29. a. 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 rules using apriori algorithm

TID	Items
T1	{M, O, N, K, E, K}
T2	{D, O, N, K, E, Y}
T3	{M, A, K, E}
T4	{M, U, C, K, Y}
T5	{C, O, O, K, I, E}

(OR)

b. Discuss in detail about the FP growth algorithm with an example.

30. a. Explain Naive Bayesian classification. Illustrate with an example of how the labels are predicted using Naïve Bayesian classification.

(OR)

b. Construct the decision tree for the basketball players dataset. Compute information gain for any three attributes

Person	Jerry colour	Offense/ defense	Injured	Play
John	Blue	Offense	No	Yes
Steve	Red	Offense	No	No
Sarah	Blue	Defense	No	Yes
Rachel	Blue	Offense	Yes	No
Richard	Red	Defense	No	No
Alex	Red	Defense	Yes	No

31. a. Briefly outline how to compute the dissimilarity between objects described by the following types of variables with examples
    (i) Interval-scaled variable
    (ii) Binary variable
    (iii) Categorical variable

(OR)

b. Explain k-means algorithm in detail. Illustrate the strength and weakness of k-means in comparison with k-mediods algorithm.

32. a. Demonstrate in detail an application of data mining for retail industry. Discuss how different forms of data mining techniques can be used in the application.

(OR)

b. Explain in detail about the types, characteristics and benefits of big data.

* * * * *

Reg. No. | | | | | | | | | | | | | | |

B.Tech. DEGREE EXAMINATION, JUNE 2019
1st to 7th 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. A _____ where each method holds the code to implement a message.
   (A) Set of variables (B) Set of methods
   (C) Set of messages (D) Set of functions

2. Data that do not comply with the general behaviour or model is _____.
   (A) Outlier analysis (B) Cluster analysis
   (C) Evolution analysis (D) Market basket analysis

3. A _____ is a measure that must be computed on the entire data set as a whole.
   (A) Distributive measure (B) Algebraic measure
   (C) Central measure (D) Holistic measure

4. Which is not a technique for handling noisy data?
   (A) Regression (B) Binning
   (C) KDD process (D) Clustering

5. _____ is a subject-oriented, integrated, time-variant, non volatile, collection of data in support of management decisions.
   (A) Data mining (B) Data warehousing
   (C) Web mining (D) Text mining

6. If a substructure occurs frequently, it is called _____.
   (A) Sequential pattern (B) Semi-structured pattern
   (C) Frequent pattern (D) Structured pattern

7. Categorical attributes are also called _____.
   (A) Nominal attributes (B) Quantitative attributes
   (C) Ordinal attributes (D) Ratio-scaled attributes

8. The apriori algorithm is used for which data mining task.
   (A) Association (B) Clustering
   (C) Classification (D) Database

9. The percentage of tuples the rule can correctly classify.
   (A) Rule coverage (B) Rule reliability
   (C) Rule accuracy (D) Rule security

10. Zero probability value can be avoided by_____.
   (A) Decision trees (B) Laplacian correlation
   (C) If-then classifiers (D) Naive Bayesian classifier

11. The ratio of the number of attributes shared by X and Y to the number of attributes possessed by X or Y is called _____.
   (A) Tanimoto coefficient (B) Jaccard coefficient
   (C) Silhouette coefficient (D) Manhattan coefficient

12. Which classifier has the minimum error rate in comparison with all other classifiers?
   (A) Zeror classifiers (B) Oner classifier
   (C) Filtered classifier (D) Bayesian classifier

13. The complexity of each iteration in k-medoids algorithm.
   (A) $O(nkt)$ (B) $O(k(n-k)^2)$
   (C) $O(n)^2$ (D) $O(t(n-t)^2)$

14. Agglomerative hierarchical clustering uses _____ strategy.
   (A) Bottom-up (B) Top-down
   (C) Relational (D) Transactional

15. The condition that holds good for k-means clustering
   (A) k << n and t >> n (B) k << n and t << n
   (C) k << n and t << n (D) k >> n and t >> n

16. Manhattan distance is also called as _____.
   (A) Euclidean distance (B) Minkowsi distance
   (C) City blocks distance (D) Similar distance

17. Which is not a characteristic of big data?
   (A) Volume (B) Variety
   (C) Visibility (D) Velocity

18. The complete application running on someone else's system is _____.
   (A) PaaS (B) SaaS
   (C) IaaS (D) CaaS

19. Google forms is the example of _____ cloud.
   (A) Public (B) Private
   (C) Hybrid (D) Public and hybrid

20. _____ provides virtual machine, virtual storage, virtual infrastructure and other hardware assets.
   (A) SaaS (B) PaaS
   (C) CaaS (D) IaaS

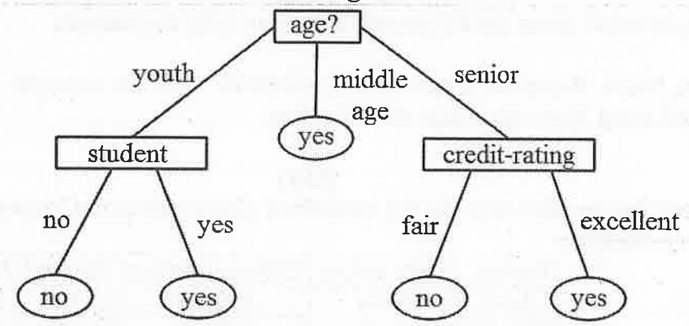
PART – B (5 × 4 = 20 Marks)
Answer **ANY FIVE** Questions

21. Define data mining? Is the word 'Data Mining' a misnomer? Why?

22. Summarize any two techniques used for data reduction.

23. Draw the box or whisker plot for the following data. Identify the outliers.
   72, 78, 79, 62, 85, 41, 64, 90, 130, 70, 46, 76, 3.

24. Predict the classification rules for the following decision tree.



25. What is dendogram? How are the cluster merged?

26. Illustrate the limitaitons of k-means clustering.

27. List any five applications of data mining.

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

28. a.i. Describe the primitives for performing data mining task.

ii. Define binning and elaborate on binning methods used for data smoothing using the following dataset
   36, 25, 38, 46, 55, 68, 72, 55, 36, 38, 67, 45.

(OR)

b.i. Explain in detail the process of knowledge discovery from databases with a diagram.

ii. Suppose that a group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their preferred type of reading material was fiction or non fiction. Find the observed frequency and construct the contingency table for the following data and find Chi square value

	Male	Female	Total
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Total	300	1200	1500