Reg. No															

B.Tech DEGREE EXAMINATION, NOVEMBER 2023

Seventh Semester

18AIE427T - DATA MINING AND ANALYTICS

(For the candidates admitted during the academic year 2020 - 2021 & 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** and **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
1.	An alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid is ---------- (A) Sampling (B) Variance (C) Centroid Distance (D) Aggregation	1	2	1
2.	What is the full form of KDD in the data mining process? (A) Knowledge data house (B) Knowledge data definition (C) Knowledge discovery data (D) Knowledge discovery database	1	1	1
3.	Which of the following process uses intelligent methods to extract data patterns? (A) Data mining (B) Text mining (C) Warehousing (D) Data selection	1	1	1
4.	$\chi 2$ (chi-square) test can be used for (A) Ordinal Attributes (B) Numeric Attributes (C) Distinct Attributes (D) Nominal Attributes	1	1	1
5.	____ property states that if a set cannot pass a test, all of its super set will fail the same test. (A) divide and conquer (B) Apriori (C) Association rule (D) Anti-Monotonic	1	2	2
6.	In what condition does the association rule can be interesting? (A) If it only satisfies min-support (B) If it satisfies min-confidence (C) If it satisfies both min-support and min-confidence (D) If it satisfies max-support	1	2	2
7.	What do you mean by support(A)? (A) Total number of transactions containing A (B) Total Number of transactions not containing A (C) Number of transactions containing A / Total number of transactions (D) Number of transactions not containing A / Total number of transactions	1	1	2
8.	Why correlation analysis is important? (A) To make apriori memory efficient (B) To find relationship between the item set (C) To find large number of interesting item set (D) To restrict the number of database iterations	1	2	2
9.	Internal node of a decision tree induction represents---- (A) outcome of the test (B) test on an attribute (C) the class label (D) the root node	1	1	3

10.	Decision trees can handle_____ (A) High dimensional data (B) low dimensional data (C) medium dimensional data (D) none of these	1	1	3
11.	------ is the statistical method that is most often used for numeric prediction (A) predictor (B) Regression analysis (C) decision tree classifier (D) Bayesian classifier	1	1	3
12.	Suppose Y is a binary valued dependent variable (1 and 0) and x1 and x2 (1) are explanatory variables. Predicting the probability Y=1 involves estimating a ------ (A) Linear regression (B) logistic regression (C) Poisson regression (D) linear probability model	1	2	3
13.	Which algorithm is used for clustering? (A) k-means (B) KNN (C) SVM (D) PCA	1	1	4
14.	Which is conclusively produced by Hierarchical Clustering? (A) final estimation of cluster centroids (B) tree showing how nearby things are to each other. (C) assignment of each point to clusters (D) all of these	1	1	4
15.	Which clustering technique requires a merging approach? (A) Partitional (B) Hierarchical (C) Naive Bayes (D) Vertical	1	1	4
16.	Consider the scenario to predict the number of newborns according to the size of storks' population by performing supervised learning is ---------- (A) Structural equation modeling (B) Clustering (C) Regression (D) Classification	1	1	4
17.	Text mining tasks does not include (A) text categorization (B) text clustering (C) concept/entity extraction (D) Regression analysis	1	1	5
18.	----------is the process of extracting useful information (e.g., user click streams) from server logs. (A) Web structure mining (B) Web content mining (C) Web usage mining (D) Web mining	1	1	5
19.	----------discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization techniques (A) Data visualization (B) Visual web mining (C) Visual text mining (D) Visual data mining	1	1	5
20.	----------discovers patterns and knowledge from spatial data. (A) Web mining (B) Text mining (C) Spatial data mining (D) Temporal data mining	1	1	5

PART - B (5 × 4 = 20 Marks)

Answer **any 5** Questions

		Marks	BL	CO
21.	Differentiate classification and prediction. Briefly explain the issues regarding Classification and Prediction.	4	3	3
22.	List the primitives that specify a data mining task	4	2	1
23.	Assume you are given a dataset about cancer detection. You have to build a classification model and achieve 96% accuracy. What makes you think that the performance of your model isn't satisfactory? Is there anything you can do about it?	4	3	2
24.	Suppose that the data mining task is to cluster points into three clusters, where the points are A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9). Use an appropriate algorithm to divide the given dataset into different clusters.	4	3	4

25. Explain in detail about mining spatial data. 4 2 5

26. From the Following Data Compute Spearman Coefficient of Correlation. 4 3 2

x	2	4	6	8	10
y	12	14	16	18	20

27. Compare and contrast between Heterogeneous and legacy database. 4 3 1

PART - C (5 × 12 = 60 Marks) **Marks BL CO**

Answer **all** Questions

28. (a) Suppose a biologist claims that an equal number of four different species of deer enters a certain wooded area in a forest each week. 12 4 1
To test this hypothesis, she records the number of each species of deer that enter the wooded area over the course of one week:
- **Species #1:** 22
- **Species #2:** 20
- **Species #3:** 23
- **Species #4:** 35

Using Chi Square Test determine if the distribution of the deer species that enter the wooded area in the forest each week is consistent with his hypothesized distribution.

(OR)

(b) Researchers have conducted a survey of 1600 coffee drinkers asking how much coffee they drink in order to confirm previous studies. Previous studies have indicated that 72% of Americans drink coffee. The results of previous studies (left) and the survey (right) are below. At $\alpha = 0.05$, is there enough evidence to conclude that the distributions are the same?

Response	% of Coffee Drinkers
2 cups per week	15%
1 cup per week	13%
1 cup per day	27%
2+ cups per day	45%

Response	Frequency
2 cups per week	206
1 cup per week	193
1 cup per day	462
2+ cups per day	739

29. (a) Suppose we are interested in analyzing transactions at All Electronics with respect to the purchase of computer games and videos. Let game refer to the transactions containing computer games, and video refer to those containing videos. Of the 10,000 transactions analyzed, the data show that 6000 of the customer transactions included computer games, while 7500 included videos, and 4000 included both computer games and videos. Suppose that a data mining program for discovering association rules runs on the data, using a minimum support of, say, 30% and a minimum confidence of 60%. Write an association rule for this scenario and find out support value and confidence value. 12 4 2

(OR)

(b) Illustrate Apriori algorithm and implement it for finding the frequent pattern for the given transactional database. Min.support:2

T-id	Itemset
I1	1,2,5
I2	2,4
I3	2,3
I4	1,2,4
I5	1,3
I6	2,3
I7	1,3
I8	1,2,3,5
I9	1,2,3

30. (a) Construct the decision tree for the given transactional database. 12 4 3

Outlook	Temperat	Humidity	Windy	play
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

(OR)

(b) Classify the fruit, which is more yellow, sweeter, and longer using an appropriate classifier.

Fruit	Yellow	Sweet	Long	Total
Orange	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	150
Total	800	850	400	1200

31. (a) Both k-means and k-medoids algorithms can perform effective clustering. 12 3 4
(i) Illustrate the strength and weakness of k-means in comparison with k-medoids.
(ii) Illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme.

(OR)

(b) Present conditions under which density-based clustering is more suitable than partitioning-based clustering and hierarchical clustering. Give examples to support your argument.

32. (a) Explain in detail about Multimedia analytics with an example. 12 2 5

(OR)

(b) Explain in detail about Social network analytics with an example.

* * * * *