27. a. Consider the following transaction table:

Trans_ID	Itemsets
11	(B,D) (B,E) (D,E) (B,D,E)
12	(A,B) (A,C) (A,E) (B,C) (B,E) (C,E) (A,B,C) (A,B,E) (A,C,E) (B,C,E)
13	(B,C) (B,D) (B,E) (C,D) (C,E) (D,E) (B,C,D) (B,C,E) (B,D,E) (C,D,E)

(i) Calculate the frequent itemset with minimum support of 67% using Apriori algorithm. 5 4 2 4

(ii) Calculate the confidence value for the rule "IF B THEN C" and derive possible association rules from the above frequent itemset. 5 4 2 4

(OR)

b.i. Explain about the procedure to convert horizontal frequent itemset to vertical frequent itemset with example. 5 3 2 3

ii. Compare strong and weak association rules with example. 5 3 2 3

28. a. Construct the decision tree for the given table using ID_3 algorithm. Write the stepwise procedure used to construct tree structure. 10 4 3 3

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	–	T	F
4	+	F	F
5	–	F	T
6	–	F	T

(OR)

b.i. Explain about ensemble method in data mining with examples. 5 3 3 1

ii. List out various classifier evaluation metrics in detail. 5 3 3 1

29. a. Consider the following sample data, to calculate two cluster values using K-means algorithm [Euclidean distance function]. 10 3 4 2

Height (H)	185	170	168	179	182	188	180	180	183	180	180	177
Weight (W)	72	56	60	68	72	77	71	70	84	88	67	76

(OR)

b. Explain about the following clustering techniques in detail

(i) DBSACN 5 3 4 3

(ii) BIRCH 5 3 4 3

30. a. Explain about various outlier detection approaches with example. 10 3 5 3

(OR)

b. Explain about the following

(i) Recommender system in data mining 5 3 6 3

(ii) Intrusion detection using data mining 5 3 6 3

* * * * *

Reg. No.														

B.Tech. DEGREE EXAMINATION, MAY 2022

Fifth & Sixth 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) Marks BL CO PO

Answer **ALL** Questions

1. Select the most appropriate way of handling missing data 1 1 1 1
   (A) Data integration (B) Data reduction
   (C) Use of global constant (D) Data cube technique

2. _____ used to measure the linear relationship between the objects. 1 2 1 1
   (A) Mean value (B) Standard deviation
   (C) Covariance (D) Correlation

3. A E-commerce company wants to segment their customers into distinct groups to promote offers, select appropriate method 1 2 1 2
   (A) Unsupervised learning (B) Supervised learning
   (C) Data transformation (D) Reinforcement learning

4. Data visualization techniques are used to _____. 1 1 1 1
   (A) Calculate accuracy (B) Detect outliers
   (C) Improve training accuracy (D) Integrate data

5. Maximizing intra-class similarity and minimizing inter-class similarity are the objectives of _____. 1 2 1 1
   (A) Classification (B) Outlier analysis
   (C) Clustering (D) Sequence pattern

6. In _____ ensemble method, each new model is influenced by the performance of those built previously. 1 1 2 2
   (A) Bagging (B) Boosting
   (C) Stacking (D) Learning

7. _____ used to minimize the misclassification error in decision tree 1 2 3 2
   (A) Boosting (B) Overfitting
   (C) Pruning (D) Bagging

8. How do you calculate confidence (X→Y)? 1 2 2 2
   (A) Support (X∩Y)/ support(X) (B) Support (X∪Y)/ support(X)
   (C) Support (X∩Y)/ support(Y) (D) Support (X∪Y)/ support(Y)

9.	_____ algorithm used to identify the frequent pattern by mining the conditional frequent pattern trees. (A) Apriori algorithm (B) Naïve bayes algorithm (C) FP growth algorithm (D) Clustering algorithm	1	2	2	2
10.	For a given scenario, 22 out of 100 people are predicted as positive of having a tumor, although they don't have a tumor. This is considered as _____. (A) True Positive (TP) (B) True Negative (TN) (C) False Negative (FN) (D) False Positive (FP)	1	?	?	?
11.	_____ used to measure the fraction of positive patterns that are classified correctly. (A) Error rate (B) Precision (C) Recall (D) F-measure	1	2	2	1
12.	_____ is a statistical classifier, which has comparable performance with decision tree. (A) Rule based classifier (B) Naive Bays classifier (C) Sequential classifier (D) Entropy based classifier	1	2	3	1
13.	In web mining, _____ is used to know the order in which urls tend to be accessed. (A) Clustering (B) Associations (C) Classifications (D) Sequential analysis	1	2	3	2
14.	Which one of the clustering technique needs the merging approach? (A) Partitioned (B) Naive Bayes (C) Hierarchical (D) Decision tree	1	1	4	2
15.	_____ used to handle the classification problems (A) Linear regression (B) Logistic regression (C) k-means (D) Preprocessing techniques	1	2	3	2
16.	_____ algorithm extracts rules directly from training data (A) k-means (B) Partition around medoids (PAM) (C) Sequential covering (D) Frequent pattern	1	2	3	1
17.	_____ example for partitioning based clustering algorithm (A) DBSCAN (B) DIANA (C) Optics (D) k-medoids	1	2	4	2
18.	Decompose data object into several levels of nested partitioning called _____. (A) Boosting (B) Dendrogram (C) Tree pruning (D) Histogram	1	2	4	3
19.	_____ method used to identify clustering structure through ordering points (A) Birch (B) Optics (C) Agnes (D) Clarans	1	4	4	4

20.	CF-Tree used under _____ type of clustering. (A) Density based (B) Grid based (C) Hierarchical based (D) Model based	1	3	4	4
21.	Speech recognition technique with single background noise, is an example for _____. (A) Global outliers (B) Collective outliers (C) Contextual outliers (D) Large outliers	1	2	5	4
22.	Select the type of outlier deviates significantly from most of other dataset (A) Graph based outlier (B) Proximity based outlier (C) Grid based outlier (D) Mean based outlier	1	2	5	4
23.	Item based recommendation system using _____ to predict user preferences. (A) Likes and dislikes measure (B) Mean adjusted matrix (C) k-mediods (D) Recall value	1	2	5	3
24.	_____ type of attacks can be identified using datamining intrusion detection (A) Information attacks (B) Denial of service (DOS) attacks (C) Password attacks (D) SQL injection attacks	1	2	5	3
25.	Which is used to perform inference on the current data to make predictions? (A) Data mining (B) Data pattern (C) Predictive (D) Descriptive	1	2	5	4

PART – B (5 × 10 = 50 Marks)

Answer **ALL** Questions

		Marks	BL	CO	PO
26. a.	Explain about various stages of KDD process in detail with proper diagram.	10	3	1	3

(OR)

b.i. Consider the following shopping mall customers details dataset.

S.No	Name	Occupation	Branch	Date	Price
1	Ramesh	Govt	DD	11-Jan	
2	Vivek	Self	CC	12-Jan	2500
3	Kiran				
4	Suresh	Private	FF	14-Jul	–300
5	Hemanth	Business	KK	14-Jan	1100
6	Sai	Govt	VV	12-Feb	500

(i)	How to handle negative, missing value in the above table and fill the missing value using the measure of central tendency?	5	3	1	3
(ii)	Explain about need for Data-pre-processing and data quality in mining operations.	5	3	1	3