No. of Printed Pages : **3**

Sl. No.

C5-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1. **Answer question 1 and any FOUR questions from 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time : 3 Hours **Total Marks : 100**

1. (a) Discuss various steps to knowledge discovery process using suitable diagram.

(b) Define the following :

(i) Metadata

(ii) Data mart

(c) What is normalization ? Explain it's types. The minimum and maximum values for the attribute income are 14,000 and 96,000 respectively. Map income to the range [0.0, 1.0]. Find out the transformed value of income 64,600 using min-max normalization.

(d) How is a data warehouse different from a database ? Also discuss the similarities between them.

(e) Explain 3-tier architecture of Data warehouse with a suitable diagram.

(f) What is ETL process ? Why it is used for ?

(g) Differentiate between following :

(i) Relational database and transactional database.

(ii) Temporal database and legacy database. **(7x4)**

2. (a) What is concept hierarchy ? Make concept hierarchy for Product and Location dimension.

(b) What are the major issues in data mining ? Discuss the issues regarding mining methodology and user interaction and performance.

(c) Define the various OLAP operations with suitable diagram.

(d) What do you mean by data cleaning ? Describe various methods for handling "missing values" in data set. **(5+5+4+4)**

3. (a) Draw a box-and-whisker plot for the following data set :

126,132,138,140,141,141,142,143,144,144,144,146,147,148,148,149,149, 150,150,150,154,155,158,158. Also find outliers.

(b) What is data cube ? Why it is important in data warehousing ?

(c) List down the differences between OLTP and OLAP.

(d) Classify the tuple X={color='RED', type="SUV", Origin="Domestic"} using Naïve Bayesian Classification. Training data is given in the following table where class label is 'Stolen'.

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

(4+5+4+5)

4. (a) Explain Chi-square test method. Show using Chi-square test whether gender and preferred reading are independent or not from given observed counts in the table below.

	Male	Female	Total
Fiction	250	200	450
Non- Fiction	50	1000	1050
Total	300	1200	1500

(b) Define data reduction process. Write down the various strategies for data reduction.

(c) What is hierarchical method for clustering ? Discuss its types.

(d) Find out the mean, variance and standard deviation for the height of animals 600 mm, 470 mm, 170 mm, 430 mm and 3005 mm. **(4+4+4+6)**

5. (a) Differentiate among ROLAP, MOLAP and HOLAP.

(b) What is web mining ? Differentiate between web content mining, web structure mining and web usage mining.

(c) Give the difference between the star and fact constellation multidimensional model with suitable diagram.

(d) Consider five points {X1, X2, X3, X4, X5} with the following coordinates as a two dimensional sample for clustering: X1=(0, 2.25); X2=(0, 0.25); X3=(1.25, 0); X4=(4.5, 0); X5=(4.5, 2.5); Illustrate the K-means clustering algorithm using the above data set. **(3+4+5+6)**

6. (a) Differentiate between dimensionality reduction and numerosity reduction. Explain sampling method for numerosity reduction.

(b) Discuss issues that are important to consider when employing a decision tree based classification algorithm. Explain the decision tree induction algorithm with appropriate examples. Discuss the disadvantages of this approach ? What is over fitting, and how can it be prevented for decision trees ?

(c) Explain Density-based clustering and discuss DBSCAN algorithm using a suitable diagram. **(6+6+6)**

7. (a) Write short note on following :

(i) Attribute subset selection

(ii) Principal Component Analysis

(iii) Histograms

(b) Explain apriori algorithm. Calculate frequent 1-item set, 2-item set and 3-item set for given transaction database. (Consider minimum support count 3)

Transaction ID	ITEM
1	{A, C, D}
2	{B, C, D}
3	{A, B, C, D}
4	{B, D}
5	{A, B, C, D}

(9+9)