

BACHELOR OF INFORMATION TECHNOLOGY ENGG. EXAMINATION, 20254th year, 1st semester**Data Warehousing and Data Mining**

Time: 3 Hours

Full Marks: 100

Read the questions carefully and answer accordingly.

CO1: Answer any 3 questionsFull marks: **15**

1. a) What is a data warehouse? How is it differ from DBMS? Compare the similarities of Data Mining with Machine Learning.
- b) Describe the steps involved in data mining process of knowledge discovery.
- c) Describe 3-tier Architecture of Data Warehouse with a neat sketch.
- d) Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? (3x5)

CO2: Answer any 2 questionsFull marks: **20**

2. a) Describe two methods for handling missing values for some attributes in real-world data. Discuss about the various types of patterns which can be mined for some business applications. Describe the measures of multidimensional data model. (3+3+4)
- b) Discuss about discretization and concept hierarchy generation for numerical data. (5+5)
- c) Give examples for defining star, snowflake and fact constellation schemas. With illustrative examples explain various OLAP operations. (6+4)

CO3: Answer any 2 questionsFull marks: **20**

3. a) Explain in detail about association rules mining task. Define confidence of an association rule. Discuss about the 'anti-monotone property of Support'. (4+3+3)
- b) Discuss about maximal frequent item set. Explain mining association rules FP-growth algorithm with minimum support threshold of 3.
 $T1 = \{F1, C1, S1, N1, B1\}$
 $T2 = \{C1, S1, B1\}$
 $T3 = \{C1, F1, N1\}$
 $T4 = \{B1, F1\}$
 $T5 = \{F1, C1, B1, N1\}$
 $T6 = \{C1, F1, N1\}$ (2+8)
- c) Illustrate the Aprori algorithm with a suitable example. Are there any limitations of Apriori algorithm? (7+3)

CO4: Answer any 2 questionsFull marks: **20**

4. a) Explain different classification techniques. Explain k-nearest neighbor algorithm with an example. (3+7)

[Turn over

- b) Discuss the significance of information gain in decision tree. In the above dataset 'Object Table', there are five discrete attributes: Object Shape, Object Size, Object Weight, Object Color, and Target Location, and the Target Location is the target variable we want to predict. Apply Decision Tree based method to identify class label (Target Location) to a new sample < Triangle, Medium, Light, Blue>.

(3+7)

Object Table				
Shape	Size	Weight	Color	Target Location
Square	Small	Light	Red	Shelf 1
Circle	Medium	Heavy	Blue	Shelf 2
Rectangle	Small	Light	Green	Shelf 3
Triangle	Large	Heavy	Red	Shelf 4
Circle	Small	Light	Blue	Shelf 2
Square	Medium	Heavy	Green	Shelf 3
Rectangle	Small	Heavy	Red	Shelf 4
Triangle	Large	Light	Blue	Shelf 2
Circle	Small	Heavy	Green	Shelf 3
Square	Medium	Light	Red	Shelf 1

- c) Explain about the measures for selecting the best split. In the above dataset 'Object Table', calculate the entropy of the 'Target Location' attribute in the entire dataset?

(4+6)

CO5: Answer any 2 questions

Full marks: 20

5. a) What is the goal of clustering? What is outlier detection? Suppose that the data mining task is to cluster the following ten books into three clusters, the distance function is Manhattan. Assign record 1, 2, and 3 as the centroid of each cluster respectively. Use the k-means algorithm to show the final three clusters. (2+3+5)

RecordID	No_of_page	Price(Rs)
1	245	135
2	335	255
3	170	910
4	146	670
5	170	250
6	365	375
7	456	440
8	250	655
9	322	260
10	167	190

- b) List out all hierarchical methods for clustering data. Define and explain any one hierarchical clustering methods.
- c) What is the drawback of k-means algorithm? Explain the DBSCAN clustering methods in detail.

(3+7)

(2+8)

CO6: Answer all questions

Full marks: 5

6. Write Short notes on (any one):
- a) Text mining
- b) Sequential Pattern Mining