(74)

Total no. of Pages-1

Roll no. .....

FIFTH SEMESTER B. Tech. (Information Technology)
END-SEMESTER EXAMINATION, December- 2021
Course Code- ITITE05
Course Title- Data warehouse and data mining

Time- 3.00 Hours

Max. Marks- 40

Note: - All questions are compulsory. Attempt only two out of three parts from every question. Missing data/ information if any, maybe suitably assumed & mentioned in the answer.

No	Question	Marks	CO
1a	What are the various principles of Data Warehouse?	4	CO1
1b	What are the advantages of using Data Marts?	4	CO1
1c	How is data cleaning different from data transformation?	4	CO3
2a	Explain multi-dimensional modelling with a suitable example of your choice.	4	CO3
2b	What is a data cube? How can it be used in dimensional modelling?	4	CO3
2c	Using "Sales" as a factor, differentiate between Star and Snowflake schema.	4	CO3
3a	Explain some document oriented NoSQL databases with their applications.	4	CO3
3b	What is an OLAP server? Explain the working of ROLAP server with its architecture.	4	CO3
3c	What are the advantages of using a Hybrid OLAP server?	4	CO3
4a	Use the following information to solve for Apiori algorithm with Support=50% and Confidence=60%: **Transaction** — **List of 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	4	CO5
4b	Using the given information, apply KNN algorithm to classify data point(X1=3,X2=7).	4	CO5

	X1 = Acid Durability (seconds) X2 = Strength (kg/square meter) Y = Classification 7 7 Bad 7 4 Bad 3 4 Good 1 4 Good		
4c	Explain the process of KDD using a suitable example and its diagram.	4	CO2
5a	Differentiate between: (i) KNN and K-means clustering (ii) Genetic Algorithm and Neural Networks	2+2=4	CO4
5b	Write short notes on the following: (i) Supervised Learning (ii) Decision Tree	2+2=4	CO2/ CO4
5c	Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9) Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).	4	CO5

END SEMESTER EXAMINATION December 2021

Course Code: COCSC16
Course Title: Data Mining

Time: 3 Hours

Max. Marks : 40

Note: - **Attempt all the five questions. Missing data/ information if any, maybe suitably assumed & mentioned in the answer.**

o.	Question	Marks	CO
	Attempt any 2 parts of the following.		
	Elaborate various stages of Data Mining Process.	4	CO1
	Differentiate classification and Regression for predictive analysis. Explain clustering.	4	CO1
	Suppose the fraction of undergraduate students who play football is 15% and the fraction of graduate students who play football is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who plays football is a graduate student? Also, Suppose 30% of the graduate students live in hostel but only 10% of the undergraduate students live in hostel. If a student plays football and lives in hostel, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in hostel and those who play football.	4	CO1
2	**Attempt any 2 parts of the following.**		
a	Discuss four techniques to deal with missing data in dataset along with suitable examples.	4	CO1
b	Calculate Dissimilarity matrix for given dataset? Object: 1, 2, 3, 4, 5, 6 Values: 40, 50, 42, 21, 30	4	CO2
c	If two data objects are given as x ={3,2,0,5,0,0,0,2,0,0} and y= {1,0,0,0,0,0,0,1,0,2}. Calculate its Cosine Similarity.	4	CO1
3	**Attempt any 2 parts of the following.**		
a	Consider market basket dataset shown in the following table. T. ID | Items Purchased | T. ID | Items Purchased | T. ID | Items Purchased 1 | {a,d,e} | 3 | {a,d,b} | 5 | {b,c} 2 | {a,d,b,c} | 4 | {a,e} | 6 | {a,d,b,e,c} Compute frequent pattern generated by FP growth algorithm if Minimum Support is 2.	4	CO4
b	For following given dataset, generate association rules using Apriori Algorithm. Consider Min Support as 50% and Confidence as 75%. Transaction ID | Items Purchased | Transaction ID | Items Purchased T1 | {bread,egg,cheese} | T3 | {bread} T2 | {egg, juice} | T4 | {bread,egg}	4	CO4

3c The following contingency table summarizes supermarket transaction data, where *sandwiches* refers to the transactions containing sandwiches, *-sandwiches* refers to the transactions that do not contain sandwiches, *burgers* refers to the transactions containing burgers, and *-burgers* refers to the transactions that do not contain burgers. [4]

	sandwiches	*-sandwiches*	Sum row
burgers	2000	500	2500
-burgers	1000	1500	2500
Sum col	3000	2000	5000

a. Suppose that the association rule "*sandwiches* → *burgers*" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

b. Compare the results using lift and χ^2 correlation measures.

