PSG COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS)

MSc DEGREE EXAMINATION MAY 2024
(First Semester)

Branch – **COMPUTER SCIENCE**

DATA MINING AND ANALYTICS

Time: Three Hours

Maximum: 75 Marks

SECTION-A (10 Marks)

Answer **ALL** questions

ALL questions carry **EQUAL** marks (10 × 1 = 10)

Module No.	Question No.	Question	K Level	CO
1	1	In the example of predicting number of babies based on storks' population size, number of babies is a. Outcome b. Feature c.Attribute d. Observation	K1	CO1
	2	Cluster is a. Group of similar objects that differ significantly from other objects b. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm c. Symbolic representation of facts or ideas from which information can potentially be extracted d. None of these	K2	CO2
2	3	Assume you want to perform supervised learning and to predict number of newborns according to size of storks' population , it is an example of a. Classification b. Regression c. Clustering d. Structural equation modeling	K1	CO1
	4	Mention a good alternative to the star schema? a. Snow flake schema b. Star schema c. Star snow flake schema d. Fact constellation	K2	CO2
3	5	Dimensionality reduction reduces the data set size by removing _____ Select one: a. Composite attributes b. Derived attributes c. Relevant attributes d. Irrelevant attributes	K1	CO1
	6	Incorrect or invalid data is known as _____ Select one: a. Missing data b. Outlier c. Changing data d. Noisy data	K2	CO2
4	7	Specify the data mining task is known as Market Basket Analysis? Select one: a. Association Analysis b. Regression c. Classification d. Outlier Analysis	K1	CO1
	8	Which one of the following statements about the K-means clustering is incorrect? a. The goal of the k-means clustering is to partition (n) observation into (k) clusters b. K-means clustering can be defined as the method of quantization c. The nearest neighbor is the same as the K-means d. All of the above	K2	CO2

Cont...

5	9	Which of the following statements about hierarchal clustering is incorrect? a. The hierarchal clustering can primarily be used for the aim of exploration b. The hierarchal clustering should not be primarily used for the aim of exploration c. Both A and B d. None of the above	K1	CO1
	10	Which one of the clustering technique needs the merging approach? a. Partitioned b. Naïve Bayes c. Hierarchical d. Both A and C	K2	CO2

SECTION - B (35 Marks)

Answer **ALL** questions

ALL questions carry **EQUAL** Marks $(5 \times 7 = 35)$

Module No.	Question No.	Question	K Level	CO
1	11.a.	Write brief notes on methods of clustering.	K3	CO3
		(OR)		
	11.b.	What are issues in data mining?		
2	12.a.	With the help of a diagram, explain Decision Tree Algorithm.	K4	CO4
		(OR)		
	12.b.	Discuss Association Algorithm In Data Mining.		
3	13.a.	Explain hierarchical clustering techniques stating their pros and cons.	K3	CO3
		(OR)		
	13.b.	Differentiate OLTP and OLAP.		
4	14.a.	List and discuss the basic features that are provided by reporting and query tools used for business analysis.	K4	CO4
		(OR)		
	14.b.	Explain K-Means partitioning algorithm in detail.		
5	15.a.	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.	K3	CO3
		(OR)		
	15.b.	What are the differences between the three main types of data warehouse usage: information processing, Analytical processing, data mining.		

Cont...

SECTION -C (30 Marks)

Answer **ANY THREE** questions

ALL questions carry **EQUAL** Marks (3 × 10 = 30)

<table>
<tr><th>Module No.</th><th>Question No.</th><th>Question</th><th>K Level</th><th>CO</th></tr>
<tr><td>1</td><td>16</td><td>Deduce Rule Based Classification with example IF-THEN Rules.</td><td>K3</td><td>CO3</td></tr>
<tr><td>2</td><td>17</td><td>With real time scenario, demonstrate the concept of Outlier analysis.</td><td>K4</td><td>CO4</td></tr>
<tr><td>3</td><td>18</td><td>Define data warehouse. Explain its features. Diagrammatically illustrate and discuss the data warehouses architecture.</td><td>K3</td><td>CO3</td></tr>
<tr><td>4</td><td>19</td><td>Discuss about mining association rules using the apriori algorithm in detail]
Find the frequent itemsets: the sets of items that have minimum support
A subset of a frequent itemset must also be a frequent itemset
i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset
Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
Use the frequent itemsets to generate association rules
The Apriori Algorithm : Pseudocode
Join Step: Ck is generated by joining Lk-1 with itself
Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
<table>
<tr><th>Trans ID</th><th>Item purchased</th></tr>
<tr><td>101</td><td>strawberry, litchi, oranges</td></tr>
<tr><td>102</td><td>strawberry, butterfruit</td></tr>
<tr><td>103</td><td>butterfruit, vanilla</td></tr>
<tr><td>104</td><td>strawberry, litchi, oranges</td></tr>
<tr><td>105</td><td>banana, oranges</td></tr>
<tr><td>106</td><td>Banana</td></tr>
<tr><td>107</td><td>banana, butterfruit</td></tr>
<tr><td>108</td><td>strawberry, litchi, apple, oranges</td></tr>
<tr><td>109</td><td>apple, vanilla</td></tr>
<tr><td>110</td><td>strawberry, litchi,</td></tr>
</table>
The set of item is {strawberry, litchi, oranges, butterfruit, vanilla, Banana, apple}. Use 0.3 for the minimum support value.</td><td>K4</td><td>CO4</td></tr>
<tr><td>5</td><td>20</td><td>What is grid based clustering? With an example explain an algorithm for grid based clustering.</td><td>K3</td><td>CO3</td></tr>
</table>

Z-Z-Z END