Sample Questions

Information Technology

Subject Name: Data Mining and Business Intelligence **Semester: VI**

Multiple Choice Questions

	Choose the correct option for following questions. All the Questions carry equal marks
1.	Which of the following can be considered as the correct process of Data Mining?
Option A:	Infrastructure, Exploration, Analysis, Interpretation, Exploitation
Option B:	Exploration, Infrastructure, Analysis, Interpretation, Exploitation
Option C:	Exploration, Infrastructure, Interpretation, Analysis, Exploitation
Option D:	Exploration, Infrastructure, Analysis, Exploitation, Interpretation
2.	Which of the following is an essential process in which the intelligent methods are applied to extract data patterns?
Option A:	Warehousing
Option B:	Data Mining
Option C:	Text Mining
Option D:	Data Selection
3.	What is KDD in data mining?
Option A:	Knowledge Discovery Database
Option B:	Knowledge Discovery Data
Option C:	Knowledge Data definition
Option D:	Knowledge data house
4.	What are the functions of Data Mining?
Option A:	Association and correctional analysis classification
Option B:	Prediction and characterization
Option C:	Cluster analysis and Evolution analysis
Option D:	All of the above
5.	Which one of the following statements about the K-means clustering is incorrect?
Option A:	The goal of the k-means clustering is to partition (n) observation into (k) clusters
Option B:	K-means clustering can be defined as the method of quantization
Option C:	The nearest neighbor is the same as the K-means
Option D:	All of the above
6.	Which one of the following can be defined as the data object which does not comply with the general behavior (or the model of available data)?
Option A:	Evaluation Analysis
Option B:	Outliner Analysis

Option C:	Classification
Option D:	Prediction
7.	Which one of the following correctly refers to the task of the classification?
Option A:	A measure of the accuracy, of the classification of a concept that is given by a certain theory
Option B:	The task of assigning a classification to a set of examples
Option C:	A subdivision of a set of examples into a number of classes
Option D:	None of the above
8.	Euclidean distance measure is can also defined as _____
Option A:	The process of finding a solution for a problem simply by enumerating all possible solutions according to some predefined order and then testing them
Option B:	The distance between two points as calculated using the Pythagoras theorem
Option C:	A stage of the KDD process in which new data is added to the existing selection
Option D:	All of the above
9.	Which of the following is a good alternative to the star schema?
Option A:	snow flake schema
Option B:	star schema
Option C:	star snow flake schema
Option D:	fact constellation
10.	Efficiency and scalability of data mining algorithms" issues come under?
Option A:	Mining Methodology and User Interaction Issues
Option B:	Performance Issues
Option C:	Diverse Data Types Issues
Option D:	None of the above
11.	_____ is the clustering technique which needs the merging approach.
Option A:	Naïve Bayes
Option B:	Hierarchical
Option C:	Partitioned
Option D:	All of the above
12.	_____ are the Data mining Application?
Option A:	Market Basket Analysis.
Option B:	Fraud Detection.
Option C:	Both A and B
Option D:	None of the above
13.	KDD process is consists of _____ steps.
Option A:	4
Option B:	9
Option C:	7
Option D:	5

14.	Which among the following is a Data Mining Algorithm?
Option A:	K-mean Algorithm
Option B:	Apriori Algorithm.
Option C:	Naive Bayes Algorithm
Option D:	All of the above
15.	Data mining requires
Option A:	Large quantities of operational data stored over a period of time
Option B:	Lots of tactical data
Option C:	Several tape drives to store archival data
Option D:	Large mainframe computers
16.	Which of the following is NOT example of ordinal attributes?
Option A:	Zip codes
Option B:	Ordered numbers
Option C:	Ascending or descending names
Option D:	Military ranks
17.	Identify the example of Nominal attribute
Option A:	Temperature
Option B:	Mass
Option C:	Salary
Option D:	Gender
18.	Which of the following is not a data pre-processing methods?
Option A:	Data Visualization
Option B:	Data Discretization
Option C:	Data Cleaning
Option D:	Data Reduction
19.	A data warehouse
Option A:	must import data from transactional systems whenever significant changes occur in the transactional data
Option B:	works on live transactional data to provide up to date and valid results
Option C:	takes regular copies of transaction data
Option D:	takes preprocessed transaction data and stores in a way that is optimized for analysis
20.	In a snowflake schema which of the following types of tables is considered?
Option A:	Fact
Option B:	Dimension
Option C:	Both (a) and (b)
Option D:	None of the above
21.	When you ____ the data, you are aggregating the data to a higher level
Option A:	Slice
Option B:	Roll Up
Option C:	Roll Down

Option D:	Drill Down
22.	Which type of data storage architecture gives fastest performance?
Option A:	ROLAP
Option B:	MOLAP
Option C:	HOLAP
Option D:	DOLAP
23.	_____supports basic OLAP operations, including slice and dice, drill-down, roll-up and pivoting.
Option A:	Information processing
Option B:	Analytical processing
Option C:	Data processing
Option D:	Transaction processing
24.	Data mining is_____?
Option A:	time variant non-volatile collection of data
Option B:	The actual discovery phase of a knowledge
Option C:	The stage of selecting the right data
Option D:	None of these
25.	Business intelligence (BI) is a broad category of application programs which includes _____
Option A:	Decision support
Option B:	Data mining
Option C:	OLAP
Option D:	All of the mentioned
26.	_____ is a performance management tool that recapitulates an organization's performance from several standpoints on a single page.
Option A:	Balanced Scorecard
Option B:	Data Cube
Option C:	Dashboard
Option D:	All of the mentioned
27.	Prediction is_____
Option A:	The result of the application of a theory or a rule in a specific case
Option B:	One of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table.
Option C:	Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.
Option D:	None of these
28.	Decision support systems (DSS) is _____
Option A:	A family of relational database management systems marketed by IBM
Option B:	Interactive systems that enable decision makers to use databases and models on a computer in order to solve ill-structured problems
Option C:	It consists of nodes and branches starting from a single root node. Each node represents a test, or decision

