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

# MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL
*(A constituent unit of MAHE, Manipal)*

## III SEMESTER MCA

## END SEMESTER EXAMINATIONS NOV/DEC 2018

## SUBJECT: DATA WAREHOUSING AND DATA MINING [MCA 5102]

**REVISED CREDIT SYSTEM**
**(22/11/2018)**

Time: 3 Hours MAX. MARKS: 50

**Instructions to Candidates:**

- Answer **ALL FIVE FULL** questions.
- Missing data may be suitable assumed.

|  |  |  |
|---|---|---|
| **1A.** | Define data mining. Explain the sequence of steps in the Knowledge Discovery process with a neat diagram. | **5** |
| **1B.** | What are the various OLAP operations supported in data cubes? Give appropriate examples | **3** |
| **1C.** | How can a nominal attribute be "value mapped" into a set of binary attributes? Give an example. | **2** |
|  |  |  |
| **2A.** | Consider the table specified below, describing CARS and answer the following questions:<br>i. Create a summary table, grouping by cylinders and display count of cars, average MPG, minimum weight and maximum displacement.<br>ii. Create a contingency table with Number of Cylinders and Model Year.<br>iii. Find the correlation between Horse power and Weight and comment on the relationship between the variables.<br>iv. Visualize the relationship between Horse power and Weight using a scatter plot. | **5** |

**CARS DATASET**

| Names | Cylinders | Displacement | Horse-power | Weight | Acceleration | Model Year | Origin | MPG |
|---|---|---|---|---|---|---|---|---|
| Chevrolet Chevelle | 8 | 307 | 130 | 3504 | 12 | 1970 | 1 | 18 |
| Plymouth Duster | 6 | 198 | 95 | 2833 | 15.5 | 1978 | 1 | 20 |
| Chevrolet Vega (SW) | 4 | 140 | 72 | 2408 | 19 | 1971 | 1 | 22 |
| Fiat 124B | 4 | 88 | 76 | 2065 | 14.5 | 1971 | 2 | 30 |
| Datsun 1200 | 4 | 72 | 69 | 1613 | 18 | 1975 | 3 | 35 |
| Buick Skylark 320 | 8 | 350 | 165 | 3693 | 11.5 | 1972 | 1 | 15 |
| Ford Maverick | 6 | 200 | 85 | 2587 | 16 | 1975 | 1 | 21 |
| Volkswagen 1131 | 4 | 97 | 46 | 1835 | 20.5 | 1970 | 2 | 19 |
| Toyota Corolla | 4 | 71 | 65 | 1773 | 19 | 1973 | 3 | 31 |
| Ford Torino | 8 | 302 | 140 | 3449 | 10.5 | 1970 | 1 | 17 |

2B. What is the need for concept hierarchies? Create a concept hierarchy for the attribute “Location”. 3

2C. What strategies can be adopted to detect redundancy during data integration? 2

3A. For the following transaction data set, find all frequent item sets for minimum support of 25%. 5

**TRANSACTIONS DATA SET**

| Transaction Id | I1 | I2 | I3 | I4 | I5 | I6 | I7 | I8 | I9 |
|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| T2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| T3 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| T4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| T5 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| T6 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| T7 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| T8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

**3B.** Given two data points X= (12, 4, 16, 7) and Y= (11, 5 , 23, 5) .Represent them as a distance matrix using

i. Euclidean distance between the data points

ii. Manhattan distance between the data points.

iii. Minkowski distance between the data points using q = 3.

**3**

**3C.** What are the advantages of the k-means clustering technique? **2**

**4A.** Consider a table of the following observations.

i. Draw a scatter plot to illustrate the relationship between x & y.

ii. Use the method of "simple non-linear regression" to predict y value for x = 14.6

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x | 3 | 6 | 9 | 8 | 10 | 11 | 12 | 13 | 13.5 | 14 | 14.5 | 15 | 15.2 | 15.3 |
| y | 4 | 5 | 7 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 22 | 28 | 35 | 42 |

**5**

**4B.** What strategies could be adopted for separation of test and training set for classifiers? **3**

**4C.** Differentiate between the following, with suitable examples.

i. Classification tree vs. Regression tree

ii. Sensitivity vs. Specificity

**2**

**5A.** Attributes of the car, such as color, type of car and its origin is recorded. The class label indicates whether the car was stolen or not. Predict the class label "Stolen" for a car with attributes – color – yellow, type – SUV and origin – Imported using the Naïve Bayesian method

| Colour | Type | Origin | Stolen |
|---|---|---|---|
| Red | Sports | Domestic | Yes |
| Red | Sports | Domestic | No |
| Red | Sports | Domestic | Yes |
| Yellow | Sports | Domestic | No |
| Yellow | Sports | Imported | Yes |
| Yellow | SUV | Imported | No |
| Yellow | SUV | Imported | Yes |
| Yellow | SUV | Domestic | No |
| Red | SUV | Imported | No |
| Red | Sports | Imported | Yes |

**5**

**5B.** Differentiate between global outliers and contextual outliers using appropriate examples. **3**

**5C.** List any two measures that quantify the accuracy of prediction algorithms. **2**