1Data selection is:
A. The actual discovery phase of a knowledge discovery process
B. The stage of selecting the right data for a KDD process
C. A subject-oriented integrated time variant non-volatile collection of data in support of management
D. None of these
Answer: B

2Discovery is:
A. It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
B. The process of executing implicit previously unknown and potentially useful information from data
C. An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
D. None of these
Answer: B

3Data mining is:
A. The actual discovery phase of a knowledge discovery process
B. The stage of selecting the right data for a KDD process
C. A subject-oriented integrated time variant non-volatile collection of data in support of management
D. None of these
Answer: A

4Knowledge engineering is:
A. The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system
B. It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks.
C. A process where an individual learns how to carry out a certain task when making a transition from a situation in which the task cannot be carried out to a situation in which the same task under the same circumstances can be carried out.
D. None of these
Answer: A

5KDD (Knowledge Discovery in Databases) is referred to:
A. Non-trivial extraction of implicit previously unknown and potentially useful information from data
B. Set of columns in a database table that can be used to identify each record within this table uniquely.
C. collection of interesting and useful patterns in a database
D. none of these

6Knowledge is referred to:
A. Non-trivial extraction of implicit previously unknown and potentially useful information from data
B. Set of columns in a database table that can be used to identify each record within this table uniquely
C. collection of interesting and useful patterns in a database
D. none of these
Answer: C

7Operational database is:
A. A measure of the desired maximal complexity of data mining algorithms

B. A database containing volatile data used for the daily operation of an organization
C. Relational database management system
D. None of these
Answer: B

8Which of the following is not a data mining functionality?
A. Characterization and Discrimination
B. Classification and regression
C. Selection and interpretation
D. Clustering and Analysis
Answer: C

9The various aspects of data mining methodologies is/are .....
i. Mining various and new kinds of knowledge
ii. Mining knowledge in multidimensional space
iii. Pattern evaluation and pattern or constraint-guided mining.
iv) Handling uncertainty, noise, or incompleteness of data

10The full form of KDD is .....
A. Knowledge Database
B. Knowledge Discovery Database
C. Knowledge Data House
D. Knowledge Data Definition
Answer: B

11The output of KDD is .....
A. Data
B. Information
C. Query
D. Useful information/Knowledge
Answer: D

12The process of removing the deficiencies and loopholes in the data is called as
A. Aggregation of data
B. Extracting of data
C. Cleaning up of data.
D. Loading of data
Answer: C

13Which of the following process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evolution and knowledge presentation?
A. KDD process
B. ETL process
C. KTL process
D. MDX process
Answer: A

14Data mining application domains are

A. Biomedical
B. DNA data analysis
C. Financial data analysis
D. Retail industry and telecommunication industry
E. All (a), (b), (c) and (d) above.
Answer: E

15Which of the following is/are the Data mining tasks?
A. Regression
B. Classification
C. Clustering
D. inference of associative rules
E. All (a), (b), (c) and (d) above.
Answer: E

16Which of the following is not an ETL tool?
A. Informatica
B. Oracle warehouse builder
C. Datastage
D. Visual studio
Answer: D

17_____ is not a data mining functionality?
A. Clustering and Analysis
B. Selection and interpretation
C. Classification and regression
D. Characterization and Discrimination
ANSWER: B

18To remove noise and inconsistent data ____ is needed.

(A)
Data Cleaning

(B)
Data Transformation

(C)
Data Reduction

(D)
Data Integration
Answer:A

19Multiple data sources may be combined is called as _____

(A)
Data Reduction

(B)

Data Cleaning

(C)
Data Integration

(D)
Data Transformation
Answer:C

20What is the use of data cleaning?

A. to remove the noisy data
B. correct the inconsistencies in data
C. transformations to correct the wrong data.
D. All of the above
Answer:D

21Data set {brown, black, blue, green , red} is example of Select one:
A. Continuous attribute
B. Ordinal attribute
C. Numeric attribute
D. Nominal attribute

Answer:D

22Binary attribute are

A.
This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit

B.
The natural environment of a certain species

C.
Systems that can be used without knowledge of internal operations

D.
None of these
Answer:A

23Euclidean distance measure is

A.
A stage of the KDD process in which new data is added to the existing selection.

B.
The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them

C.

