C.U.SHAH UNIVERSITY
Winter Examination-2015

Subject Name: Data Warehouse & Data Mining

Subject Code: 5CS03DWD1   **Branch:** M.Sc.I.T.

Semester: 3   **Date:** 30/12/2015   **Time:** 02:30 **To** 05:30   **Marks:** 70

<u>Instructions:</u>
   (1) Use of Programmable calculator and any other electronic instrument is prohibited.
   (2) Instructions written on main answer book are strictly to be obeyed.
   (3) Draw neat diagrams and figures (if necessary) at right places.
   (4) Assume suitable data if needed.

SECTION – I

Q-1	**Define following terms/phenomena**	**(07)**
a.	Dimensional Tables	**1**
b.	Characteristics of data warehouse	**1**
c.	Time-Series databases	**1**
d.	Steps in the data Mining process	**1**
e.	Linear Regression	**1**
f.	Base cube ID	**1**
g.	Metadata repository	**1**
Q-2	**Attempt all questions**	**(14)**
a.	Explain Star Schema with suitable Examples.	**5**
b.	Explain Data Integration & Transformation with Different Method.	**5**
c.	What is Data Warehouse? Briefly explain the key words used in the definition.	**4**
	OR	
Q-2	**Attempt all questions**	**(14)**
a.	Define KDD process with its steps to generate knowledge.	**5**
b.	Explain Snowflake Schema with suitable Examples.	**5**
c.	Explain Attribute Subset Selection Method.	**4**
Q-3	**Attempt all questions**	**(14)**
a.	Explain OLAP operations with suitable examples.	**7**
b.	Explain Data Mining Coupling with examples.	**7**
	OR	
Q-3	**Attempt all questions**	**(14)**
a.	What is Data Mining? Explain the Data Mining Architecture in details.	**7**
b.	What Indexing OLAP Data? Explain Different type of Indexing with examples.	**7**

SECTION – II

Q-4	**Define Following terms**	**(07)**
	a. Predictive model.	**1**
	b. Cluster analysis?	**1**
	c. Hierarchical method	**1**
	d. Association rule	**1**
	e. Bayesian classification	**1**
	f. Market Basket Analysis	**1**
	g. Accuracy in classification	**1**

Q-5   **Attempt all questions**   **(14)**

a. Apply Apriori on following data-set and find out frequent combination/Item-set for 3 products where minimum support =2.   **5**

Tid	1	2	3	4
Items Name	A,C,D	B.C,E	A,B,C,E	B,E

b. A 2X2 contingency table summarizing the transactions with respect to game and video purchases.   **5**

	Game	Game Not	Sum of Row
Video	3000	2500	5500
Not Like Video	2000	2500	4500
Sum of Column	5000	5000	10000

Find Correlation analysis using $\chi2$.

c. Explain K-medoids method with suitable examples   **4**

<div align="center">OR</div>

Attempt all questions

Q-5   **a.** The following table shows the midterm and final exam grades obtained for students in a database course.   **5**

x	Mid Exam	72	50	81	74	94	86	59	83	65	33	88	81
y	Final Exam	84	63	77	78	90	75	49	79	77	52	74	90

b) Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

(c) Predict the final exam grade of a student who received an 86 on the midterm exam.

b. Correlation with Lift (A, B). Define the Lift (A, B) measurement between 2 products TV and GAME where there are total 5000 transactions carried out. Out of which TV is sold in 3000 transactions. DVD is sold in 2500 transactions. TV and GAME together sold in 2000 transactions. Find out the Lift (TV, GAME) correlation.   **5**

c. Explain different variable used in Cluster.   **4**

Q-6   **Attempt all questions**   **(14)**

a. Discuss the Application of Financial Data analysis.   **7**

b. What is K-means clustering? Create 2 clusters for the given data-set using K-Means method and Euclidean Distance. Assign City C1 and C2 as initial Centroid point for the 2 Clusters after first iteration.   **7**

City	C1	C2	C3	C4
X	64	60	59	68
Y	60	61	70	71

OR

Q-6   **Attempt all Questions**

a. Discuss the Application of Telecommunication Industries. **7**

b. Cluster the following eight points (with (x, y) representing locations) into three clusters: **7**
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
Assume that initial cluster centers are: A2 (2, 5), A5 (5, 8) and A8 (1, 2) Use Manhattan
distance for the distance function between two points after the first iteration.