Q4 **Attempt any 2 parts of the following.**

4a Consider the following 1-D data. [4]

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Label	N	N	P	P	P	N	N	P	N	N

a. Classify the data point X = 5.0 according to its 3-, and 5- nearest neighbors.

b. What would be the class label if distance-weighted voting approach is used?

4b Consider the dataset shown in the following table having three attributes A1, A2, and A3. Predict the class label for a test sample (A1 = 0, A2 = 1, A3 = 0) using the Naïve Bayes Algorithm. [4]

Instance	A1	A2	A3	Class Label
1.	0	0	0	Class 1
2.	0	0	1	Class 2
3.	0	1	1	Class 2
4.	0	1	1	Class2
5.	0	0	1	Class 1
6.	1	0	1	Class 1
7.	1	0	1	Class 2
8.	1	0	1	Class 2
9.	1	1	1	Class 1
10.	1	0	1	Class 1

4c Explain following: GINI Index, Entropy, Average Information Entropy and Information gain. [4]

Q5 **Attempt any 2 parts of the following.**

5a Expectation Maximization clustering works on which parameter and how it is done. What is the role of Parsing and soft parsing. [4]

5b Given points five points C1, C2, C3, C4 and C5 where $d(C_1,C_2)$: 7, $d(C_1,C_3)$: 4, $d(C_1,C_4)$: 1, $d(C_1,C_5)$: 2, $d(C_2,C_3)$: 8, $d(C_2,C_4)$: 5, $d(C_2,C_5)$: 11, $d(C_3,C_4)$: 9, $d(C_3,C_5)$: 10 and $d(C_4,C_5)$: 3. Perform Agglomerative clustering (Single Linkage) on these points and find Dendrogram. [4]

5c Divide the given sample data in three (3) clusters using k-means Algorithm. [4]
Height: 183, 171, 167, 176, 180, 177, 180, 180, 182, 183, 185, 185
Weight: 70, 56, 60, 72, 84, 76, 71, 68, 69, 77.72, 74

Total no. of pages: 2 Roll no.....

MID-SEMESTER EXAMINATION, SEPTEMBER - 2022

Course Code – ITITE05

Course Title – Data Warehouse and Data Mining

Time – 1.5 Hours Max. Marks – 15

Note – Attempt all questions. Missing data/information (if any), may be suitably assumed & mentioned in the answer.

Q. No	Questions	Marks	CO
1a	What is a Data Warehouse? Explain three-tier Data Warehouse architecture with a neat and labeled diagram.	2	CO1
1b	How is a data warehouse different from an operational database? Distinguish between them w.r.t. OLAP and OLTP.	1	CO3
2a	What can you accomplish with data gateway? Explain with diagrammatic illustration the working of data gateway. How is standard mode different from personal mode of data gateway?	2	CO1
2b	What are the steps involved in data cleaning? Briefly explain each one.	1	CO1
3a	What is Database synchronization? Discuss the concepts of Unidirectional and Bidirectional Database synchronization taking suitable example.	2	CO1
3b	Why is market moving from ETL to ELT in the cloud data warehouse? Discuss the pros and cons of ETL vs. ELT.	1	CO1
4a	What do you understand by Dimensional Modeling? Highlight and explain the key steps decided during the design of a Dimensional Model.	2	CO3
4b	State the objectives of JAD session. What is the Business Analyst's role in one?	1	CO3
5a	Suppose that a data warehouse for *Big-University* consists of the following four dimensions: *student, course, semester, instructor* and two measures *count, avg_grade*. When 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. (i) Draw *a snowflake schema* diagram for the data warehouse.	1.5	CO3

Total no. of Pages 1

Roll no. .....

SEMESTER-UG 5th

MID-SEMESTER EXAMINATION, September, 2022

Course Code- **COCSC16, CDCSC16**
Course Title- **Data Mining**

Time- 1.5 Hours

Max. Marks- 15

Note: - Attempt all questions. Missing data/ information (if any) may be suitably assumed & mentioned in the answer.

