UNIT ONE	SUB : 410244 (D) DMW					
Sr. No.	**Questions**	**a**	**b**	**c**	**d**	**Ans**
1	Which of the following applied on warehouse?	write only	read only	both a & b	none	**B**
2	Data can be store , retrive and updated in …	SMTOP	OLTP	FTP	OLAP	**B**
3	Data mining is Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases	TRUE	FALSE			**A**
4	Data in the real world is	incomplete	inconsitent	noisy	all	**D**
5	What are Measure of Data Quality	Accuracy	Completeness	Consistency	all	**D**
6	Data cleaning is fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies	TRUE	FALSE			**A**
7	Data integration is Integration of multiple databases, data cubes, or files	TRUE	FALSE			**A**
8	Data transformation is Normalization and aggregation	TRUE	FALSE			**A**
9	Data reduction Obtains reduced representation in volume but produces the same or similar analytical results	TRUE	FALSE			**A**
10	Data discretization is Part of data reduction but with particular importance, especially for numerical data	TRUE	FALSE			**A**

11	Missing data may be due to	equipment malfunction	inconsistent with other recorded data and thus deleted data not entered due to misunderstanding	certain data may not be considered important at the time of entry	all	**D**
12	Incorrect attribute values may due to	faulty data collection instruments	data entry problems	data transmission problems	all	**D**
13	data cleaning is not required for duplicate records	TRUE	FALSE			**B**
14	Binning method first sort data and partition into (equi-depth) bins	TRUE	FALSE			**A**
15	Data can be smoothed by fitting the data to a function, such as with regression.	TRUE	FALSE			**A**
16	Linear regression - involves finding the___________line to fit two attributes (or variables)	best	average	worst		**A**
17	Data cleaning is fill in __________ values	existing	missing			**B**
18	Data integration is Integration of multiple databases, data cubes, or files	TRUE	FALSE			**A**

19	Data transformation is ________________and aggregation	Normalization	Denormalization			A
20	Data reduction Obtains reduced representation in volume but produces the_________ or similar analytical results	same	different			A
21	Data discretization is Part of data reduction but with particular importance, especially for _____________data	Character	numerical			B
22	Redundant data occur often when integration of multiple databases	TRUE	FALSE			A
23	The same attribute may have different names in different databases	TRUE	FALSE			A
24	Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies	TRUE	FALSE			A
25	Correlation coefficient is also called Pearson's product moment coefficient	TRUE	FALSE			A
26	Min-max normalization performs a linear transformation on the original data.	TRUE	FALSE			A
27	The values for an attribute, A, are normalized based on the mean and standard deviation of A	Min-max normalization	z-score normalization			B
28	The values for an attribute, A, are normalized based on the mean and standard deviation of A in z-score normalization	TRUE	FALSE			A

29	z-score normalization is useful when the actual minimum and maximum of attribute A are unknown	TRUE	FALSE			A
30	Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A.	TRUE	FALSE			A
31	Data reduction obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results	TRUE	FALSE			A
32	Run Length Encoding is lossless	TRUE	FALSE			A
33	Jpeg compression is	lossy	lossless			A
34	Wavelet Transform Decomposes a signal into different frequency subbands	TRUE	FALSE			A
35	Principal Component Analysis (PCA) is used for dimensionality reduction	TRUE	FALSE			A
36	Normalization by______________ scaling normalizes by moving the decimal point of values of attribute A.	binary	octal	decimal		C
37	Data cube aggregation is normalization	TRUE	FALSE			B
38	ordinal attribute have values from an ___________set	ordered	unordered			A
39	Run Length Encoding is	lossy	lossless			B

40	Nominal attribute have values from an ___________set	ordered	unordered			**B**

UNIT TWO	SUB : 410244 (D) DMW					
Sr. No.	**Questions**	**a**	**b**	**c**	**d**	**Ans**
1	What is the type of relationship in star schema?	many-to-many.	one-to-one	many-to-one	one-to-many	**d**
2	Fact tables are ___________.	completely demoralized.	partially demoralized.	completely normalized.	partially normalized.	**c**
3	Data warehouse is volatile, because obsolete data are discarded	TRUE	FALSE			**b**
4	Which is NOT a basic conceptual schema in Data Modeling of Data Warehouses?	Star schema	Tree schema	Snowflake schema	Fact constellations	**b**
5	Which is NOT a valid OLAP Rule by E.F.Codd?	Accessibility	Transparency	Flexible reporting	Reliability	**d**
6	Which is NOT a valid layer in Three-layer Data Warehouse Architecture in Conceptual View?	Processed data layer	Real-time data layer	Derived data layer	Reconciled data layer	**a**
7	Among the types of fact tables which is not a correct type ?	Fact-less fact table	Transaction fact tables	Integration fact tables	Aggregate fact tables	**c**
8	Among the followings which is not a characteristic of Data Warehouse?	Integrated	Volatile	Time-variant	Subject oriented	**b**
9	what is not considered as isssues in data warehousing?	optimization	data transformation	extraction	inter mediation	**d**

10	which one is NOT considering as a standard query technique?	Drill-up	Drill-across	DSS	Pivoting	c
11	Among the following which is not a type of business data ?	Real time data	Application data	Reconciled data	Derived data	b
12	A data warehouse is which of the following?	Can be updated by end users.	Contains numerous naming conventions and formats.	Organized around important subject areas.	Contains only current data.	c
13	A snowflake schema is which of the following types of tables?	Fact	Dimension	Helper	All of the above	d
14	The extract process is which of the following?	Capturing all of the data contained in various operational systems	Capturing a subset of the data contained in various operational systems	Capturing all of the data contained in various decision support systems	Capturing a subset of the data contained in various decision support systems	b
15	The generic two-level data warehouse architecture includes which of the following?	At least one data mart	Data that can extracted from numerous internal and external sources	Near real-time updates	All of the above.	b
16	Which one is **correct** regarding MOLAP ? A.Data is stored and fetched from the main data warehouse. B.Use complex SQL queries to fetch data from the main warehouse C.Large volume of data is used.	All are incorrect	A and B is correct.	Only C	Only A	a

17	In terms of data warehouse,metadata can be define as, A.Metadata is a road-map to data warehouse B.Metadata in data warehouse defines the warehouse objects. C.Metadata acts as a directory.	A and B is correct	A and C is correct	B is correct	All are incorrect	d
18	In terms of RLOP model, choose the most suitable answer A.The warehouse stores atomic data. B.The application layer generates SQL for the two dimensional view. C.The presentation layer provides the multidimensional view.	A and B is correct	A and C is correct	B & C is correct	All are incorrect	d
19	In the OLAP model, the _____________ provides the multidimensional view.	C. Data layer	D. Data link layer	B. Presentation layer	A. Application layer	c
20	Which of the following is **not true** regarding characteristics of warehoused data?	Changed data will be added as new data	Data warehouse can contains historical data	Obsolete data are discarded	Users can change data once entered into the data warehouse	d
21	ETL is an abbreviation for Elevation, Transformation and Loading	TRUE	FALSE			b
22	which is the core of the multidimensional model that consists of a large set of facts and a number of dimensions?	Multidimensional cube	Data model	Data cube	None of the above	c

23	Which of the following statements is incorrect	ROLAPs have large data volumes	Data form of ROLAP is large multidimentional array made of cubes	MOLAP uses sparse matrix technology to manage data sparcity	Access for MOLAP is faster than ROLAP	**b**
24	Which of the following standard query techniques increase the granularity	roll-up	dril-down	slicing	dicing	**b**
25	The full form of OLAP is	Online Analytical Processing	Online Advanced Processing	Online Analytical Performance	Online Advanced Preparation	**a**
26	Which of the following statements is/are incorrect about ROLAP A) ROLAP fetched data from datawarehouse. B) ROLAP data store as data cubes. C) ROLAP use sparse matrix to manage data sparsity.	A and B	B and C	A and C	A	**b**
27	________ is a standard query technique that can be used within OLAP to zoom in to more detailed data by changing dimensions.	Drill-up	Drill-down	Pivoting	Drill-across	**b**
28	Which of the following statements is/are correct about Fact constellation schema A) Fact constellation schema can be seen as a combination of many star schemas. B) It is possible to cerate fact constellation schema, for each star schema or snowflake schema. C) Can be identified as a flexible schema for implementation.	A	B	A and C	All of the above	**d**

29	How to describe the data contained in the data warehouse?	Relational data	Operational data	Meta data	Informational data	c
30	The output of an OLAP query is displayed as a A.Pivot B.Matrix C.Excel	A	A,B	B,C	All of the above	c
31	One can perform Query operations in the data present in Data Warahouse	TRUE	FALSE			a
32	A ____________ combines facts from multiple processes into a single fact table and eases the analytic burden on BI applications.	Aggregate fact table	Consolidated fact table	Transaction fact table	Accumulating snapshot fact table	b
33	In OLAP operations, Slicing is the technique of ______________	Selecting one particular dimension from a given cube and providing a new sub-cube	Selecting two or more dimensions from a given cube and providing a new sub-cube	Rotating the data axes in order to provide an alternative presentation of data	Performing aggregation on a data cube	a
34	Standalone data marts built by drawing data directly from operational or external sources of data or both are known as independent data marts	TRUE	FALSE			a
35	Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing is known as	Integrated	Time-variant	Subject oriented	Non-volatile	c

36	Most of the time data ware house is A Read B Write	A	B	A and B	None of the above	a
37	Data granularity is ------------------- of details of data ? A.summarization B.transformation C.level	A & B	B & C	A , B & C	C	d
38	Which one is not a type of fact?	Fully Addictive	Cumulative addictive	Semi Addictive	Non Addictive	c
39	When the level of details of data is reducing the data granularity goes higher	TRUE	FALSE			b
40	Data Warehouses are having summarized and reconciled data which can be used by decision makers	TRUE	FALSE			a
41	_______ refers to the currency and lineage of data in a data warehouse	Operational metadata	Business metadata	Technical metadata	End-User meatdata	a

UNIT THREE	SUB : 410244 (D) DMW					
Sr. No.	**Questions**	**a**	**b**	**c**	**d**	**Ans**
1	Euclidean distance measure is	A stage of the KDD process in which new data is added to the existing selection.	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	The distance between two points as calculated using the Pythagoras theorem	None of these	**c**
2	Hidden knowledge referred to	A set of databases from different vendors, possibly using different database paradigms	An approach to a problem that is not guaranteed to work but performs well in most cases	Information that is hidden in a database and that cannot be recovered by a simple SQL query.	None of these	**c**

3	Enrichment is	A stage of the KDD process in which new data is added to the existing selection	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	The distance between two points as calculated using the Pythagoras theorem	None of these	a
4	A dissimilarity coefficient is metric if it meets the four metric properties, including the triangular inequality for all possible triplets of points in the D matrix	TRUE	FALSE	Both a & b	None of these	a
5	A dissimilarity coefficient is semimetric if it violates the triangular inequality for all possible triplets of point in D matrix	TRUE	FALSE	Both a & b	None of these	b
6	A D coefficient is Euclidean if it always produces D matrices that can be fully represented in Euclidean space without distortion	TRUE	FALSE	Both a & b	None of these	a
7	A non- Euclidean dissimilarity matrix is identified by the criterion that principal coordinate analysis (PCoA) of that matrix produces some negative eigenvalues	TRUE	FALSE	Both a & b	None of these	a
8	Ecologists prefer to remove double zeros from the calculation of (dis)similarity coefficients because double zeros have no clear, unambiguous ecological interpretation	TRUE	FALSE	Both a & b	None of these	a

9	In double-zerosymmetrical coefficients, like the simple matching coefficient, double zeros affect the S or D value	TRUE	FALSE	Both a & b	None of these	a
10	Plane which have set of points satisfying certain relationships, expressible in terms of distance and angle is known as	Euclidean Plane	Dihedral plane	one dimensional plane	zero plane	a
11	Which of the following distance metric can not be used in k-NN?	Manhattan	Minkowski	Tanimoto	All of these	d
12	Which of the following is true about Manhattan distance?	It can be used for continuous variables	It can be used for categorical variables	It can be used for categorical as well as continuous	None of these	a
13	Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?	1	2	4	8	a
14	Suppose, you want to predict the class of new data point x=1 and y=1 using eucludian distance in 3-NN. In which class this data point belong to?	+ Class	– Class	cant say	None of these	a
15	Which of the following would be the leave on out cross validation accuracy for k=5?	(2/14)	(4/14)	(6/14)	None of these	d

16	What is Manhattan distance?	The distance between two points in a vector data layer calculated as the length of the line between them.	The distance between two points in a raster data layer calculated as the number of cells crossed by a straight line between them.	The distance between two points in a raster data layer calculated as the sum of the cell sides intersected by a straight line between them.	None of these	c
17	Which of the following combination is incorrect ?	Continuous – euclidean distance	Continuous – correlation similarity	Binary – manhattan distance	None of the Mentioned	d
18	The two-dimensional Euclidean plane is known as	Euclidean Plane	Dihedral plane	one dimensional plane	zero plane	a
19	The standardised form of Euclidean distance is called as	Manhattan distance	Mahalanobis distance	Dendogram	none of these	b
20	The distance between two points calculated using Pythagoras theorem is	Manhattan	Minkowski	Tanimoto	Euclidean	d
21	Identify the example of Nominal attribute	Temprature	salary	mass	gender	d
22	Nominal and ordinal attributes can be collectively referred to as_________ attributes	Perfect	Qualitative	Consistant	Optimized	b
23	A similarity S and a dissimilarity D matrix have zeros on the main diagonal	TRUE	FALSE	Both a & b	None of these	b

24	The distance between species profilesand the Hellinger, chord,and chi-square distancesare Euclidean indices	TRUE	FALSE	Both a & b	None of these	a
25	In most cases, sqrt(D) or sqrt(1–S) turns a non-Euclidean matrix to Euclidean	TRUE	FALSE	Both a & b	None of these	a
26	For descriptors withdifferent physical units, the Euclidean distance computed on standardized descriptors makes sense;the distances then have no physical units	TRUE	FALSE	Both a & b	None of these	a
27	A non-Euclidean dissimilarity matrix is identified by the criterion hat principal coordinate analysis (PCoA)of that matrix produces some negative eigenvalues	TRUE	FALSE	Both a & b	None of these	a
28	A D coefficient is Euclidean if it always produces D matrices that can be fully represented in Euclidean space without distortion	TRUE	FALSE	Both a & b	None of these	a
29	A similarity S and a dissimilarity D matrix have zeros on the main diagonal	TRUE	FALSE	Both a & b	None of these	b
30	.....are the different types of attributes	nominal	ordinal	interval	All of these	d
31	... are the types of data sets	graph	record	ordered	All of these	d

32	…are the types of ordered data	spatial data	temporal data	sequential data	All of these	d
33	… are the example of data quality problems	missing value	wrong data	duplicate data	All of these	d
34	Numerical measure of how different two data objects are…	similarity measure	dissimilarity measure	Both a & b	none of these	b
35	Numerical measure of how same two data objects are…	similarity measure	dissimilarity measure	Both a & b	none of these	a
36	Combining two or more attributes (or objects) into a single attribute (or object)	Aggregation	Sampling	Transformation	none of these	a
37	Which of the following is true about Manhattan distance?	It can be used for continuous variables	It can be used for categorical variables	It can be used for categorical as well as continuous	None of these	a
38	Sampling is the main technique employed for data reduction	Aggregation	Sampling	Transformation	none of these	b
39	…is the process of converting a continuous attribute into an ordinal attribute	Discretization	Sampling	Transformation	none of these	a

UNIT FOUR	SUB : 410244 (D) DMW					
Sr. No.	**Questions**	**a**	**b**	**c**	**d**	**Ans**
1	What does Apriori algorithm do?	It mines all frequent patterns through pruning rules with lesser	It mines all frequent patterns through pruning rules with higher suppor	Both a and b	None of these	**a**
2	What does FP growth algorithm do?	It mines all frequent patterns through pruning rules with lesser support	It mines all frequent patterns through pruning rules with higher support	It mines all frequent patterns by constructing a FP tree	All of these	**c**
3	What techniques can be used to improve the efficiency of apriori algorithm?	hash based techniques	transaction reduction	Partitioning	All of these	**d**
4	What do you mean by support(A)?	Total number of transactions containing A	Total Number of transactions not containing A	Number of transactions containing A / Total number of transactions	Number of transactions not containing A / Total number of transactions	**c**
5	Which of the following is direct application of frequent itemset mining?	Social Network Analysis	Market Basket Analysis	outlier detection	intrusion detection	**b**

6	What is not true about FP growth algorithms?	It mines frequent itemsets without candidate generation	There are chances that FP trees may not fit in the memory	FP trees are very expensive to build	It expands the original database to build FP trees	**d**
7	When do you consider an association rule interesting?	If it only satisfies min_support	If it only satisfies min_confidence	If it satisfies both min_support and min_confidence	There are other measures to check so	**c**
8	What is the difference between absolute and relative support?	Absolute -Minimum support count threshold and Relative-Minimum support	Absolute-Minimum support threshold and Relative-Minimum support count threshold	Both a and b	None of these	**a**
9	What is the relation between candidate and frequent itemsets?	A candidate itemset is always a frequent itemset	A frequent itemset must be a candidate itemset	No relation between the two	None of these	**b**
10	Which technique finds the frequent itemsets in just two database scans?	Patitioning	sampling	hashing	None of these	**a**
11	Which of the following is true?	Both apriori and FP-Growth uses horizontal data format	Both apriori and FP-Growth uses vertical data format	Both a and b	None of these	**a**
12	What is the principle on which Apriori algorithm work?	If a rule is infrequent, its specialized rules are also infrequent	If a rule is infrequent, its generalized rules are also infrequent	Both a and b	None of these	**a**

13	Which of these is not a frequent pattern mining algorithm	Apriori	FP growth	Decision trees	Eclat	**c**
14	Which algorithm requires fewer scans of data?	Apriori	FP growth	Both a and b	None of these	**b**
15	What are Max_confidence, Cosine similarity, All_confidence?	Frequent pattern mining algorithms	Measures to improve efficiency of apriori	Pattern evaluation measure	None of these	**c**
16	What are closed itemsets?	An itemset for which at least one proper supert itemset has same support	An item setwhose no proper super-itemset has same support	Both a and b	None of these	**b**
17	What are closed frequent itemsets?	A closed itemset	A frequent itemset	An itemset which is both closed and frequent	None of these	**c**
18	What are maximal frequent itemsets?	A frequent itemsetwhose no super-itemset is frequent	A frequent itemset whose super-itemset is also frequent	Both a and b	None of these	**a**
19	Why is correlation analysis important?	To make apriori memory efficient	To weed out uninteresting frequent itemsets	To find large number of interesting itemsets	To restrict the number of database iterations	**b**

20	What will happen if support is reduced?	Number of frequent itemsets remains same	Some itemsets will add to the current set of frequent itemsets	Some itemsets will become infrequent while others will become frequent	Can not say	**b**
21	Can FP growth algorithm be used if FP tree cannot be fit in memory?	Yes	No	Both a and b	None of these	**b**
22	What is association rule mining?	Same as frequent itemset mining	Finding of strong association rules using frequent itemsets	Both a and b	None of these	**b**
23	What is frequent pattern growth?	Same as frequent itemset mining	Use of hashing to make discovery of frequent itemsets more efficient	Mining of frequent itemsets without candidate generation	None of these	**c**
24	When is sub-itemset pruning done?	A frequent itemset 'P' is a proper subset of another frequent itemset 'Q'	Support (P) = Support(Q)	When both a and b is true	When a is true and b is not	**c**
25	Which of the following is not null invariant measure(that does not considers null transactions)?	all_confidence	max_confidence	cosine measure	lift	**d**
26	The apriori algorithm works in a ..and ..fashion?	top-down and depth-first	top-down and breath-first	bottom-up and depth-first	bottom-up and breath-first	**d**

27	Our use of association analysis will yield the same frequent itemsets and strong association rules whether a specific item occurs once or three times in an individual transaction	TRUE	FALSE	Both a and b	None of these	**a**
28	In association rule mining the generation of the frequent itermsets is the computational intensive step.	TRUE	FALSE	Both a and b	None of these	**a**
29	The number of iterations in apriori ___________	increases with the size of the data	decreases with the increase in size of the data	increases with the size of the maximum frequent set	decreases with increase in size of the maximum frequent set	**c**
30	Which of the following are interestingness measures for association rules?	recall	lift	accuracy	compactness	**b**
31	Frequent item sets is	Superset of only closed frequent item sets	Superset of only maximal frequent item sets	Subset of maximal frequent item sets	Superset of both closed frequent item sets and maximal frequent item sets	**d**