The distance between two points as calculated using the Pythagoras theorem

D.None of These

24If there is a very strong correlation between two variables then the correlation coefficient must be
a. any value larger than 1
b. much smaller than 0, if the correlation is negative
c. much larger than 0, regardless of whether the correlation is negative or positive
d. None of these alternatives is correct.

Answer:B

Which of the following is a good alternative to the star schema?
A. Snowflake schema
B. Star schema
C. Star snowflake schema
D. Fact constellation
ANSWER: D

Patterns that can be discovered from a given database are which type
A. More than one type
B. Multiple types always
C. One type only
D. No specific type
ANSWER: A

A star schema has what type of relationship between a dimension and fact table?
A. Many-to-many
B. One-to-one
C. One-to-many
D. All of the above.
ANSWER: C

A snowflake schema is which of the following types of tables?
A. Fact
B. Dimension
C. Helper
D. All of the above
ANSWER: D

Euclidean distance measure is
A. A stage of the KDD process in which new data is added to the existing selection.
B. The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
C. The distance between two points as calculated using the Pythagoras theorem
D. None of these
ANSWER: C

Which one manages both current and historic transactions?
A. OLTP
B. OLAP
C. Spread sheet
D. XML
Answer: B

The data Warehouse is_____.
A. ReadOnly
B. WriteOnly
C. Read and write only
D. None of these
ANSWER: A

Expansion for DSS in DW is_____.
A. Decision Support system
B. Decision Single System
C. Data Storable System
D. Data support system
ANSWER: A

The time horizon in Data warehouse is usually _____.
A. 1-2 years
B. 3-4 years
C. 5-6 years
D. 5-10 years
ANSWER: D

_____describes the data contained in the data warehouse
A. Relational data
B. Operational Data
C. Meta Data
D. Informational Data
ANSWER: C

Treating incorrect or missing data is called as _____.
A. Selection.
B. Preprocessing
C. Transformation
D. Interpretation
ANSWER: B

Converting data from different sources into a common format for processing is called as_____.
A. Selection.
B. Preprocessing
C. Transformation
D. Interpretation
ANSWER: C

Which is not a property of data warehouse?
A. Subject oriented
B. Time varient
C. Volatile
D. collection from heterogeneous sources
ANSWER: C

Data warehousing is used in_____
A. Transaction System
B. Database management system
C. Decision support system
D. Expert system
ANSWER: C

What are the characeristics of OLAP systems?
A. Query driven
B. More users
C. Integrated

D. Store current data
ANSWER: C

Data warehouse is based on_____
A. two dimensional model
B. three dimensional model
C. Multi dimensional model
D. Unidimensional model
ANSWER: C

Data warehousing is related to_____
A. delete data
B. Update data
C. Write new data
D. scan and load data for analysis
ANSWER: D

Multidimensional model of data warehouse called as_____
A. data structure
B. table
C. tree
D. data cube
ANSWER: D

OLAP usage is____
A. Repetative
B. Adhoc
C. Frequently
D. Daily
ANSWER: B

In data warehousing what is time-variant data?
A. Data in the warehouse is only accurate and valid at some point in time or over time interval
B. Data in the warehouse is always accurate and valid
C. Data in the warehouse is not accurate
D. Data in the warehouse is only accurate sometimes
ANSWER: A

Is the data in a data warehouse generally updated in real-time?
A. YES
B. NO
ANSWER: B

What is a Star Schema?
A. A star schema consists of a fact table with a single table for each dimension
B. A star schema is a type of database system
C. A star schema is used when exporting data from the database
D. None of these
ANSWER: A

