QUESTION BANK
BSC301 - Data Warehousing and Mining Question Bank

Part A – Multiple choice questions

1. __________ is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.

A. Data Mining
B. **Data Warehousing**
C. Web Mining
D. Text Mining

2. The data Warehouse is__________.

A. **Read only**
B. Write only
C. Read write only
D. None

3. Expansion for DSS in DW is__________.

A. **Decision Support system**
B. Decision Single System
C. Data Storable System
D. Data Support System

4. The important aspect of the data warehouse environment is that data found within the data warehouse is___________.

A. Subject-oriented
B. Time-variant
C. Integrated
D. **All of the above**

5. The time horizon in Data warehouse is usually __________.

A. 1-2 years
B. 3-4years
C. 5-6 years
D. **5-10 years**

6. The data is stored, retrieved & updated in ____________.

A. OLAP
B. **OLTP**
C. SMTP

D. FTP

7. __________describes the data contained in the data warehouse.
A. Relational data
B. Operational data
C. **Metadata**
D. Informational data

8. ____________predicts future trends & behaviors, allowing business managers to make proactive, knowledge-driven decisions.
A. Data warehouse.
B. **Data mining**
C. Datamart
D. Metadata

9. __________ is the heart of the warehouse.
A. Data mining database servers
B. **Data warehouse database servers**
C. Data mart database servers
D. Relational data base servers

10. ________________ is the specialized data warehouse database.
A. Oracle
B. DBZ
C. Informix
D. **Redbrick**

11. ________________defines the structure of the data held in operational databases and used by operational applications.
A. User-level metadata
B. Data warehouse metadata
C. **Operational metadata**
D. Data mining metadata

12. ________________ is held in the catalog of the warehouse database system.
A. Application level metadata
B. **Algorithmic level metadata**
C. Departmental level metadata
D. Core warehouse metadata

13. __________maps the core warehouse metadata to business concepts, familiar and useful to end users.

A. **Application level metadata**

B. User level metadata

C. Enduser level metadata

D. Core level metadata

14. ______consists of formal definitions, such as a COBOL layout or a database schema.

A. **Classical metadata**

B. Transformation metadata

C. Historical metadata

D. Structural metadata

15. _____________consists of information in the enterprise that is not in classical form.

A. Mushy metadata

B. Differential metadata

C. Data warehouse

D. Data mining

16. ______________databases are owned by particular departments or business groups.

A. Informational

B. **Operational**

C. Both informational and operational

D. Flat

17. The star schema is composed of __________ fact table.

A. **one**

B. two

C. three

D. four

18. The time horizon in operational environment is ___________.

A. 30-60 days

B. **60-90 days**

C. 90-120 days

D. 120-150 days

19. The key used in operational environment may not have an element of__________.

A. Time

B. Cost

C. Frequency
D. Quality

20. Data can be updated in _____environment.
A. Data warehouse
B. Data mining
C. **Operational**
D. Informational

21. Record cannot be updated in _____________.
A. OLTP
B. Files
C. RDBMS
D. **Data warehouse**

22. The source of all data warehouse data is the____________.
A. **Operational environment**
B. Informal environment
C. Formal environment
D. Technology environment

23. Data warehouse contains_____________data that is never found in the operational environment.
A. Normalized
B. Informational
C. **Summary**
D. Denormalized

24. The modern CASE tools belong to _______ category.
A. **Analysis**
B. Development
C. Coding
D. Delivery

25. Bill Inmon has estimated___________of the time required to build a data warehouse, is consumed in the conversion process.
A. 10 percent
B. 20 percent
C. 40 percent
D. **80 percent**

26. Detail data in single fact table is otherwise known as__________.
A. Monoatomic data
B. Diatomic data
C. **Atomic data**
D. Multiatomic data

27. _______test is used in an online transactional processing environment.
A. MEGA
B. MICRO
C. MACRO
D. **ACID**

28. ___________ is a good alternative to the star schema.
A. Star schema
B. Snowflake schema
C. **Fact constellation**
D. Star-snowflake schema