32	Assume that we have a dataset containing information about 200 individuals. A supervised data mining session has discovered the following rule: IF age < 30 & credit card insurance = yes THEN life insurance = yes Rule Accuracy: 70% and Rule Coverage: 63% How many individuals in the class life insurance= no have credit card insurance and are less than 30 years old?	63	30	38	70	**c**
33	In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are	100	4950	200	5000	**b**
34	Significant Bottleneck in the Apriori algorithm is	Finding frequent itemsets	pruning	Candidate generation	Number of iterations	**c**
35	Which Association Rule would you prefer	High support and medium confidence	High support and low confidence	Low support and high confidence	Low support and low confidence	**c**
36	The apriori property means	If a set cannot pass a test, its supersets will also fail the same test	To decrease the efficiency, do level-wise generation of frequent item sets	To improve the efficiency, do level-wise generation of frequent item	If a set can pass a test, its supersets will fail the same test	**a**
37	If an item set 'XYZ' is a frequent item set, then all subsets of that frequent item set are	undefined	not frequent	frequent	cant say	**c**

38	To determine association rules from frequent item sets	Only minimum confidence needed	Neither support not confidence needed	Both minimum support and confidence are needed	Minimum support is needed	**c**
39	If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is	C –> A	D –> ABCD	A –> BC	B –> ADC	**b**
40	What is frequent pattern growth?	Same as frequent itemset mining	Use of hashing to make discovery of frequent itemsets more efficient	Mining of frequent itemsets without candidate generation	None of these	**c**

UNIT FIVE SUB : 410244 (D) DMW

Sr. No.	Questions	a	b	c	d	Ans
1	Data set {brown, black, blue, green , red} is example of Select one:	a. Continuous attribute	b. Ordinal attribute	c. Numeric attribute		B
2	Which of the following activities is NOT a data mining task? Select one:	a. Predicting the future stock price of a company using historical records	b. Monitoring and predicting failures in a hydropower plant	c. Extracting the frequencies of a sound wave	d. Monitoring the heart rate of a patient for abnormalities	C
3	The difference between supervised learning and unsupervised learning is given by	a. unlike unsupervised learning, supervised learning needs labeled data	b. unlike unsupervised learning, supervised learning can be used to detect outliers	c. there is no difference	d. unlike supervised leaning, unsupervised learning can form new classes	A
4	Regression analysis is a form of predictive modelling technique	TRUE	FALSE			A
5	Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).	TRUE	FALSE			A

6	Logistic regression should be used when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.	TRUE	FALSE			**A**
7	Decision Tree is used to build classification and regression models.	TRUE	FALSE			**A**
8	Sequential Covering Algorithm can be used to extract___________rules form the training data	Do_WHILE	IF-THEN			**B**
9	Bayesian Belief Network or Bayesian Network or Belief Network is a Probabilistic Graphical Model (PGM) that represents conditional dependencies between random variables through a Directed Acyclic Graph (DAG).	TRUE	FALSE			**A**
10	In *k-NN classification*, the output is a class membership.	TRUE	FALSE			**A**
11	Associative classification integrates _______________ and association rule discovery to build classification models (classifiers).	Regression	classification			**B**
12	BAYESIAN BELIEF NETWORKS To represent the probabilistic relationships between different classes.	TRUE	FALSE			**A**
13	Regression analysis is a form of _______________ modelling technique	definate	predictive			**B**
14	Regression analysis is used for forecasting, time series modelling and finding the causal effect relationship between the variables.	TRUE	FALSE			**A**
15	Logistic regression should be used when the ___________ variable is binary (0/ 1, True/ False, Yes/ No) in nature.	independent	dependent			**B**
16	In *k-NN regression*, the output is the property value for the object.	TRUE	FALSE			**A**
17	Decision Tree Mining belongs to supervised class learning.	TRUE	FALSE			**A**

18	A regression equation is a polynomial regression equation if the power of independent variable is more than 1.	TRUE	FALSE			A
19	Decision Tree is used to create data models that will predict class labels or values for the decision-making process.	TRUE	FALSE			A
20	Regression indicates the significant relationships between dependent variable and independent variable.	TRUE	FALSE			A
21	Decision Tree Mining belongs to ________________class learning.	supervised	unsupervised			A
22	A decision tree works for both discrete and continuous variables.	TRUE	FALSE			A
23	Decision tree induction is the method of learning the decision trees from the training set.	TRUE	FALSE			A
24	Case-Based Reasoning (CBR) is used to solve problems by finding similar, past cases and adapting their solutions.	TRUE	FALSE			A
25	The Case-based reasoning CBR process can be described as a cyclic procedure	TRUE	FALSE			A
26	Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).	TRUE	FALSE			A
27	Case-based reasoning (CBR) is the process of solving new problems based on the solutions of similar past problems.[	TRUE	FALSE			A
28	Sequential Covering Algorithm can be used to extract IF-THEN rules form the training data	TRUE	FALSE			A
29	The Case-based reasoning CBR process can be described as a _________procedure	cyclic	Random	acyclic	none	A

30	Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold.	TRUE	FALSE			**A**
31	An associative classifier (AC) is a kind of supervised learning model that uses association rules to assign a target value.	TRUE	FALSE			**A**
32	Regression indicates the significant relationships between dependent variable and independent variable.	TRUE	FALSE			**A**
33	Decision Tree Mining belongs to ________________class learning.	supervised	unsupervised			**A**
34	Regression analysis is used for forecasting, time series modelling and finding the causal effect relationship between the variables.	TRUE	FALSE			**A**
35	Logistic regression should be used when the __________ variable is binary (0/ 1, True/ False, Yes/ No) in nature.	independent	dependent			**B**
36	Case-based reasoning (CBR) is an experience-based approach to solving new problems by adapting previously successful solutions to similar problems.	TRUE	FALSE			**A**
37	Regression and classification are categorized under the same umbrella of supervised machine learning.	TRUE	FALSE			**A**
38	Regression and classification are categorized under____________ machine learning.	supervised	unsupervised			**A**
39	K-NN is a lazy learner because it doesn't learn a discriminative function from the training data but “memorizes” the training dataset instead	TRUE	FALSE			**A**
40	In machine learning, lazy learning is a learning method in which generalization of the training data is, in theory, delayed until a query is made to the system	TRUE	FALSE			**A**

UNIT SIX SUB : 410244 (D) DMW

Sr. No.	Questions	a	b	c	d	Ans
1	The problem of finding hidden structure in unlabeled data is called	Supervised learning	Unsupervised learning	Reinforcement learning	none of the above	b
2	Which of the following is true for Classification?	A subdivision of a set	A measure of the accuracy	The task of assigning a classification	All of these	a
3	Classification and regression are the properties of…	data manipulation	data mining	both A & B	none of the above	b
4	We define a ______ as a subdivison of a set of examples into a number of classes	kingdom	tree	classification	array	c
5	What is inductive learning?	learning by hypothesis	learning by analyzing	learning by generalizing	none of these	c
6	In a multiclass classification problem, Bayes classifier assigns an instance to the class corresponding to:	Highest aposteriori probability	Highest apriori probability	Lowest aposteriori probability	none of these	c
7	Multiclass classifiers are also known as:	Mutlilabel classifiers	Multinomial classifiers	Multioutput classifiers	none of these	b
8	Task of inferring a model from labeled training data is called	Unsupervised learning	Supervised learning	Reinforcement learning	none of these	b
9	The problem of finding hidden structure in unlabeled data is called unsupervised learning	TRUE	FALSE			a
10	The problem of finding hidden structure in unlabeled data is called supervised learning	TRUE	FALSE			b
11	Multiclass classifiers are also known as Multinomial classifiers	TRUE	FALSE			a

12	Task of inferring a model from labeled training data is called Supervised learning	TRUE	FALSE			a
13	Classification is	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	a
14	Classification is A subdivision of a set of examples into a number of classes	TRUE	FALSE			a
15	Task of inferring a model from labeled training data is called Unsupervised learning	TRUE	FALSE			b
16	Classification accuracy is	A subdivision of a set of examples into a number of classes	Measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	b
17	Classification task referred to	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	c
18	Hybrid learning is	Machine-learning involving different techniques	The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned	Learning by generalizing from examples	None of these	a

19	Incremental learning referred to	Machine-learning involving different techniques	The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned	Learning by generalizing from examples	None of these	**b**
20	Learning is	The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system	It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks.	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.	None of these	**c**
21	Classification accuracy is Measure of the accuracy, of the classification of a concept that is given by a certain theory	TRUE	FALSE			**a**

22	Learning algorithm referrers to	An algorithm that can learn	A sub-discipline of computer science that deals with the design and implementation of learning algorithms	A machine-learning approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.	None of these	a
23	Inductive learning is	Machine-learning involving different techniques	The learning algorithmic analyzes the examples on a systematic basis and makes incremental adjustments to the theory that is learned	Learning by generalizing from examples	None of these	c

24	Bayesian classifiers is	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.	Any mechanism employed by a learning system to constrain the search space of a hypothesis	An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.	None of these	a
25	Reinforcement learning is based on goal-directed learning from interaction	TRUE	FALSE			a
26	Classification and regression are the properties Of data mining	TRUE	FALSE			a
27	Multi-perspective learning is needed for multi-perspective decision making.	TRUE	FALSE			a
28	Types of Learning	Supervised learning	Unsupervised learning	both A & B	none of these	c
29	In reinforcement learning,a reward function that is used to define goal in a reinforcement learning problem.	TRUE	FALSE			a
30	In reinforcement learning,a value function that is used to define goal in a reinforcement learning problem.	TRUE	FALSE			b
31	In Supervised learning the decision is made on the initial input or the input given at the start	TRUE	FALSE			a

32	Chess game is example of reinforcement learning	TRUE	FALSE			a
33	Chess game is example of supervised learning	TRUE	FALSE			b
34	In Reinforcement learning decision is dependent	TRUE	FALSE			a
35	Supervised learning the decisions are independent of each other	TRUE	FALSE			a
36	Supervised learning the decisions are independent of each other so labels are given to each decision	TRUE	FALSE			a
37	Supervised learning the decisions are --------of each other so labels are given to each decision.	independent	dependent	both A & B	none of these	a
38	Reward and value function is sub elements of reinforcement learning	TRUE	FALSE			a
39	Reward and value function is not sub elements of reinforcement learning	TRUE	FALSE			b
40	Object recognition is example of supervised learning	TRUE	FALSE			a

SUB : 410244(D) DMW

Data Mining and Warehouse MCQS with Answer

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.
ANSWER: B
2. The data Warehouse is___________.
A. read only.
B. write only.
C. read write only.
D. none.
ANSWER: A
3. Expansion for DSS in DW is___________.
A. Decision Support system.
B. Decision Single System.
C. Data Storable System.
D. Data Support System.
ANSWER: A
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.
ANSWER: D
5. The time horizon in Data warehouse is usually ___________.
A. 1-2 years.
B. 3-4years.
C. 5-6 years.
D. 5-10 years.
ANSWER: D
6. The data is stored, retrieved & updated in _____________.
A. OLAP.
B. OLTP.
C. SMTP.
D. FTP.
ANSWER: B
7. ___________describes the data contained in the data warehouse.
A. Relational data.
B. Operational data.
C. Metadata.
D. Informational data.
ANSWER: C
8. _____________predicts future trends & behaviors, allowing business managers to make proactive, knowledge-driven decisions.
A. Data warehouse.

B. Data mining.
C. Datamarts.
D. Metadata.
ANSWER: B
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.
ANSWER: B
10. _________________ is the specialized data warehouse database.
A. Oracle.
B. DBZ.
C. Informix.
D. Redbrick.
ANSWER: D
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.
ANSWER: C
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.
ANSWER: B
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.
ANSWER: A
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.
ANSWER: A
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.
ANSWER: A
16. . _______________databases are owned by particular departments or business groups.
A. Informational.
B. Operational.
C. Both informational and operational.
D. Flat.

ANSWER: B
17. The star schema is composed of ___________ fact table.
A. one.
B. two.
C. three.
D. four.
ANSWER: A
18. The time horizon in operational environment is ____________.
A. 30-60 days.
B. 60-90 days.
C. 90-120 days.
D. 120-150 days.
ANSWER: B
19. The key used in operational environment may not have an element of___________.
A. time.
B. cost.
C. frequency.
D. quality.
ANSWER: A
20. Data can be updated in _____environment.
A. data warehouse.
B. data mining.
C. operational.
D. informational.
ANSWER: C
21. Record cannot be updated in ______________.
A. OLTP
B. files
C. RDBMS
D. data warehouse
ANSWER: D
22. The source of all data warehouse data is the_____________.
A. operational environment.
B. informal environment.
C. formal environment.
D. technology environment.
ANSWER: A
23. Data warehouse contains______________data that is never found in the operational environment.
A. normalized.
B. informational.
C. summary.
D. denormalized.
ANSWER: C
24. The modern CASE tools belong to ________ category.
A. a. analysis.
B. b.Development
C. c.Coding
D. d.Delivery
ANSWER: A
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.
ANSWER: D
26. Detail data in single fact table is otherwise known as__________.
A. monoatomic data.
B. diatomic data.
C. atomic data.
D. multiatomic data.
ANSWER: C
27. _______test is used in an online transactional processing environment.
A. MEGA.
B. MICRO.
C. MACRO.
D. ACID.
ANSWER: D
28. ___________ is a good alternative to the star schema.
A. Star schema.
B. Snowflake schema.
C. Fact constellation.
D. Star-snowflake schema.
ANSWER: C
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.
ANSWER: C
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.
ANSWER: C
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.
ANSWER: B
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.
ANSWER: C
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.
ANSWER: D

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.
ANSWER: B
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.
ANSWER: A
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.
ANSWER: B
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
ANSWER: D
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.
ANSWER: B
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.
ANSWER: A
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.
ANSWER: A
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.
ANSWER: C
42. Fact tables are ____________.
A. completely demoralized.
B. partially demoralized.

C. completely normalized.
D. partially normalized.
ANSWER: C
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.
ANSWER: A
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.
ANSWER: D
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.
ANSWER: D
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.
ANSWER: A
47. ________ are responsible for running queries and reports against data warehouse tables.
A. Hardware.
B. Software.
C. End users.
D. Middle ware.
ANSWER: C
48. Query tool is meant for __________.
A. data acquisition.
B. information delivery.
C. information exchange.
D. communication.
ANSWER: A
49. Classification rules are extracted from _____________.
A. root node.
B. decision tree.
C. siblings.
D. branches.
ANSWER: B
50. Dimensionality reduction reduces the data set size by removing ____________.
A. relevant attributes.
B. irrelevant attributes.
C. derived attributes.
D. composite attributes.
ANSWER: B
51. ___________ is a method of incremental conceptual clustering.
A. CORBA.

B. OLAP.
C. COBWEB.
D. STING.
ANSWER: C
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.
ANSWER: A
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.
ANSWER: C
54. Multidimensional database is otherwise known as____________.
A. RDBMS
B. DBMS
C. EXTENDED RDBMS
D. EXTENDED DBMS
ANSWER: B
55. Data warehouse architecture is based on ______________.
A. DBMS.
B. RDBMS.
C. Sybase.
D. SQL Server.
ANSWER: B
56. Source data from the warehouse comes from _______________.
A. ODS.
B. TDS.
C. MDDB.
D. ORDBMS.
ANSWER: A
57. ________________ is a data transformation process.
A. Comparison.
B. Projection.
C. Selection.
D. Filtering.
ANSWER: D
58. The technology area associated with CRM is _______________.
A. specialization.
B. generalization.
C. personalization.
D. summarization.
ANSWER: C
59. SMP stands for _______________.
A. Symmetric Multiprocessor.
B. Symmetric Multiprogramming.
C. Symmetric Metaprogramming.
D. Symmetric Microprogramming.
ANSWER: A

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.
ANSWER: C
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.
ANSWER: C
62. MDDB stands for ____________.
A. multiple data doubling.
B. multidimensional databases.
C. multiple double dimension.
D. multi-dimension doubling.
ANSWER: B
63. _______________ is data about data.
A. Metadata.
B. Microdata.
C. Minidata.
D. Multidata.
ANSWER: A
64. ____________ is an important functional component of the metadata.
A. Digital directory.
B. Repository.
C. Information directory.
D. Data dictionary.
ANSWER: C
65. EIS stands for _______________.
A. Extended interface system.
B. Executive interface system.
C. Executive information system.
D. Extendable information system.
ANSWER: C
66. ____________ is data collected from natural systems.
A. MRI scan.
B. ODS data.
C. Statistical data.
D. Historical data.
ANSWER: A
67. ________________ is an example of application development environments.
A. Visual Basic.
B. Oracle.
C. Sybase.
D. SQL Server.
ANSWER: A
68. The term that is not associated with data cleaning process is ______.

A. domain consistency.
B. deduplication.
C. disambiguation.
D. segmentation.
ANSWER: D
69. _____________ are some popular OLAP tools.
A. Metacube, Informix.
B. Oracle Express, Essbase.
C. HOLAP.
D. MOLAP.
ANSWER: A
70. Capability of data mining is to build ____________ models.
A. retrospective.
B. interrogative.
C. predictive.
D. imperative.
ANSWER: C
71. ______________ is a process of determining the preference of customer's majority.
A. Association.
B. Preferencing.
C. Segmentation.
D. Classification.
ANSWER: B
72. Strategic value of data mining is _______________.
A. cost-sensitive.
B. work-sensitive.
C. time-sensitive.
D. technical-sensitive.
ANSWER: C
73. _____________ proposed the approach for data integration issues.
A. Ralph Campbell.
B. Ralph Kimball.
C. John Raphlin.
D. James Gosling.
ANSWER: B
74. The terms equality and roll up are associated with _____________.
A. OLAP.
B. visualization.
C. data mart.
D. decision tree.
ANSWER: C
75. Exceptional reporting in data warehousing is otherwise called as ___________.
A. exception.
B. alerts.
C. errors.
D. bugs.
ANSWER: B
76. _____________ is a metadata repository.
A. Prism solution directory manager.
B. CORBA.
C. STUNT.
D. COBWEB.
ANSWER: A

77. _________________ is an expensive process in building an expert system.
A. Analysis.
B. Study.
C. Design.
D. Information collection.
ANSWER: D
78. The full form of KDD is __________.
A. Knowledge database.
B. Knowledge discovery in database.
C. Knowledge data house.
D. Knowledge data definition.
ANSWER: B
79. The first International conference on KDD was held in the year ______________.
A. 1996.
B. 1997.
C. 1995.
D. 1994.
ANSWER: C
80. Removing duplicate records is a process called ______________.
A. recovery.
B. data cleaning.
C. data cleansing.
D. data pruning.
ANSWER: B
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.
ANSWER: A
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.
ANSWER: B
83. Discovery of cross-sales opportunities is called ________________.
A. segmentation.
B. visualization.
C. correction.
D. association.
ANSWER: D
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.
ANSWER: B
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.
ANSWER: D
86. The power of self-learning system lies in __________.
A. cost.
B. speed.
C. accuracy.
D. simplicity.
ANSWER: C
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.
ANSWER: A
88. How many components are there in a data warehouse?
A. two.
B. three.
C. four.
D. five.
ANSWER: D
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.
ANSWER: D
90. ________ is data that is distilled from the low level of detail found at the current detailed leve.
A. Highly summarized data.
B. Lightly summarized data.
C. Metadata.
D. Older detail data.
ANSWER: B
91. Highly summarized data is _______.
A. compact and easily accessible.
B. compact and expensive.
C. compact and hardly accessible.
D. compact.
ANSWER: A
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.
ANSWER: C
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.
ANSWER: D
94. Which of the following is not a old detail storage medium?
A. Phot Optical Storage.

B. RAID.
C. Microfinche.
D. Pen drive.
ANSWER: D
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.
ANSWER: A
96. The data in current detail level resides till _________ event occurs.
A. purge.
B. summarization.
C. archieved.
D. all of the above.
ANSWER: D
97. The dimension tables describe the __________.
A. entities.
B. facts.
C. keys.
D. units of measures.
ANSWER: B
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.
ANSWER: C
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.
ANSWER: B
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.
ANSWER: C
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.
ANSWER: C
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.

