	Given the Confusion matrix , the accuracy is

Classes	YES	NO
YES	90	210
NO	140	9560

Option A:	60%
Option B:	100%
Option C:	96.5%
Option D:	35%
9.	Data collection is done by crawling through number of web pages in
Option A:	Data Mining
Option B:	Web mining
Option C:	Text Mining
Option D:	Spatial data mining
10.	What is a Dendrogram?
Option A:	A tree diagram used to illustrate the arrangement of clusters in hierarchical clustering
Option B:	A tree diagram used to illustrate the arrangement of clusters in partitional clustering
Option C:	A type of hierarchical clustering
Option D:	A type of bar chart diagram to visualize K-means clusters.

Question 2	Solve any Two out of Three 10 marks each
A	The college wants to record the marks for the courses completed by students using the dimensions: a) Course b) Student c) Time and a measure of Aggregate marks. Create a cube and describe following operations: i) roll up ii) Drill down iii) Slice iv) Dice
B	Discuss the different steps involved in data preprocessing.
C	Consider the following dataset S, which contains observations of several cases of sunburn:

	Name	Hair	Height	Weight	Dublin	Result
	Sarah	Blonde	Average	Light	No	Sunburned
	Dana	Blonde	Tall	Average	Yes	None
	Alex	Brown	Short	Average	Yes	None
	Annie	Blonde	Short	Average	No	Sunburned
	Emily	Red	Average	Heavy	No	Sunburned
	Pete	Brown	Tall	Heavy	No	None
	John	Brown	Average	Heavy	No	None
	Katie	Brown	Short	Light	Yes	None

Unseen sample X= <brown,tall,average,No> Predict the result value as sunburned or None.

Question 3	**Solve any Two out of Three**	**10 marks each**

A

The table below shows the six data points. Apply Agglomerative clustering to find clusters. Use Euclidian distance measure. consider single linkage.

	X	y
D_1	0.4	0.53
D_2	0.22	0.38
D_3	0.35	0.32
D_4	0.26	0.19
D_5	0.08	0.41
D_6	0.45	0.30

B

A database has four transactions .Let min sup =60% and min conf=80%.

TID	Date	Items purchased
T100	21/04/2022	{K,A,D,B}
T200	21/04/2022	{D,A,C,E,B}
T300	22/04/2022	{C,A,B,E}
T400	23/04/2022	{B,A,D}

	Find all the frequent item sets using apriori algorithm and also list all the strong association rules.
C	What is web structure mining? List the approaches used to structure the web pages to improve on the effectiveness of search engines and crawlers. Explain page rank technique in detail

Question 4	Solve any Four Questions out of Six 5 marks each
A	Explain major issues in data mining.
B	Differentiate between OLTP and OLAP.
C	Explain web usage mining in detail.
D	What are the various methods for estimating classifiers accuracy.
E	Use k means algorithm to create 3-clusters for given set of values:{2,3,6,8,9,12,15,18,22}
F	What are the various Issues regarding Classification and Prediction?

2/2

(3 Hours) [Total Marks: 80]

Note: 1. Question no.1 is compulsory.
2. Attempt any three out of remaining five.
3. Assumptions made should be clearly indicated.
4. Figures to the right indicates full marks.
5. Assume suitable data whenever necessary.

Q. 1 Solve any four. **(20)**

A Every data structure in the data warehouse contains the time element. Why?

B In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

C What are the various methods for estimating a classifier's accuracy?

D Explain market basket analysis with an example.

E Describe K medoids algorithm.

F Explain CLARANS extension in web mining.

Q. 2 A Consider the quarterly sales of four companies C1, C2, C3.C4. The dimensions are
a) Time
b) Shopping category (Men's, Women's, Electronics, Home)
c) Company
Create a cube and describe all five OLAP operation. **(10)**

B Apply the Naïve Bayes classifier to classify the tuple <Red, SUV, Domestic> For the given dataset below. **(10)**

Instance no.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	sports	Imported	Yes

Q.3 A Discuss the different types of attributes. (10)

B Suppose that the data mining task is to cluster the following points into 3 clusters .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 we initially assign A1,B1,C1 as the center of each cluster respectively, Use the k means algorithm to show only a) the three cluster centers after the first round of execution b) The final three clusters. (10)

Q.4 A For a supermarket chain, consider the dimensions namely Product, Store, time,promotion. The schema contains the three facts namely units_sales, dollar_sales, and cost_ dollars.
Design a star schema and calculate the maximum number of base fact table records for the values given below:
Time period: 5 years
Stores: 300 reporting daily sales
Product: 40000 products in each store (about 4000 sell daily in each store
Promotion: a sold item may be in only one promotion in a store on a given day. (10)

B A database has five transactions. (10)

T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, K, I, E}

Let minimum support =3, Find all frequent itemsets using FP-growth algorithm.

Q.5 A What is web structure mining? Describe page ranking technique with the help of example. (10)

B Use agglomerative algorithm using the following data and plot a dendrogram using single link approach. The following figure contains sample data items indicating the distance between the elements. (10)

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

Q. 6 A Apply apriori algorithm on the following dataset to find strong association rules. Minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%) **(10)**

Transaction ID	Items
T1	Hot dogs, Buns, Ketchup
T2	Hot dogs, Buns
T3	Hot dogs, Coke, Chips
T4	Coke, Chips
T5	Chips, Ketchup
T6	Hotdogs ,Coke, Chips

B Is Web mining different from classical data mining? Justify your answer. Describe types of web mining. **(10)**