29. The biggest drawback of the level indicator in the classic star-schema is that it limits_________.
A. Quantify
B. Qualify
C. **Flexibility**
D. Ability

30. A data warehouse is _____________.
A. Updated by end users
B. Contains numerous naming conventions and formats
C. **Organized around important subject areas**
D. Contains only current data

31. An operational system is _____________.
A. Used to run the business in real time and is based on historical data
B. **Used to run the business in real time and is based on current data**
C. Used to support decision making and is based on current data
D. Used to support decision making and is based on historical data

32. The generic two-level data warehouse architecture includes __________.
A. At least one data mart

B. Data that can extracted from numerous internal and external sources
C. **Near real-time updates**
D. Far real-time updates

33. The active data warehouse architecture includes __________
A. At least one data mart
B. Data that can extracted from numerous internal and external sources
C. Near real-time updates
D. **All of the above**

34. Reconciled data is ___________.
A. Data stored in the various operational systems throughout the organization
B. **Current data intended to be the single source for all decision support systems**
C. Data stored in one operational system in the organization
D. Data that has been selected and formatted for end-user support applications

35. Transient data is _____________.
A. Data in which changes to existing records cause the previous version of the records to be eliminated
B. Data in which changes to existing records do not cause the previous version of the records to be eliminated
C. Data that are never altered or deleted once they have been added
D. Data that are never deleted once they have been added

36. The extract process is ______.
A. Capturing all of the data contained in various operational systems
B. **Capturing a subset of the data contained in various operational systems**
C. Capturing all of the data contained in various decision support systems
D. Capturing a subset of the data contained in various decision support systems

37. Data scrubbing is _____________.
A. A process to reject data from the data warehouse and to create the necessary indexes
B. A process to load the data in the data warehouse and to create the necessary indexes
C. A process to upgrade the quality of data after it is moved into a data warehouse
D. A process to upgrade the quality of data before it is moved into a data warehouse

38. The load and index is ______________.
A. A process to reject data from the data warehouse and to create the necessary indexes
B. A process to load the data in the data warehouse and to create the necessary indexes
C. A process to upgrade the quality of data after it is moved into a data warehouse

D. A process to upgrade the quality of data before it is moved into a data warehouse

39. Data transformation includes __________.
A. **A process to change data from a detailed level to a summary level**
B. A process to change data from a summary level to a detailed level
C. Joining data from one source into various sources of data
D. Separating data from one source into various sources of data

40. ____________ is called a multifield transformation.
A. **Converting data from one field into multiple fields**
B. Converting data from fields into field
C. Converting data from double fields into multiple fields
D. Converting data from one field to one field

41. The type of relationship in star schema is __________________.
A. Many-to-many
B. One-to-one
C. **One-to-many**
D. Many-to-one

42. Fact tables are ___________.
A. Completely demoralized
B. Partially demoralized
C. **Completely normalized**
D. Partially normalized

43. _______________ is the goal of data mining.
A. **To explain some observed event or condition**
B. To confirm that data exists.
C. To analyze data for expected relationships.
D. To create a new data warehouse.

44. Business Intelligence and data warehousing is used for ________.
A. Forecasting
B. Data Mining
C. Analysis of large volumes of product sales data
D. All of the above

45. The data administration subsystem helps you perform all of the following, except__________.

A. Backups and recovery

B. Query optimization

C. Security management

D. Create, change, and delete information

46. The most common source of change data in refreshing a data warehouse is _______.

A. Queryable change data

B. Cooperative change data

C. Logged change data

D. Snapshot change data

47. ________ are responsible for running queries and reports against data warehouse tables.

A. Hardware

B. Software

C. End users

D. Middle ware

48. Query tool is meant for __________.

A. Data acquisition

B. Information delivery

C. Information exchange

D. Communication

49. Classification rules are extracted from _____________.

A. Root node

B. Decision tree

C. Siblings

D. Branches

50. Dimensionality reduction reduces the data set size by removing ____________.

A. Relevant attributes

B. Irrelevant attributes

C. Derived attributes

D. Composite attributes

51. ___________ is a method of incremental conceptual clustering.