Q. No.	Question	Marks	CO
1a	What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining? Discuss the motivation behind OLAP mining (OLAM).	2	CO1
1b	Given the attribute name and values of attributes, classify its type (i) Gender-Male, Female (ii) Height-5.8,6.2 (iii) Color-Black, Brown, Grey	1	
2a	A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck. In a test, you counted 1600 cards, and observed the following: Spades 404 , Hearts 420 , Diamonds 400 , Clubs 376 Could it be that the suits are equally likely? Or are these discrepancies too much to be random?	2	CO2
2b	Starting with the base cuboid [day,doctor,patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010?	1	
3a	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 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) What is the mean of the data? What is the median? (b) What is standard deviation? What is the mode of the data?	2	CO
3b	Suppose that x and y are the first two term-frequency vectors i.e., x = (5,0,3,0,2,0,0,2,0,0) and y = (3,0,2,0,1,1,0,1,0,1). How similar are x and y? compute the cosine similarity between the two vectors.	1	
4a	Apply discretization on following data objects using binning: a) [7,5,7,7,7,7,6,9,8] b) [1,2,4,6,7,8,9,8,4]	2	CO
4b	Given two sets of s1 and s2, S1 = {1, 2, 3, 4, 5}, s2 = {4, 5, 6, 7, 8, 9, 10}. find the Jaccard Index and the Jaccard Distance between the two sets.	1	
5a	In a class mark of students in two subjects are enlisted as follows: {55, 11, 36, 72, 41, 91, 96, 63, 29, 44} and {32, 67, 89, 23, 17, 48, 59, 77, 61, 7}. Find the covariance matrix for this data.	2	CO
5b	Explain the techniques to deal with noise in the dataset.	1	

Total No. of Pages: 2

Roll No. _____

V/VII SEM - B. TECH END SEMESTER EXAMINATION, NOV-DEC-2022

Course Code: ITITE05

Course Title: Data Warehouse and Data Mining

Time: 3 Hours

Max. Marks: 40

Note: - Attempt all the five questions. Missing data/ information (if any), may be suitably assumed & mentioned in the answer.

Q. No.	Questions	Marks	CO
Q1	**Attempt any two parts of the following:**		
1a	Data warehousing is the only viable means to resolve the information crisis and to provide strategic information. List four reasons to support this assertion and explain them.	4	CO1
1b	What is a Data Warehouse? Highlight and discuss the four key characteristics of a Data Warehouse.	4	CO1
1c	Explain the major building blocks/components of a Data Warehouse with a well-labeled diagram.	4	CO1
Q2	**Attempt any two parts of the following:**		
2a	Elaborately discuss the concepts of the following schemas in Data Warehouse modelling (taking suitable examples with well-labeled diagram for each one): (i) *Snowflake schema* (ii) *Galaxy schema*	4 (2 + 2)	CO3
2b	Suppose that a data warehouse for *Gross sales* consists of the four dimensions: *date, spectator, location* and *movie*, and the two measures, *count* and *charge*, where *charge* is the fare that a spectator pays when watching a movie on a given date. Spectators may be students, adults or seniors, with each category/status having its own charge rate. (i) Draw *a star schema* diagram for the data warehouse. (ii) Starting with the base cuboid [*date, spectator, location, movie*], what specific *OLAP operations* should be performed in order to list the total charge paid by *student* spectators at PVR_Vegas in the year 2021?	4 (2 + 2)	CO3
2c	Suppose a company would like to design a data warehouse to facilitate the analysis of moving vehicles in an online analytical processing manner. The company registers huge amounts of auto movement data in the format of (*Auto_ID, location, speed, time*). Each *Auto_ID* represents a vehicle associated with information, such as *vehicle_category*, *driver_category*, etc., and each location may be associated with a street in a city. Assume that a street map is available for the city. (i) Design such a data warehouse to facilitate effective online analytical processing in multidimensional space. (ii) The movement data may contain noise. Discuss how you would develop a method to automatically discover data records that were likely erroneously registered in the data repository.	4 (2 + 2)	CO3

