

Reg. No.																			
-----------------	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

B.Tech. DEGREE EXAMINATION, MAY 2024

Fifth to Seventh Semester

18CSE355T – DATA MINING AND ANALYTICS

(For the candidates admitted during the academic year 2018-2019 to 2021-2022)

Note:

- (i) **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.
- (ii) **Part - B & Part - C** should be answered in answer booklet.

Time: 3 hours

Max. Marks: 100

PART – A (20 × 1 = 20 Marks)

Answer **ALL** Questions

- | | Marks | BL | CO | PO |
|---|-------|----|----|----|
| 1. Data mining is best described as the process of _____.
(A) Identifying patterns in data (B) Deducing relationships data
(C) Representing data (D) Simulating trends in data | 1 | 1 | 1 | 1 |
| 2. Given the mean of the dataset is 13 and the standard deviation is 5. Using Z-score normalization to transform the value 8, gives a new value of _____.
(A) 0 (B) 1
(C) -1 (D) -1.5 | 1 | 2 | 1 | 2 |
| 3. _____ allows the system to identify hidden pattern within unlabeled datasets on its own.
(A) Supervised learning (B) Unsupervised learning
(C) Regression (D) Pattern recognition | 1 | 1 | 1 | 1 |
| 4. Identify which is not the ratio scaled attribute among the given attribute
(A) Money (B) Weight
(C) Temperature in kelvin (D) Grade of students | 1 | 1 | 1 | 1 |
| 5. Which of the following are interesting measures for association rules?
(A) Recall (B) Lift
(C) Accuracy (D) Compactness | 1 | 1 | 2 | 2 |
| 6. Assume the minimum support is 60% and the number of transactions in the database is 5, then find the support value
(A) 3 (B) 4
(C) 5 (D) 6 | 1 | 4 | 2 | 2 |
| 7. Which classifier has the minimum error rate in comparison with all other classifiers?
(A) Bayesian classifier (B) Zero R classifier
(C) Filtered classifier (D) One R classifier | 1 | 1 | 2 | 2 |

8. Maximal frequent itemset is a _____ 1 1 2 2
 (A) Frequent itemset for which none of its immediate supersets are frequent (B) Frequent itemset for which none of its immediate subsets are frequent
 (C) Frequent itemset for which none of its immediate supersets have the same count as itself (D) Frequent itemset for which its immediate supersets have the same support count as itself
9. The Grid based clustering algorithm is _____. 1 1 4 4
 (A) k-means (B) STING
 (C) COBWEB (D) DEN CLUE
10. Zero probability value can be avoided using which technique? 1 1 3 1
 (A) If-then rules (B) Decision tree
 (C) Naive Bayes classification (D) Laplacian correction
11. Iterative dichotomizer 3, uses _____ attribute selection measure. 1 1 3 1
 (A) Information gain (B) Gini index
 (C) Gain ratio (D) Lift
12. Decision tree suffers _____. 1 1 3 1
 (A) Low bias and high variance (B) High bias and high variance
 (C) Low bias and low variance (D) High bias and low variance
13. Let say $P_1 = (5,3)$ and $P_2 = (2,1)$, represents two objects, find the Euclidean distance among them. 1 3 4 2
 (A) 5.0 (B) 3.65
 (C) 2.23 (D) 2
14. The merge approach is used to construct cluster in _____ clustering. 1 1 4 1
 (A) Partitional (B) k-means
 (C) Hierarchical (D) Naive Bayes
15. The k-means clustering algorithm fails to provide good results when there is _____. 1 1 4 4
 (A) No outliers (B) Data points are in convex shape
 (C) Outliers and density spread of data point varies in size (D) No variation of density of data points
16. A hierarchical agglomerative algorithm is _____. 1 1 4 1
 (A) BIRCH (B) DBSCAN
 (C) STING (D) k-means clustering
17. The _____ outlier requires the specification of time and location to detect it. 1 1 5 1
 (A) Conditional outlier (B) Collective outlier
 (C) Global outlier (D) No such outlier
18. There is no requirement for class labels in training data is known as _____. 1 2 5 6
 (A) Supervised learning (B) Unsupervised learning
 (C) Machine learning (D) Natural language processing