ANSWER: A
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.
ANSWER: B
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.
ANSWER: C
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.
ANSWER: A
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.
ANSWER: B
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.
ANSWER: C
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.
ANSWER: D
109. Which of the following is a predictive model?
A. Clustering.
B. Regression.
C. Summarization.
D. Association rules.
ANSWER: B
110. Which of the following is a descriptive model?
A. Classification.
B. Regression.
C. Sequence discovery.
D. Association rules.
ANSWER: C
111. A ___________ model identifies patterns or relationships.
A. Descriptive.

B. Predictive.
C. Regression.
D. Time series analysis.
ANSWER: A
112. A predictive model makes use of ________.
A. current data.
B. historical data.
C. both current and historical data.
D. assumptions.
ANSWER: B
113. _____________ maps data into predefined groups.
A. Regression.
B. Time series analysis
C. Prediction.
D. Classification.
ANSWER: D
114. ___________ is used to map a data item to a real valued prediction variable.
A. Regression.
B. Time series analysis.
C. Prediction.
D. Classification.
ANSWER: B
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.
ANSWER: B
116. In _________ the groups are not predefined.
A. Association rules.
B. Summarization.
C. Clustering.
D. Prediction.
ANSWER: C
117. Link Analysis is otherwise called as ____________.
A. affinity analysis.
B. association rules.
C. both A & B.
D. Prediction.
ANSWER: C
118. __________ is a the input to KDD.
A. Data.
B. Information.
C. Query.
D. Process.
ANSWER: A
119. The output of KDD is ___________.
A. Data.
B. Information.
C. Query.
D. Useful information.
ANSWER: D
120. The KDD process consists of _________ steps.

A. three.
B. four.
C. five.
D. six.
ANSWER: C
121. Treating incorrect or missing data is called as ___________.
A. selection.
B. preprocessing.
C. transformation.
D. interpretation.
ANSWER: B
122. Converting data from different sources into a common format for processing is called as ________.
A. selection.
B. preprocessing.
C. transformation.
D. interpretation.
ANSWER: C
123. Various visualization techniques are used in ___________ step of KDD.
A. selection.
B. transformaion.
C. data mining.
D. interpretation.
ANSWER: D
124. Extreme values that occur infrequently are called as _________.
A. outliers.
B. rare values.
C. dimensionality reduction.
D. All of the above.
ANSWER: A
125. Box plot and scatter diagram techniques are _______.
A. Graphical.
B. Geometric.
C. Icon-based.
D. Pixel-based.
ANSWER: B
126. __________ is used to proceed from very specific knowledge to more general information.
A. Induction.
B. Compression.
C. Approximation.
D. Substitution.
ANSWER: A
127. Describing some characteristics of a set of data by a general model is viewed as ____________
A. Induction.
B. Compression.
C. Approximation.
D. Summarization.
ANSWER: B
128. _____________ helps to uncover hidden information about the data.
A. Induction.
B. Compression.
C. Approximation.

D. Summarization.
ANSWER: C
129. ________ are needed to identify training data and desired results.
A. Programmers.
B. Designers.
C. Users.
D. Administrators.
ANSWER: C
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.
ANSWER: B
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.
ANSWER: D
132. Incorrect or invalid data is known as __________.
A. changing data.
B. noisy data.
C. outliers.
D. missing data.
ANSWER: B
133. ROI is an acronym of _________.
A. Return on Investment.
B. Return on Information.
C. Repetition of Information.
D. Runtime of Instruction
ANSWER: A
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.
ANSWER: B
135. ____________ data are noisy and have many missing attribute values.
A. Preprocessed.
B. Cleaned.
C. Real-world.
D. Transformed.
ANSWER: C
136. The rise of DBMS occurred in early ____________.
A. 1950's.
B. 1960's
C. 1970's
D. 1980's.
ANSWER: C
137. SQL stand for __________.
A. Standard Query Language.

B. Structured Query Language.
C. Standard Quick List.
D. Structured Query list.
ANSWER: B
138. Which of the following is not a data mining metric?
A. Space complexity.
B. Time complexity.
C. ROI.
D. All of the above.
ANSWER: D
139. Reducing the number of attributes to solve the high dimensionality problem is called as ________.
A. dimensionality curse.
B. dimensionality reduction.
C. cleaning.
D. Overfitting.
ANSWER: B
140. 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.
ANSWER: C
141. ______ are effective tools to attack the scalability problem.
A. Sampling.
B. Parallelization
C. Both A & B.
D. None of the above.
ANSWER: C
142. Market-basket problem was formulated by __________.
A. Agrawal et al.
B. Steve et al.
C. Toda et al.
D. Simon et al.
ANSWER: A
143. Data mining helps in __________.
A. inventory management.
B. sales promotion strategies.
C. marketing strategies.
D. All of the above.
ANSWER: D
144. The proportion of transaction supporting X in T is called _________.
A. confidence.
B. support.
C. support count.
D. All of the above.
ANSWER: B
145. The absolute number of transactions supporting X in T is called ___________.
A. confidence.
B. support.
C. support count.
D. None of the above.
ANSWER: C

146. 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.
ANSWER: A
147. 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%
ANSWER: A
148. 7 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%
ANSWER: D
149. The left hand side of an association rule is called ___________.
A. consequent.
B. onset.
C. antecedent.
D. precedent.
ANSWER: C
150. The right hand side of an association rule is called _____.
A. consequent.
B. onset.
C. antecedent.
D. precedent.
ANSWER: A
151. 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.
ANSWER: D
152. 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.
ANSWER: B
153. 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.
ANSWER: A
154. 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.
ANSWER: B
155. Any superset of an infrequent set is an infrequent set. This is _______.
A. Maximal frequent set.
B. Border set.
C. Upward closure property.
D. Downward closure property.
ANSWER: C
156. If an itemset is not a frequent set and no superset of this is a frequent set, then it is _______.
A. Maximal frequent set
B. Border set.
C. Upward closure property.
D. Downward closure property.
ANSWER: B
157. A priori algorithm is otherwise called as __________.
A. width-wise algorithm.
B. level-wise algorithm.
C. pincer-search algorithm.
D. FP growth algorithm.
ANSWER: B
158. The A Priori algorithm is a ___________.
A. top-down search.
B. breadth first search.
C. depth first search.
D. bottom-up search.
ANSWER: D
159. The first phase of A Priori algorithm is _______.
A. Candidate generation.
B. Itemset generation.
C. Pruning.
D. Partitioning.
ANSWER: A
160. The second phaase of A Priori algorithm is ____________.
A. Candidate generation.
B. Itemset generation.
C. Pruning.
D. Partitioning.
ANSWER: C
161. 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.
ANSWER: B
162. The a priori frequent itemset discovery algorithm moves _______ in the lattice.

A. upward.
B. downward.
C. breadthwise.
D. both upward and downward.
ANSWER: A
163. 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.
ANSWER: B
164. 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.
ANSWER: A
165. MFCS is the acronym of _____.
A. Maximum Frequency Control Set.
B. Minimal Frequency Control Set.
C. Maximal Frequent Candidate Set.
D. Minimal Frequent Candidate Set.
ANSWER: C
166. Dynamuc Itemset Counting Algorithm was proposed by ____.
A. Bin et al.
B. Argawal et at.
C. Toda et al.
D. Simon et at.
ANSWER: A
167. Itemsets in the ______ category of structures have a counter and the stop number with them.
A. Dashed.
B. Circle.
C. Box.
D. Solid.
ANSWER: A
168. The itemsets in the _______category structures are not subjected to any counting.
A. Dashes.
B. Box.
C. Solid.
D. Circle.
ANSWER: C
169. Certain itemsets in the dashed circle whose support count reach support value during an iteration
move into the ______.
A. Dashed box.
B. Solid circle.
C. Solid box.
D. None of the above.
ANSWER: A
170. Certain itemsets enter afresh into the system and get into the _______, which are essentially the
supersets of the itemsets that move from the dashed circle to the dashed box.
A. Dashed box.

B. Solid circle.
C. Solid box.
D. Dashed circle.
ANSWER: D
171. The itemsets that have completed on full pass move from dashed circle to ________.
A. Dashed box.
B. Solid circle.
C. Solid box.
D. None of the above.
ANSWER: B
172. The FP-growth algorithm has ________ phases.
A. one.
B. two.
C. three.
D. four.
ANSWER: B
173. A frequent pattern tree is a tree structure consisting of ________.
A. an item-prefix-tree.
B. a frequent-item-header table.
C. a frequent-item-node.
D. both A & B.
ANSWER: D
174. The non-root node of item-prefix-tree consists of ________ fields.
A. two.
B. three.
C. four.
D. five.
ANSWER: B
175. The frequent-item-header-table consists of __________ fields.
A. only one.
B. two.
C. three.
D. four.
ANSWER: B
176. The paths from root node to the nodes labelled 'a' are called __________.
A. transformed prefix path.
B. suffix subpath.
C. transformed suffix path.
D. prefix subpath.
ANSWER: D
177. The transformed prefix paths of a node 'a' form a truncated database of pattern which co-occur with a
is called _______.
A. suffix path.
B. FP-tree.
C. conditional pattern base.
D. prefix path.
ANSWER: C
178. 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.

ANSWER: C
179. Which of the following is a clustering algorithm?
A. A priori.
B. CLARA.
C. Pincer-Search.
D. FP-growth.
ANSWER: B
180. ________ 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.
ANSWER: A
181. ___________ 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.
ANSWER: B
182. Which of the following is a data set in the popular UCI machine-learning repository?
A. CLARA.
B. CACTUS.
C. STIRR.
D. MUSHROOM.
ANSWER: D
183. In _________ algorithm each cluster is represented by the center of gravity of the cluster.
A. k-medoid.
B. k-means.
C. STIRR.
D. ROCK.
ANSWER: B
184. 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.
ANSWER: A
185. Pick out a k-medoid algoithm.
A. DBSCAN.
B. BIRCH.
C. PAM.
D. CURE.
ANSWER: C
186. Pick out a hierarchical clustering algorithm.
A. DBSCAN
B. BIRCH.
C. PAM.

D. CURE.
ANSWER: A
187. 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.
ANSWER: C
188. BIRCH is a ________.
A. agglomerative clustering algorithm.
B. hierarchical algorithm.
C. hierarchical-agglomerative algorithm.
D. divisive.
ANSWER: C
189. 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.
ANSWER: A
190. 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.
ANSWER: D
191. The partition algorithm uses _______ scans of the databases to discover all frequent sets.
A. two.
B. four.
C. six.
D. eight.
ANSWER: A
192. 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.
ANSWER: A
193. ________is the most well known association rule algorithm and is used in most commercial products.
A. Apriori algorithm.
B. Partition algorithm.
C. Distributed algorithm.
D. Pincer-search algorithm.
ANSWER: A
194. An algorithm called________is used to generate the candidate item sets for each pass after the first.
A. apriori.
B. apriori-gen.

C. sampling.
D. partition.
ANSWER: B
195. The basic partition algorithm reduces the number of database scans to ________ & divides it into partitions.
A. one.
B. two.
C. three.
D. four.
ANSWER: B
196. ___________and prediction may be viewed as types of classification.
A. Decision.
B. Verification.
C. Estimation.
D. Illustration.
ANSWER: C
197. ___________can be thought of as classifying an attribute value into one of a set of possible classes.
A. Estimation.
B. Prediction.
C. Identification.
D. Clarification.
ANSWER: B
198. Prediction can be viewed as forecasting a_________value.
A. non-continuous.
B. constant.
C. continuous.
D. variable.
ANSWER: C
199. _________data consists of sample input data as well as the classification assignment for the data.
A. Missing.
B. Measuring.
C. Non-training.
D. Training.
ANSWER: D
200. Rule based classification algorithms generate ______ rule to perform the classification.
A. if-then.
B. while.
C. do while.
D. switch.
ANSWER: A
201. ____________ are a different paradigm for computing which draws its inspiration from neuroscience.
A. Computer networks.
B. Neural networks.
C. Mobile networks.
D. Artificial networks.
ANSWER: B
202. The human brain consists of a network of ___________.
A. neurons.
B. cells.

C. Tissue.
D. muscles.
ANSWER: A
203. Each neuron is made up of a number of nerve fibres called _____________.
A. electrons.
B. molecules.
C. atoms.
D. dendrites.
ANSWER: D
204. The ___________is a long, single fibre that originates from the cell body.
A. axon.
B. neuron.
C. dendrites.
D. strands.
ANSWER: A
205. A single axon makes ___________ of synapses with other neurons.
A. ones.
B. hundreds.
C. thousands.
D. millions.
ANSWER: C
206. _____________ is a complex chemical process in neural networks.
A. Receiving process.
B. Sending process.
C. Transmission process.
D. Switching process.
ANSWER: C
207. __________ is the connectivity of the neuron that give simple devices their real power. a. b. c. d.
A. Water.
B. Air.
C. Power.
D. Fire.
ANSWER: D
208. ___________ are highly simplified models of biological neurons.
A. Artificial neurons.
B. Computational neurons.
C. Biological neurons.
D. Technological neurons.
ANSWER: A
209. The biological neuron's _________ is a continuous function rather than a step function.
A. read.
B. write.
C. output.
D. input.
ANSWER: C
210. The threshold function is replaced by continuous functions called ________ functions.
A. activation.
B. deactivation.
C. dynamic.
D. standard.
ANSWER: A
211. The sigmoid function also knows as __________functions.
A. regression.

B. logistic.
C. probability.
D. neural.
ANSWER: B
212. MLP stands for ______________________.
A. mono layer perception.
B. many layer perception.
C. more layer perception.
D. multi layer perception.
ANSWER: D
213. In a feed- forward networks, the conncetions between layers are ___________ from input to output.
A. bidirectional.
B. unidirectional.
C. multidirectional.
D. directional.
ANSWER: B
214. The network topology is constrained to be __________________.
A. feedforward.
B. feedbackward.
C. feed free.
D. feed busy.
ANSWER: A
215. RBF stands for _____________.
A. Radial basis function.
B. Radial bio function.
C. Radial big function.
D. Radial bi function.
ANSWER: A
216. RBF have only _______________ hidden layer.
A. four.
B. three.
C. two.
D. one.
ANSWER: D
217. RBF hidden layer units have a receptive field which has a ____________; that is, a particular input
value at which they have a maximal output.
A. top.
B. bottom.
C. centre.
D. border.
ANSWER: C
218. ___________ training may be used when a clear link between input data sets and target output values
does not exist.
A. Competitive.
B. Perception.
C. Supervised.
D. Unsupervised.
ANSWER: D
219. ___________ employs the supervised mode of learning.
A. RBF.

B. MLP.
C. MLP & RBF.
D. ANN.
ANSWER: C
220. _________________ design involves deciding on their centres and the sharpness of their Gaussians.
A. DR.
B. AND.
C. XOR.
D. RBF.
ANSWER: D
221. ____________ is the most widely applied neural network technique.
A. ABC.
B. PLM.
C. LMP.
D. MLP.
ANSWER: D
222. SOM is an acronym of ________________.
A. self-organizing map.
B. self origin map.
C. single organizing map.
D. simple origin map.
ANSWER: A
223. _____________ is one of the most popular models in the unsupervised framework.
A. SOM.
B. SAM.
C. OSM.
D. MSO.
ANSWER: A
224. The actual amount of reduction at each learning step may be guided by __________.
A. learning cost.
B. learning level.
C. learning rate.
D. learning time.
ANSWER: C
225. The SOM was a neural network model developed by _________.
A. Simon King.
B. Teuvokohonen.
C. Tomoki Toda.
D. Julia.
ANSWER: B
226. SOM was developed during _____________.
A. 1970-80.
B. 1980-90.
C. 1990 -60.
D. 1979 -82.
ANSWER: D
227. Investment analysis used in neural networks is to predict the movement of __________ from previous
data.
A. engines.
B. stock.
C. patterns.

D. models.
ANSWER: B
228. SOMs are used to cluster a specific ______________ dataset containing information about the patient's
drugs etc.
A. physical.
B. logical.
C. medical.
D. technical.
ANSWER: C
229. GA stands for ________________.
A. Genetic algorithm
B. Gene algorithm.
C. General algorithm.
D. Geo algorithm.
ANSWER: A
230. GA was introduced in the year ___________.
A. 1955.
B. 1965.
C. 1975.
D. 1985.
ANSWER: C
231. Genetic algorithms are search algorithms based on the mechanics of natural________.
A. systems.
B. genetics.
C. logistics.
D. statistics.
ANSWER: B
232. GAs were developed in the early ______________.
A. 1970.
B. 1960.
C. 1950.
D. 1940.
ANSWER: A
233. The RSES system was developed in ____________.
A. Poland.
B. Italy.
C. England.
D. America.
ANSWER: A
234. Crossover is used to ________.
A. recombine the population's genetic material.
B. introduce new genetic structures in the population.
C. to modify the population's genetic material.
D. All of the above.
ANSWER: A
235. The mutation operator _______.
A. recombine the population's genetic material.
B. introduce new genetic structures in the population.
C. to modify the population's genetic material.
D. All of the above.
ANSWER: B
236. Which of the following is an operation in genetic algorithm?

A. Inversion.
B. Dominance.
C. Genetic edge recombination.
D. All of the above.
ANSWER: D
237. . ____________ is a system created for rule induction.
A. RBS.
B. CBS.
C. DBS.
D. LERS.
ANSWER: D
238. NLP stands for __________.
A. Non Language Process.
B. Nature Level Program.
C. Natural Language Page.
D. Natural Language Processing.
ANSWER: D
239. Web content mining describes the discovery of useful information from the _______contents.
A. text.
B. web.
C. page.
D. level.
ANSWER: B
240. Research on mining multi-types of data is termed as _______ data.
A. graphics.
B. multimedia.
C. meta.
D. digital.
ANSWER: B
241. _______ mining is concerned with discovering the model underlying the link structures of the web.
A. Data structure.
B. Web structure.
C. Text structure.
D. Image structure.
ANSWER: B
242. __________ is the way of studying the web link structure.
A. Computer network.
B. Physical network.
C. Social network.
D. Logical network.
ANSWER: C
243. The _________ propose a measure of standing a node based on path counting.
A. open web.
B. close web.
C. link web.
D. hidden web.
ANSWER: B
244. In web mining, _______ is used to find natural groupings of users, pages, etc.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.