<table>
<tr><td>Q3</td><td>Attempt any two parts of the following:</td><td></td><td></td></tr>
<tr><td>3a</td><td>What are the differences between these three main types of data warehouse usage: information processing, analytical processing and data mining? Discuss the motivation behind OLAM (OLAP mining).</td><td>4</td><td>CO2, CO4</td></tr>
<tr><td>3b</td><td>Differentiate between ROLAP, MOLAP and HOLAP. Explain the following OLAP operations in the multidimensional data model (taking suitable example with diagram).
(i) Roll-up
(ii) Drill-down
(iii) Slice for
(iv) Dice for
(v) Pivot</td><td>4
-
(1½ + 2½)</td><td>CO3</td></tr>
<tr><td>3c</td><td>Give an introduction to NoSQL databases. Compare the applicabilities of CouchDB and MongoDB.</td><td>4</td><td>CO2, CO3</td></tr>
<tr><td>Q4</td><td>Attempt any two parts of the following:</td><td></td><td></td></tr>
<tr><td>4a</td><td>Explain how decision tree induction algorithm works. Elucidate the concept of its key factors: Entropy and Information Gain. How are they used to build decision trees?</td><td>4</td><td>CO5</td></tr>
<tr><td>4b</td><td>(i) Explain Classification and Prediction in Data Mining.
(ii) Differentiate between eager learners and lazy learners. Give examples for both of them.</td><td>4
(2 + 2)</td><td>CO2, CO4</td></tr>
<tr><td>4c</td><td>Consider the following dataset for a classification task. Using k-Nearest Neighbor (KNN) algorithm with Euclidean Distance as the distance metric, predict the class label Fruit Taste for data sample Strawberry with k=1, 3 and 5. Write down your calculations clearly.
<table>
<tr><th>Fruit</th><th>Sweetness</th><th>Sourness</th><th>Fruit Taste</th></tr>
<tr><td>Lemon</td><td>1</td><td>9</td><td>Sour</td></tr>
<tr><td>Grapefruit</td><td>2</td><td>8</td><td>Sour</td></tr>
<tr><td>Orange</td><td>3</td><td>7</td><td>Sour</td></tr>
<tr><td>Cherry</td><td>6</td><td>4</td><td>Sweet</td></tr>
<tr><td>Banana</td><td>9</td><td>1</td><td>Sweet</td></tr>
<tr><td>Grapes</td><td>8</td><td>2</td><td>Sweet</td></tr>
<tr><td>Strawberry</td><td>5</td><td>5</td><td>?</td></tr>
</table></td><td>4</td><td>CO5</td></tr>
<tr><td>Q5</td><td>Attempt any two parts of the following:</td><td></td><td></td></tr>
<tr><td>5a</td><td>What is Cluster Analysis? Describe the different types of Clustering methods/techniques with examples.</td><td>4</td><td>CO5</td></tr>
<tr><td>5b</td><td>For the following transactional dataset, find the frequent itemsets and generate strong association rules using Apriori algorithm. Consider the thresholds as min. support = 40% and min. confidence = 70%. Calculate the confidence of each rule and identify all the strong association rules.
<table>
<tr><th>T_ID</th><th>Items bought</th></tr>
<tr><td>1</td><td>{Bread, Butter, Milk}</td></tr>
<tr><td>2</td><td>{Bread, Butter}</td></tr>
<tr><td>3</td><td>{Beer, Cookies, Diapers}</td></tr>
<tr><td>4</td><td>{Milk, Diapers, Bread, Butter}</td></tr>
<tr><td>5</td><td>{Beer, Diapers}</td></tr>
</table></td><td>4</td><td>CO5</td></tr>
<tr><td>5c</td><td>How FP-tree is better than Apriori Algorithm? Using FP-Growth algorithm, demonstrate construction of FP-tree for the transactional dataset given above in (b) part of Question 5.</td><td>4</td><td>CO5</td></tr>
</table>

VIII 7/3/20

160

No of pages: 1 **MID SEMESTER EXAMINATION FEB 2020** Roll No. .....

Course code: ITD08 Course: DATA WAREHOUSE AND DATA MINING Sem: 8

Branch: COE & IT& IC Max. Marks : 15

Note: All questions carry 5 marks. Time: 1.30 hour

Q.1 Define the data warehousing. Explain the need and purpose of developing data warehouse. Discuss 3-tier architecture of data warehouse?

OR

Why ETL is consider an essential process in data warehousing. Explain the requirement of ETL process. How ETL process is optimized in term of processing speed?

Q.2 What is Meta data? Discuss the various types of the Meta data and their utilities in the data warehouse?

OR

What are the objectives of the dimensional modeling? List and explain various dimensional modeling schemas in data warehouses. Also Create the star schema for loan department of a banking organization?

Q.3 What do you mean by the OLAP? Give the classification of the OLAP. Also list the OLAP operations with supporting example?

OR

Given the marks of four students as 8, 10, 15 and 20. How to Calculate the Z-Score normalization and Min-max normalization of the given data?

