15.4

TB246911B

Reg. No :.....

Name :.....

BACHELOR'S DEGREE (C.B.C.S) EXAMINATION, MARCH 2024

2021 ADMISSIONS REGULAR

SEMESTER VI - B.Sc. Computer Applications & B.Voc S.W.D

CA6B12AB18 - Data Mining

Time : 3 Hours **Maximum Marks : 80**

Part A

I. Answer any Ten questions. Each question carries 2 marks **(10x2=20)**

1. Interpret Frequent Patterns and Frequent Item Set.
2. Suppose that the data for analysis includes the attribute age. The age values for the 10 data tuples are (in increasing order) 13, 15, 20,21, 22, 22, 25, 25, 33, 35, 52, 70. What is the mean of the data? What is the median?
3. The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transaction containing hot dogs, No hot dog refers to the transaction that do not contain hot dogs. Hamburgers refers to the transaction containing hamburgers , No hamburgers refers to the transaction that do not contain hamburgers. Generate an Expected Frequency table.

	Hot Dogs	No Hot Dogs	Total
Hamburgers	2000	500	2500
No Hamburgers	1000	1500	2500
Total	3000	2000	5000

4. Discuss concept hierarchy.
5. Recall Snowflake schema.
6. Define Outlier.
7. How do you choose the best split while constructing a decision tree?
8. Discuss the need of Association rule mining.
9. Explain the general characteristics of density – based methods.
10. Explain the general characteristics of hierarchical methods.
11. Analyse WWW mining.
12. Explain Biological sequence data.

Part B

II. Answer any Six questions. Each question carries 5 marks **(6x5=30)**

13. Illustrate classic problems in data mining that are highly related to data mining.
14. In a certain sample of 2000 families, 1200 students are consumers of milk, out of 1600 males , 1096 male students consumes milk. Use chi-square test and state whether there is any significant difference between consumption of milk among male and female students. [Table value of χ^2 at 5% level of significance for degrees of freedom 1 is 3.84]
15. Discuss multidimensional data model schema.
16. Discuss different distance and similarity measures.
17. Formulate the application of Market Basket analysis in Marketing.
18. Compare and contrast partitioning and grid-based clustering methods.
19. Illustrate the two statistical distance methods used in K-means clustering algorithm.

20. Discuss Mining Cyber-Physical System Data.

21. Illustrate the applications of Data Mining in Science and Engineering.

Part C

III. Answer any Two questions. Each question carries 15 marks **(2x15=30)**

22. Explain data mining as a step in the process of knowledge discovery.

23. Discuss the concept of Support Vector Machines.

24. Find the association rule for the following Database D using Apriori algorithm.
Min support and confidence = 50%

Database D

TID	Onion	Potato	Burger	Milk	Beer
T1	1	1	1	0	0
T2	0	1	1	1	0
T3	0	0	0	1	1
T4	1	1	0	1	0
T5	1	1	1	0	1
T6	1	1	1	1	1

25. Find the value of decision class for the given Combination (Green, Hard, Y, Wrinkled) from the given training data using Bayesian Classification.

Example	**Colour**	**Toughness**	**Fungus**	**Appearance**	**Poisonous**
1	Green	Hard	N	Smooth	Not Poi
2	Green	Hard	Y	Smooth	Not Poi
3	Green	Soft	N	Wrinkled	Not Poi
4	Orange	Hard	N	Wrinkled	Poi
5	Green	Soft	Y	Smooth	Poi
6	Orange	Hard	Y	Wrinkled	Poi
7	Orange	Hard	N	Wrinkled	Poi
8	Orange	Soft	Y	Wrinkled	Poi
9	Green	Hard	Y	Wrinkled	?