A. CORBA

B. OLAP

C. COBWEB
D. STING

52. Effect of one attribute value on a given class is independent of values of other attribute is called _________.
A. **Value independence**
B. Class conditional independence
C. Conditional independence
D. Unconditional independence

53. The main organizational justification for implementing a data warehouse is to provide ______.
A. Cheaper ways of handling transportation
B. Decision support
C. **Storing large volume of data**
D. Access to data

54. Multidimensional database is otherwise known as____________.
A. RDBMS
B. DBMS
C. EXTENDED RDBMS
D. EXTENDED DBMS

55. Data warehouse architecture is based on ______________.
A. DBMS
B. **RDBMS**
C. Sybase
D. SQL Server

56. Source data from the warehouse comes from _______________.
A. **ODS**
B. TDS
C. MDDB
D. ORDBMS

57. _______________ is a data transformation process.
A. Comparison
B. Projection
C. Selection
D. **Filtering**

58. The technology area associated with CRM is _______________.
A. Specialization
B. Generalization
C. **Personalization**
D. Summarization

59. SMP stands for _______________.
A. Symmetric Multiprocessor
B. Symmetric Multiprogramming
C. Symmetric Metaprogramming
D. Symmetric Microprogramming

60. __________ are designed to overcome any limitations placed on the warehouse by the nature of the relational data model.
A. Operational database
B. Relational database
C. Multidimensional database
D. Data repository

61. __________ are designed to overcome any limitations placed on the warehouse by the nature of the relational data model.
A. Operational database
B. Relational database
C. Multidimensional database
D. Data repository

62. MDDB stands for ___________.
A. Multiple Data Doubling
B. Multidimensional Databases
C. Multiple Double Dimension
D. Multi-dimension Doubling

63. ______________ is data about data.
A. Metadata
B. Microdata
C. Minidata
D. Multidata

64. ___________ is an important functional component of the metadata.
A. Digital directory
B. Repository
C. Information directory
D. Data dictionary

65. EIS stands for ______________.
A. Extended Interface System
B. Executive Interface System
C. Executive Information System
D. Extendable Information System

66. ___________ is data collected from natural systems.
A. MRI scan
B. ODS data
C. Statistical data
D. Historical data

67. _______________ is an example of application development environments.
A. Visual Basic
B. Oracle
C. Sybase
D. SQL Server

68. The term that is not associated with data cleaning process is ______.
A. Domain consistency
B. Deduplication
C. Disambiguation
D. Segmentation

69. ____________ are some popular OLAP tools.
A. **Metacube, Informix**
B. Oracle Express, Essbase
C. HOLAP
D. MOLAP

70. Capability of data mining is to build ___________ models.
A. Retrospective
B. Interrogative
C. Predictive

D. Imperative

71. _____________ is a process of determining the preference of customer's majority.
A. Association
B. Preferencing
C. Segmentation
D. Classification

72. Strategic value of data mining is ______________.
A. Cost-sensitive
B. Work-sensitive
C. Time-sensitive
D. Technical-sensitive

73. ____________ proposed the approach for data integration issues.
A. Ralph Campbell
B. Ralph Kimball
C. John Raphlin
D. James Gosling

74. The terms equality and roll up are associated with ____________.
A. OLAP
B. Visualization
C. Data mart
D. Decision tree

75. Exceptional reporting in data warehousing is otherwise called as __________.
A. Exception
B. Alerts
C. Errors
D. Bugs

76. ____________ is a metadata repository.
A. Prism solution directory manager
B. CORBA
C. STUNT
D. COBWEB

77. ________________ is an expensive process in building an expert system.
A. Analysis

B. Study
C. Design
D. Information collection

78. The full form of KDD is _________.
A. Knowledge database
B. Knowledge discovery in database
C. Knowledge data house
D. Knowledge data definition