ANSWER: A
245. In web mining, __________ is used to know the order in which URLs tend to be accessed.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: C
246. In web mining, __________ is used to know which URLs tend to be requested together.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: B
247. ___________ describes the discovery of useful information from the web contents.
A. Web content mining.
B. Web structure mining.
C. Web usage mining.
D. All of the above.
ANSWER: A
248. ________ is concerned with discovering the model underlying the link structures of the web.
A. Web content mining.
B. Web structure mining.
C. Web usage mining.
D. All of the above.
ANSWER: B
249. The ____________ engine for a data warehouse supports query-triggered usage of data
A. NNTP
B. SMTP
C. OLAP
D. POP
ANSWER: C
250. _________ 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
ANSWER: B

ZEAL EDUCATION SOCIETY'S

ZEAL COLLEGE OF ENGINEERING AND RESEARCH

NARHE | PUNE -41 | INDIA

DEPARTMENT OF COMPUTER ENGINEERING

Name of the Teacher: V. R. Vasekar

Class: BE **Subject: Data Mining and Warehousing**
AY: 2020-21 **SEM: I**

UNIT-1

1)	Binary 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
Ans:	**a**
Explanation:	All statement are true about Machine Learning.
2)	"Efficiency and scalability of data mining algorithms" issues come under?
	a) Mining Methodology and User Interaction Issues b) Performance Issues c) Diverse Data Types Issues d) None of the above
Ans:	**b**
Explanation:	**In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.**
3)	——- is not a data mining functionality?
	a) Clustering and Analysis b) Selection and interpretation c) Classification and regression Characterization and Discrimination
Ans:	**b**
Explanation:	Selection and interpretation
4)	——— is the output of KDD
	a) Query b) Data c) Useful Information d) information
Ans:	**c**
Explanation:	Useful Information
5)	Which of the following is not belong to data mining?t is unsupervised learning ?
	a) Knowledge extraction b) Data archaeology c) Data exploration d) Data transformation

Ans:	**d**
Explanation:	Data transformation
6)	Which of the following is the right approach to Data Mining?
	e) Infrastructure, exploration, analysis, exploitation, interpretation f) Infrastructure, exploration, analysis, interpretation, exploitation g) Infrastructure, analysis, exploration, interpretation, exploitation None of these
Ans:	**b**
Explanation:	Infrastructure, exploration, analysis, interpretation, exploitation
7)	**Background knowledge referred to**
	a) Additional acquaintance used by a learning algorithm to facilitate the learning process b) A neural network that makes use of a hidden layer c) It is a form of automatic learning. d) None of these
Ans:	**a**
Explanation:	Additional acquaintance used by a learning algorithm to facilitate the learning process
8)	**Data 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
Ans:	**a**
Explanation:	The actual discovery phase of a knowledge discovery process
09)	**Data 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
Ans:	**b**
Explanation:	The stage of selecting the right data for a KDD process
10)	The Example of nominal attribute is
	a) Hair_color b) smoker c) temperature d) drink size

Ans:	**a**
Explanation:	Nominal means "relating to names." The values of a nominal attribute are symbols or names of things
11)	The Example of binary attribute is
	a) gender b) drink_size c) tempertaure d) professionl_rank
Ans	**b**
Explanation:	A binary attribute is a nominal attribute with only two categories or states:0 or1
12)	The Example of ordinary attribute is
	a) Years_of_experience b) age c) occupation d) customer_id
Ans:	**b**
Explanation:	An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them
13)	Data cleaning includes____
	a. Handling missing values and noisy data b. Reduction of attributes c. Relevant attribute selection d. Sample data selection
Ans:	**a**
Explanation:	Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data**.**
14)	To deal with missing values, the following strategy is used__
	e. Use a measure of central tendency f. Reduction of attribute g. Sample data selection h. Data converted into other form
Ans:	**a**
Explanation:	measures of central tendency, which indicate the "middle" value of a data distribution
15)	Noise is ___
	a) Missing value from dataset b) Inaccurate data c) a random error or variance in a measured variable d) the data whose value known to user
Ans:	**c**
Explanation:	

16)	At the time of data integration following problem ocuures___
	a) Selection of proper values b) Raw data conversion c) Entity identification d) Attribute subset selction
Ans:	**c**
Explanation:	Schema integration and object matching can be tricky.
17)	Which of the following is not example of data reduction strategy?
	a) Outlier detection b) Principal Component Analysis c) Attribute subset selection d) Wavelet transforms
Ans:	**a**
Explanation:	Outlier detection
18)	Data Transformation Strategies includes____
	a) smoothing b) Attribute construction c) Normalization d) All of the above
Ans:	**d**
Explanation:	Smoothing, attribute construction and normalization includes in data transformation
19)	Data Discretization is used for____
	a) transforms numeric data by mapping values to interval or concept labels b) smoothing c) Attribute construction d) Normalization
Ans:	**a**
Explanation:	transforms numeric data by mapping values to interval or concept labels
20)	KDD stands for
	a) K data values b) Knowledge discovery from dataset c) K dataset d) None of the above
Ans.	b
Explaination	Knowledge discovery from dataset
21)	**Data transformation includes:**

	a) data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations b) an essential process where intelligent methods are applied to extract data patterns c) data relevant to the analysis task are retrieved from the database d) it is used for knowledge representation.
Ans	**a**
Explanation	data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations
22)	**Pattern evaluation includes__**
	a) data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations b) an essential process where intelligent methods are applied to extract data patterns c) data relevant to the analysis task are retrieved from the database d) Identify the truly interesting patterns representing knowledge based on interestingness measures
Ans	**d**
Explanation	To identify the truly interesting patterns representing knowledge based on interestingness measures
23)	**In KDD, the knowledge representation term used for__**
	a) data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations b) an essential process where intelligent methods are applied to extract data patterns c) visualization and knowledge representation techniques are used to present mined knowledge to users d) Identify the truly interesting patterns representing knowledge based on interestingness measures
Ans	**c**
Explanation	visualization and knowledge representation techniques are used to present mined knowledge to users
24)	Data mining functionalities are used to___
	a) to specify the kinds of patterns or knowledge to be found in data mining tasks b) to select data c) to find missing values d) to analyze the mining result
Ans	a
Explanation	a) Data mining functionalities are used to specify the kinds of patterns or

	knowledge to be found in data mining tasks
25)	The challenging issues in data mining research____
	a) efficiency and scalability b) dealing with diverse data types c) user interaction d) all of the above
Ans	d
Explanation	There are many challenging issues in data mining research. Areas include mining methodology, user interaction, efficiency and scalability, and dealing with diverse data types. Data mining research has strongly impacted society and will continue to do so in the future

Name and Sign of Subject Teacher

ZEAL EDUCATION SOCIETY'S

ZEAL COLLEGE OF ENGINEERING AND RESEARCH

NARHE | PUNE -41 | INDIA

DEPARTMENT OF COMPUTER ENGINEERING

Name of the Teacher: V. R. Vasekar

Class: BE | **Subject: Data Mining and Warehousing**
AY: 2020-21 | **SEM: II**

UNIT-2 Data Warehouse

1. ___ is a subject oriented, integrated, time variant, non-volatile collection of data in support of management decisions.	
	a) Data Mining b) Data Warehousing c) Web mining d) Text mining
Ans:	**b**
Explanation:	Data Warehousing
2. Data Warehouse is	
	a) Read only b) Write only c) Read and write only d) none
Ans:	**a**
Explanation:	Because of historical data storage
3. Expansion for DSS in DW is___	
	a) Decision Single System b) Decision storable system c) Decision Support System d) Data Support System
Ans:	**c**
Explanation:	Decision support system
4. The important aspect of 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
Ans:	**d**
Explanation:	All are correct
5. The time horizon in Data warehouse is usually__	
	a) 1-2 year b) 3-4 year c) 5-6 years d) 5-10 years
Ans:	**d**
Explanation:	5 to 10 years
6. The data is stored , retrieved and updated in___	

	a) OLAP b) OLTP c) SMTP d) FTP
Ans:	**b**
Explanation:	Online Analytical Transaction processing
7. ____describes the data oriented in the data warehouse	
	a) Relational data b) Operational data c) Metadata d) Informational data
Ans:	**c**
Explanation:	metadata
8. ____ predicts the future trends and behaviours, allowing business managers to make proactive knowledge-driven decisions	
	a) Data warehouse b) Data mining c) Datamarts d) metadata
Ans:	**b**
Explanation:	
9. ____ is the heart of Datawarehouse	
	a) Data mining database server b) Data warehouse database servers c) Data mart database servers d) Relational database servers
Ans:	**b**
Explanation:	Data warehouse database servers
10. ____is the specialized data warehouse database	
	a) Oracle b) DBZ c) Informix d) Redbricks
Ans:	**d**
Explanation:	Redbricks
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
Ans	**c**
Explanation:	Operational metadata

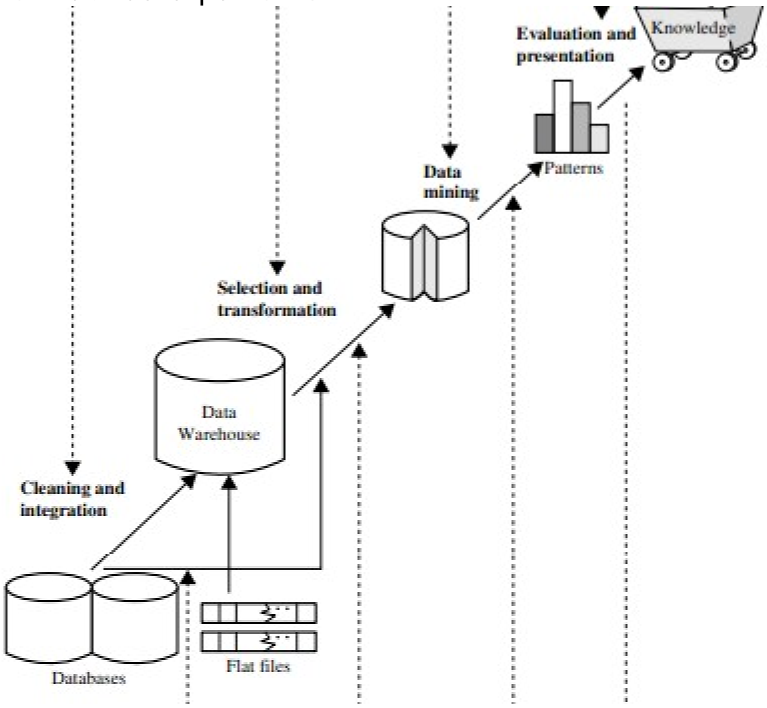
12.----helds the catelog of the warehouse database system	
	a) Application level metadata b) Algorithmic level metadata c) Departmental level metadata d) Core warehouse metadata
Ans:	**b**
Explanation:	Algorithmic level metadata
13. ___maps the core warehouse metadata to business concepts, familiar and useful to end-users	
	a) Application level metadata. b) User level metadata.C. c) Enduser level metadata. d) Core level metadata
Ans:	**a**
Explanation:	
14. The star schema is composed of ___________ fact table.	
	a) One b) Two c) Three d) four
Ans:	**a**
Explanation:	Only one fact table
15. The source of all data warehouse data is the__	
	a) operational environment b) informal environment c) formal environment. d) technology environmen
Ans:	**a**
Explanation:	
16.The @active data warehouse architecture includes which of the following?	
	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.
Ans:	**d**
Explanation:	
17.An operational system is which of the following?	
	a) A system that is used to run the business in real time and is based on historical data. b) A system that is used to run the business in real time and is based on current data. c) A system that is used to support decision making and is based on

	current data. d) A system that is used to support decision making and is based on historical data.
Ans:	**b**
Explanation:	
18.A data warehouse is which of the following?	
	a) Can be updated by end users. b) Contains numerous naming conventions and formats. c) Organized around important subject areas d) Contains only current data.
Ans:	**c**
Explanation:	Data warehouse is subject oriented
19. Good performance can be achieved in a data mart environment by extensive use of	
	a) Indexes b) creating profile records c) volumes of data d) all of the above
Ans:	**d**
Explanation:	
20.	Warehouse administrator responsible for
	a) Administrator b) Maintenance c) both a and b d) none of the above
Ans	**c**
Explaination	
21. What is data cube?	
	a) allows data to be modeled and viewed in multiple dimensions b) data with dimensions c) data values **d)** description about data
Ans.	**a**
23 . Which of the following is not a multidimensional data model?	
	a) Star schema b) Fact constellation c) Snowflake schemas d) Entity-relationship model
Ans	d
Explanation	Three models of data warehouse: star, snowflake and fact constellation
24. Snowflake schema consists of ___fact tables	
	a) One b) Two c) Three

	d) four
Ans	a
Explanation	Having only one fact table and many dimension tables
25. Fact constellation consists of __ fact tables	
	a) one b) two c) three d) many
Ans	d
Explanation	Many fact tables and many dimension tables

Name and Sign of Subject Teacher

1. .... is an essential process where intelligent methods are applied to extract data patterns.



A) Data warehousing

B) Data mining

C) Text mining

D) Data selection

2. Data mining can also applied to other forms such as ....

i) Data streams

ii) Sequence data

iii) Networked data

iv) Text data

v) Spatial data

A) i, ii, iii and v only

B) ii, iii, iv and v only

C) i, iii, iv and v only

D) All i, ii, iii, iv and v

3. Which 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

4. .... is a summarization of the general characteristics or features of a target class of data.

A) Data Characterization

B) Data Classification

C) Data discrimination

D) Data selection

5. .... is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

A) Data Characterization

B) Data Classification

C) Data discrimination

D) Data selection

6. Strategic value of data mining is ....

A) cost-sensitive

B) work-sensitive

C) time-sensitive

D) technical-sensitive

7. .... is the process of finding a model that describes and distinguishes data classes or concepts.

A) Data Characterization

B) Data Classification

C) Data discrimination

D) Data selection

8. The 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

A) i, ii and iv only

B) ii, iii and iv only

C) i, ii and iii only

D) All i, ii, iii and iv

9. The full form of KDD is ....

A) Knowledge Database

B) Knowledge Discovery Database

C) Knowledge Data House

D) Knowledge Data Definition

10. The out put of KDD is ....

A) Data

B) Information

C) Query

D) Useful information

Data Warhouse & Data Mining 700 - MCQ’s

TOPIC ONE – INTRODUCTION TO DATA MINING

EASY QUESTIONS

1.Data mining is an integral part of _________.
A. SE.
B. DBMS.
C. KDD.
D. OS.
ANSWER: C

2. __________ is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management decisions.
A. Data Mining.
B. Data Warehousing.
C. Web Mining.
D. Text Mining.
ANSWER: B

3. KDD describes the _________.
A. whole process of extraction of knowledge from data
B. extraction of data
C. extraction of information
D. extraction of rules
ANSWER: A

4. The data Warehouse is__________.
A. read only.
B. write only.
C. read write only.
D. none.
ANSWER: A

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

6. 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.
ANSWER: D

7. The data is stored, retrieved & updated in ____________.
A. OLAP.
B. OLTP.
C. SMTP.
D. FTP.
ANSWER: B

8. __________describes the data contained in the data warehouse.
A. Relational data.
B. Operational data.
C. Metadata.
D. Informational data.
ANSWER: C

9. ____________predicts future trends &behaviors, allowing business managers to make proactive,knowledge-driven decisions.
A. Data warehouse.
B. Data mining.
C. Datamarts.
D. Metadata.
ANSWER: B

10. __________ 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.
ANSWER: B

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

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.
ANSWER: B

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.
ANSWER: A

14. Data can be updated in _____environment.
A. data warehouse.
B. data mining.
C. operational.
D. informational.
ANSWER: C

15. Record cannot be updated in _____________.
A. OLTP
B. files
C. RDBMS
D. data warehouse
ANSWER: D

16. Detail data in single fact table is otherwise known as__________.
A. monoatomic data.
B. diatomic data.
C. atomic data.
D. multiatomic data.
ANSWER: C

17. 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.
ANSWER: C

18. ______________ is data about data.
A. Metadata.
B. Microdata.
C. Minidata.
D. Multidata.
ANSWER: A

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

20. The term that is not associated with data cleaning process is ______.
A. domain consistency.
B. deduplication.
C. disambiguation.
D. segmentation.
ANSWER: D

21. Capability of data mining is to build ___________ models.
A. retrospective.
B. interrogative.
C. predictive.
D. imperative.

ANSWER: C

22. ______________ is a process of determining the preference of customer's majority.
A. Association.
B. Preferencing.
C. Segmentation.
D. Classification.
ANSWER: B

23. Exceptional reporting in data warehousing is otherwise called as ___________.
A. exception.
B. alerts.
C. errors.
D. bugs.
ANSWER: B

24. The full form of KDD is __________.
A. Knowledge database.
B. Knowledge discovery in database.
C. Knowledge data house.
D. Knowledge data definition.
ANSWER: B

25. Removing duplicate records is a process called ______________.
A. recovery.
B. data cleaning.
C. data cleansing.
D. data pruning.
ANSWER: B

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

27. Discovery of cross-sales opportunities is called _________________.
A. segmentation.
B. visualization.
C. correction.
D. association.
ANSWER: D

28. 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.
ANSWER: B

29. 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.
ANSWER: C

30. Which of the following is not an old detail storage medium?
A. Phot Optical Storage.
B. RAID.
C. Microfinche.
D. Pen drive.
ANSWER: D

31. The dimension tables describe the _________.
A. entities.
B. facts.
C. keys.
D. units of measures.
ANSWER: B

32. Which of the following is not the other name of Data mining?
A. Exploratory data analysis.
B. Data driven discovery.
C. Deductive learning.
D. Data integration.
ANSWER: D

33. Which of the following is a predictive model?
A. Clustering.
B. Regression.
C. Summarization.
D. Association rules.
ANSWER: B

34. Which of the following is a descriptive model?
A. Classification.
B. Regression.
C. Sequence discovery.
D. Association rules.
ANSWER: C

35. A ___________ model identifies patterns or relationships.
A. Descriptive.
B. Predictive.
C. Regression.
D. Time series analysis.
ANSWER: A

36. A predictive model makes use of ________.
A. current data.
B. historical data.
C. both current and historical data.
D. assumptions.

ANSWER: B

37. _____________ maps data into predefined groups.
A. Regression.
B. Time series analysis
C. Prediction.
D. Classification.
ANSWER: D

38. ___________ is used to map a data item to a real valued prediction variable.
A. Regression.
B. Time series analysis.
C. Prediction.
D. Classification.
ANSWER: B

39. In _____________, the value of an attribute is examined as it varies over time.
A. Regression.
B. Time series analysis.
C. Sequence discovery.
D. Prediction.
ANSWER: B

40. In _________ the groups are not predefined.
A. Association rules.
B. Summarization.
C. Clustering.
D. Prediction.
ANSWER: C

41. __________ is the input to KDD.
A. Data.
B. Information.
C. Query.
D. Process.
ANSWER: A

42. The output of KDD is ___________.
A. Data.
B. Information.
C. Query.
D. Useful information.
ANSWER: D

43. The KDD process consists of _________ steps.
A. three.
B. four.
C. five.
D. six.
ANSWER: C

44. Treating incorrect or missing data is called as ___________.
A. selection.
B. preprocessing.
C. transformation.
D. interpretation.
ANSWER: B

45. Converting data from different sources into a common format for processing is called as ________.
A. selection.
B. preprocessing.
C. transformation.
D. interpretation.
ANSWER: C

46. Various visualization techniques are used in ___________ step of KDD.
A. selection.
B. transformaion.
C. data mining.
D. interpretation.
ANSWER: D

47. Extreme values that occur infrequently are called as _________.
A. outliers.
B. rare values.
C. dimensionality reduction.
D. Inliers
ANSWER: A

48. Box plot and scatter diagram techniques are _______.
A. Graphical.
B. Geometric.
C. Icon-based.
D. Pixel-based.
ANSWER: B

49. __________ is used to proceed from very specific knowledge to more general information.
A. Induction.
B. Compression.
C. Approximation.
D. Substitution.
ANSWER: A

50. Describing some characteristics of a set of data by a general model is viewed as
A. Induction.
B. Compression.
C. Approximation.
D. Summarization.
ANSWER: B

51. _____________ helps to uncover hidden information about the data.
A. Induction.
B. Compression.
C. Approximation.

