R20 Regulation *Subject code:3P7FA*

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY

(Autonomous, Accredited by NAAC with 'A+' Grade)

B.Tech VII Semester Regular/Supplementary Examinations, December2024

Data Warehousing and Data Mining

(IT)

Maximum Marks: 70 Date:10.01.2025 Duration: 3 hours

Note: **1.This question paper contains two parts A and B.**
2. Part A is compulsory which carries 20 marks. Answer all questions in Part A.
3. Part B consists of 5 Units. Answer any one full question from each unit which carries 10M.
4. Each question carries 10 marks and may have a, b, c, d as sub questions.

Part-A

	All the following questions carry equal marks (10X2M=20 Marks)	**CO**	**Bloom Tx**
1	Give a brief note on Fact-Less-Facts.	1	2
2	Differentiate Data warehouse and DBMS	1	4
3	What is Data Discretization?	2	1
4	Discuss about data cleaning.	2	2
5	Differentiate multiple and multilevel association rule mining	3	4
6	What is the purpose of Apriori Algorithm?	3	3
7	Define frequent sets, confidence and support	4	2
8	State Bayes theorem	4	2
9	Differentiate OPTICS and DBSCAN	5	4
10	Define the centroid of the cluster	5	2

Part-B

	Answer All the following questions.**(5X10M=50Marks)**	**CO**	**Bloom Tx**
11	A) Define metadata and explain the types of metadata [5M] B) Explain star schema with an example [5M]	1	3
	OR		
12	Define Data Cube computation. Explain the various methods for Data Cube Computation. Discuss Construction of Multi-dimensional model. And its operations [10M]	1	2
13	A) Discuss about Data Mining task primitives with examples. [5M] B) List and define the measures of Similarity and Dissimilarity [5M]	2	2
	OR		
14	A) Discuss the issues to be considered during data integration. [5M] B) How can we handle missing values? Illustrate with an example [5M]	2	4
15	A)What is more efficient method for Generalizing association rule? [5M]		

<table>
<tr><td></td><td>B)Describe a data set for which sampling would actually increase the amount of work. In other words it would be faster to work on full data set. [5M]</td><td>3</td><td>3</td></tr>
<tr><td></td><td>OR</td><td></td><td></td></tr>
<tr><td>16</td><td>Apply FP-growth algorithm on the following database to find all of the strong association rules with min_sup = 60% and min_conf = 80% [10M]

<table><tr><td>TID</td><td>Items</td></tr><tr><td>T100</td><td>I1,I2,I3,I4,I5,I6</td></tr><tr><td>T200</td><td>I2,I3,I4,I5,I6,I7</td></tr><tr><td>T300</td><td>I1,I4,I5,I8</td></tr><tr><td>T400</td><td>I1,I4,I6,I9,I11</td></tr><tr><td>T500</td><td>I2,I4,I5,I9,I10</td></tr></table></td><td>3</td><td>5</td></tr>
<tr><td>17</td><td>Perform KNN classification for the data given below for X= {P1= 3,P2=7} where k=3. [10M]
<table><tr><td>P1</td><td>P2</td><td>CLASS</td></tr><tr><td>7</td><td>7</td><td>FALSE</td></tr><tr><td>7</td><td>4</td><td>FALSE</td></tr><tr><td>3</td><td>4</td><td>TRUE</td></tr><tr><td>1</td><td>4</td><td>TRUE</td></tr></table></td><td>4</td><td>5</td></tr>
<tr><td></td><td>OR</td><td></td><td></td></tr>
<tr><td>18</td><td>A) Explain in detail the Naive-Bayes Classifier. [5M]
B) List the characteristics of K- Nearest neighbor classification [5M]</td><td>4</td><td>2</td></tr>
<tr><td>19</td><td>Discuss various Hierarchical clustering methods. (Agglomerative, divisive, Chameleon and BIRCH). [10M]</td><td>5</td><td>2</td></tr>
<tr><td></td><td>OR</td><td></td><td></td></tr>
<tr><td>20</td><td>A) What is outlier detection? Explain distance based outlier detection [5M]
B)What is the drawback of k-means algorithm? How can we modify the algorithm to diminish that problem? [5M]</td><td>5</td><td>3</td></tr>
</table>