19. Identify, which of the following is not a data mining application? 1 1 5 2
 (A) Fraud detection (B) Image recognition
 (C) Customer segmentation (D) Speech recognition

20. Which of the following is not a type of recommendation system? 1 1 5 2
 (A) Content-based filtering (B) Collaborative filtering
 (C) Hybrid filtering (D) Clustering

PART – B (5 × 4 = 20 Marks)

Answer ANY FIVE Questions

Marks BL CO PO

21. Discuss the activities involved in data preprocessing. 4 2 1 1
22. Examine the following age attribute data and find out the five number summary. The age values of the data tuples are 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 35, 35, 36, 60, 72, 80. 4 4 1 2
23. Define maximum frequent itemset and closed frequent itemset. Give examples. 4 1 2 2
24. What do you mean by market basket analysis and how it can help in a supermarket? 4 1 2 1
25. List out the advantages and disadvantages of decision tree over other classification methods. 4 1 3 1
26. For the one dimensional dataset {8,11,21,29,36}, perform hierarchical clustering and plot the dendrogram to visualize it using complete linkage. 4 3 4 3
27. Justify how data mining applications are used for financial data analysis. 4 2 6 4

PART – C (5 × 12 = 60 Marks)

Answer ALL Questions

Marks BL CO PO

28. a. With a neat sketch, explain the steps in the process of knowledge discovery in databases. 12 4 1 1

(OR)

- b. Can call kinds of data be mined. Discuss about what kinds of data can be mined. 12 2 1 1

29. a. A database has five transactions. Let the minimum support and confidence, min-sup = 60%, min-confidence = 80%. Find the frequent itemset and generate the association rules using Apriori algorithm. 12 3 2 1

TID	Items
T ₁	{1,2,3,4,5,6}
T ₂	{7,2,3,4,5,6}
T ₃	{1,8,4,5}
T ₄	{1,9,0,4,6}
T ₅	{0,2,2,4,5}

(OR)

- b. A database has eight transactions. Let the minimum support, min-sup = 30%. Find the frequent item sets using FP growth algorithm. 12 3 2 1

TID	Items
T ₁	{E,A,D,B}
T ₂	{D,A,C,E,B}
T ₃	{C,A,B,E}
T ₄	{B,A,D}
T ₅	{D}
T ₆	{D,B}
T ₇	{A,D,E}
T ₈	{B,C}

30. a. Describe Naive Bayesian classification. Illustrate with an example of how the labels are predicted using Naïve Bayesian classification. 12 4 3 1

(OR)

- b. What is decision tree induction? Determine the root for the given dataset using information gain. 12 4 3 5

S.no	Outlook	Company	Sailboat	Class sail
1	Sunny	Big	Small	Yes
2	Sunny	Medium	Small	Yes
3	Sunny	Medium	Big	Yes
4	Sunny	No	Small	Yes
5	Sunny	Big	Big	Yes
6	Rainy	No	Small	No
7	Rainy	Medium	Small	Yes
8	Rainy	Big	Big	Yes
9	Rainy	No	Big	No
10	Rainy	Medium	Big	No

31. a. Explain k-means algorithm in detail and apply k-means algorithm for the following five points (with (x,y) representing locations) into two clusters:
A1 (3, 10), A2 (7, 5), A3 (10, 4), A4 (5, 9), A5(8, 5) 12 2 4 4

(OR)

- b. Illustrate the DBSCAN techniques with an example. 12 3 4 1
32. a. Discuss in detail about the supervised, semi supervised and unsupervised method for detecting the outlier. 12 2 5 2

(OR)

- b. Explain in detail, how data mining algorithm can be used for intrusion detection and prevention. 12 2 6 4

* * * * *