No. of Printed Pages : **3**

Sl. No.

BE6-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.**

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

1. (a) Define data mining. Write at least four application of it.

(b) What is data warehouse ? Write differences between operational database system and data warehouse.

(c) Find mean, median, mode, and range of the data : 23, 29, 22, 19, 23, 24, 12

(d) What is unsupervised learning ? Which are its applications ?

(e) What is a neural network ? What are the components of neural networks ? Draw a diagram of a sample neural network.

(f) Discuss issues of data mining in object oriented databases.

(g) What is web mining ? How is it useful for web analytics ? **(7x4)**

2. (a) What is missing values in data ? Why data cleaning is required ? What are the ways to handle missing values in data ?

(b) The FRUIT_BASKET company wants to classify apple and orange with two features: sweetness (X1) and acidity (X2). Build a K-Nearest Neighbors (KNN) classifier to predict the type of fruit (either "Apple" or "Orange") based on these features. Consider the following data :

Sweetness (X1)	Acidity (X2)	Type
8	3	Apple
6	2	Apple
4	7	Orange
7	5	Orange
3	6	Orange

Choose K=3, What type (Apple or Orange) would be predicted for a new data point with sweetness=5 and acidity=4 ? Use Euclidean distance.

(c) What is density-based clustering? What are its advantages and disadvantages ? Define core-point, border point and noise point. **(6+6+6)**

3. (a) Explain Knowledge Discovery in Databases (KDD) process with a suitable diagram.

(b) What is frequent item sets? Explain Apriori Algorithm with the help of a suitable example. **(9+9)**

4. (a) What is data normalization ? What are the methods to normalize the data ? Explain at least one with example.

(b) How data mining can be applied on text databases ? Explain.

(c) Imagine you are analyzing a dataset related to the performance of students in high school. You have data on the number of hours students spend studying and their scores on a test. The dataset includes the following information :

Study Hours	Test Scores
3	80
6	92
2	75
5	88
7	96

- What is the dependent variable in this example ? What is the independent variable or feature ?
- Predict the test score of a student who study 4 hours using linear regression method. **(6+6+6)**

5. (a) You want to cluster customers of an e-commerce website based on their purchasing behaviour. You have collected data on two features: the total amount spent by each customer and the number of items purchased. Group customers into two segments using K-means clustering algorithm. Consider Euclidean distance. Initialize the centroids with the points: (1000, 20) and (2000, 30). The dataset includes the following information :

Customer ID	Total Amount Spent	Number of Items
1	1000	20
2	2000	30
3	500	10
4	3000	40
5	1500	25

(b) Write at least two data mining tools and explain the functionality of these tools in brief. **(9+9)**

6. (a) Find the value of the following evaluation measures from the confusion matrix.

	Actual Positive	**Actual Negative**
Predicated Positive	150	50
Predicated Negative	30	130

(i) TP, TN, FP, FN

(ii) Accuracy

(iii) Precision

(iv) Recall

(v) F1 score

(b) Differentiate among: ROLAP, MOLAP and HOLAP.

(c) Write a short note on Attribute Oriented Induction (AOI). **(6+6+6)**

7. (a) Explain three tier architecture of Data Warehouse using a suitable diagram.

(b) Describe the various OLAP operations on multidimensional data with example. **(9+9)**

- o O o -