b. Elaborate supervised and semi-supervised outlier detection with an example.	12	4	5	3
31. a. Explain in detail about ensemble methods of classification.	12	4	3	1
(OR)				
b. Describe Naive Bayesian classification. Illustrate with an example of how the labels are predicted using Naive Bayesian classification.	12	4	3	1
32. a. Explain the following algorithm in detail. (i) K-means (ii) K-medoids	12	3	4	6
(OR)				
b. Illustrate the DB-scan algorithm with example.	12	3	3	1

* * * * *

Reg. No.														

B.Tech. DEGREE EXAMINATION, JUNE 2023
Fifth & Sixth 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. The process of removing the deficiencies and loop holes in the data is called as (A) Aggregation of data (B) Extracting of data (C) Cleaning up of data (D) Loading of data	1	1	1	1
2. Which of the following is the most important when deciding on the data structure of a data mart? (A) XML data exchange standards (B) Data access tools to be used (C) Meta data naming conventions (D) Extract, transform and load (ETL) tools to be used	1	2	1	1
3. Which one of the following does not sequence patterns for prediction? (A) A word prediction (B) Weather prediction application (C) Face detection application (D) Stock market prediction	1	2	1	1
4. Firms that are engaged in mining the emotions of the users? (A) Social media sites (B) In-depth interviews (C) Focus groups (D) Experiments	1	1	1	1
5. What techniques can be used to improve the efficiency of Apriori algorithm? (A) Hash-based techniques (B) Transaction increases (C) Sampling (D) Cleaning	1	2	2	2
6. How do you calculate confidence (A → B)? (A) Support (A∩B) / Support (A) (B) Support (A∩B) / Support (B) (C) Support (A∪B) / Support (A) (D) Support (A∪B) / Support (B)	1	1	2	2
7. You are a data-scientist in an e-commerce company you are analyzing all the transactions happened past 1 week. You observe that of 500 transactions that happened, 200 had a mobile phone in them. What is the support for mobile phones in the last one week? (A) 0.3 (B) 0.4 (C) 0.5 (D) 0.6	1	1	2	2

8. If {A, B, C, D} is a frequent item set, candidate rule which is no possible — 1 1 2 2
   (A) C → A (B) D → ABCD
   (C) A → BC (D) B → ADC

9. Which of the following statement is true about the classification? — 1 2 3 3
   (A) Market basket analysis (B) Similar group generation
   (C) Find unknown class with trained model (D) Identify the data object which does not comply with the general behaviour

10. Zero probability value can be avoided using _____. — 1 1 3 4
   (A) Decision trees (B) If-then classification
   (C) Laplacian smoothing (D) Naive-Bayesian classification

11. _____ refers to the level of understanding and insights that is provided by the classifier or predictor. — 1 1 3 4
   (A) Robustness (B) Scalability
   (C) Speed (D) Interpretability

12. _____ can be used to identity whether any two given attributes are statistically related. — 1 1 3 3
   (A) Relevance analysis (B) Regression analysis
   (C) Attribute subset selection (D) Correlation analysis

13. Which of the following is finally produced by hierarchical clustering? — 1 1 4 4
   (A) Final estimate of cluster centroids (B) Tree showing how close things are to each other
   (C) Assignment of each point to clusters (D) Estimate of centroids

14. Point out the wrong statement — 1 2 4 5
   (A) k-means clustering is a method of vector quantization (B) k-means clustering aims to partition N observations into K clusters
   (C) k-nearest neighbor is same as k-means (D) Clusters are group of similar data items

15. Which of the following algorithm is most sensitive to outliers? — 1 1 4 5
   (A) k-means clustering algorithm (B) k-medians clustering algorithm
   (C) k-modes clustering algorithm (D) k-mediods clustering algorithm

16. A good clustering method will produce high quality clusters with. — 1 2 4 6
   (A) High inter class similarity (B) Low intra class similarity
   (C) High intra class similarity (D) No inter class similarity

17. Which one of the following can be defined as the data object which does not comply with the general behaviour (or model of the available data) — 1 2 5 7
   (A) Evaluation analysis (B) Outliner analysis
   (C) Classification (D) Prediction

18. _____ helps improve machine learning results by combining several models. — 1 2 5 6
   (A) Machine learning (B) Bagging
   (C) Entropy (D) Ensemble learning

19. Which of the following is not true in detecting outliers? — 1 2 5 4
   (A) Proximity-base approaches (B) Clustering-base approaches
   (C) Time-base approaches (D) Classification approaches

20. The class labels of training data is unknown in _____. — 1 2 5 6
   (A) Supervised learning (B) Unsupervised learning
   (C) Machine learning (D) NLP

PART – B (5 × 4 = 20 Marks) — Marks BL CO PO
Answer **ANY FIVE** Questions

21. Define "Data Integration" in data Pre-processing. — 4 2 1 1

22. Explain market-basket analysis with an example. — 4 4 2 1

23. Compare supervised and unsupervised learning. — 4 4 3 2

24. Explain "Partitioning method" in clustering. — 4 2 4 1

25. List and explain the four challenges of outlier detection. — 4 2 5 4

26. Explain about information gain in decision tree induction algorithm. — 4 4 3 6

27. Define support and confidence in data mining. — 4 2 2 3

PART – C (5 × 12 = 60 Marks) — Marks BL CO PO
Answer **ALL** Questions

28. a. Explain the different steps involved in KDD process with diagram. — 12 4 1 1

(OR)

b. Explain about the data pre-processing techniques with examples. — 12 4 1 4

29. a. Explain the different methods to improve efficiency of Apriori algorithm. — 12 2 2 6

(OR)

b. A database has 9 transactions. Let the minimum support = 2, find all the frequent itemset and generate all the valid association rules using frequent pattern (FP) growth approach. — 12 4 2 1

TID	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
List of item ID's	$i_1\ i_2\ i_5$	$i_2\ i_4$	$i_2\ i_3$	$i_1\ i_2\ i_4$	$i_1\ i_3$	$i_2\ i_3$	$i_1\ i_3$	$i_1\ i_2\ i_3\ i_5$	$i_1\ i_2\ i_3$

30. a. Analyze how data mining can be used to improve financial data analysis service. — 12 5 5 1

(OR)