79. The first International conference on KDD was held in the year _____________.
A. 1996
B. 1997
C. 1995
D. 1994

80. Removing duplicate records is a process called _____________.
A. Recovery
B. Data cleaning
C. Data cleansing
D. Data pruning

81. ____________ contains information that gives users an easy-to-understand perspective of the information stored in the data warehouse.
A. Business metadata
B. Technical metadata
C. Operational metadata
D. Financial metadata

82. _______________ helps to integrate, maintain and view the contents of the data warehousing system.
A. Business directory
B. Information directory
C. Data dictionary
D. Database

83. Discovery of cross-sales opportunities is called ________________.
A. Segmentation
B. Visualization
C. Correction

D. Association

84. Data marts that incorporate data mining tools to extract sets of data are called ______.
A. Independent data mart
B. **Dependent data marts**
C. Intra-entry data mart
D. Inter-entry data mart

85. ____________ can generate programs itself, enabling it to carry out new tasks.
A. Automated system
B. Decision making system
C. Self-learning system
D. Productivity system

86. The power of self-learning system lies in __________.
A. Cost
B. Speed
C. Accuracy
D. Simplicity

87. Building the informational database is done with the help of _______.
A. Transformation or propagation tools
B. Transformation tools only
C. Propagation tools only
D. Extraction tools

88. How many components are there in a data warehouse?
A. Two
B. Three
C. Four
D. **Five**

89. Which of the following is not a component of a data warehouse?
A. Metadata.
B. Current detail data
C. Lightly summarized data
D. **Component Key**

90. __________ is data that is distilled from the low level of detail found at the current detailed level.
A. Highly summarized data
B. Lightly summarized data
C. Metadata
D. Older detail data

91. Highly summarized data is _______.
A. Compact and easily accessible
B. Compact and expensive
C. Compact and hardly accessible
D. Compact

92. A directory to help the DSS analyst locate the contents of the data warehouse is seen in _________.
A. Current detail data
B. Lightly summarized data
C. Metadata
D. Older detail data

93. Metadata contains atleast _________.
A. The structure of the data
B. The algorithms used for summarization
C. The mapping from the operational environment to the data warehouse
D. All of the above

94. Which of the following is not a old detail storage medium?
A. Phot optical storage
B. Raid
C. Microfinche
D. Pen drive

95. The data from the operational environment enter _______ of data warehouse.
A. Current detail data
B. Older detail data
C. Lightly summarized data
D. Highly summarized data

96. The data in current detail level resides till ________ event occurs.

A. Purge
B. Summarization
C. Archieved
D. All of the above

97. The dimension tables describe the _________.

A. Entities
B. Facts
C. Keys
D. Units of measures

98. The granularity of the fact is the _____ of detail at which it is recorded.

A. Transformation
B. Summarization
C. **Level**
D. Transformation and summarization

99. Which of the following is not a primary grain in analytical modeling?

A. Transaction
B. Periodic snapshot
C. Accumulating snapshot
D. All of the above

100. Granularity is determined by ______.

A. Number of parts to a key
B. Granularity of those parts
C. **Both A and B**
D. None of the above

101. ___________ of data means that the attributes within a given entity are fully dependent on the entire primary key of the entity.

A. Additivity
B. Granularity
C. Functional dependency
D. Dimensionality

102. A fact is said to be fully additive if ___________.

A. It is additive over every dimension of its dimensionality
B. Additive over atleast one but not all of the dimensions
C. Not additive over any dimension

D. None of the above

103. A fact is said to be partially additive if ___________.

A. It is additive over every dimension of its dimensionality

B. Additive over atleast one but not all of the dimensions

C. Not additive over any dimension

D. None of the above

104. A fact is said to be non-additive if ___________.

A. It is additive over every dimension of its dimensionality

B. Additive over atleast one but not all of the dimensions

C. Not additive over any dimension

D. None of the above

105. Non-additive measures can often combined with additive measures to create new _________.

A. Additive measures

B. Non-additive measures

C. Partially additive

D. All of the above

106. A fact representing cumulative sales units over a day at a store for a product is a _________.

A. Additive fact

B. Fully additive fact

C. Partially additive fact

D. Non-additive fact

107. ___________ of data means that the attributes within a given entity are fully dependent on the entire primary key of the entity.

