

Register No.

--	--	--	--	--	--	--	--

BE Degree Examination May 2022

Sixth Semester

Computer Science and Engineering

18CSE05 – DATA WAREHOUSING AND DATA MINING

(Regulation 2018)

Time: Three hours

Maximum: 100 marks

Answer all Questions

Part – A ($10 \times 2 = 20$ marks)

1. Differentiate data mining and data warehousing. [CO1,K2]
2. What are patterns? How are they useful in decision making? [CO1,K1]
3. In real-world data, tuples with missing values for some attributes are a common issue. How will you handle this problem? [CO2,K1]
4. What will happen if data pre-processing is not done? [CO2,K1]
5. How support and confidence values are useful in association rule mining? [CO3,K1]
6. What are the shortcomings of Apriori algorithm? [CO3,K1]
7. Distinguish between classification and prediction. [CO4,K2]
8. What is the use of cross validation in classification? [CO4,K1]
9. State the categories of clustering methods. [CO5,K1]
10. Outlier detection and clustering analysis are two highly related tasks. Justify the reason. [CO5,K1]

Part – B ($5 \times 16 = 80$ marks)

11. a. i) Explain the three-tier architecture of a data warehouse with a neat (10) [CO1,K2]
diagram.
ii) Recall about any three major issues that occur in data mining. (6) [CO1,K1]
- (OR)
- b. i) Illustrate the various steps involved in knowledge discovery process with a (10) [CO1,K2]
neat diagram.
ii) Where is star and snowflake schema used? Give an example. (6) [CO1,K1]
12. a. Briefly outline how to compute the dissimilarity between objects described by (16) [CO2,K2]
the following types of variables:
(i) Numerical (interval-scaled) variables
(ii) Asymmetric binary variables
(iii) Categorical variables
(iv) Ratio-scaled variables
(v) Nonmetric vector objects

(OR)

- b. i) Outline the major tasks in data preprocessing. (10) [CO2,K2]
 ii) Summarize the different methods to handle missing values in a data. (6) [CO2,K2]

13. a. Apply Apriori algorithm to the following data set and find the frequent patterns (16) [CO3,K3] and generate association rules. Use the minimal support count value as 2 and the minimum confidence value is 70%.

TID	List of item IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

(OR)

- b. Find frequent patterns by applying FP-growth algorithm. Minimum support (16) [CO3,K3] count is 2.

Transaction ID	Items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

14. a. The following table represents the ten days climate data. Construct decision tree (16) [CO4,K3] for the purpose of making the decision on playing.

Day	Weather	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	rainy	Mild	High	Strong	No

(OR)

- b. Consider the following dataset where the column Stolen is the class label. We (16) [CO4,K3] need to classify a (Red, Domestic, SUV). Use Naïve Bayes classifier for the classification purpose.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

15. a. Categorize the typical requirements of clustering in data mining and compare (16) [CO5,K4] the general characteristics of the different clustering methods.

(OR)

- b. Compare each of the following clustering algorithms in terms of the following (16) [CO5,K4] criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations.

- 1) CLARA
- 2) BIRCH
- 3) ROCK
- 4) Chameleon
- 5) DBSCAN

Bloom's Taxonomy Level	Remembering (K1)	Understanding (K2)	Applying (K3)	Analysing (K4)	Evaluating (K5)	Creating (K6)
Percentage	16	31	35	18	-	-