D. Summarization.
ANSWER: C

52. Incorrect or invalid data is known as _________.
A. changing data.
B. noisy data.
C. outliers.
D. missing data.
ANSWER: B

53. 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.
ANSWER: B

54. ___________ data are noisy and have many missing attribute values.
A. Preprocessed.
B. Cleaned.
C. Real-world.
D. Transformed.
ANSWER: C

55. __________ describes the discovery of useful information from the web contents.
A. Web content mining.
B. Web structure mining.
C. Web usage mining.
D. Web development.
ANSWER: A

56. _______ is concerned with discovering the model underlying the link structures of the web.
A. Web content mining.
B. Web structure mining.
C. Web usage mining.
D. Web development.
ANSWER: B

57. A _____ algorithm takes all the data at once and tries to create a hypothesis based on this data.
A. supervised.
B. batch learning.
C. unsupervised.
D. incremental learning.
ANSWER: B

58. A ________ algorithm takes a new piece of information at each learning cycle and tries to revise the theory using
new data.
A. supervised.
B. batch learning.
C. unsupervised.
D. incremental learning.
ANSWER: B

59. ________ is used to find the vaguely known data.
A. SQL.
B. KDD.
C. Data mining.
D. Sybase.
ANSWER: C

60. The easiest way to gain access to the data and facilitate effective decision making is to set up a _______.
A. database.
B. data mart.
C. data warehouse.
D. operational.
ANSWER: C

61. Smaller local data warehouse is called as ____.
A. data mart.
B. database.
C. data model.
D. meta data.
ANSWER: B

62. The _______ data are stored in data warehouse.
A. operational.
B. historical.
C. transactional.
D. optimized.
ANSWER: B

63. A decision support system is a system that ________.
A. can constantly change over time.
B. cannot change.
C. copies the data.
D. supports the system.
ANSWER: A

64. Metadata is used by the end users for ______.
A. managing database.
B. structuring database.
C. querying purposes.
D. making decisions.
ANSWER: C

65. The _________ techniques are used to load information from operational database to data warehouse.
A. reengineering.
B. reverse.
C. transfer.
D. replication.
ANSWER: D

66. In machine learning _________ phase try to find the patterns from observations.
A. observation
B. theory
C. analysis
D. prediction
ANSWER: C

67. Information content is closely related to ______ and transparency.
A. algorithm.
B. search space.
C. learning.
D. statistical significance.
ANSWER: D

68. The _________ is used to express the hypothesis describing the concept.
A. computer language.
B. algorithm.
C. definition.
D. theory
ANSWER: A

69. A definition of a concept is complete if it recognizes __________.
A. all the information.
B. all the instances of a concept.
C. only positive examples.
D. negative examples.
ANSWER: B

70. The results of machine learning algorithms are always have to be checked for their __________.
A. observations.
B. calculations
C. programs.
D. statistical relevance.
ANSWER: D

71. A _________ is necessary condition for KDDs effective implement.
A. data set.
B. database.
C. data warehouse.
D. data.
ANSWER: C

72. KDD is a _________.
A. new technology that is use to store data.
B. multidisciplinary field of research.
C. database technology.
D. expert system.
ANSWER: B

INTERMEDIATE QUESTIONS

73. 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.
ANSWER: C

74. 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.
ANSWER: B

75. 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.
ANSWER: A

76. 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.
ANSWER: B

77. 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.
ANSWER: A

78. ________________ 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.
ANSWER: A

79. Business Intelligence and data warehousing is not used for ________.
A. Forecasting.
B. Data Mining.
C. Analysis of large volumes of product sales data.
D. Discarding data.
ANSWER: D

80. Classification rules are extracted from _____________.
A. root node.
B. decision tree.
C. siblings.
D. branches.
ANSWER: B

81. Reducing the number of attributes to solve the high dimensionality problem is called as ________.
A. dimensionality curse.
B. dimensionality reduction.
C. cleaning.
D. Overfitting.
ANSWER: B

82. 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.
ANSWER: C

83. Data mining helps in __________.
A. inventory finalisation.
B. sales.
C. marketing products.
D. Debt collection.
ANSWER: A

84. 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.
ANSWER: D

85. 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.
ANSWER: B

86. Metadata describes __________.
A. contents of database.
B. structure of contents of database.
C. structure of database.
D. database itself.
ANSWER: B

87. The partition of overall data warehouse is _______.
A. database.
B. data cube.
C. data mart.

D. operational data.
ANSWER: C

88. The information on two attributes is displayed in _____________ in scatter diagram.
A. visualization space.
B. scatter space.
C. cartesian space.
D. interactive space.
ANSWER: C

89. OLAP is used to explore the ____________ knowledge.
A. shallow.
B. deep.
C. multidimensional.
D. hidden.
ANSWER: C

90. Hidden knowledge can be found by using _________.
A. searching algorithm.
B. pattern recognition algorithm.
C. searching algorithm.
D. clues.
ANSWER: B

91. The next stage to data selection in KDD process _______.
A. enrichment.
B. coding.
C. cleaning.
D. reporting.
ANSWER: C

92. Enrichment means _____.
A. adding external data.
B. deleting data.
C. cleaning data.
D. selecting the data.
ANSWER: A

93. The decision support system is used only for ________.
A. cleaning.
B. coding.
C. selecting.
D. queries.
ANSWER: D

94. Which of the following is closely related to statistical significance and transparency?
A. Classification Accuracy.
B. Transparency.
C. Statistical significance.
D. Search Complexity.
ANSWER: B

95. ________ is the technique which is used for discovering patterns in dataset at the beginning of data mining process.
A. Kohenon map.
B. Visualization.
C. OLAP.
D. SQL.
ANSWER: B

96. _______ is the heart of knowledge discovery in database process.
A. Selection.
B. Data ware house.
C. Data mining.
D. Creative coding.
ANSWER: D

97. In KDD and data mining, noise is referred to as ________.
A. repeated data.
B. complex data.
C. meta data.
D. random errors in database.
ANSWER: D

98. The technique of learning by generalizing from examples is ________.
A. incremental learning.
B. inductive learning.
C. hybrid learning.
D. generalized learning.
ANSWER: B

99. The _______ plays an important role in artificial intelligence.
A. programming skill.
B. scheduling.
C. planning.
D. learning capabilities.
ANSWER: D

100. Data mining is used to refer ______ stage in knowledge discovery in database.
A. selection.
B. retrieving.
C. discovery.
D. coding.
ANSWER: C

101. ______ could generate rule automatically.
A. KDD.
B. machine learning.
C. artificial intelligence.
D. expert system.
ANSWER: B

102. A good introduction to machine learning is the idea of ______.
A. concept learning.
B. content learning.
C. theory of falsification.

D. Poppers law.
ANSWER: A

103. The algorithms that are controlled by human during their execution is _______ algorithm.
A. unsupervised.
B. supervised.
C. batch learning.
D. incremental.
ANSWER: B

104. Background knowledge depends on the form of ______________.
A. theoretical knowledge.
B. hypothesis.
C. formulae.
D. knowledge representation.
ANSWER: D

ADVANCED QUESTIONS

105. Dimensionality reduction reduces the data set size by removing ____________.
A. relevant attributes.
B. irrelevant attributes.
C. derived attributes.
D. composite attributes.
ANSWER: B

106. 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.
ANSWER: C

107. Multidimensional database is otherwise known as____________.
A. RDBMS
B. DBMS
C. EXTENDED RDBMS
D. EXTENDED DBMS
ANSWER: B

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

109. 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.

ANSWER: A

110. 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.
ANSWER: C

111. Rule based classification algorithms generate ______ rule to perform the classification.
A. if-then.
B. while.
C. do while.
D. switch.
ANSWER: A

112. ___________ training may be used when a clear link between input data sets and target output valuesdoes not exist.
A. Competitive.
B. Perception.
C. Supervised.
D. Unsupervised.
ANSWER: D

113. Web content mining describes the discovery of useful information from the _______contents.
A. text.
B. web.
C. page.
D. level.
ANSWER: B

114. Research on mining multi-types of data is termed as _______ data.
A. graphics.
B. multimedia.
C. meta.
D. digital.
ANSWER: B

115. _________ is the way of studying the web link structure.
A. Computer network.
B. Physical network.
C. Social network.
D. Logical network.
ANSWER: C

116. In web mining, _______ is used to find natural groupings of users, pages, etc.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: A

117. In web mining, _________ is used to know which URLs tend to be requested together.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: B

118. The ___________ engine for a data warehouse supports query-triggered usage of data
A. NNTP
B. SMTP
C. OLAP
D. POP
ANSWER: C

119. ________ 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
ANSWER: B

120. Which of the following are the important qualities of good learning algorithm.
A. Consistent, Complete.
B. Information content, Complex.
C. Complete, Complex.
D. Transparent, Complex.
ANSWER: A

TOPIC TWO – GETTING TO KNOW YOUR DATA

EASY QUESTIONS

121. The _______ is a symbolic representation of facts or ideas from which information can potentially be extracted.
A. knowledge.
B. data.
C. algorithm.
D. program.
ANSWER: B

122. A collection of interesting and useful patterns in database is called _______.
A. knowledge.
B. information.
C. data.
D. algorithm.
ANSWER: A

123. 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.
ANSWER: C

124. The process of finding the right formal representing of a certain body of knowledge in order to represent it inknowledge based system is___________.
A. re-engineering.
B. replication.
C. knowledge engineering.
D. reverse engineering.
ANSWER: C

125. OR methods deals with ________type of data.
A. quantitative.
B. qualitative.
C. standard.
D. predict.
ANSWER: A

126. _________analysis divides data into groups that are meaningful, useful, or both.
A. Cluster.
B. Association.
C. Classifiction.
D. Relation.
ANSWER: A

127. A representation of data objects as columns and attributes as rows is called__________.
A. matrix.
B. data matrix.
C. table.
D. file.
ANSWER: B

128. Which of the following is not a data mining attribute?
A. nominal.
B. ordinal.
C. interval.
D. multiple.
ANSWER: D

129. Patterns of machine-language program are__________.
A. definitive theories.
B. hypothesis.
C. not-definitive theories.
D. quantitative.
ANSWER: B

130. Nominal and ordinal attributes are collectively referred to as__________ attributes.
A. qualitative.
B. perfect.
C. consistent.
D. optimized.
ANSWER: A

131. A data set can often be viewed as a collection of _______.
A. data mart.
B. data.

C. data object.
D. template.
ANSWER: C

132. An important element in machine learning is _________.
A. flow.
B. knowledge.
C. observation.
D. language.
ANSWER: C

133. ________ is the closeness of repeated measurements to one another.
A. Precision.
B. Bias.
C. Accuracy.
D. non-scientific.
ANSWER: A
ANSWER: B

134. Which of the following is not a data mining attribute?
A. nominal.
B. ordinal.
C. interval.
D. multiple.
ANSWER: D

135. Patterns of machine-language program are__________.
A. definitive theories.
B. hypothesis.
C. not-definitive theories.
D. quantitative.
ANSWER: B

136. Nominal and ordinal attributes are collectively referred to as__________ attributes.
A. qualitative.
B. perfect.
C. consistent.
D. optimized.
ANSWER: A

137. A data set can often be viewed as a collection of _______.
A. data mart.
B. data.
C. data object.
D. template.
ANSWER: C

138. An important element in machine learning is _________.
A. flow.
B. knowledge.
C. observation.
D. language.
ANSWER: C

139. ____________ is used for discrete target variable.
A. Nominal.
B. Classification.
C. Clustering.
D. Association.
ANSWER: B

140. A goal of data mining includes which of the following?
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
ANSWER: A

141. is a subject-oriented, integrated, time-variant, nonvolatile collection of data in supportof management decisions.
A. Data Mining.
B. Data Warehousing.
C. Web Mining.
D. Text Mining.
ANSWER: B

142. Collection, analysis, interpretation or explanation of data.
A. Statistics
B. Information retrieval
C. Data mining
D. Cluster analysis
Answer: A

143. Data objects represesents
A. Values
B. Entity
C. Data
D. Attributes
Answer : B

INTERMEDIATE QUESTIONS

144. The term that is not associated with data cleaning process is _______.
A. domain consistance.
B. de-duplication.
C. disambiguation.
D. segmentation.
ANSWER: D

The ______ is a useful method of discovering patterns at the beginning of data mining process.
A. calculating distance.
B. visualization techniques.
C. decision trees.
D. association rules.
ANSWER: B

145. Data mining methodology states that in optimal situation data mining is an ______.
A. standard process.
B. complete process.
C. creative process.
D. ongoing process.
ANSWER: D

146. ____________ is a knowledge discovery process.
A. Data cleaning.
B. Data warehousing.
C. Data mining.
D. Data transformation.
ANSWER: A

147. OLAP is used for ___________.
A. online application processing.
B. online analytical processing.
C. online aptitude processing.
D. online administration and processing.
ANSWER: B

148. Which of the following is not an issue related to concept learning
A. Supervised learning.
B. Unsupervised learning.
C. Self learning.
D. Concept learning.
ANSWER: D

149. Removing duplicate records is a process called_________.
A. recovery.
B. data cleaning.
C. data cleansing.
D. data pruning.
ANSWER: B

150. Data marts that incorporate data mining tools to extract sets of data is called_______.
A. independent data mart.
B. dependent data marts.
C. intra-entry data mart.
D. inter-entry data mart.
ANSWER: B

151. The problem of finding hidden structure in unlabelled data is called…
A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning
D. Semisupervised learning
ANSWER : B

152. Task of inferring a model from labelled training data is called
A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning

D. Semisupervised learning
ANSWER : B

153. Self-organizing maps are an example of…
A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning
D. Missing data imputation
ANSWER : A

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

155. Classification rules are extracted from
A. root node
B. decision tree.
C. siblings.
D. branches.
ANSWER: B

156. Which one of the following is not a part of empirical cycle in scientific research?
A. Observation
B. Theory.
C. Self learning.
D. Prediction.
ANSWER: C

157. In machine learning _________ phase try to find the patterns from observations.
A. observation
B. theory
C. analysis
D. prediction
ANSWER: C

158. ANSWER: D
Data warehouse architecture is based on _______________.
A. DBMS.
B. RDBMS.
C. Sybase.
D. SQL Server.
ANSWER: B

ADVANCED QUESTIONS

159. The ___ algorithm can be applied in cleaning data.
A. search.
B. pattern recognition.
C. learning.
D. clustering.
ANSWER: B

160. ________ is the type of pollution that is difficult to trace.
A. Duplication of records.
B. Ambiguition.
C. Lack of domain consistency.
D. Lack of information.
ANSWER: C

161. The statement that is true about data mining is ______.
A. data mining is not a single technique.
B. it finds the hidden patterns from data set.
C. it is a real discovery process.
D. all forms of pollutions are found during the data mining stage itself.
ANSWER: D

162. The first step in data mining project is ________.
A. rough analysis of data set using traditional query tools.
B. cleaning the data.
C. recognizing the patterns.
D. visualizing the patterns.
ANSWER: A

163. SQL can find ________ type of data.
A. narrow data.
B. multidimensional data.
C. shallow data.
D. hidden data.
ANSWER: C

164. _______ is used to find relationship between multidimensional data.
A. K-nearest neighbor.
B. Decision trees.
C. Association rules.
D. OLAP tools.
ANSWER: D

165. Which one of the following is not true about OLAP?
A. They create no new knowledge.
B. OLAP is powerful that data mining tool.
C. They cannot search for new solution.
D. OLAP tool store their data in special multidimensional format.
ANSWER: B

166. Genetic algorithm is viewed as a kind of______.
A. meta learning strategy.
B. machine learning.
C. evolution.
D. OLAP tool.
ANSWER: A

167. The _________is a knowledge that can be found by using pattern recognition algorithm.
A. hidden knowledge.
B. deep.
C. shallow.

D. multidimensional.
ANSWER: A

168. Shannons notation of information content of message is_______.
A. Log 1divided by n equals log n.
B. log n equals log 1divided by n.
C. log 1divided by n equals minus log n.
D. log minus n =log 1divided by n.
ANSWER: C

169. Which of the following features usually applies to data in a data warehouse
A. Data are often deleted.
B. Most applications consist of transactions.
C. Data are rarely deleted.
D. Relatively few records are processed by applications.
ANSWER: C

170. Which of the following is true
A. The data warehouse consists of data marts and operational data
B. The Data Warehouse consists of data marts and application data.
C. The Data Warehouse is used as a source for the operational data.
D. The operational data are used as a source for the data warehouse
ANSWER: D

171. How do you better define a data warehouse as
A. Can be updated by end users.
B. Contains numerous naming conventions and formats.
C. Organized around important subject areas.
D. Contains only current data.
ANSWER: C

172. Which of the following is an operational system
A. A system that is used to run the business in real time and is based on historical data
B. A system that is used to run the business in real time and is based on current data.
C. A system that is used to support decision making and is based on current data.
D. A system that is used to support decision making and is based on historical data.
ANSWER: B

173. 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 off-time updates.
D. historic data.
ANSWER: B

174. Which of the following is reconciled data
A. Current data intended to be the single source for all decision support systems
B. Data stored in the various operational systems throughout the organization.
C. Data stored in one operational system in the organization.
D. Data that has been selected and formatted for end-user support applications.
ANSWER: A

175. Which of the following is an extract process
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.
ANSWER: B

176. Which of the following is the not a types of clustering?
A. K-means.
B. Hiearachical.
C. Partitional.
D. Splitting.
ANSWER: D

177. 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.
ANSWER: A

178. The ______________ is called a multi field transformation.
A. conversion of data from one field into multiple fields.
B. conversion of data from fields into field.
C. conversion of data from double fields into multiple fields
D. conversion of data from one field to one field.
ANSWER: A

179. Which of the given technology is not well-suited for data mining
A. Expert system technology.
B. Data visualization.
C. Technology limited to specific data types such as numeric data types.
D. Parallel architecture.
ANSWER: C

180. What is true about the multidimensional model?
A. It typically requires less disk storage.
B. It typically requires more disk storage.
C. Typical business queries requiring aggregate functions take more time.
D. Typical business queries requiring aggregate functions take more time.
ANSWER: B

181. Which of the following function involves data cleaning, data standardizing and summarizing
A. Storing data.
B. Transforming data.
C. Data acquisition.
D. Data Access.
ANSWER: B

182. Which of the following problems bog down the development of data mining projects
A. Financial problem.
B. Lack of technical assistance.
C. Lack of long-term vision.

D. Legal and privacy restrictions.
ANSWER: C

183. ________ is the closeness of repeated measurements to one another.
A. Precision.
B. Bias.
C. Accuracy.
D. non-scientific.
ANSWER: A

184. Which of the following matrix consist asymmetric data?
A. Sparse data matrix.
B. Indentity matrix.
C. Confusion matrix.
D. Cross matrix.
ANSWER: A

185. Which of the following matrix consist asymmetric data?
A. Sparse data matrix.
B. Indentity matrix.
C. Confusion matrix.
D. Cross matrix.
ANSWER: A

186. You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is an example of
Supervised learning
Unsupervised learning
Serration
Dimensionality reduction
ANSWER: A

187. Algoritm is
A. It uses machine-learning technique. Here a program can learn from past experience.
B. Computational procedure that takes some values as input and procedure takes some value as output
C. Science of making machines perform tasks that would require intelligence when performed by humans
D. Processing procedure
ANSWER: A

188. The information on two attributes is displayed in _____________ in scatter diagram.
A. visualization space.
B. scatter space.
C. cartesian space.
D. interactive space.
ANSWER: C

189. K-nearest neighbor is one of the ________.
A. learning technique.
B. OLAP tool.
C. purest search technique.
D. data warehousing tool.