What is a Snowflake Schema?
A. Each dimension table is normalized, which may create additional

tables attached to the dimension tables
B. A Snowflake schema is a type of database system
C. A Snowflake schema is used when exporting data from the database
D. None of these
ANSWER: A

What does the acronym ETL stands for?
A. Explain,Transfer and Lead
B. Extract,Transform and Load
C. Extract,Transfer and Load
D. Effect,Transfer and Load
ANSWER: B

What is the system of data warehousing mostly used for?
A. Data integration and Data Mining
B. Data Mining and Data Storage
C. Reporting and Data Analysis
D. Data Cleaning and Data Storage
ANSWER: C

Which small logical units do data warehouses hold large amounts of information?
A. Data Storage
B. Data Marts
C. Access layers
D. Data Miners
ANSWER: B

Why do we need ODS?
A. To update data periodically
B. To prepare data for ETL
C. To back up data
D. To prepare data for regression
ANSWER: B

Which one is correct for data warehousing?
A. It can be updated by end users
B. It can solve all business questions
C. It is designed for focus subject areas
D. It contains only current data
ANSWER: C

Why do we apply in snowflake schema?
A. Aggregation
B. Normalization
C. Specialization
D. Generalization
ANSWER: B

The data collected in data warehouse can be used for analyzing purposes.
A. TRUE
B. FALSE
ANSWER: A

A snowflake schema is a normalized star schema
A. TRUE
B. FALSE
ANSWER: A

A fact table is related to dimensional table as a ___ relationship
A. 1:M
B. M:N
C. M:1
D. 1:1
ANSWER: C

Data warehouse contains_____data that is never found in the operational environment
A. normalized.
B. Informational
C. Summary
D. Denormalized
ANSWER: C

Identify correct type of attribute.
A. nominal
B. binary
C. ordinal
D. All of these
ANSWER: D

Minkowski distance is a function used to find the distance between two
A. Binary vectors
B. Boolean-valued vectors
C. Real-valued vectors
D. Categorical vectors
ANSWER: C

Which distance measure is similar to Simple Matching Coefficient (SMC)?
A. Euclidean
B. Hamming
C. Jaccard
D. Manhattan
ANSWER: B

Data set of designation {Professor, Assistant Professor, Associate Professor} is example of_____attribute.
A. Continuous
B. Ordinal
C. Numeric
D. Nominal
ANSWER: D

Identify correct example of ordinal attributes?
A. Price of product
B. Age of person
C. Car colors
D. Students Grade
ANSWER: D

Identify the correct example of Nominal Attributes.
A. Weight of person in Kg
B. Income categories - HIGH, MEDIUM, LOW
C. Mobile number
D. All above
ANSWER: B

Consider the two objects i and j with nominal attributes, the dissimilarity between these objects are calculated using below equation:
d(i,j)= (p-m)/p. In this formula what p and m represents?
A. m is the number of matches, p is the total number of rows in the dataset
B. m is the number of matches, p is the total number of variables/features
C. m is the matrix, p is the total number of variables/features

D. All are wrong
ANSWER: B

When objects are represented using single attribute, the proximity value 1 indicates :
A. Objects are similar
B. Objects are dissimilar
C. Not equal
D. Reflexive
ANSWER: A

The name of the table used for measuring similarity between objects represenred using 2 or more binary attributes is:
A. Sqaure Matrix
B. Contegency Table
C. Triangular Matrix
D. None of the above
ANSWER: B

Gender is the example of Asymmetric Binary Attribute.
A. TRUE
B. FALSE
ANSWER: B

Identity correct equation of Jacard Coefficient:
A. J= f11/f01+f10+f11
B. J= f11+f00/f01+f10+f11
C.J= f11+f00/f01+f10
D. None of these
ANSWER: A

If distance d is given we can calculate similarity using equation s= d-1. (True/ False)
A. True
B. False
ANSWER: A

What equation we get when r parameter =2 in Minskowski Distance formula?
A. Manhattan distance
B. Euclidean distance
C. LMaximum Distance
D. All
ANSWER: B

