Reg. No.																

Question Paper Code	**13489**

B.E. / B.Tech. - DEGREE EXAMINATIONS, APRIL / MAY 2025

Sixth Semester

Computer Science and Business Systems

20CBEL601 - DATA MINING AND ANALYTICS WITH LABORATORY

Regulations - 2020

Duration: 3 Hours Max. Marks: 100

PART - A (MCQ) (10 × 1 = 10 Marks)

Answer ALL Questions

Marks K – Level CO

1. Which of the following process uses intelligent methods to extract data patterns? *1 K1 CO1*
(a) Data mining (b) Text mining (c) Warehousing (d) Data selection

2. What is the full form of KDD in the data mining process? *1 K1 CO1*
(a) Knowledge data house (b) Knowledge data definition
(c) Knowledge discovery data (d) Knowledge discovery database

3. How the two attributes are defined in Covariance? *1 K1 CO2*
(a) Identical (b) Different (c) Binary (d) Nominal

4. How the class information is used during discretization process? *1 K1 CO2*
(a) Supervised discretization (b) Unsupervised discretization
(c) Clustered discretization (d) Disorganized discretization

5. What distance metric is commonly used in KNN? *1 K1 CO3*
(a) Manhattan Distance (b) Euclidean Distance
(c) Cosine Similarity (d) All of the above

6. What does K represent in K-Nearest Neighbors (KNN)? *1 K1 CO3*
(a) The number of decision trees used
(b) The number of neighbors considered for classification
(c) The number of independent variables
(d) The number of hidden layers in a neural network

7. Which link function is used to model a binomial response in logistic regression? *1 K1 CO4*
(a) Logic function (b) Log function (c) Identity function (d) Square root function

8. What is the primary purpose of logistic regression? *1 K1 CO4*
(a) Predicting continuous values (b) Modeling binary or categorical outcomes
(b) Clustering similar data points (d) Reducing dimensionality in datasets

9. How to measure the effectiveness on K-Nearest Neighbors (KNN) ? *1 K1 CO5*
(a) The number of independent variables (b) The choice of the number of neighbors (k)
(c) The intercept value (d) The size of the residuals

10. How is the auto-correlation function (ACF) defined for a time series? *1 K1 CO6*
(a) It is the ratio of auto-covariance to variance
(b) It is the sum of all past observations
(c) It is the squared difference between observations
(d) It is the moving average of a time series

PART - B (12 × 2 = 24 Marks)

Answer ALL Questions

11. Define Left skewness and right skewness with example. *2 K1 CO1*

12. List the Euclidean distance and Manhattan distance. *2 K1 CO1*

13. Show how outlier is detected in data mining and data analytics. *2 K1 CO2*

14. Define Exploratory Data Analysis. *2 K1 CO2*

15. What is Bayesian network? *2 K1 CO3*

16. Compare and contrast attribute relevance and attribute generalization. *2 K2 CO3*

17. Compare time-series forecasting and predictive modeling. *2 K2 CO4*

18. What do you understand from the terms correlation and regression? *2 K1 CO4*

19. What do you mean by semi parametric regression models and additive regression models? *2 K1 CO5*

20. When does Newton-Raphson fail? *2 K1 CO5*

21. Define the terms Exploratory time series analysis. *2 K1 CO6*

22. Define contrast Autoregressive, and Moving Average Models. *2 K1 CO6*

PART - C (6 × 11 = 66 Marks)

Answer ALL Questions

23. a) (i) Explain the following: (a) Binning (b) regression (c) Clustering (d) Smoothing (e) Generalization (f) Aggregation. *5 K2 CO1*

(ii) Summarize OLAP And OLTP. *6 K2 CO1*

OR

b) (i) Explain the steps involved in KDD with a neat diagram and also describe data cleaning process. *5 K2 CO1*

(ii) Explain the various applications of data mining. *6 K2 CO1*

24. a) Explain the various data preprocessing steps: data cleaning, transformation, and reduction with examples. *11 K2 CO2*

OR

b) The mean of the data set X? *11 K2 CO2*

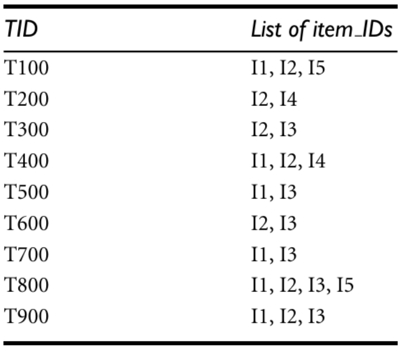
(a) solve A data set for analysis includes only one attribute X:

X = {7,12,5,8,5,9,13,12,19,7,12,12,13,3,4,5,13,8,7,6}

(b) Calculate the median?

(c) Find the standard deviation for X.

25. a) Consider the below given AllElectronics transaction database, D. *11 K2 CO3*



TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Generate candidate itemsets and frequent itemsets using Apriori algorithm, where the minimum support count is 2.

OR

b) Summarize the nearest neighbor classification algorithm with suitable examples. *11 K2 CO3*

26. a) Identify how Logistic regression differs from Linear regression with suitable graphical representations. *11 K3 CO4*

OR

b) A researcher wants to understand the relationship between the number of hours a student studies and their score in a statistics exam. The following data was collected from a sample of 8 students: *11 K3 CO4*

Student	Hours Studied (X)	Exam Score (Y)
1	2	65
2	3	70
3	5	75
4	4	72
5	6	78
6	8	85
7	7	82
8	9	88

Predict the exam score of the students when she studies 12 Hours using Logistic Regression.

27. a) Explain in detail Marquardt Method. *11 K2 CO5*

OR

b) Explain grid search and randomized search with suitable python code. *11 K2 CO5*

28. a) Illustrate the steps in building an ARIMA model for forecasting. *11 K2 CO6*

OR

b) Explain Holt-Winters smoothing and show how it is used for forecasting. *11 K2 CO6*

K1 – Remember; K2 – Understand; K3 – Apply; K4 – Analyze; K5 – Evaluate; K6 – Create