A. Additivity

B. Granularity

C. Functional Dependency

D. Dependency

108. Which of the following is the other name of Data mining?

A. Exploratory data analysis

B. Data driven discovery

C. Deductive learning

D. All of the above

109. Which of the following is a predictive model?

A. Clustering
B. Regression
C. Summarization
D. Association rules

110. Which of the following is a descriptive model?

A. Classification
B. Regression
C. Sequence discovery
D. Association rules

111. A ___________ model identifies patterns or relationships.

A. **Descriptive**
B. Predictive
C. Regression
D. Time series analysis

112. A predictive model makes use of ________.

A. Current data
B. Historical data
C. Both A and B
D. Assumptions

113. ____________ maps data into predefined groups.

A. Regression
B. Time series analysis
C. Prediction
D. **Classification**

114. __________ is used to map a data item to a real valued prediction variable.

A. Regression
B. Time series analysis
C. Prediction
D. Classification

115. In ____________, the value of an attribute is examined as it varies over time.

A. Regression
B. Time series analysis

C. Sequence discovery
D. Prediction

116. In ________ the groups are not predefined.
A. Association rules
B. Summarization
C. Clustering
D. Prediction

117. Link Analysis is otherwise called as ___________.
A. Affinity analysis
B. Association rules
C. Both A and B
D. Prediction

118. _________ is a the input to KDD.
A. **Data**
B. Information
C. Query
D. Process

119. The output of KDD is __________.
A. Data
B. Information
C. Query
D. Useful information

120. The KDD process consists of ________ steps.
A. Three
B. Four
C. **Five**
D. Six

121. Treating incorrect or missing data is called as ___________.
A. Selection
B. **Preprocessing**
C. Transformation
D. Interpretation

122. Converting data from different sources into a common format for processing is called as ________.

A. Selection

B. Preprocessing

C. Transformation

D. Interpretation

123. Various visualization techniques are used in ___________ step of KDD.

A. Selection

B. Transformaion

C. Data mining

D. Interpretation

124. Extreme values that occur infrequently are called as _________.

A. **Outliers**

B. Rare values

C. Dimensionality reduction

D. All of the above

125. Box plot and scatter diagram techniques are _______.

A. Graphical

B. **Geometric**

C. Icon-based

D. Pixel-based

126. __________ is used to proceed from very specific knowledge to more general information.

A. Induction

B. Compression

C. Approximation

D. Substitution

127. Describing some characteristics of a set of data by a general model is viewed as ____________

A. Induction

B. Compression

C. Approximation

D. Summarization

128. ____________ helps to uncover hidden information about the data.

A. Induction
B. Compression
C. Approximation
D. Summarization

129. _______ are needed to identify training data and desired results.

A. Programmers
B. Designers
C. **Users**
D. Administrators

130. Overfitting occurs when a model _________.

A. Does fit in future states
B. Does not fit in future states
C. Does fit in current state
D. Does not fit in current state

131. The problem of dimensionality curse involves ___________.

A. The use of some attributes may interfere with the correct completion of a data mining task
B. The use of some attributes may simply increase the overall complexity
C. Some may decrease the efficiency of the algorithm
D. All of the above

132. Incorrect or invalid data is known as _________.

A. Changing data
B. Noisy data
C. Outliers
D. Missing data

133. ROI is an acronym of ________.

A. Return on Investment
B. Return on Information
C. Repetition of Information
D. Runtime of Instruction

134. The ____________ of data could result in the disclosure of information that is deemed to be confidential.

A. Authorized use
B. Unauthorized use
C. Authenticated use
D. Unauthenticated use

135. ___________ data are noisy and have many missing attribute values.