ANSWER: C

190. In K- nearest neighbor the input is translated to ___________.
A. values
B. points in multidimensional space
C. strings of characters
D. nodes
ANSWER: B

191. What is a tag cloud?
A. Is a visualization of statistics of user-preferred order.
B. Collection of data objects.
C. Data analysis
D. Data mining application
Answer: A

192. Analysis of variance is a statistical method of comparing the _________ of several populations.
A. standard deviations
B. variances
C. means
D. proportions
Answer: A

193. _________________ is the specialized data warehouse database.
A. Oracle.
B. DBZ.
C. Informix.
D. Redbrick.
ANSWER: D

194. The source of all data warehouse data is the_____________.
A. operational environment.
B. informal environment.
C. formal environment.
D. technology environment.
ANSWER: A

195. Which of the following is a descriptive model?
A. Classification.
B. Regression.
C. Sequence discovery.
D. Association rules.
ANSWER: C

196. A ____________ model identifies patterns or relationships.
A. Descriptive.
B. Predictive.
C. Regression.
D. Time series analysis.
ANSWER: A

Dr.G.R.Damodaran College of Science

(Autonomous, affiliated to the Bharathiar University, recognized by the UGC)Re-accredited at the 'A' Grade Level by the **NAAC** and ISO 9001:2008 Certified
CRISL rated 'A' (TN) for MBA and MIB Programmes

II M.Sc(IT) [2012-2014]
Semester III
Core: Data Warehousing and Mining - 363U1
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.
ANSWER: B

2. The data Warehouse is__________.
A. read only.
B. write only.
C. read write only.
D. none.
ANSWER: A

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

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.
ANSWER: D

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

6. The data is stored, retrieved & updated in ____________.
A. OLAP.
B. OLTP.
C. SMTP.

D. FTP.
ANSWER: B

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

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

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.
ANSWER: B

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

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.
ANSWER: C

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.
ANSWER: B

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.
ANSWER: A

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.
ANSWER: A

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.
ANSWER: A

16. . ______________databases are owned by particular departments or business groups.
A. Informational.
B. Operational.
C. Both informational and operational.
D. Flat.
ANSWER: B

17. The star schema is composed of __________ fact table.
A. one.
B. two.
C. three.
D. four.
ANSWER: A

18. The time horizon in operational environment is ___________.
A. 30-60 days.
B. 60-90 days.
C. 90-120 days.
D. 120-150 days.
ANSWER: B

19. The key used in operational environment may not have an element of__________.
A. time.
B. cost.
C. frequency.
D. quality.
ANSWER: A

20. Data can be updated in _____environment.
A. data warehouse.
B. data mining.
C. operational.
D. informational.
ANSWER: C

21. Record cannot be updated in _____________.
A. OLTP
B. files
C. RDBMS

D. data warehouse
ANSWER: D

22. The source of all data warehouse data is the____________.
A. operational environment.
B. informal environment.
C. formal environment.
D. technology environment.
ANSWER: A

23. Data warehouse contains_____________data that is never found in the operational environment.
A. normalized.
B. informational.
C. summary.
D. denormalized.
ANSWER: C

24. Data redundancy between the environments results in less than ____________percent.
A. one.
B. two.
C. three.
D. four.
ANSWER: A

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.
ANSWER: D

26. Detail data in single fact table is otherwise known as__________.
A. monoatomic data.
B. diatomic data.
C. atomic data.
D. multiatomic data.
ANSWER: C

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

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

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.
ANSWER: C

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.
ANSWER: C

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.
ANSWER: B

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.
ANSWER: C

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.
ANSWER: D

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.
ANSWER: B

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.
ANSWER: A

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.
ANSWER: B

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
ANSWER: D

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.
ANSWER: B

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.
ANSWER: A

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.
ANSWER: A

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.
ANSWER: C

42. Fact tables are ___________.
A. completely demoralized.
B. partially demoralized.
C. completely normalized.
D. partially normalized.
ANSWER: C

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.
ANSWER: A

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.
ANSWER: D

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.
ANSWER: D

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.
ANSWER: A

47. ________ are responsible for running queries and reports against data warehouse tables.
A. Hardware.
B. Software.
C. End users.
D. Middle ware.
ANSWER: C

48. Query tool is meant for __________.
A. data acquisition.
B. information delivery.
C. information exchange.
D. communication.
ANSWER: A

49. Classification rules are extracted from _____________.
A. root node.
B. decision tree.
C. siblings.
D. branches.
ANSWER: B

50. Dimensionality reduction reduces the data set size by removing ____________.
A. relevant attributes.
B. irrelevant attributes.
C. derived attributes.
D. composite attributes.
ANSWER: B

51. ___________ is a method of incremental conceptual clustering.
A. CORBA.
B. OLAP.
C. COBWEB.
D. STING.
ANSWER: C

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.
 ANSWER: A

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.
 ANSWER: C

54. Maintenance of cache consistency is the limitation of __________________.
   A. NUMA.
   B. UNAM.
   C. MPP.
   D. PMP.
 ANSWER: C

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

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

57. ________________ is a data transformation process.
   A. Comparison.
   B. Projection.
   C. Selection.
   D. Filtering.
 ANSWER: D

58. The technology area associated with CRM is _______________.
   A. specialization.
   B. generalization.
   C. personalization.
   D. summarization.
 ANSWER: C

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

D. Symmetric Microprogramming.
ANSWER: A

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.
ANSWER: C

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.
ANSWER: C

62. MDDB stands for ___________.
A. multiple data doubling.
B. multidimensional databases.
C. multiple double dimension.
D. multi-dimension doubling.
ANSWER: B

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

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

65. EIS stands for ______________.
A. Extended interface system.
B. Executive interface system.
C. Executive information system.
D. Extendable information system.
ANSWER: C

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

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

68. The term that is not associated with data cleaning process is ______.
A. domain consistency.
B. deduplication.
C. disambiguation.
D. segmentation.
ANSWER: D

69. _____________ are some popular OLAP tools.
A. Metacube, Informix.
B. Oracle Express, Essbase.
C. HOLAP.
D. MOLAP.
ANSWER: A

70. Capability of data mining is to build ____________ models.
A. retrospective.
B. interrogative.
C. predictive.
D. imperative.
ANSWER: C

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

72. Strategic value of data mining is _______________.
A. cost-sensitive.
B. work-sensitive.
C. time-sensitive.
D. technical-sensitive.
ANSWER: C

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

74. The terms equality and roll up are associated with _____________.
A. OLAP.
B. visualization.
C. data mart.
D. decision tree.

ANSWER: C

75. Exceptional reporting in data warehousing is otherwise called as __________.
A. exception.
B. alerts.
C. errors.
D. bugs.
ANSWER: B

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

77. ________________ is an expensive process in building an expert system.
A. Analysis.
B. Study.
C. Design.
D. Information collection.
ANSWER: D

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

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

80. Removing duplicate records is a process called _____________.
A. recovery.
B. data cleaning.
C. data cleansing.
D. data pruning.
ANSWER: B

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.
ANSWER: A

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.
ANSWER: B

83. Discovery of cross-sales opportunities is called ________________.
A. segmentation.
B. visualization.
C. correction.
D. association.
ANSWER: D

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.
ANSWER: B

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.
ANSWER: D

86. The power of self-learning system lies in __________.
A. cost.
B. speed.
C. accuracy.
D. simplicity.
ANSWER: C

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.
ANSWER: A

88. How many components are there in a data warehouse?
A. two.
B. three.
C. four.
D. five.
ANSWER: D

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.
ANSWER: D

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

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

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.
ANSWER: C

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.
ANSWER: D

94. Which of the following is not a old detail storage medium?
A. Phot Optical Storage.
B. RAID.
C. Microfinche.
D. Pen drive.
ANSWER: D

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.
ANSWER: A

96. The data in current detail level resides till ________ event occurs.
A. purge.
B. summarization.
C. archieved.
D. all of the above.
ANSWER: D

97. The dimension tables describe the _________.
A. entities.
B. facts.
C. keys.

D. units of measures.
ANSWER: B

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.
ANSWER: C

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.
ANSWER: B

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.
ANSWER: C

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.
ANSWER: C

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.
ANSWER: A

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.
ANSWER: B

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.
ANSWER: C

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.
ANSWER: A

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.
ANSWER: B

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.
ANSWER: C

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.
ANSWER: D

109. Which of the following is a predictive model?
A. Clustering.
B. Regression.
C. Summarization.
D. Association rules.
ANSWER: B

110. Which of the following is a descriptive model?
A. Classification.
B. Regression.
C. Sequence discovery.
D. Association rules.
ANSWER: C

111. A ___________ model identifies patterns or relationships.
A. Descriptive.
B. Predictive.
C. Regression.
D. Time series analysis.
ANSWER: A

112. A predictive model makes use of ________.
A. current data.
B. historical data.
C. both current and historical data.
D. assumptions.

ANSWER: B

113. ____________ maps data into predefined groups.
A. Regression.
B. Time series analysis
C. Prediction.
D. Classification.
ANSWER: D

114. __________ is used to map a data item to a real valued prediction variable.
A. Regression.
B. Time series analysis.
C. Prediction.
D. Classification.
ANSWER: B

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.
ANSWER: B

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

117. Link Analysis is otherwise called as ___________.
A. affinity analysis.
B. association rules.
C. both A & B.
D. Prediction.
ANSWER: C

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

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

120. The KDD process consists of ________ steps.
A. three.
B. four.

C. five.
D. six.
ANSWER: C

121. Treating incorrect or missing data is called as ___________.
A. selection.
B. preprocessing.
C. transformation.
D. interpretation.
ANSWER: B

122. Converting data from different sources into a common format for processing is called as ________.
A. selection.
B. preprocessing.
C. transformation.
D. interpretation.
ANSWER: C

123. Various visualization techniques are used in ___________ step of KDD.
A. selection.
B. transformaion.
C. data mining.
D. interpretation.
ANSWER: D

124. Extreme values that occur infrequently are called as _________.
A. outliers.
B. rare values.
C. dimensionality reduction.
D. All of the above.
ANSWER: A

125. Box plot and scatter diagram techniques are _______.
A. Graphical.
B. Geometric.
C. Icon-based.
D. Pixel-based.
ANSWER: B

126. __________ is used to proceed from very specific knowledge to more general information.
A. Induction.
B. Compression.
C. Approximation.
D. Substitution.
ANSWER: A

127. Describing some characteristics of a set of data by a general model is viewed as ____________
A. Induction.
B. Compression.
C. Approximation.
D. Summarization.
ANSWER: B

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

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

129. _______ are needed to identify training data and desired results.
A. Programmers.
B. Designers.
C. Users.
D. Administrators.
ANSWER: C

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.
ANSWER: B

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.
ANSWER: D

132. Incorrect or invalid data is known as _________.
A. changing data.
B. noisy data.
C. outliers.
D. missing data.
ANSWER: B

133. ROI is an acronym of ________.
A. Return on Investment.
B. Return on Information.
C. Repetition of Information.
D. Runtime of Instruction
ANSWER: A

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.
ANSWER: B

135. ___________ data are noisy and have many missing attribute values.
A. Preprocessed.
B. Cleaned.
C. Real-world.
D. Transformed.

ANSWER: C

136. The rise of DBMS occurred in early ___________.
A. 1950's.
B. 1960's
C. 1970's
D. 1980's.
ANSWER: C

137. SQL stand for _________.
A. Standard Query Language.
B. Structured Query Language.
C. Standard Quick List.
D. Structured Query list.
ANSWER: B

138. Which of the following is not a data mining metric?
A. Space complexity.
B. Time complexity.
C. ROI.
D. All of the above.
ANSWER: D

139. Reducing the number of attributes to solve the high dimensionality problem is called as ________.
A. dimensionality curse.
B. dimensionality reduction.
C. cleaning.
D. Overfitting.
ANSWER: B

140. 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.
ANSWER: C

141. ______ are effective tools to attack the scalability problem.
A. Sampling.
B. Parallelization
C. Both A & B.
D. None of the above.
ANSWER: C

142. Market-basket problem was formulated by __________.
A. Agrawal et al.
B. Steve et al.
C. Toda et al.
D. Simon et al.
ANSWER: A

143. Data mining helps in __________.
A. inventory management.
B. sales promotion strategies.

C. marketing strategies.
D. All of the above.
ANSWER: D

144. The proportion of transaction supporting X in T is called _________.
A. confidence.
B. support.
C. support count.
D. All of the above.
ANSWER: B

145. The absolute number of transactions supporting X in T is called ___________.
A. confidence.
B. support.
C. support count.
D. None of the above.
ANSWER: C

146. 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.
ANSWER: A

147. 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%
ANSWER: A

148. 7 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%
ANSWER: D

149. The left hand side of an association rule is called __________.
A. consequent.
B. onset.
C. antecedent.
D. precedent.
ANSWER: C

150. The right hand side of an association rule is called _____.
A. consequent.
B. onset.
C. antecedent.
D. precedent.

ANSWER: A

151. 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.
ANSWER: D

152. 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.
ANSWER: B

153. 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.
ANSWER: A

154. 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.
ANSWER: B

155. Any superset of an infrequent set is an infrequent set. This is _______.
A. Maximal frequent set.
B. Border set.
C. Upward closure property.
D. Downward closure property.
ANSWER: C

156. If an itemset is not a frequent set and no superset of this is a frequent set, then it is _______.
A. Maximal frequent set
B. Border set.
C. Upward closure property.
D. Downward closure property.
ANSWER: B

157. A priori algorithm is otherwise called as __________.
A. width-wise algorithm.
B. level-wise algorithm.
C. pincer-search algorithm.
D. FP growth algorithm.
ANSWER: B

158. The A Priori algorithm is a ___________.
A. top-down search.

B. breadth first search.
C. depth first search.
D. bottom-up search.
ANSWER: D

159. The first phase of A Priori algorithm is _______.
A. Candidate generation.
B. Itemset generation.
C. Pruning.
D. Partitioning.
ANSWER: A

160. The second phaase of A Priori algorithm is ____________.
A. Candidate generation.
B. Itemset generation.
C. Pruning.
D. Partitioning.
ANSWER: C

161. 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.
ANSWER: B

162. The a priori frequent itemset discovery algorithm moves _______ in the lattice.
A. upward.
B. downward.
C. breadthwise.
D. both upward and downward.
ANSWER: A

163. 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.
ANSWER: B

164. 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.
ANSWER: A

165. MFCS is the acronym of _____.
A. Maximum Frequency Control Set.
B. Minimal Frequency Control Set.
C. Maximal Frequent Candidate Set.
D. Minimal Frequent Candidate Set.
ANSWER: C

166. Dynamuc Itemset Counting Algorithm was proposed by ____.
A. Bin et al.
B. Argawal et at.
C. Toda et al.
D. Simon et at.
ANSWER: A

167. Itemsets in the ______ category of structures have a counter and the stop number with them.
A. Dashed.
B. Circle.
C. Box.
D. Solid.
ANSWER: A

168. The itemsets in the _______category structures are not subjected to any counting.
A. Dashes.
B. Box.
C. Solid.
D. Circle.
ANSWER: C

169. Certain itemsets in the dashed circle whose support count reach support value during an iteration move into the ______.
A. Dashed box.
B. Solid circle.
C. Solid box.
D. None of the above.
ANSWER: A

170. Certain itemsets enter afresh into the system and get into the _______, which are essentially the supersets of the itemsets that move from the dashed circle to the dashed box.
A. Dashed box.
B. Solid circle.
C. Solid box.
D. Dashed circle.
ANSWER: D

171. The itemsets that have completed on full pass move from dashed circle to ________.
A. Dashed box.
B. Solid circle.
C. Solid box.
D. None of the above.
ANSWER: B

172. The FP-growth algorithm has ________ phases.
A. one.
B. two.
C. three.
D. four.
ANSWER: B

173. A frequent pattern tree is a tree structure consisting of ________.
A. an item-prefix-tree.

B. a frequent-item-header table.
C. a frequent-item-node.
D. both A & B.
ANSWER: D

174. The non-root node of item-prefix-tree consists of ________ fields.
A. two.
B. three.
C. four.
D. five.
ANSWER: B

175. The frequent-item-header-table consists of __________ fields.
A. only one.
B. two.
C. three.
D. four.
ANSWER: B

176. The paths from root node to the nodes labelled 'a' are called __________.
A. transformed prefix path.
B. suffix subpath.
C. transformed suffix path.
D. prefix subpath.
ANSWER: D

177. The transformed prefix paths of a node 'a' form a truncated database of pattern which co-occur with a is called _______.
A. suffix path.
B. FP-tree.
C. conditional pattern base.
D. prefix path.
ANSWER: C

178. 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.
ANSWER: C

179. Which of the following is a clustering algorithm?
A. A priori.
B. CLARA.
C. Pincer-Search.
D. FP-growth.
ANSWER: B

180. _______ 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.

ANSWER: A

181. __________ 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.
ANSWER: B

182. Which of the following is a data set in the popular UCI machine-learning repository?
A. CLARA.
B. CACTUS.
C. STIRR.
D. MUSHROOM.
ANSWER: D

183. In ________ algorithm each cluster is represented by the center of gravity of the cluster.
A. k-medoid.
B. k-means.
C. STIRR.
D. ROCK.
ANSWER: B

184. 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.
ANSWER: A

185. Pick out a k-medoid algoithm.
A. DBSCAN.
B. BIRCH.
C. PAM.
D. CURE.
ANSWER: C

186. Pick out a hierarchical clustering algorithm.
A. DBSCAN
B. BIRCH.
C. PAM.
D. CURE.
ANSWER: A

187. 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.
ANSWER: C

188. BIRCH is a ________.

A. agglomerative clustering algorithm.
B. hierarchical algorithm.
C. hierarchical-agglomerative algorithm.
D. divisive.
ANSWER: C

189. 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.
ANSWER: A

190. 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.
ANSWER: D

191. The partition algorithm uses _______ scans of the databases to discover all frequent sets.
A. two.
B. four.
C. six.
D. eight.
ANSWER: A

192. 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.
ANSWER: A

193. ________is the most well known association rule algorithm and is used in most commercial products.
A. Apriori algorithm.
B. Partition algorithm.
C. Distributed algorithm.
D. Pincer-search algorithm.
ANSWER: A

194. An algorithm called________is used to generate the candidate item sets for each pass after the first.
A. apriori.
B. apriori-gen.
C. sampling.
D. partition.
ANSWER: B

195. The basic partition algorithm reduces the number of database scans to ________ & divides it into partitions.

A. one.
B. two.
C. three.
D. four.
ANSWER: B

196. ___________and prediction may be viewed as types of classification.
A. Decision.
B. Verification.
C. Estimation.
D. Illustration.
ANSWER: C

197. ___________can be thought of as classifying an attribute value into one of a set of possible classes.
A. Estimation.
B. Prediction.
C. Identification.
D. Clarification.
ANSWER: B

198. Prediction can be viewed as forecasting a_________value.
A. non-continuous.
B. constant.
C. continuous.
D. variable.
ANSWER: C

199. _________data consists of sample input data as well as the classification assignment for the data.
A. Missing.
B. Measuring.
C. Non-training.
D. Training.
ANSWER: D

200. Rule based classification algorithms generate ______ rule to perform the classification.
A. if-then.
B. while.
C. do while.
D. switch.
ANSWER: A

201. ____________ are a different paradigm for computing which draws its inspiration from neuroscience.
A. Computer networks.
B. Neural networks.
C. Mobile networks.
D. Artificial networks.
ANSWER: B

202. The human brain consists of a network of ___________.
A. neurons.
B. cells.
C. Tissue.

D. muscles.
ANSWER: A

203. Each neuron is made up of a number of nerve fibres called _____________.
A. electrons.
B. molecules.
C. atoms.
D. dendrites.
ANSWER: D

204. The ___________is a long, single fibre that originates from the cell body.
A. axon.
B. neuron.
C. dendrites.
D. strands.
ANSWER: A

205. A single axon makes ___________ of synapses with other neurons.
A. ones.
B. hundreds.
C. thousands.
D. millions.
ANSWER: C

206. _____________ is a complex chemical process in neural networks.
A. Receiving process.
B. Sending process.
C. Transmission process.
D. Switching process.
ANSWER: C

207. _________ is the connectivity of the neuron that give simple devices their real power. a. b. c. d.
A. Water.
B. Air.
C. Power.
D. Fire.
ANSWER: D

208. __________ are highly simplified models of biological neurons.
A. Artificial neurons.
B. Computational neurons.
C. Biological neurons.
D. Technological neurons.
ANSWER: A

209. The biological neuron's _________ is a continuous function rather than a step function.
A. read.
B. write.
C. output.
D. input.
ANSWER: C

210. The threshold function is replaced by continuous functions called ________ functions.
A. activation.

B. deactivation.
C. dynamic.
D. standard.
ANSWER: A

211. The sigmoid function also knows as __________functions.
A. regression.
B. logistic.
C. probability.
D. neural.
ANSWER: B

212. MLP stands for ______________________.
A. mono layer perception.
B. many layer perception.
C. more layer perception.
D. multi layer perception.
ANSWER: D

213. In a feed- forward networks, the conncetions between layers are ___________ from input to output.
A. bidirectional.
B. unidirectional.
C. multidirectional.
D. directional.
ANSWER: B

214. The network topology is constrained to be __________________.
A. feedforward.
B. feedbackward.
C. feed free.
D. feed busy.
ANSWER: A

215. RBF stands for _____________.
A. Radial basis function.
B. Radial bio function.
C. Radial big function.
D. Radial bi function.
ANSWER: A

216. RBF have only _______________ hidden layer.
A. four.
B. three.
C. two.
D. one.
ANSWER: D

217. RBF hidden layer units have a receptive field which has a ____________; that is, a particular input value at which they have a maximal output.
A. top.
B. bottom.
C. centre.
D. border.