Option D:	None of these
29.	Association analysis is used to discover patterns that describe _____associated features in the data.
Option A:	largely
Option B:	fewer
Option C:	strongly
Option D:	moderately
30.	Binary attribute are _____
Option A:	This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit
Option B:	The natural environment of a certain species
Option C:	Systems that can be used without knowledge of internal operations
Option D:	None of these

Descriptive Questions

10 marks each

1. Explain role of Business intelligence in any one of following domain: Fraud Detection, Market Segmentation, retail industry, and telecommunications industry. Explain how data mining can be helpful in any of these cases.
2. Explain Star, Snowflake, and Fact Constellation Schema for Multidimensional Database
3. Explain Data warehouse architecture
4. What is clustering? Explain K-means clustering algorithm. Suppose the data for clustering- {2, 4, 10, 12, 3, 20, 11, 25} Consider k-2, cluster the given data using above algorithm.
5. Explain multilevel association & multidimensional association rules with example.
6. Define support, confidence. Also generate association rules. A database has four transitions. Let minimum support and confidence is 50%

D=

Tid	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

7. Define support, confidence. Also generate association rules. A database has four transitions. Let minimum support = 2 and confidence is 80%

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

8. Explain Business Intelligence and decision support system.
9. Short note on Outlier analysis and describe the methods that can be used for outliers.
10. Explain KDD process using figure.
11. Define outlier analysis? Why outlier mining is important? Briefly describe the different approaches: statistical-based outlier detection, distance-based outlier detection and deviation- based outlier detection.
12. What is noise? Explain data smoothing methods as noise removal technique to divide given data into bins of size 3 by bin partition (equal frequency), by bin means, by bin medians and by bin boundaries. Consider the data:10, 2, 19, 18, 20, 18, 25, 28, 22
13. State the Apriori Property. Generate candidate itemsets, frequent itemsets and association rules using Apriori algorithm on the following data set with minimum

support count is 2.

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

14. Consider a data warehouse for a hospital where there are three dimensions:

    a) Doctor b) Patient c) Time
    Consider two measures
    i) Count
    ii) Charge where charge is the fee that the doctor charges a patient for a visit.
    For the above example create a cube and illustrate the following OLAP operations.

1. Rollup 2) Drill down 3) Slice 4) Dice 5) Pivot.

15. Consider the following data points: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.

a) What is the mean of the data? What is the median?
b) What is the mode of data?
c) What is the midrange of the data?
d) Can you find Q1,Q3?
e) Show a boxplot of the data.

16. Explain different methods that can be used to evaluate and compare the accuracy of different classification algorithms?

17. Predict the class for X={age=youth, income=medium, student=yes, credit_rating=fair} using Naive Bayes Classification

Id	Age	Income	Student	Credit-rating	buys computer
1	Young	High	No	Fair	No
2	Young	High	No	Good	No
3	Middle	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Good	No
7	Middle	Low	Yes	Good	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Good	Yes
12	Middle	Medium	No	Good	Yes
13	Middle	High	Yes	Fair	Yes
14	Old	Medium	No	Good	No

18. Short note on DBSCAN clustering algorithm with example.

19. Consider the following data points:
    11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,72,73,75.
    (a) Find Mean, Median and Mode.
    (b) Show a box plot of the data. Clearly indicating the five-number summary.

20. Why is Data Preprocessing required? Explain the different steps involved in data preprocessing.

21. Illustrate any one classification technique for the above data set. Show how we can classify a new tuple. With (Homeowner=Yes; status=Employed; Income=Average).

Id	Homeowner	Status	Income	Defaulted
1	Yes	Employed	High	No
2	No	Business	Average	No
3	No	Employed	Low	No
4	Yes	Business	High	No
5	No	Unemployed	Average	Yes
6	No	Business	Low	No
7	Yes	Unemployed	High	No
8	No	Employed	Average	Yes

	9	No	Business	Low	No	
	10	No	Employed	Average	Yes	

5 marks each
1) Explain why data warehouses are needed for developing business solutions from today's perspective. Discuss the role of data marts.
2) Explain various features of Data Warehouse?
3) Discuss the application of data warehousing and data mining
4) A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data – Justify.
5) Give differences between OLAP and OLTP.
6) Explain various OLAP operations
7) Differentiate Fact table vs. Dimension table
8) Define the term "data mining". Discuss the major issues in data mining
9) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem
10) Explain the following data normalization techniques: (i) min-max normalization and (ii) decimal scaling.
11) Describe various methods for handling missing data values
12) What are the limitations of the Apriori approach for mining? Briefly describe the techniques to improve the efficiency of Apriori algorithm
13) What is market basket analysis? Explain the two measures of rule interestingness: *support* and *confidence* with suitable example.
14) Explain measures for finding rule interestingness (support, confidence) with example.
15) Compare association and classification. Briefly explain associative classification with suitable example.
16) What is an attribute selection measure? Explain different attribute selection measures with example.
17) Do feature wise comparison between classification and prediction.
18) Explain Linear regression with example.
19) Explain data mining application for fraud detection.
20) Discuss applications of data mining in Banking and Finance.
21) How K-Mean clustering method differs from K-Medoid clustering method?
22) How FP tree is better than Apriori algorithm- Justify
23) Define information gain, entropy, gini index