1. (a) Describe the steps in the KDD process with a suitable block diagram.  5
   (b) Compare between OLTP and OLAP.  5
   (c) What will be the effect of performing attribute oriented Induction (AOI) on the initial working relation **student** with attributes such as name, gender, birth-date, birth place, address, phone-no, and gpa.  10

2. (a) Using the table given below, create a classification model using decision true technique. Indicate how to utilize the model to estimate the risk category of the customer with (**Credit-History** – bad, **Debt** – high, **Collateral** – none, **Income** – (15-35k)).  10

| Sr. No. | Debt | Collateral | Income | Credit - History | Risk |
|---|---|---|---|---|---|
| 1 | high | none | 0-15 k | bad | high risk |
| 2 | high | none | 15-35 k | unknown | high risk |
| 3 | low | none | 15-35 k | unknown | Moderate risk |
| 4 | low | none | 0-15 k | unknown | high risk |
| 5 | low | none | over 35 k | unknown | low risk |
| 6 | low | adequate | over 35 k | unknown | low risk |
| 7 | low | none | 0-15 k | bad | high risk |
| 8 | low | adequate | over 35 k | bad | Moderate risk |
| 9 | low | none | over 35 k | good | low risk |
| 10 | high | adequate | over 35 k | good | low risk |
| 11 | high | none | 0-15 k | good | high risk |
| 12 | high | none | 15-35 k | good | Moderate risk. |

   (b) Define a data warehouse. Explain the architecture of data warehouse with suitable block diagram.  10

3. (a) Consider the data set given. Create the adjacency matrix. Use single link agglomerative technique to cluster the given data. Draw the dendogram.  4+4+2
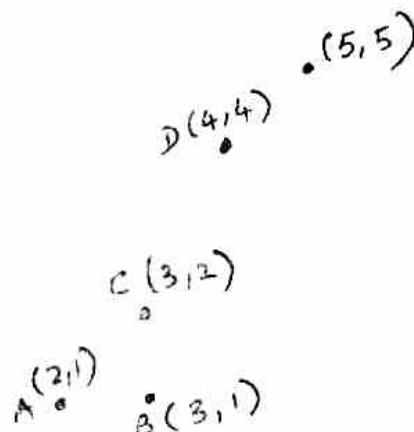


   (b) What are the different ways of finding the distance between two clusters ?  5
   (c) Define Factless Fact tables with a suitable example.  5

[ TURN OVER

4. (a) What is Association Rule Mining ? Give the Apriori algorithm. Apply AR Mining **2+4+4**
to find all frequent itemsets from the following table :—

| Transcation – ID | Items |
|---|---|
| 100 | 1, 2, 5 |
| 200 | 2, 4 |
| 300 | 2, 3 |
| 400 | 1, 2, 4 |
| 500 | 1, 3 |
| 600 | 1, 3 |
| 700 | 1, 3, 2, 5 |
| 800 | 1, 3 |
| 900 | 1, 2, 3 |

Minimum Support Count = 2
Minimum – Confidence – 70%.

(b) Explain the major steps in the ETL process with a suitable diagram and an example.     **10**

5. (a) The college wants to record the grades for the courses completed by students.
There are four dimensions :—
(i) Course                    (iii) Student
(ii) Professor                (iv) Period.
The only fact that is to be recorded in the table is course-grade :–
(i) Design star schema.                                                      **5**
(ii) Write DMQL for the above star schema.                                   **5**
(b) Using the above example describe the following OLAP operations :—         **10**
Slice, Dice, Roll-up, Drill-down, Pivot.

6. (a) What are crawlers ? How do periodic crawlers differ from incremental crawlers ? **2+2+6**
Give the architecture of focussed crawlers and explain how it is used.
(b) Explain how HITS algorithm finds hubs and authoritative pages.           **10**

7. Write short notes on (any **four**) :—                                     **20**
(a) Outlier Mining
(b) Applications of Web Usage Mining
(c) Snowflake Schema
(d) Generalized Association Rules
(e) Top-down and Bottom-up approaches in data warehousing.