ANSWER: C

218. ___________ training may be used when a clear link between input data sets and target output values does not exist.
A. Competitive.
B. Perception.
C. Supervised.
D. Unsupervised.
ANSWER: D

219. ___________ employs the supervised mode of learning.
A. RBF.
B. MLP.
C. MLP & RBF.
D. ANN.
ANSWER: C

220. ________________ design involves deciding on their centres and the sharpness of their Gaussians.
A. DR.
B. AND.
C. XOR.
D. RBF.
ANSWER: D

221. ___________ is the most widely applied neural network technique.
A. ABC.
B. PLM.
C. LMP.
D. MLP.
ANSWER: D

222. SOM is an acronym of _______________.
A. self-organizing map.
B. self origin map.
C. single organizing map.
D. simple origin map.
ANSWER: A

223. ____________ is one of the most popular models in the unsupervised framework.
A. SOM.
B. SAM.
C. OSM.
D. MSO.
ANSWER: A

224. The actual amount of reduction at each learning step may be guided by _________.
A. learning cost.
B. learning level.
C. learning rate.
D. learning time.
ANSWER: C

225. The SOM was a neural network model developed by ________.
A. Simon King.

B. Teuvokohonen.
C. Tomoki Toda.
D. Julia.
ANSWER: B

226. SOM was developed during ____________.
A. 1970-80.
B. 1980-90.
C. 1990 -60.
D. 1979 -82.
ANSWER: D

227. Investment analysis used in neural networks is to predict the movement of _________ from previous data.
A. engines.
B. stock.
C. patterns.
D. models.
ANSWER: B

228. SOMs are used to cluster a specific _____________ dataset containing information about the patient's drugs etc.
A. physical.
B. logical.
C. medical.
D. technical.
ANSWER: C

229. GA stands for _______________.
A. Genetic algorithm
B. Gene algorithm.
C. General algorithm.
D. Geo algorithm.
ANSWER: A

230. GA was introduced in the year __________.
A. 1955.
B. 1965.
C. 1975.
D. 1985.
ANSWER: C

231. Genetic algorithms are search algorithms based on the mechanics of natural_______.
A. systems.
B. genetics.
C. logistics.
D. statistics.
ANSWER: B

232. GAs were developed in the early _____________.
A. 1970.
B. 1960.
C. 1950.
D. 1940.

ANSWER: A

233. The RSES system was developed in ___________.
A. Poland.
B. Italy.
C. England.
D. America.
ANSWER: A

234. Crossover is used to _______.
A. recombine the population's genetic material.
B. introduce new genetic structures in the population.
C. to modify the population's genetic material.
D. All of the above.
ANSWER: A

235. The mutation operator ______.
A. recombine the population's genetic material.
B. introduce new genetic structures in the population.
C. to modify the population's genetic material.
D. All of the above.
ANSWER: B

236. Which of the following is an operation in genetic algorithm?
A. Inversion.
B. Dominance.
C. Genetic edge recombination.
D. All of the above.
ANSWER: D

237. . ___________ is a system created for rule induction.
A. RBS.
B. CBS.
C. DBS.
D. LERS.
ANSWER: D

238. NLP stands for _________.
A. Non Language Process.
B. Nature Level Program.
C. Natural Language Page.
D. Natural Language Processing.
ANSWER: D

239. Web content mining describes the discovery of useful information from the _______contents.
A. text.
B. web.
C. page.
D. level.
ANSWER: B

240. Research on mining multi-types of data is termed as _______ data.
A. graphics.
B. multimedia.

C. meta.
D. digital.
ANSWER: B

241. _______ mining is concerned with discovering the model underlying the link structures of the web.
A. Data structure.
B. Web structure.
C. Text structure.
D. Image structure.
ANSWER: B

242. _________ is the way of studying the web link structure.
A. Computer network.
B. Physical network.
C. Social network.
D. Logical network.
ANSWER: C

243. The ________ propose a measure of standing a node based on path counting.
A. open web.
B. close web.
C. link web.
D. hidden web.
ANSWER: B

244. In web mining, _______ is used to find natural groupings of users, pages, etc.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: A

245. In web mining, _________ is used to know the order in which URLs tend to be accessed.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: C

246. In web mining, _________ is used to know which URLs tend to be requested together.
A. clustering.
B. associations.
C. sequential analysis.
D. classification.
ANSWER: B

247. __________ describes the discovery of useful information from the web contents.
A. Web content mining.
B. Web structure mining.
C. Web usage mining.
D. All of the above.
ANSWER: A

248. _______ is concerned with discovering the model underlying the link structures of the web.

A. Web content mining.
B. Web structure mining.
C. Web usage mining.
D. All of the above.
ANSWER: B

249. A link is said to be _________ link if it is between pages with different domain names.
A. intrinsic.
B. transverse.
C. direct.
D. contrast.
ANSWER: B

250. A link is said to be _______ link if it is between pages with the same domain name.
A. intrinsic.
B. transverse.
C. direct.
D. contrast.
ANSWER: A

Staff Name
LAXMI.SREE.B.R.

	marks	question	A	B	C	D	ans
0	1	To integrate heterogeneous databases, how many approaches are there in Data Warehousing?	2	3	4	5	Data warehousing involves data cleaning, data integration, and data consolidations. To integrate heterogeneous databases, we have the following two approaches: Query Driven Approach, Update Driven Approach
1	1	___________ refers to the description and model regularities or trends for objects whose behavior changes over time.	Evolution Analysis	Outlier Analysis	Prediction	Classification	Evolution Analysis: Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.
2	1	The mapping or classification of a class with some predefined group or class is known as?	Data Discrimination	Data Characterization	Data Set	Data Sub Structure	Data Discrimination: It refers to the mapping or classification of a class with some predefined group or class
3	1	In which step of Knowledge Discovery, multiple data sources are combined?	Data Integration	Data Cleaning	Data Selection	Data Transformation	Data Integration: multiple data sources are combined.
4	1	What is the strategic value of data mining?	Time-sensitive	Work-sensitive.	Cost-sensitive	Technical-sensitive.	Time-Sensitive is the strategic value of data mining.
5	2	The first step involved in knowledge discovery is?	Data Cleaning	Data Selection	Data Transformation	Data Integration	The first step involved in the knowledge discovery is Data Integration.
6	2	Which of the following is not a data mining functionality?	Selection and interpretation	Classification and regression	Characterization and Discrimination	Clustering and Analysis	Selection and interpretation is not a function of data mining
7	2	In Data Characterization, the class under study is called as?	Target Class	Initial Class	Study Class	Final Class	Data Characterization: This refers to summarizing data of class under study. This class under study is called Target Class.
8	2	Capability of data mining is to build ____________ models.	Predictive.	Interrogative.	Retrospective.	Imperative.	The predictive model has the capability of data mining

	marks	question	A	B	C	D	ans
9	2	"Handling of relational and complex types of data" issue comes under?	Diverse Data Types Issues	Performance Issues	Mining Methodology and User Interaction Issues	None	The database may contain complex data objects, multimedia data objects, spatial data, temporal data, etc. One system can't mine all this kind of data.
10	2	What is true about data mining?	All	Data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation	Data mining is the procedure of mining knowledge from data.	Data Mining is defined as the procedure of extracting information from huge sets of data	Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge is extracted so that it can be used.
11	2	What is KDD	Knowledge Discovery Database	Knowledge Database	Knowledge Data House	Knowledge Data Definition	The KDD stands for Knowledge Discovery Database.
12	2	Which of the following is the correct application of data mining?	All	Corporate Analysis & Risk Management	Fraud Detection	Market Analysis and Management	Data mining is highly useful in the following domains: Market Analysis and Management, Corporate Analysis & Risk Management, Fraud Detection
13	2	Which of the following is not a data mining metric?	All	Time complexity.	ROI	Space complexity.	All of the above are algorithm metrics.
14	2	DMQL stands for?	Data Mining Query Language	Dataset Mining Query Language	DBMiner Query Language	Data Marts Query Language	The Data Mining Query Language (DMQL) was proposed by Han, Fu, Wang, et al. for the DBMiner data mining system.

	marks	question	A	B	C	D	ans
15	2	The analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs is called?	Mining of Correlations	Mining of Clusters	Mining of Association	None	Mining of Correlations: It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative, or no effect on each other.
16	2	What is the use of data cleaning?	All	Correct the inconsistencies in data	Transformations to correct the wrong data.	To remove the noisy data	Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data preprocessing step while preparing the data for a data warehouse.
17	2	"Efficiency and scalability of data mining algorithms" issues come under?	Performance Issues	Mining Methodology and User Interaction Issues	Diverse Data Types Issues	None	In order to effectively extract the information from a huge amount of data in databases, the data mining algorithm must be efficient and scalable.
18	3	Data mining helps in ___________.	All	Sales promotion strategies.	Marketing strategies.	Inventory management.	All are the properties of data mining
19	3	___________ may be defined as the data objects that do not comply with the general behavior or model of the data available.	Outlier Analysis	Evolution Analysis	Prediction	Classification	Outlier Analysis: Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
20	3	……………………….. is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.	Data discrimination	Data Classification	Data Characterization	Data selection	Data discrimination is the feature

	marks	question	A	B	C	D	ans
21	3	A sequence of patterns that occur frequently is known as?	Frequent Subsequence	. Frequent Item Set	Frequent Sub Structure	All of the above	Frequent Subsequence: A sequence of patterns that occur frequently such as purchasing a camera is followed by a memory card.
22	3	-------- is an essential process where intelligent methods are applied to extract data patterns.	Data mining	Data warehousing	Text mining	Data selection	Data mining is an essential process where AI is used.
23	3	Does the pattern evaluation issue come under?	Mining Methodology and User Interaction Issues	Performance Issues	Diverse Data Types Issues	None of the above	Pattern evaluation: The patterns discovered should be interesting because either they represent common knowledge or lack of novelty.
24	3	What predicts future trends & behaviors, allowing business managers to make proactive,knowledge-driven decisions.	Data mining.	Data warehouse.	Datamarts.	Metadata.	Data mining predicts future trends.
25	3	Which of the following is the other name of Data mining?	All	Data-driven discovery.	Deductive learning.	Exploratory data analysis.	All the above are the name of data mining
26	3	How many categories of functions involved in Data Mining?	2	3	4	5	There are two categories of functions involved in Data Mining: 1. Descriptive, 2. Classification and Prediction
27	3	Does Data Mining System Classification consist of?	All	Machine Learning	Information Science	Database Technology	A data mining system can be classified according to the following criteria: Database Technology, Statistics, Machine Learning, Information Science, Visualization, Other Disciplines
28	3	Which of the following is the correct disadvantage of the Query-Driven Approach in Data Warehousing?	All	It is very inefficient and very expensive for frequent queries.	This approach is expensive for queries that require aggregations.	The Query Driven Approach needs complex integration and filtering processes.	All statements are a disadvantage of the Query-Driven Approach in Data Warehousing.
29	3	Which of the following is the correct advantage of the Update-Driven Approach in Data Warehousing?	Both A and B	The data can be copied, processed, integrated, annotated, summarized, and restructured in the semantic data store in advance.	This approach provides high performance.	None	Both A and B are the advantages of the Update-Driven Approach in Data Warehousing.

	marks	question	A	B	C	D	ans
30	1	SELECT item name, color, clothes SIZE, SUM(quantity)\nFROM sales\nGROUP BY rollup(item name, color, clothes SIZE);\nHow many grouping is possible in this rollup?\n	4	8	2	1	{ (item name, color, clothes size), (item name, color), (item name), () }.
31	1	The operation of changing the dimensions used in a cross-tab is called as ________	Pivoting	Alteration	Piloting	Renewing	We can change the dimensions used in a cross tab. The operation of changing a dimension used in a cross-tab is called pivoting.
32	1	OLAP stands for	Online analytical processing	Online analysis processing	Online transaction processing	Online aggregate processing	OLAP is the manipulation of information to support decision making.
33	1	State true or false: In OLAP, analysts cannot view a dimension in different levels of detail.	"False"	"True"	None	None	In OLAP, analysts cannot view a dimension in different levels of detail. The different levels of detail are classified into a hierarchy.
34	1	Data that can be modeled as dimension attributes and measure attributes are called _______ data.	Multidimensional	Singledimensional	Measured	Dimensional	Given a relation used for data analysis, we can identify some of its attributes as measure attributes, since they measure some value, and can be aggregated upon.
35	1	Business Intelligence and data warehousing is used for ________.	All	Data Mining.	Analysis of large volumes of product sales data.	Forecasting	All are used in data ware house
36	1	The operation of moving from coarser granular data to finer granular data is called _______	Drill down	Increment	Rollback	Reduction	OLAP systems permit users to view the data at any level of granularity. The process of moving from finer granular data to coarser granular data is called as drill-down.
37	2	The operation of moving from finer-granularity data to a coarser granularity (using aggregation) is called a ________	Rollup	Drill down	Dicing	Pivoting	The opposite operation—that of moving from coarser-granularity data to finer-granularity data—is called a drill down.

	marks	question	A	B	C	D	ans
38	2	State true or false: OLAP systems can be implemented as client-server systems	"True"	"False"	None	None	OLAP systems can be implemented as client-server systems. Most of the current OLAP systems are implemented as client-server systems.
39	2	Data that can be modelled as dimension attributes and measure attributes are called ___________	Multi-dimensional data	Mono-dimensional data	Measurable data	Efficient data	Data that can be modeled as dimension attributes and measure attributes are called multi-dimensional data.
40	2	The process of viewing the cross-tab (Single dimensional) with a fixed value of one attribute is	Slicing	Dicing	Pivoting	Both Slicing and Dicing	The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Dice selects two or more dimensions from a given cube and provides a new sub-cube.
41	2	The time horizon in Data warehouse is usually __________.	5-10 years.	3-4 years	5-6 years.	1-2 years.	5 to 10 years is the horizon time
42	2	How many dimensions of multi-dimensional data do cross tabs enable analysts to view?	2	1	3	None	Cross-tabs enables analysts to view two dimensions of multi-dimensional data, along with the summaries of the data.
43	2	What do data warehouses support?	OLAP	OLTP	OLAP and OLTP	Operational databases	OLAP support data warehouses
44	2	Data warehouse architecture is based on ______________.	RDBMS	DBMS	Sybase.	SQL Server	RDBMS is the data warehouse architecture.
45	2	What does collector_type_id stands for in the following code snippet? core.sp_remove_collector_type [@collector_type_uid =] ‘collector_type_uid’	uniqueidentifier	membership role	directory	None	collector_type_uid is the GUID for the collector type.
46	2	The generalization of cross-tab which is represented visually is ___________ which is also called as a data cube.	Two-dimensional cube	Multidimensional cube	N-dimensional cube	Cuboid	Each cell in the cube is identified for the values for the three-dimensional attributes.
47	2	The source of all data warehouse data is the____________.	Operational environment.	Informal environment.	Formal environment.	Technology environment	Operational environment is the source of data warehouse

	marks	question	A	B	C	D	ans
48	3	What is the sum of all components of a normalized histogram?	1	-1	0	None	A normalized histogram. p(rk) = nk / n\nWhere, n is total number of pixels in image, rk the kth gray level and nk total pixels with gray level rk.\nHere, p(rk) gives the probability of occurrence of rk.\n
49	3	Which of the following OLAP systems do not exist?	None	MOLAP	ROLAP	HOLAP	HOLAP means Hybrid OLAP, MOLAP means multidimensional OLAP, ROLAP means relational OLAP. This means all of the above OLAP systems exist.
50	3	We want to add the following capabilities to Table2: show the data for 3 age groups (20-39, 40-60, over 60), 3 revenue groups (less than $10,000, $10,000-$30,000, over $30,000) and add a new type of account: Money market. The total number of measures will be:	More than 100	4	Between 10 and 30 (boundaries includeD.	Between 40 and 60 (boundaries includeD.	More than 100 is the capabilities to Table2
51	3	The _______ function allows substitution of values in an attribute of a tuple	Decode	Unknown	Cube	Substitute	The decode function allows substitution of values in an attribute of a tuple. The decode function does not always work as we might like for null values because predicates on null values evaluate to unknown.
52	3	The operation of moving from finer granular data to coarser granular data is called _______	Roll up	Increment	Reduction	Drill down	OLAP systems permit users to view the data at any level of granularity. The process of moving from finer granular data to coarser granular data is called as a roll-up.
53	3	The ___________ engine for a data warehouse supports query-triggered usage of data	OLAP	SMTP	NNTP	POP	OLAP is the engine of data warehouse
54	3	In SQL the cross-tabs are created using	Slice	Dice	Pivot	All	Pivot (sum(quantity) for color in ('dark','pastel','white')).

	marks	question	A	B	C	D	ans
55	3	Which one of the following is the right syntax for DECODE?	DECODE (expression, search, result [, search, result]… [, default])	DECODE (expression, result [, search, result]… [, default], search)	DECODE (search, result [, search, result]… [, default], expression)	DECODE (search, expression, result [, search, result]… [, default])	The right synatax for DECODE is DECODE (expression, search, result [, search, result]… [, default])
56	3	The value at the intersection of the row labeled "India" and the column "Savings" in Table2 should be:	800000	300000	200000	300000	800,000 is value at the intersection of the row labeled "India" and the column "Savings" in Table2
57	3	__________ is the heart of the warehouse.	Data warehouse database servers	Data mining database servers.	Data mart database servers.	Relational data base servers.	The heart of data warehouse is Data warehouse database servers.
58	3	{ (item name, color, clothes size), (item name, color), (item name, clothes size), (color, clothes size), (item name), (color), (clothes size), () }	None	Group by the cubic	Group by	Group by rollup	'Group by cube' is used.
59	3	The data Warehouse is__________.	Read-only.	Write only.	Read write only	None	The data warehouse is read-only
60	1	Cluster analysis is a type of ... ?	Unsupervised data mining	Supervised data mining	Depends on the data	Can not say	Unsupervised data mining is the cluster analysis
61	1	Challenges of clustering includes?	All	Scalability	Noisy data	High dimensionality of data	All are the challenges of clustering
62	1	Which of the following combination is incorrect?	None	Continuous – correlation similarity	Binary – manhattan distance	Continuous – euclidean distance	You should choose a distance/similarity that makes sense for your problem.
63	1	Hierarchical clustering should be primarily used for exploration.	"True"	"False"	None	None	Hierarchical clustering is deterministic.
64	1	In clustering high dimensional data comes with problems like?	All	Reduction of algorithm performance	Reduction in algorithm efficiency	Increase in complexity	All mention are the problems od clustering
65	1	Which of the following clustering requires merging approach?	Hierarchical	Partitional	Naive Bayes	None	Hierarchical clustering requires a defined distance as well.
66	1	Which of the following is required by K-means clustering?	All	Number of clusters	Initial guess as to cluster centroids	Defined distance metric	K-means clustering follows the partitioning approach.
67	1	Which clustering procedure is characterized by the formation of a tree like structure?	Hierarchical clustering	Optimizing partitioning	Partition based clustering	Density clustering	Hierarchical clustering is tree like structure.
68	1	Point out the wrong statement.	k-nearest neighbor is same as k-means	none er	k-means clustering aims to partition n observations into k clusters	k-means clustering is a method of vector quantization	k-nearest neighbor has nothing to do with k-means.
69	2	What is dissimilarity?	Both a and b	A metric that is used to measure the closeness of objects.	A metric that is used in clustering.	None	Dissimilarity means metric used in clustering and closeness of objects.