----------End----------

Total No. of Page: 2

Roll No. _____

BTech Semester V

END SEMESTER EXAMINATION DECEMBER 2022

Course Code: COCSC16, CDCSC16
Course Title: Data Mining

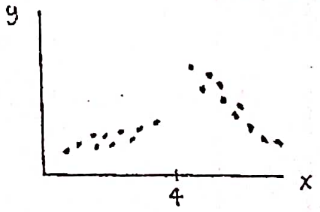
Time: 3 Hours

Max. Marks: 40

Note: - **Attempt all the five questions. Missing data/ information if any, maybe suitably assumed & mentioned in the answer**

Q. No.	Question	Marks	CO
Q1	**Attempt any 2 parts of the following.**		
1a	Explain various stages of data mining with diagram. Marks of 10 students in Data Mining subject are given as follows: 78, 51, 86, 91, 33, 27, 25 46, 55, 59. Calculate normalized values of these using Min-Max normalization.	2+2	CO1+ CO3
1b	Given the confusion matrix below, (a) Calculate the accuracy. Is this a good result? Why? (b) Explain 3 alternative performance measures to support your opinion.	2+2	CO2+ CO4
1c	Regression. Given the graph below, predict y by x, if MAE < 0.75 then it's a good result. Use one sentence for each regression method to explain whether it would have good results or not. (a) kNN for regression, k = 3 (b) binary tree regression (c) linear regression (d) local regression, k = 3 (e) local regression, k = 20	4	CO4
Q2	**Attempt any 2 parts of the following.**		
2a	What do you mean by a factless fact table in the context of data warehousing? Differentiate between fact table and dimension table.	4	CO5
2b	Price of a few items in a store is given as follows. Calculate Dissimilarity matrix for this.	4	CO2+ CO4
c	Predict that an Average build person having high Weight, having good experience and having fair performance will win in wrestling or not.	4	CO3

	True fraud	True non-fraud
Predicted fraud	1	4
Predicted non-fraud	9	986



Object No	1	2	3	4	5	6	7
Price	28	45	64	93	59	67	38

Build	Weight	Experience	Performance	Win?
Average	Medium	Good	Fair	No
Good	High	Moderate	Excellent	Yes
Normal	Low	Moderate	Fair	No
Good	Medium	Good	Fair	No
Good	High	Moderate	Fair	Yes
Average	High	Moderate	Excellent	Yes
Normal	Medium	Good	Fair	No
Normal	High	Moderate	Excellent	Yes
Average	Low	Good	Fair	No
Normal	Medium	Good	Fair	No
Average	High	Moderate	Excellent	No

Q3	**Attempt any 2 parts of the following.**		
3a	Explain following with one example with respect to itemset mining: A. Frequent itemset B. Minimum support C. Histogram D. Interesting measures	4	CO1+ CO5
3b	Trace the results of using the Apriori algorithm on the grocery store example with support threshold s=33.34% and confidence threshold c=60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence. Transaction ID | Items T1 | HotDogs, Buns, Ketchup T2 | HotDogs, Buns T3 | HotDogs, Coke, Chips T4 | Chips, Coke T5 | Chips, Ketchup T6 | HotDogs, Coke, Chips	4	CO2
3c	Use the following transactional with support threshold=3 Item | Frequency T1 | {E,K, M,N, O,Y} T2 | {D,E,K,N, O,Y} T3 | {A,E,K,M} T4 | {C,K,M,U,Y} T5 | {C,E,I,K,O,O} a.) Build a frequent pattern tree (FP-Tree). Show for each transaction how the tree evolves. b.) Use Fp-Growth to discover the frequent itemsets from this FP-tree	4	CO2
Q4	**Attempt any 2 parts of the following.**		
4a	Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. What advantage does (a) have over (b)?	4	CO2
4b	State 2 differences and 2 similarities between bagging and boosting. Suppose you are given 'n' predictions on test data by 'n' different models (M1, M2, ... Mn) respectively. Which method(s) can be used to combine the predictions of these models?	2+2	CO2+ CO4
4c	Dataset collected for whether information over a course of period is given below. Using the ID3 algorithm calculate the outcome.	4	CO3

Build	Weight	Experience	Performance	Win?
Average	Medium	Good	Fair	No
Good	High	Moderate	Excellent	Yes
Normal	Low	Moderate	Fair	No
Good	Medium	Good	Fair	No
Good	High	Moderate	Fair	Yes
Average	High	Moderate	Excellent	Yes
Normal	Medium	Good	Fair	No
Normal	High	Moderate	Excellent	Yes
Average	Low	Good	Fair	No
Normal	Medium	Good	Fair	No
Average	High	Moderate	Excellent	No