A. Preprocessed
B. Cleaned
C. Real-world
D. Transformed

136. The rise of DBMS occurred in early ___________.

A. 1950's
B. 1960's
C. 1970's
D. 1980's.

137. Which of the following is not a data mining metric?

A. Space complexity
B. Time complexity
C. ROI
D. All of the above

138. Reducing the number of attributes to solve the high dimensionality problem is called as ________.

A. Dimensionality curse
B. Dimensionality reduction
C. Cleaning
D. Overfitting

139. Data that are not of interest to the data mining task is called as ______.

A. Missing data
B. Changing data
C. Irrelevant data
D. Noisy data

140. __________ are effective tools to attack the scalability problem.

A. Sampling.
B. Parallelization
C. Both A and B
D. None of the above

141. Market-basket problem was formulated by __________.

A. Agrawal et al.
B. Steve et al.
C. Toda et al.
D. Simon et al.

142. Data mining helps in __________.

A. Inventory management
B. Sales promotion strategies
C. Marketing strategies
D. All of the above

143. The proportion of transaction supporting X in T is called _________.

A. Confidence
B. Support
C. Support count
D. All of the above

144. The absolute number of transactions supporting X in T is called ___________.

A. Confidence
B. Support
C. Support count
D. None of the above

145. The value that says that transactions in D that support X also support Y is called ______________.

A. Confidence
B. Support
C. Support count
D. None of the above

146. If T consist of 500000 transactions, 20000 transaction contain bread, 30000 transaction contain jam, 10000 transaction contain both bread and jam. Then the support of bread and jam is _______.

A. 2%

B. 20%

C. 3%

D. 30%

147. If T consist of 500000 transactions, 20000 transaction contain bread, 30000 transaction contain jam, 10000 transaction contain both bread and jam. Then the confidence of buying bread with jam is _______.

A. 33.33%

B. 66.66%

C. 45%

D. 50%

148. The left hand side of an association rule is called __________.

A. Consequent

B. Onset

C. Antecedent

D. Precedent

149. The right hand side of an association rule is called _____.

A. Consequent

B. Onset

C. Antecedent

D. Precedent

150. Which of the following is not a desirable feature of any efficient algorithm?

A. To reduce number of input operations

B. To reduce number of output operations

C. To be efficient in computing

D. To have maximal code length

151. All set of items whose support is greater than the user-specified minimum support are called as_____________.

A. Border set

B. Frequent set

C. Maximal frequent set

D. Lattice

152. If a set is a frequent set and no superset of this set is a frequent set, then it is called ________.

A. Maximal frequent set
B. Border set
C. Lattice
D. Infrequent sets

153. Any subset of a frequent set is a frequent set. This is ___________.

A. Upward closure property
B. Downward closure property
C. Maximal frequent set
D. Border set

154. A priori algorithm is otherwise called as __________.

A. Width-wise algorithm
B. Level-wise algorithm
C. Pincer-search algorithm
D. FP growth algorithm

155. The A Priori algorithm is a ___________.

A. Top-down search
B. Breadth first search
C. Depth first search
D. Bottom-up search

156. The first phase of A Priori algorithm is _______.

A. Candidate generation
B. Itemset generation
C. Pruning
D. Partitioning

157. The second phaase of A Priori algorithm is ____________.

A. Candidate generation
B. Itemset generation
C. Pruning
D. Partitioning

158. The _______ step eliminates the extensions of (k-1)-itemsets which are not found to be frequent, from being considered for counting support.

A. Candidate generation
B. **Pruning**
C. Partitioning
D. Itemset eliminations

159. The a priori frequent itemset discovery algorithm moves _______ in the lattice.

A. **Upward**
B. Downward
C. Breadthwise
D. Both upward and downward

160. After the pruning of a priori algorithm, _______ will remain.

A. Only candidate set
B. No candidate set
C. Only border set
D. No border set

161. The number of iterations in a priori ___________.

A. Increases with the size of the maximum frequent set
B. Decreases with increase in size of the maximum frequent set
C. Increases with the size of the data
D. Decreases with the increase in size of the data

162. Itemsets in the ______ category of structures have a counter and the stop number with them.

A. **Dashed**
B. Circle
C. Box
D. Solid

163. The goal of _____ is to discover both the dense and sparse regions of a data set.

A. Association rule
B. Classification
C. **Clustering**
D. Genetic Algorithm

164. Which of the following is a clustering algorithm?

A. A priori

B. **CLARA**

C. Pincer-Search

D. FP-growth

165. _________ clustering technique start with as many clusters as there are records, with each cluster having only one record.

