Avinashilingam Institute for Home Science and Higher Education for Women

Deemed to be University Estd. u/s 3 of UGC Act 1956, Category A by MHRD (now MoE)
Re-accredited with A++ Grade by NAAC. CGPA 3.65/4, Category I by UGC
Coimbatore - 641 043, Tamil Nadu, India

Master's Degree Examination – May 2025
II Semester

Class : I M.C.A. **Time: 3 Hours**
Major : Computer Applications **Max. Marks: 100**

23MCAC12 Data Mining and Warehousing

Course Outcomes:

CO1: Apply the various steps of the KDD process and apply the relevant preprocessing techniques in large datasets
CO2: Delineate the processes involved in the construction of a data warehouse
CO3: Analyze and comprehend the concepts of multidimensional databases.
CO4: Propose solutions using data mining tools and demonstrate application of association, classification and clustering algorithms.
CO5: Develop strategies for BI and other applications using Weka.

Part A **10 x 1 = 10**
Choose the Correct Answer

1. The process of removing noise and inconsistent data is called CO1K1
   a. Data cleaning b. Data transformation c. Data reduction d. Data integration

2. Which of the following schemas is commonly used for organizing multidimensional databases in data warehousing? CO1K1
   a. Hierarchical Schema b. Network Schema c. Star Schema d. Relational Schema

3. The number of transactions that contain the item set is called CO2K1
   a. Maximum frequency b. Minimum frequency
   c. Absolute frequency d. Frequency

4. Which of the following best describes Constraint-Based Association Mining? CO2K1
   a. A method that finds all possible associations in a dataset without any restrictions
   b. A technique that applies user-specified constraints to guide the discovery of association rules
   c. A process that removes redundant association rules after mining is complete
   d. A method that only considers frequent item sets without applying any constraints

5. The difference between the original information requirement and the new information requirement after partitioning is known as CO3K2
   a. Information gain b. Gini index c. Gain ratio d. Minimum description length

6. Which of the following statements is true about non-linear regression? CO3K1
   a. Non-linear regression models the relationship between variables using a straight line
   b. Non-linear regression can model more complex relationships than linear regression
   c. Non-linear regression cannot be used for predictive modelling
   d. Non-linear regression assumes the residuals must always be normally distributed

7. Which of the following methods finds mutually exclusive clusters of spherical shape? CO4K2
   a. Density-based methods b. Grid-based methods
   c. Hierarchical methods d. Partitioning methods

8. What is the primary impact of outliers in data analysis? CO4K2
   a. They always improve the accuracy of predictive models
   b. They have no effect on statistical measures like mean and standard deviation
   c. They can distort statistical summaries and influence model performance
   d. They are always due to errors in data collection

9. The primary goal of sequential pattern mining is to CO5K1
   a. find associations between unrelated data items
   b. discover frequent subsequences in a sequence database
   c. cluster similar sequences together
   d. compress time-series data

10. Which of the following is NOT a major component of the WEKA tool? CO5K1
   a. Explorer b. Experimenter c. Knowledge Flow d. Tensor Board

Part B **5 x 6 = 30**

Answer ALL questions

Each answer should not exceed 400 words or two pages

11.a. What is a data warehouse? Explain its design principles in detail. CO1K2

(or)

11.b. What are the advantages and challenges of using a Multidimensional Database? CO1K2

12.a.i. Define Association Rule Mining.

ii. Explain the advantages and disadvantages of Association Rule Mining with an example. CO2K3

(or)

12.b. Define the Apriori algorithm with an example and explain the importance of the Apriori property. CO2K2

13.a. Describe Naire Balyesian classification with an example. CO3K3

(or)

13.b. Define decision tree induction. List advantages and limitations of decision trees. CO3K3

14.a. Explain the working of DBSCAN. How does DBSCAN handle outliers? CO4K2

(or)

14.b. What are the different types of constraints used in clustering? Give an example where constraint-based clustering is useful. CO4K2

15.a. Explain any two methods for mining time series data. CO5K3

(or)

15.b. Describe the primary uses of WEKA tool in data mining. CO5K2

Part C **5 x 12 = 60**

Answer ALL questions

Each answer should not exceed 800 words or four pages

16.a. Explain the difference between data warehousing and data mining. CO1K3

(or)

16.b.i. Discuss how computations can be performed efficiently on data cubes.

ii. Write short notes on meta data. CO1K3

17.a. Compare FP-Growth and Apriori in terms of performance. CO2K3

(or)

17.b.i. Define association rule mining.

ii. Discuss the advantages and limitations of association rule mining. CO2K2

18.a.i. Explain rule-based classification with IF-THEN rules**.**

ii. Example of rule-based classification in medical diagnosis. CO3K3

(or)

18.b.i. Discuss the confusion matrix and its components.

ii. Discuss the various metrics used for classification. CO3K3

19.a. Discuss different clustering techniques: partitioning, hierarchical, density-based, model-based, constraint-based**.** CO4K4

(or)

19.b. Compare hierarchical clustering with K-Means. CO4K3

20.a. Describe different approaches for similarity search in text mining. Give examples of their applications. CO5K3

(or)

20.b. Compare and contrast different classification techniques available in WEKA. Which one is best suited for biological data analysis? CO5K4

****