R20 Regulation *Subject code: 3P6GB*

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY

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

B.Tech VI Semester Supplementary Examinations, February 2024

DATA WAREHOUSING AND DATA MINING

(CSE(AI&ML))

Maximum Marks: 70 Date:17.02.2024 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	How is data ware house different from a database?	CO1	L1
2	Why data transformation is essential in the process of knowledge discovery?	CO1	L1
3	Compare between data cleaning and noisy data.	CO2	L2
4	What do you mean by data reduction techniques?	CO2	L1
5	What is the purpose of Apriori algorithm?	CO3	L1
6	Define strong association rule.	CO3	L1
7	What are association rules?	CO4	L1
8	Define Bayes theorem.	CO4	L1
9	What are the cons of decision trees?	CO5	L1
10	How can you make the K means algorithm more scalable?	CO5	L1

Part-B

Answer All the following questions. (5X10M=50Marks)			
11	A. Discuss multidimensional data model and explain various schemas for multidimensional data model. 5M B. "Predicting the outcomes of tossing a fair pair of dice". Is it a data mining task? Why or Why not? 5M	CO1	L2
	OR		
12	A. How data in the Data Warehouse is stored? What are the various OLAP operations that can be performed on the Data warehouse data? 5M B. Discuss the basic characteristics of a data warehouse. 5M	CO1	L3
13	A. What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set) 5M B. Discuss the activities of data cleaning with the process associated with it. 5M	CO2	L3
	OR		

<table>
<tr><td>14</td><td>A. Illustrate the data preprocessing steps that may be applied to the data to help improve the accuracy, efficiency and scalability of the classification process. 5M
B. What are the major challenges of mining a huge amount of data in comparison with mining a small amount of data. 5M</td><td>CO2</td><td>L3</td></tr>
<tr><td>15</td><td>A. Compare the advantages of FP growth algorithm over apriori algorithm.5M
B. Explain how frequent itemset mining leads to discovery of associations and correlations in market basket analysis. 5M</td><td>CO3</td><td>L3</td></tr>
<tr><td></td><td>OR</td><td></td><td></td></tr>
<tr><td>16</td><td>A. Apply the frequent item sets for the following data using Apriori algorithm with minimum support count = 2 and minimum confidence =60% . 5M
<table>
<tr><td>TID</td><td>Items</td></tr>
<tr><td>T100</td><td>I1,I2,I5</td></tr>
<tr><td>T200</td><td>I2,I4</td></tr>
<tr><td>T300</td><td>I2,I3</td></tr>
<tr><td>T400</td><td>I1,I2,I4</td></tr>
<tr><td>T500</td><td>I1,I3</td></tr>
<tr><td>T600</td><td>I2,I3</td></tr>
<tr><td>T700</td><td>I1,I3</td></tr>
<tr><td>T800</td><td>I1,I2,I3,I5</td></tr>
<tr><td>T900</td><td>I1,I2,I3</td></tr>
</table>
B. What is the Apriori property? How is it used by the apriori algorithm? What are the drawbacks of the Apriori algorithm. 5M</td><td>CO3</td><td>L3</td></tr>
<tr><td>17</td><td>A. Analyze classification rules from decision tree? If so how? What are the enhancements to the basic decision tree? 5M
B. Write an algorithm for K-nearest neighbor classification given k and n, the number of attributes describing each tuple. 5M</td><td>CO4</td><td>L4</td></tr>
<tr><td></td><td>OR</td><td></td><td></td></tr>
<tr><td>18</td><td>A. Why is tree pruning useful in decision tree induction? What is the drawback of using a separate set of tuples to evaluate pruning. 5M
B. Why naive Bayesian classification is called “naive”? Briefly outline the major ideas of naive Bayesian classification. 5M
C.</td><td>CO4</td><td>L3</td></tr>
<tr><td>19</td><td>A. Construct a dendogram for the given data (Similarity matrix) using single and complete link for hierarchical clustering. 5M
<table>
<tr><td></td><td>P1</td><td>P2</td><td>P3</td><td>P4</td><td>P5</td></tr>
<tr><td>P1</td><td>1.0</td><td>0.10</td><td>0.41</td><td>0.55</td><td>0.35</td></tr>
<tr><td>P2</td><td>0.10</td><td>1.00</td><td>0.64</td><td>0.47</td><td>0.98</td></tr>
<tr><td>P3</td><td>0.41</td><td>0.64</td><td>1.00</td><td>0.44</td><td>0.85</td></tr>
<tr><td>P4</td><td>0.55</td><td>0.47</td><td>0.44</td><td>1.0</td><td>0.76</td></tr>
<tr><td>P5</td><td>0.35</td><td>0.98</td><td>0.85</td><td>0.76</td><td>1.0</td></tr>
</table></td><td>CO5</td><td>L3</td></tr>
</table>

	B. Explain briefly the differences between "classification" and "clustering" and give an informal example of an application that would benefit from each technique. 5M		
	OR		
20	A. Discuss in detail about the various detection techniques in outlier. 5M B. Consider five points { X1, X2, X3, X4, X5} with the following coordinates as a two dimensional sample for clustering: X1 = (0,2.5); X2 = (0,0); X3= (1.5,0); X4 = (5,0); X5 = (5,2) Illustrate the K-means partitioning algorithm using the above data set. 5M	CO5	L2