A. **Agglomerative**

B. divisive

C. Partition

D. Numeric

166. ___________ clustering techniques starts with all records in one cluster and then try to split that cluster into small pieces.

A. Agglomerative

B. **Divisive**

C. Partition

D. Numeric

167. Which of the following is a data set in the popular UCI machine-learning repository?

A. CLARA

B. CACTUS

C. STIRR

D. MUSHROOM

168. In ________ algorithm each cluster is represented by the center of gravity of the cluster.

A. K-medoid

B. K-means

C. STIRR

D. ROCK

169. In ___________ each cluster is represented by one of the objects of the cluster located near the center.

A. K-medoid

B. K-means

C. STIRR

D. ROCK

170. Pick out a K-medoid algoithm.
A. DBSCAN
B. BIRCH
C. **PAM**
D. CURE

171. Pick out a hierarchical clustering algorithm.
A. DBSCAN
B. BIRCH.
C. PAM.
D. CURE.

172. CLARANS stands for _______.
A. Clara net server
B. Clustering large application range network search
C. Clustering large applications based on randomized search
D. Clustering Application Randomized Search

173. The cluster features of different subclusters are maintained in a tree called ___________.
A. **CF tree**
B. FP tree
C. FP growth tree
D. B tree

174. The ________ algorithm is based on the observation that the frequent sets are normally very few in number compared to the set of all itemsets.
A. A priori
B. Clustering
C. Association rule
D. Partition

175. The partition algorithm uses _______ scans of the databases to discover all frequent sets.
A. Two
B. Four
C. Six
D. Eight

176. The basic idea of the apriori algorithm is to generate_________ item sets of a particular size & scans the database.

A. **Candidate**
B. Primary
C. Secondary
D. Superkey

177. __________data consists of sample input data as well as the classification assignment for the data.

A. Missing
B. Measuring
C. Non-training
D. Training

178. Rule based classification algorithms generate ______ rule to perform the classification.

A. If-then
B. While
C. Do while
D. Switch

179. _________ displays of data such as maps, charts and other graphical representation allow data to be presented compactly to the users.

A. Hidden
B. Visual
C. Obscured
D. Concealed

Part B – Descriptive Questions

1. Define Data mining? Explain about data mining on what kind of data? 10 M

2. a) What is KDD? Explain about data mining as a step in the process of knowledge discovery. 6 M

b) How to classify data mining systems? Discuss 4 M

3. Discuss about the following

a) What motivated Data mining? Explain 5 M

b) Data mining as a step in the process of knowledge discovery. 5 M

4. Discuss about Data Mining Task primitives with examples? 10 M

5. Explain in detail about Data mining functionalities? 10 M

6. Write a note on statistical description of data. 10 M

7. Describe about Major issues in Data mining. 10 M

8. Define the following,

a) Data cleaning 2M
b) Data integration 2 M
c) Data reduction 2 M
d) Data transformation 2 M
e) Data discretization 2 M

