Reg. No.

B.Tech. DEGREE EXAMINATION, NOVEMBER 2019
Third to Seventh Semester

15CS331E – DATA MINING AND ANALYTICS
(For the candidates admitted during the academic year 2015 – 2016 to 2017 – 2018)

Note:
(i) **Part - A** should be answered in OMR sheet within first 45 minutes and OMR sheet should be handed over to hall invigilator at the end of 45th minute.
(ii) **Part - B** and **Part - C** should be answered in answer booklet.

Time: Three Hours Max. Marks: 100

PART – A (20 × 1 = 20 Marks)
Answer **ALL** Questions

1. The full form of KDD is
   (A) Knowledge data base discovery (B) Knowledge data discovery
   (C) Knowledge data definition (D) Knowledge discovery database

2. _____ is a comparison of the general features of target class data objects with general features of objects from set of contrasting classes
   (A) Data classification (B) Data discrimination
   (C) Data selection (D) Data characterization

3. The degree to which numerical data tend to spread is called
   (A) Mode of the data (B) Median of the data
   (C) Variance of the data (D) Central tendency of the data

4. Outlier can be detected using _____ mining technique
   (A) Classification (B) Association rule
   (C) Linear regression (D) Prediction

5. Which of the following is direct application of frequent itemset mining?
   (A) Social network analysis (B) Intrusion detection
   (C) Outlier detection (D) Market basket analysis

6. The technique that finds the frequency itemsets in first two database scans
   (A) Partitioning (B) Sampling
   (C) Hashing (D) Dynamic itemset counting

7. The support $(A \Rightarrow B)$ in association mining is represented as
   (A) $P(A \cup B)$ (B) $P(A \cap B)$
   (C) $P(A / B)$ (D) $P(B / A)$

8. This approach is best when we are interested in finding all possible interactions among a s of attributes
   (A) Decision tree (B) K-means algorithm
   (C) Association rules (D) Genetic learning

9. The association rule buys (X, 'laptop') ⇒ buys (X, "HP-Printer") is called as
   (A) Multilevel association rule (B) Quantitative association rule
   (C) Boolean association rule (D) Closed item association rule

10. A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory is known as
   (A) Baye's rules (B) Bayesian classifier
   (C) Neural network (D) SVM

11. The group of similar objects that differ significantly from other objects is called as
   (A) Class (B) Category
   (C) Cluster (D) Set

12. Multidimensional association rules with no repeated predicates are called
   (A) Hybrid dimensional association rules (B) Interdimensional association rules
   (C) Multilevel association rules (D) Intradimensional association rules

13. A bank loan officer wants to analyze the data inorder to know which customers are risky or which are safe. What mining task is suitable for him?
   (A) Classification (B) Prediction
   (C) Association (D) Clustering

14. The minimum number of variables or features required to perform clustering.
   (A) 0 (B) 1
   (C) 2 (D) 3

15. Which of the following clustering algorithm suffers from the problem of convergence at local optima?
   (A) k-means clustering (B) Divisive clustering
   (C) Agglomerative clustering (D) Conceptual clustering

16. Which method of analysis does not classify variables as dependent or independent
   (A) Regression analysis (B) Discriminant analysis
   (C) Analysis of variance (D) Cluster analysis

17. The most commonly used measure of similarity is the _____.
   (A) Chebychev's distance (B) Euclidean distance
   (C) City-block distance (D) Manhattan distance

18. FaaS is a _____ model in cloud computing.
   (A) Architecture (B) Deployment
   (C) Service (D) Engineering

19. The algorithms that gives the feedback in forms of positive or negative in a dynamic environment is called as
   (A) Active learning (B) Meta learning
   (C) Reinforcement learning (D) Supervised learning

20. The property that refers to data quality and the data value in big data is
   (A) Veracity (B) Velocity
   (C) Variety (D) Volume

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

21. List and describe the five primitives for specifying a data mining task.

22. Using the data for age in increasing order 13, 15, 16, 33, 35, 40, 70 perform min-max normalization to transform the value 40 for age onto the range [0.0, 1.0].

23. What are the disadvantages of Apriori algorithm and how it can be improved?

24. What are the preprocessing steps applied to improve the efficiency of classification process?

25. Outline how to compute dissimilarity for categorical and symmetric binary variable.

26. What are the advantages of wavelet transformation when used in clustering?

27. Write any two examples of data mining in the retail industry.

PART – C (5 × 12 = 60 Marks)
Answer **ALL** Questions

28. a. Describe the data mining functionalities and the kinds of pattern they discover.

(OR)

b. Explain the ways of handling missing values and noisy data.

29. a. A database has five transactions. Let min-support = 3. The database is as follows

TID	Items bought
T_{100}	{M,O,N,K,E,Y}
T_{200}	{D,O,N,K,E,Y}
T_{300}	{M,A,K,E}
T_{400}	{M,U,C,K,Y}
T_{500}	{C,O,K,I,E}

Construct FP growth tree and find the frequent patterns using FP growth.

(OR)

b.i. With an example, explain Apriori algorithm. (8 Marks)

ii. Compare FP growth and Apriori algorithm. (4 Marks)

30. a. Compare classification and prediction. Explain the ID3 algorithm to construct a decision tree.

(OR)

b. Discuss the working principle of linear and non-linear regression methods.

31. a. Explain the following clustering algorithms with an example
   (i) Agglomerative clustering
   (ii) Divisive clustering

(OR)

b. Explain k-medoids clustering algorithm with example.

32. a. Elaborate the applications of data mining in
    (i) Finance
    (ii) Telecommunication industry

(OR)

b.i. What is big data? List and explain the characteristics of big data.

ii. What are the key characteristics of cloud computing?

* * * * *