Identify the distance measure to calculate distance between two objects:
A. Manhattan
B. L2
C. L1
D. Contgency Matrix
ANSWER: A

_____is a generalization of Manhattan, Euclidean and Max Distance

A. Euclidean Distance
B. Minkowski Distance
C. Manhattan distance
D. Jaccard Distance
ANSWER: B

_____ distance is based on L2 norm.
A. Euclidean Distance
B. Minkowski Distance
C. Manhattan distance
D. Jaccard Distance
ANSWER: A

_____ distance is based on L1 norm.
A. Euclidean Distance
B. Minkowski Distance
C. Manhattan distance
D. Jaccard Distance
ANSWER: C

_____ refers to a similarity or dissimilarity
A. Distance
B. Proximity
C. Enclidean
D. Manhattan
ANSWER: B

Which is not the type of attribute used in distance measure?
A. Ordinal
B. Nominal
C. Binay
D. Rank
ANSWER: D

_____ method is used to find the distance between two objects represented by Nominal attributes.
A. Euclidean Distance
B. Minkowski Distance
C. Manhattan distance
D. Simple Matching
ANSWER: D

_____ method is used to find the distance between two objects represented by numerical attributes.
A. Euclidean Distance
B. Minkowski Distance
C. Manhattan distance
D. All of these
ANSWER: D

_____ method is used to find the distance between two objects represented by Binary attributes.
A. Euclidean Distance
B. Minkowski Distance

C. Manhattan distance
D. Jaccard coefficient
ANSWER: D

Contingency table is prepared for _____ attribute data.
A. Ordinal
B. Nominal
C. Binay
D. Integer
ANSWER: C

Which is not the property of distance?
A. Distance is nonnegative number
B. Distance of an object to itself is 0
C. Distance is a symmetric function
D. Distance is negative number
ANSWER: D

If d1 and d2 are two vectors, identify correct equation of cosine similarity.
A. Cos(d1, d2)= (d1.d2)/ ||d1|| ||d2||
B. Cos(d1, d2)= |d1|| ||d2|| / (d1.d2)
C. Cos(d1, d2)= (d1.d2)
D. Cos(d1, d2)= (d1.d2)/ ||d1||
ANSWER: A

Which are the applications of proximity measures?
A. Classification
B. Clustering
C. KNN classifier
D. All of these
ANSWER: D

If o1 and o2 are two objects and distance between these objects is 1 then o1 and o2 are totally similar (True/false)
A. True
B. False
ANSWER: B

If o1 and o2 are two objects and distance between these objects is 1 then o1 and o2 are totally dissimilar (True/false)
A. True
B. False
ANSWER: A

_____ matrix represents the distance between all objects in the dataset
A. Confusion
B. Dissimilarity
C. Similarity
D. Square
ANSWER: B

If o1 and o2 are two objects and distance between these objects is 1

then it means_____
A. o1 and o2 are totally similar
B. o1 and o2 are totally dissimilar
C. o1 and o2 are similar
D. o1 and o2 are partially dissimilar
ANSWER: B

If o1 and o2 are two objects and distance between these objects is zero then o1 and o2 are totally dissimilar (True/false)
A. True
B. False
ANSWER: B

If o1 and o2 are two objects and distance between these objects is zero then it means_____
A. o1 and o2 are totally similar
B. o1 and o2 are totally dissimilar
C. o1 and o2 are similar
D. o1 and o2 are partially dissimilar
ANSWER: A

If o1 and o2 are two objects and distance between these objects is zero then o1 and o2 are totally similar (True/false)
A. True
B. False
ANSWER: A

Identify the correct subtype of Binary attribute.
A. Ordinal
B. Asymmetric
C. Symmetric
D. Both B and C
ANSWER: D

_____ Is higher when objects are more alike
A. Dissimilarity
B. Distance
C. Similarity
D. Accuracy
ANSWER: C

_____ Lower when objects are more alike.
A. Dissimilarity
B. Recall
C. Similarity
D. Accuracy
ANSWER: A