28. a. Explain rule-based classification in detail with example. 10 4 3 5

(OR)

b. Construct the decision tree for the basketball players data. Compute information gain for any two attributes. 10 4 3 5

Person	Jersey color	Offense/ Defense	Injured	Play
John	Blue	Offense	No	Yes
Steve	Red	Offense	No	Yes
Sarah	Blue	Defense	No	Yes
Rachel	Blue	Offense	Yes	No
Richard	Red	Defense	No	No
Alex	Red	Defense	Yes	No

29. a.i. How can you make the k-means algorithm more scalable? 5 3 4 1

ii. Differentiate agglomerative and divisive hierarchical clustering. 5 3 4 1

(OR)

b. Illustrate DBSCAN algorithm with example. 10 4 4 7

30. a. Define outlier detection. Elaborate any four outlier detection techniques. 10 4 5 4

(OR)

b. Explain any two applications of data mining. 10 4 5 6

* * * * *

Reg. No.														

B.Tech. DEGREE EXAMINATION, NOVEMBER 2022

Sixth/ Seventh Semester

18CSE355T – DATA MINING AND ANALYTICS

(For the candidates admitted from the academic year 2018-2019 to 2019-2020)

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** should be answered in answer booklet.

Time: 2½ Hours Max. Marks: 75

PART – A (25 × 1 = 25 Marks)

Answer **ALL** Questions

	Marks	BL	CO	PO
1. Where is data warehousing used? (A) Transaction system (B) Logical system (C) Decision support system (D) Banking system	1	1	1	1
2. What is the process of removing the deficiencies and loop holes in the data called? (A) Cleaning up of data (B) Aggregation of data (C) Loading of data (D) Extraction of data	1	2	1	1
3. In data warehousing we used _____ kinds of model. (A) 1D model (B) 2D model (C) 3D model (D) Multi-dimensional model	1	2	1	1
4. What is KDD in data mining? (A) Knowledge discovery database (B) Knowledge discovery data (C) Knowledge data definition (D) Knowledge data driver	1	1	1	1
5. Which of the following is an essential process in which the intelligent methods are applied to extract data patters? (A) Warehousing (B) Data mining (C) Text mining (D) Data selection	1	2	1	1
6. _____ is used as the first step in the knowledge discovery process. (A) Data selection (B) Data cleaning (C) Data transfer (D) Data integration	1	2	2	2
7. Which classifier has the minimum error rate in comparison with all other classifiers? (A) Bayesian classifier (B) Zero R classifier (C) One R classifier (D) Filtered classifier	1	1	2	2
8. The algorithm used by gird based method (A) K-means (B) COBWEB (C) DENCLUE (D) STING	1	1	2	2

9. When spatial index is used as the computational complexity of DBSCAN will be — 1 2 2 2
   (A) O (log n) (B) O (n)
   (C) O (n^2) (D) O (n log n)

10. Minkowi distance is the generalization of — 1 2 2 2
   (A) Euclidean and Tanimoto distance (B) Tanimoto and Manhattan distance
   (C) Euclidean and Manhattan distance (D) Elucidean and Tanimoto, Manhattan distance

11. Which is not a category of clustering techniques? — 1 1 3 2
   (A) Model-based method (B) Constraint-based method
   (C) Grid-based method (D) Label-based method

12. Which is not an attribute selection measure? — 1 2 3 2
   (A) Gini index (B) Information gain
   (C) Information index (D) Gain ratio

13. The percentage of tuples the rule can correctly classify — 1 2 3 2
   (A) Rule coverage (B) Rule reliability
   (C) Rule accuracy (D) Rule security

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

15. Interestingness of association rule is found by — 1 2 3 5
   (A) Confidence only (B) Support only
   (C) Info gain (D) Confidence and support

16. The complexity of each iteration in K-medoids algorithm — 1 2 4 6
   (A) $O(nkt)$ (B) $O(n^2)$
   (C) $O\left(k(n-k)^2\right)$ (D) $O\left(t(n-k)^2\right)$

17. Agglomerative hierarchical clustering uses _____ strategy. — 1 1 4 7
   (A) Bottom-up (B) Top-down
   (C) Relational (D) Mesh

18. Data, that can be stored, accessed and processes in the form of fixed format is termed as — 1 2 4 6
   (A) Unstructured (B) Semi-structured
   (C) Semi-unstructured (D) Structured

19. An example of supervised learning — 1 1 4 5
   (A) Clustering (B) Dimensionality reduction
   (C) DBSCAn (D) SVM

20. _____ is both a combination of software and hardware based computing resources delivered as network service. — 1 2 4 6
   (A) Mobile computing (B) Distributed computing
   (C) Grid computing (D) Cloud computing

21. Learning from past experience is _____. — 1 2 5 7
   (A) Machine learning (B) Improvised learning
   (C) Predictive learning (D) Machine and improvised learning

22. The _____ cloud infrastructure is operated for the exclusive use of organization. — 1 1 5 6
   (A) Public (B) Hybrid
   (C) Private (D) Public and hybrid

23. _____ provides virtual machine, virtual storage, virtual infrastructure and other hardware assets. — 1 1 5 4
   (A) IaaS (B) SaaS
   (C) PaaS (D) CaaS

24. Which is not a characteristic of big data? — 1 2 5 6
   (A) Volume (B) Variety
   (C) Visibility (D) Velocity

25. Interquartile range in — 1 2 5 7
   (A) $\theta_3 - \theta_1$ (B) Max-Min
   (C) $\theta_3 - \theta_2$ (D) $\theta_2 - \theta_1$

PART – B (5 × 10 = 50 Marks) — Marks BL CO PO
Answer **ALL** Questions

26. a. Explain about pre-processing techniques. — 10 3 1 1

(OR)

   b. Define binning and elaborate on any four binning methods used for data smoothing using following dataset. — 10 3 1 2
   36, 25, 38, 46, 55, 68, 72, 55, 36, 38, 67, 45

27. a. Discuss in detail about FP growth algorithm with an example. — 10 4 1 1

(OR)

   b. A database has five transaction. Let the minimum support be min-sup=60% and minimum confidence be min-confi = 80%. Find the frequent itemset and generate all the valid association rule using Apriori algorithm. — 10 4 2 1

TID	Items
T_1	{M,O,N,K,E,Y}
T_2	{D,O,N,K,E,Y}
T_3	{M,A,K,E}
T_4	{M,U,C,K,Y}
T_5	{C,O,O,K,I,E}