9. a) Why do we preprocess the data? Discuss? 5 M

b) Write in brief about Data cleaning? 5 M

10. Explain the following?

a) Data Integration 5 M

b) Data Transformation methods 5 M

11. What is Data reduction? Discuss in detail? 10 M

12. a) Describe about Data discretization? 5 M

b) Write about Dimensionality reduction methods? 5 M

11. a) Define Data warehouse? Discuss Design principles. 5 M

b) Write in brief about schemas in multidimensional data model. 5 M

12. Explain about the Three-tier data warehouse architecture with a neat diagram. 10 M

13. Discuss the following

a) Star schema 3 M

b) Snow Flake schema 3 M

c) Fact constellation schema 4 M

14. a) What are steps in designing the data warehouse? Explain. 5 M

b) Compare OLTP and OLAP. 5 M

15. Describe in brief about Data warehouse implementation. 10 M

16. Explain the following in OLAP

a) Roll up operation 2 M

b) Drill down operation 2 M

c) Slice operation 2 M

d) Dice operation 2 M

e) Pivot operation 2 M

17. Explain about the Apriori algorithm for finding frequent item sets with an example. 10 M

18. You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order: 1 – order: 9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity we assign the meal items short names (M1 – M5) rather than the full descriptive names (e.g., Big Mac). 10 M

Meal Item	**Item IDs**	**Meal Item**	**Item IDs**
Order: 1	M1, M2, M5	Order: 6	M2, M3
Order: 2	M2, M4	Order: 7	M1, M3
Order: 3	M2, M3	Order: 8	M1, M2, M3, M5
Order: 4	M1, M2, M4	Order: 9	M1, M2, M3
Order: 5	M1, M3		

For all of the parts below the minimum support is 2/9 (.222) and the minimum confidence is 7/9 (.777). Note that you only need to achieve this level, not exceed it. Show your work for full credit (this mainly applies to part a).

a. Apply the Apriori algorithm to the dataset of transactions and identify all frequent k itemset.

b. Find all strong association rules of the form: X ^ Y ^ Z and note their confidence values. Hint: the answer is not 0 so you should have at least one frequent 3-frequent itemset.

19. a) What are the drawbacks of Apriori Algorithm? Explain. 5 M

b) Write the FP Growth Algorithm. 5 M

20. Discuss about the pattern evaluation methods. 10 M

21. Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, explain with an example 10 M

22. What are the various Constraints in Constraint based Association rule mining? Explain. 10 M

23. List and explain the steps involved in decision tree classification algorithm 10 M

24. Go through the given data and do the following,

Outlook	**Temperature**	**Humidity**	**Wind**	**Play**
Sunny	79	High	Weak	No
Sunny	56	High	Strong	No
Overcast	79	High	Weak	Yes
Rain	60	High	Weak	No
Rain	88	Normal	Weak	Yes
Rain	64	Normal	Strong	No
Overcast	88	Normal	Strong	Yes
Sunny	78	High	Weak	No
Sunny	66	Normal	Weak	Yes
Rain	68	Normal	Weak	Yes

a) Construct the rule 5 M

b) Draw the decision tree pattern based on that rule 5 M

25. What are splitting indices? Explain different splitting indices. 10 M

26. Describe the data classification process with a neat diagram. How does the Naive Bayesian classification works? Explain. 10 M

27. a) What is Bayes theorem? Explain. 5 M

b) Discuss about Naïve Bayesian Classification. 5 M

28. 14 days of information is given in the table. With respect to that information calculate the following,

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	No
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Sunny	Cool	High	Strong	No
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

5 M
3 M
2 M

f) Calculate the probability and conditional probability.

g) Calculate the play status for the condition,
Outlook = Overcast
Temperature = Mild
Humidity = Normal
Wind = Weak

h) Calculate the normalized probability values for the answers found in question b).

29. Describe in detail about Rule based Classification. 10 M

30. Write a note on model selection and evaluation with an example. 10 M

31. a) What is prediction? Explain about Linear regression method. 5 M
b) Discuss about Accuracy and Error measures. 5 M

32. Define Clustering? Explain about Types of Data in Cluster Analysis? 10 M

33. a) What is outlier detection? Explain distance based outlier detection 5 M
b) Write partitioning around mediods algorithm. 5 M

34. a) Write a note on K-means clustering algorithm. 5 M
b) Write the key issue in hierarchical clustering algorithm. 5 M

35. What are outliers? Discuss the methods adopted for outlier detection 10 M

36. Discuss in detail about Data mining Applications. 10 M