(iii) Use min-max normalization to transform the value 69 for temperature onto the range [0.0, 1.0]
(iv) Use normalization by decimal scaling to transform the value 69 for temperature.
(v) Comment on which method you would prefer to use for the given data, and why. **5**

2B Given the following sorted data:
Data: 10, 15, 24, 36, 36, 48, 59, 75, 75, 85, 90, 95, 95, 100, 100.
Apply the following binning methods and find out the optimal method using V-optimal principle. Bin Size = 3.
(i) Bins of equal-frequency
(ii) Bins of equal-width
(iii) Max Difference. **3**

2C Suppose that a small number of customers lie about their demographic profile, and this results in a mismatch between the buying behavior and the demographic profile, as suggested by comparison with the remaining data. Which data mining problem would be best suited to find such customers? **2**

3A Consider the transactional database given in Table Q. 3A.

Table Q. 3A

TID	ITEMS
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

5

Determine all the frequent patterns and maximal patterns at support values of 3 and 4. Represent the transaction database in vertical format.

3B Given the decision tree of Figure Q. 3B (a), predict the class of the following new instance of Figure Q. 3B (b), which describes a new loan applicant. The last column is the Class attribute, which shows whether each loan application was approved (denoted by Yes) or not (denoted by No) in the past. Also explain the role of pure subset formulation in a decision tree.

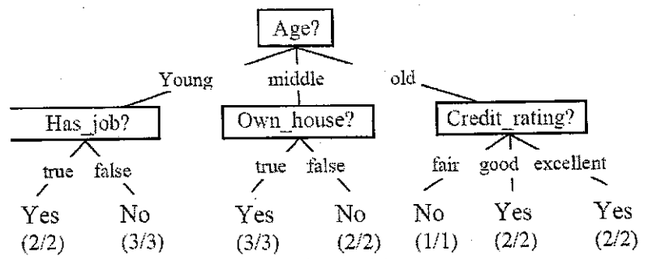

Fig Q. 3B (a)

Age	Has_job	Own_house	Credit-rating	Class
young	false	false	good	?

3

Fig Q. 3B (b)

3C Write in brief about the following in the context of rule based classifiers:
(i) Mutually Exclusive Rules
(ii) Exhaustive Rules. **2**

4A Given the transactional database of Table Q.4A, find the top three association rules using Apriori algorithm. Given minimum support = 50% and minimum confidence = 50%. **5**

Table Q.4A

TID	ITEMS
1	A, B, C, D, G, H
2	A, B, C, D, E, F, H
3	B, C, D, E, H
4	B, E, G, H
5	A, B, D, E, G, H
6	A, C, F, G, H
7	B, D, E, G, H
8	A, C, D, E, G, H
9	B, C, D, E, H
10	A, C, E, F, H
11	C, E, H
12	A, D, E, F, H
13	B, C, E, F, H
14	A, B, C, F, H
15	A, B, E, F, H

4B Describe the following challenges encountered while mining link information between objects on the Web:
(i) Logical Vs Statistical Dependencies
(ii) Feature Construction
(iii) Link Prediction **3**

4C Compare and Contrast the following sampling methods with a real-time example for each:
(i) Simple random sample without replacement (SRSWOR) of size s; and
(ii) Simple random sample with replacement (SRSWR) of size s. **2**

5A Apply the Pincer-Search algorithm on the transactional database given in Table Q. 5A and find the maximal frequent itemsets. **5**

Table Q.5A

Transaction	Products
1	Bananas, Orange Papaya
2	Papaya, Jack fruit
3	Orange, Mango
4	Papaya, Water Melon
5	Orange, Watermelon

5B Assume we have a data set D with only two classes, positive and negative. Calculate the entropy values for the following three different compositions of positive and negative examples:

(i) 50% positive and 50% negative examples
(ii) 20% positive and 80% negative examples
(iii) 100% positive examples **3**

5C Discuss how CLARA and CLARANS algorithms overcome the deficiencies of K-Means and K-Medoids algorithm. **2**

Reg. No. | | | | | | | | | |

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent institution of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY)

MAKEUP EXAMINATIONS, JUNE 2018

DATA WAREHOUSING AND DATA MINING [ICT 3202]

REVISED CREDIT SYSTEM

(18/06/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data if any may be suitable assumed.

1A Mention any five prominent steps in the design of a data warehouse with an example for each step mentioned. **5**

1B Given the contingency table of Q.1B, perform the Chi-Square test to determine the strength of association between the variables Eye Color and Hair Color.

(i) If the association rule "Black $\Longrightarrow$ Light" is mined. Given a minimum support threshold of 30% and a minimum confidence threshold of 50%, determine if this association rule is strong.

(ii) Determine whether the attribute Green is dependent on the attribute Dark? What type of correlation exists between the attributes others ad Light? **3**

Table Q. 1B

	Light	**Dark**	Σ_{row}
Black	32 (24.1)	12 (19.9)	44
Green	14 (19.7)	22 (16.3)	36
Others	6 (8.2)	9 (6.8)	15
Σ_{col}	52	43	95

1C An analyst processes Web logs in order to create records with the ordering information for Web page accesses from different users. What is the type of this data? Justify your answer with an example for the same. **2**

2A For a particular region the temperatures and the corresponding humidity values are recorded as in Table Q. 2A.

Table Q. 2A

Temperature	75	80	85	72	69	72	83	64
Humidity	70	90	85	95	70	90	78	65
Temperature	81	71	65	75	68	70	85	95
Humidity	75	80	70	80	80	96	55	95

(i) Calculate the mean, median and standard deviation of temperature and humidity.
(ii) Draw the box plots for temperature and humidity.

ICT 3202