BRAINWARE UNIVERSITY

Term End Examination 2024-2025
Programme – MCA-2022/MCA-2023
Course Name – Data Warehousing and Data Mining
Course Code - MCA305A
(Semester III)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A
(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) Which of the following is not a step in the data mining process?
a) Data preprocessing b) Data exploration
c) Data visualization d) Model evaluation

(ii) Which data mining technique is used for finding patterns in a large dataset without any specific target variable?
a) Classification b) Clustering
c) Regression d) Association

(iii) Which of the following is an example of supervised learning in data mining?
a) Clustering customers into segments based on their purchase history dataset b) Predicting the price of a house based on its features
c) Finding frequent itemsets in a transaction dataset d) Grouping similar documents based on their content

(iv) Which of the following is not a data mining task?
a) Classification b) Regression
c) Sorting data d) Clustering

(v) Which data mining technique is often used for fraud detection and anomaly detection?
a) Classification b) Clustering
c) Regression d) Outlier detection

(vi) What is the primary purpose of data preprocessing in data mining?
a) To reduce the dimensionality of the dataset b) To make the data more understandable for humans
c) To improve the quality of data and prepare it for analysis d) To generate new data from the existing dataset

(vii) What does KDD stand for in the context of data analysis?
a) Knowledge Database Design b) Knowledge Discovery in Databases
c) Key Data Definitions d) Knowledge Development and Deployment

(viii) Which step in the KDD process involves the selection of relevant data from the database?

a) Data Cleaning
b) Data Integration
c) Data Selection
d) Data Transformation

(ix) What is regression in data mining?

a) Going backward.
b) Predicting a value based on other data.
c) Running a race.
d) Drawing pictures.

(x) What is data transformation in data mining?

a) The process of converting data into gold.
b) The process of converting data into a different format for analysis.
c) The process of deleting data from a dataset.
d) The process of creating new data from scratch.

(xi) Which of the following best explains how to compare APRIORI and FP-Growth algorithms in terms of handling database scans?

a) APRIORI performs multiple passes over the database, while FP-Growth scans the database only twice.
b) Both algorithms perform multiple database scans but differ in candidate generation methods.
c) FP-Growth performs a single database scan, while APRIORI does not require any.
d) APRIORI uses fewer scans than FP-Growth due to its candidate generation technique.

(xii) When generating frequent item sets, what is the primary factor that determines if an item set is considered to be frequent?

a) Itemset length
b) Minimum support threshold
c) Data size
d) Maximum confidence level

(xiii) Choose the correct assumption that the APRIORI principle is based on regarding subsets of frequent item sets.

a) If an item set is frequent, then all its subsets must be frequent.
b) If an item set is infrequent, some of its subsets may still be frequent.
c) Subsets of infrequent item sets may also be infrequent.
d) The support of subsets is always higher than their supersets.

(xiv) Determine What is the first step in the FP-Growth algorithm?

a) Generating frequent itemsets.
b) Constructing the FP-Tree.
c) Calculating support counts.
d) Pruning infrequent items.

(xv) In a decision tree, what are the internal nodes responsible for?

a) Representing the final decision or output
b) Representing the features or attributes of the data
c) Representing the class labels of the data
d) Storing the training data points

Group-B

(Short Answer Type Questions) 3 x 5=15

2. Recall the concepts of classification and clustering in Data Mining. (3)
3. Explain Some Real-World Applications of Data Mining. (3)
4. Explain why a data warehouse is used ? (3)
5. Differentiate between Cluster and classification. (3)
6. Explain the "support" of an itemset in the Apriori algorithm. (3)

OR

How is "confidence" calculated for an association rule in the Apriori algorithm? (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Create an FP-Growth tree for a given dataset of customer transactions at a grocery shop. (5)
Transaction 1: {Apple, Banana, Cheese} Transaction 2: {Banana, Cheese, Diapers}
Transaction 3: {Apple, Banana, Cheese, Diapers} Transaction 4: {Apple, Diapers} Transaction 5: {Banana, Cheese}
8. Describe the key features of MOLAP, ROLAP, and HOLAP. How do they differ from one to another? (5)
9. Examine the importance of classification in data mining vary across different industries. (5)
10. Describe the key components of a star schema and explain how it organizes data in a data warehouse. (5)
11. Evaluate the Divisive hierarchical clustering method with the following set of points in 2D space: A(1, 2) B(2, 3) C(4, 3) D(7, 5) E(6, 6). (5)
12. Evaluate the closed itemsets and maximal itemsets from the following dataset of customer transactions at a grocery store Transaction 1: {Apple, Banana, Cheese} Transaction 2: {Banana, Cheese, Diapers} Transaction 3: {Apple, Banana, Cheese, Diapers} Transaction 4: {Apple, Diapers} Transaction 5: {Banana, Cheese}. (5)

OR

Design a comprehensive strategy to prevent overfitting in a decision tree model for a complex dataset. (5)

****