	marks	question	A	B	C	D	ans
70	2	The most important part of ... is selecting the attributes on which clustering is done?	Formulating the clustering problem	Data preprocessing for clustering	Deciding the clustering procedure	Analysing the cluster	Formulating the clustering problem is the imporatant part of clustering.
71	2	K-means is not deterministic and it also consists of number of iterations.	"True"	"False"	None	None	K-means clustering produces the final estimate of cluster centroids.
72	2	k-means clustering is also referred to as ....?	Non-hierarchical clustering	Optimizing partitioning	Divisive clustering	Agglomerative clustering	Non-hierarchical clustering is called as k-means clustering
73	2	Which is not a type of clustering?	Decision driven	Similarity based	Density based	Partition Based	All other are the type of clustering
74	2	Which of the following is finally produced by Hierarchical Clustering?	Tree showing how close things are to each other	Final estimate of cluster centroids	Assignment of each point to clusters	All	Hierarchical clustering is an agglomerative approach.
75	2	Which of the following is not clustering technique?	Derivative	Agglomerative	Partitioning	Density Based	Derivative is not a clustering technique.
76	2	Which of the following function is used for k-means clustering?	k-means	k-mean	heatmap	None	K-means requires a number of clusters.
77	2	Which of the below sentences is true with respect of clustering?	In clustering, larger the distance the more similar the object	The dendrogram is read from right to left	Clustering should be done on samples of 300 or more	Cluster analysis reduces the number of objects, not the number of variables, by grouping them into a much smaller number of clusters	In clustering, larger the distance the more similar the object is true for clustering.
78	2	Clustering is what type of learning?	Unsupervised	supervised	Semi-supervised	None	Unsupervised is a type of learning
79	2	Point out the correct statement.	All	Hierarchical clustering is also called HCA	In general, the merges and splits are determined in a greedy manner	The choice of an appropriate metric will influence the shape of the clusters	Some elements may be close to one another according to one distance and farther away according to another.
80	2	Hierarchical clustering is slower than non-hierarchical clustering?	"True"	"False"	Depends on data	Can not say	Hierarchical clustering is slower than non-hierarchical clustering
81	3	When does a model is said to do over-fitting?	It does not fit in future state	It does not fit in current state	It does not fit in both current and future state	None	It does not fit in future state is a model.
82	3	What is a cluster?	Group of similar objects with significant dissimilarity with objects of other groups	Group objects having a similar feature from a group of similar objects.	Simplification of data to make it ready for a classification algorithm.	None	The group of similar objects with significant dissimilarity with objects of other groups is called as cluster

	marks	question	A	B	C	D	ans
83	3	Which method of analysis does not classify variables as dependent or independent?	Cluster analysis	Discriminant analysis	Analysis of variance	Regression analysis	Cluster analysis is not classify variables as dependent or independent
84	3	In clustering ?	Groups are not predefined	Groups are predefined	Depends on the data	Can not say	Groups are not predefined in clustering
85	3	Which of the following are clustering techniques?	All	Density Based	Partitioning	Agglomerative	All are the clustering techniques.
86	3	What is clustering?	Process of grouping similar objects	Process of classifying new object	Both a and b	None of the above	Clustering is a group of similar objects
87	3	When is density based clustering preferred?	All	Not sure about the number of clusters present	Noise and outliers are present	Clusters are irregular or intertwined	All are the density based clustering
88	3	In the K-means clustering algorithm the distance between cluster centroid to each object is calculated using ....method.	Euclidean distance	Cluster distance	Cluster width	None	Euclidean distance is the k-means clustering algorithm.
89	1	Which technique finds the frequent itemsets in just two database scans?	Partitioning	Sampling	Hashing	Dynamic itemset counting	Partitioning is technique that finds the frequent itemsets
90	1	What is association rule mining?	Finding of strong association rules using frequent itemsets	Same as frequent itemset mining	Using association to analyse correlation rules	None	Finding of strong association rules using frequent itemsets is an assoication rule.
91	1	An itemsetwhose no proper super-itemset has same support is closed itemsets	An itemset which is both closed and frequent	A frequent itemset	A closed itemsetA closed itemset	None	An itemset which is both closed and frequent are closed frequent itemsets.
92	1	Which of the following is true?	Both apriori and FP-Growth uses horizontal data format	Both apriori and FP-Growth uses vertical data format	Apriori uses horizontal and FP-Growth uses vertical data format	Apriori uses vertical and FP-Growth uses horizontal data format	Both apriori and FP-Growth uses horizontal data format is true
93	1	What will happen if support is reduced?	Some itemsets will add to the current set of frequent itemsets	The number of frequent itemsets remains same	Some itemsets will become infrequent while others will become frequent	Can not say	Support is reduced by some itemsets will add to the current set of frequent itemsets
94	1	How do you calculate Confidence(A -> B)?	Support(A B) / Support (A)	Support(A B) / Support (B)	Support(A B) / Support (A)	Support(A B) / Support (B)	None
95	1	What is the principle on which Apriori algorithm work?	If a rule is infrequent, its specialized rules are also infrequent	If a rule is infrequent, its generalized rules are also infrequent	Both a and b	None	The Apriori algorithm works on if a rule is infrequent, its specialized rules are also infrequent
96	1	What does Apriori algorithm	It mines all frequent patterns through pruning rules with lesser support	It mines all frequent patterns through pruning rules with higher support	Both a and b	None of the above	Apriori algorithm works on It mines all frequent patterns through pruning rules with lesser support

	marks	question	A	B	C	D	ans
97	2	What are maximal frequent itemsets?	A frequent itemsetwhose no super-itemset is frequent	A frequent itemset whose super-itemset is also frequent	A non-frequent itemset whose super-itemset is frequent	None	A frequent itemsetwhose no super-itemset is frequent is maximal frequent itemsets.
98	2	What is not true about FP growth algorithms?	It expands the original database to build FP trees.	There are chances that FP trees may not fit in the memory.	FP trees are very expensive to build .	It mines frequent itemsets without candidate generation.	It expands the original database to build FP trees is not true
99	2	Which of these is not a frequent pattern mining algorithm?	Decision trees	FP growth	Apriori	Eclat	Decision trees is not a frequent pattern mining algorithm
100	2	This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration	K-Means clustering	Conceptual clustering	Expectation maximization	Agglomerative clustering	K-Means clustering is the current iteration of the algorithm.
101	2	Which of the following is not null invariant measure(that does not considers null transactions)?	lift	max_confidence	cosine measure	all_confidence	lift is not null invariant measure
102	2	What is the difference between absolute and relative support?	Absolute - Minimum support count threshold and Relative - Minimum support threshold	Absolute - Minimum support threshold and Relative - Minimum support count threshold	Both mean same	None	None
103	2	Can FP growth algorithm be used if FP tree cannot be fit in memory?	No	Yes	Both a and b	None of the above	No we cannot use FP growth algorithm
104	2	What are closed itemsets?	An itemsetwhose no proper super-itemset has same support	An itemset for which at least one proper super-itemset has same support	An itemset for which at least super-itemset has same confidence	An itemsetwhose no proper super-itemset has same confidence	An itemsetwhose no proper super-itemset has same support is closed itemsets
105	2	What does FP growth algorithm do?	It mines all frequent patterns by constructing a FP tree	It mines all frequent patterns through pruning rules with higher support	It mines all frequent patterns through pruning rules with lesser support	All	FP growth algorithm do all frequent patterns by constructing a FP tree.
106	2	What do you mean by support(A)?	A Number of transactions containing A / Total number of transactions	Total Number of transactions not containing A	Total number of transactions containing	Number of transactions not containing A / Total number of transactions Ans: Number of transactions containing A / Total number of transactions	Support (A) means Number of transactions containing A / Total number of transactions
107	2	Find all strong association rules given the support is 0.6 and confidence is 0.8.	→ I5, →	→ I5, → → I2	Null rule set	Cannot be determined	None
108	3	When do you consider an association rule interesting?	If it satisfies both min_support and min_confidence	If it only satisfies min_confidence	If it only satisfies min_support If it satisfies both min_support and min_confidence	There are other measures to check so	If it satisfies both min_support and min_confidence association rule works

	marks	question	A	B	C	D	ans
109	3	When is sub-itemset pruning done?	When both a and b is true	A frequent itemset 'P' is a proper subset of another frequent itemset 'Q'	Support (P) = Support(Q)	When a is true and b is not	both are ture when sub-itemset pruning is done
110	3	What is the effect of reducing min confidence criteria on the same?	Some association rules will add to the current set of association rules	Number of association rules remains same.	Some association rules will become invalid while others might become a rule.	Can not say	Some association rules will add to the current set of association rules is the effect of reducing min confidence criteria on the same
111	3	Which of the following is direct application of frequent itemset mining?	Market Basket Analysis	Social Network Analysis	Outlier Detection	Intrusion Detection	Market Basket Analysis is direct application of frequent itemset mining.
112	3	Why is correlation analysis important?\nFor questions given below consider the data Transactions :\n1. I1, I2, I3, I4, I5, I6\n2. I7, I2, I3, I4, I5, I6\n3. I1, I8, I4, I5\n4. I1, I9, I10, I4, I6\n5. I10, I2, I4, I11, I5	To weed out uninteresting frequent itemsets	To make apriori memory efficient	To find large number of interesting itemsets	To restrict the number of database iterations	To weed out uninteresting frequent itemsets is correlation analysis
113	3	The apriori algorithm works in a ..and ..fashion?	Bottom-up and breath-first	Top-down and breath-first	Bottom-up and depth-first	Top-down and depth-first	Apriori algorithm works in bottom-up and breath-first fashion.
114	3	Which algorithm requires fewer scans of data?	FP growth	Apriori	Both a and b	None	FP growth algorithm requires fewer scans of data
115	3	Find odd man out:	DBSCAN	K mean	PAM	K medoid	None
116	3	What techniques can be used to improve the efficiency of apriori algorithm?	All	Transaction Reduction	Partitioning	Hash-based techniques	All techniques are used to improve the efficiency of apriori algorithm
117	3	What is the relation between candidate and frequent itemsets?	A frequent itemset must be a candidate itemset	A candidate itemset is always a frequent itemset	No relation between the two	Both are same	Relation between candidate and frequent itemsets is frequent itemset must be a candidate itemset
118	3	What are Max_confidence, Cosine similarity, All_confidence?	Pattern evaluation measure	Measures to improve efficiency of apriori	Frequent pattern mining algorithms	None	Pattern evaluation measure are Max_confidence, Cosine similarity, All_confidence
119	1	End Nodes are represented by __________	Triangles	Squares	Disks	Circles	None
120	1	Multivariate split is where the partitioning of tuples is based on a combination of attributes rather than on a single attribute.	"True"	"False"	None	None	None
121	1	Self-organizing maps are an example of	Unsupervised learning	Supervised learning	Reinforcement learning	Missing data imputation	None
122	1	Assume you want to perform supervised learning and to predict number of newborns according to size of storks’ population (http://www.brixtonhealth.com/storksBabies.pdf), it is an example of	Regression	Classification	Clustering	Structural equation modeling	Regression can predict number of newborns according to size of storks’ population

	marks	question	A	B	C	D	ans
123	1	Some telecommunication company wants to segment their customers into distinct groups to send appropriate subscription offers, this is an example of	Unupervised learning	Data extraction	Serration	Supervised learning	Unsupervised learning is telecommunication company
124	1	Decision Nodes are represented by ____________	Squares	Disks	Circles	Triangles	None
125	1	In the example of predicting number of babies based on storks’ population size, number of babies is	Outcome	Feature	Attribute	Observation	Outcome is the example of predicting numbers.
126	1	Cost complexity pruning algorithm is used in?	CART	C4	ID3	All	None
127	1	Attribute selection measures are also known as splitting rules.	"True"	"False"	None	None	Attribute selection measures are also known as splitting rules
128	1	How will you counter over-fitting in decision tree?	By pruning the longer rules	By creating new rules	Both By pruning the longer rules’ and ‘ By creating new rules’	None of the options	By pruning the longer rules you can counter over-fitting in decision tree
129	1	Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the other.	"True"	"False"	None	None	Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the other.
130	2	Which of the following classifications would best suit the student performance classification systems?	If...then... analysis	Market-basket analysis	Regression analysis	Cluster analysis	If...then... analysis is the best suit the student performance classification systems
131	2	A _________ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.	Decision tree	Graphs	Trees	Neural Networks	Refer the definition of Decision tree.
132	2	Cost complexity pruning algorithm is used in?	CART	C4.5	ID3	All	CART is the cost complexity used
133	2	The problem of finding hidden structure in unlabeled data is called	Unupervised learning	Supervised learning	Reinforcement learning	Data extraction	Unsupervised learning is unlabeled data
134	2	You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is in an example of	Supervised learning	Unsupervised learning	Serration	Dimensionality reduction	None
135	2	What is the approach of basic algorithm for decision tree induction?	Greedy	Top Down	Procedural	Step by Step	Greedy approach is basic algorithm for decision tree induction
136	2	Choose from the following that are Decision Tree nodes?	All	End Nodes	Chance Nodes	Decision Nodes	None
137	2	Which of the following sentences are true?	All	A pruning set of class labelled tuples is used to estimate cost complexity.	The best pruned tree is the one that minimizes the number of encoding bits.	In pre-pruning a tree is 'pruned' by halting its construction early.	All statements are true
138	2	Which of the following is not involve in data mining?	Knowledge extraction	Data transformation	Data exploration	Data archaeology	Data transformation is not involved in data mining

	marks	question	A	B	C	D	ans
139	2	Gini index does not favour equal sized partitions.	"False"	"True"	None	None	Gini index favour equal sized partitions
140	3	What is Decision Tree?	Flow-Chart & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label	Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label	Flow-Chart	None	Refer the definition of Decision tree.
141	3	Which one of these is not a tree based learner?	Bayesian classifier	ID3	CART	Random Forest	None
142	3	Task of inferring a model from labeled training data is called	Supervised learning	Unsupervised learning	Reinforcement learning	Complax learning	Task of inferring a model from labeled training data is called Supervised learning
143	3	What are two steps of tree pruning work?	Postpruning and Prepruning	Pessimistic pruning and Optimistic pruning	Cost complexity pruning and time complexity pruning	None of the options	Postpruning and Prepruning are two steps of tree pruning.
144	3	What are tree-based classifiers?	Both	Classifiers that perform a series of condition checking with one attribute at a time	Classifiers which form a tree with each attribute at one level	None	Both are tree-based classifiers
145	3	Which one of these is a tree based learner?	Random Forest	Bayesian Belief Network	Bayesian classifier	Rule based	Random Forest is the tree-based leraner.
146	3	Which of the following are the advantage/s of Decision Trees?	All	Use a white box model, If given result is provided by a model	Worst, best and expected values can be determined for different scenarios	Possible Scenarios can be added	None
147	3	When the number of classes is large Gini index is not a good choice.	"True"	"False"	None	None	Gini index is not a good choice
148	3	Discriminating between spam and ham e-mails is a classification task, true or false?	"True"	"False"	None	None	None
149	1	Point out the wrong statement.	Simple random sampling of time series is probably the best way to resample times series data.	Three parameters are used for time series splitting	Horizon parameter is the number of consecutive values in test set sample	All	Simple random sampling of time series is probably not the best way to resample times series data.
150	1	The cluster sampling, stratified sampling or systematic samplings are types of ________	Random sampling	Indirect sampling	Direct sampling	Non random sampling	The cluster sampling, stratified sampling or systematic samplings are types of random sampling.
151	1	Which of the following can be used to generate balanced cross–validation groupings from a set of data?	createFolds	createSample	createResample	None	createResample can be used to make simple bootstrap samples.

	marks	question	A	B	C	D	ans
152	1	Which of the following is classified as unknown or exact value that represents the whole population?	Parameter	Guider	Predictor	Estimator	The unknown or exact value that represents the whole population is called as parameter. Generally parameters are defined by small Roman symbols.
153	1	Which of the following is NOT supervised learning?	PCA	Decision Tree	Linear Regression	Naive Bayesian	PCA is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss.
154	1	In which of the following types of sampling the information is carried out under the opinion of an expert?	Judgement sampling	Convenience sampling	Purposive sampling	Quota sampling	In judgement sampling is carried under an opinion of an expert. The judgement sampling often results in a bias because of the variance in the expert opinion.
155	1	Which of the following package tools are present in caret?	All	Feature selection	Model tuning	Pre-processing	There are many different modeling functions in R.
156	1	Which of the following can be used to create sub–samples using a maximum dissimilarity approach?	maxDissim	minDissim	inmaxDissim	All	Splitting is based on the predictors.
157	2	Which of the factors affect the performance of learner system does not include?	Good data structures	Training scenario	Type of feedback	Representation scheme used	Factors that affect the performance of learner system does not include good data structures.
158	2	Which of the following function can be used to create balanced splits of the data?	createDataPartition	newDataPartition	renameDataPartition	None	If the argument to this function is a factor, the random sampling occurs within each class and should preserve the overall class distribution of the data.
159	2	In language understanding, the levels of knowledge that does not include?	Empirical	Syntactic	Phonological	Logical	In language understanding, the levels of knowledge that do not include empirical knowledge.
160	2	Which of the following function can create the indices for time series type of splitting?	createTimeSlices	newTimeSlices	binTimeSlices	None	Rolling forecasting origin techniques are associated with time series type of splitting.

	marks	question	A	B	C	D	ans
161	2	The selected clusters in a clustering sampling are known as ________	Elementary units	Primary units	Secondary units	Proportional units	In Cluster the population is divided into various groups called as clusters. The selected clusters in a sample are called as elementary units.
162	2	Among the following which is not a horn clause?	p → Øq	Øp V q	p → q	p	p → Øq is not a horn clause.
163	2	High entropy means that the partitions in classification are	Not pure	Pure	Useful	Useless	Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.\nIt is a measure of disorder or purity or unpredictability or uncertainty.\n
164	2	A sample size is considered large in which of the following cases?	n > or = 30	n > or = 50	n < or = 30	n < or = 50	Generally a sample having 30 or more sample values is called a large sample. By the Central Limit Theorem such a sample follows a Normal Distribution.
165	2	The method of selecting a desirable portion from a population which describes the characteristics of whole population is called as ________	Sampling	Segregating	Dividing	Implanting	The method of selecting a desirable portion from a population that describes the characteristics of the whole population is called as Sampling.
166	2	Which of the following statements about Naive Bayes is incorrect?	Attributes are statistically dependent of one another given the class value.	Attributes are equally important.	Attributes are statistically independent of one another given the class value.	Attributes can be nominal or numeric	Attributes are statistically dependent of one another given the class value Attributes are statistically independent of one another given the class value.
167	2	Sampling error increases as we increase the sampling size.	"False"	"True"	None	None	Sampling error is inversely proportional to the sampling size. As the sampling size increases the sampling error decreases.

	marks	question	A	B	C	D	ans
168	2	Point out the correct statement.	All	Caret includes several functions to pre-process the predictor data	The function dummyVars can be used to generate a complete set of dummy variables from one or more factors	Asymptotics are used for inference usually	The function dummyVars takes a formula and a data set and outputs an object that can be used to create the dummy variables using the predict method.
169	2	If the mean of population is 29 then the mean of sampling distribution is __________	29	30	21	31	In a sampling distribution the mean of the population is equal to the mean of the sampling distribution. Hence mean of population=29. Hence mean of sampling distribution=29.
170	3	Caret stands for classification and regression training.	"True"	"False"	None	None	The caret package is a set of functions that attempt to streamline the process for creating predictive models.
171	3	Caret does not use the proxy package.	"False"	"True"	None	None	Caret uses the proxy package.
172	3	A model of language consists of the categories which does not include?	Structural units	Role structure of units	System constraints	Language units	A model of language consists of the categories which does not include structural units.
173	3	Suppose we want to make a voters list for the general elections 2019 then we require __________	Census	Sampling error	Random error	Simple error	Study of population is called a Census. Hence for making a voter list for the general elections 2019 we require Census.
174	3	Different learning methods does not include?	Introduction	Analogy	Deduction	Memorization	Different learning methods does not include the introduction.
175	3	Suppose we would like to perform clustering on spatial data such as the geometrical locations of houses. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate?	Density-based clustering	Decision Trees	Model-based clustering	K-means clustering	The density-based clustering methods recognize clusters based on the density function distribution of the data object. For clusters with arbitrary shapes, these algorithms connect regions with sufficiently high densities into clusters.

	marks	question	A	B	C	D	ans
176	3	Which of the following function can be used to maximize the minimum dissimilarities?	All	minDiss	avgDiss	sumDiss	sumDiss can be used to maximize the total dissimilarities.
177	3	In sampling distribution what does the parameter k represents ________	Sampling interval	Secondary interval	Multi stage interval	Sub stage interval	In sampling distribution the parameter k represents Sampling interval. It represents the distance between which data is taken.
178	3	A machine learning problem involves four attributes plus a class. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?	72	24	48	12	Maximum possible different examples are the products of the possible values of each attribute and the number of classes;\n3 * 2 * 2 * 2 * 3 = 72\n

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

MCQ
SUBJECT: DATA MINING AND WAREHOUSING
UNIT-I

1. ________ is as finding hidden information in a database.
   a) Data mining
   b) Database access
   c) DBMS
   d) Data warehouse.
2. KDD means ___________ discovery in databases.
   a) King
   b) Kite
   c) Knowledge
   d) Kind
3. ___________ model makes a prediction about values of data using known results found from different data.
   a) Descriptive
   b) Preference
   c) Predictive
   d) Algorithm
4. ___________ maps data into predefined grouped or classes.
   a) Classification
   b) Regression
   c) Prediction
   d) Summarization
5. __________ model identifies patterns or relationships in data.
   a) Predictive
   b) Non-predictive
   c) Descriptive
   d) Unpredictable
6. ___________ is he use of algorithm to extract the information and patterns derived by the KDD process.
   a) Data mining
   b) Data base
   c) Data access
   d) Data processing
7. ________ is he process of finding useful information and paterns in data.
   a) Data mining
   b) KDD
   c) Data warehouse
   d) Data processing
8. __________ is a type of classification where an input pattern is classified into one of several classes based on predefined classes
   a) Pattern recognition
   b) TSA
   c) Clustering

d) Prediction

9. _________is used to map data item into real valued prediction variable.
   a) Clustering
   b) Classification
   c) Regression
   d) TSA
10. ___________ is used to visualize the time series.
    a) Time series plot
    b) Watch dog
    c) Time series analysis
    d) Grouping
11. Clustering is also called as ____________.
    a) Grouping
    b) Segmentation
    c) Unsupervised learning
    d) All the above
12. Summarization is also called as______________.
    a) Characterization
    b) Generalization
    c) Simple description
    d) All the above
13. _________ maps data into sunsets with associated simple description .
    a) Summarization
    b) Association Rules
    c) Classification
    d) Clustering
14. __________ refers to the DM task of uncovering relationships among data.
    a) Link analysis
    b) Clustering
    c) TSA
    d) Summarization
15. __________ is a model that identifies specific types of data association.
    a) TSA
    b) Sequence discovery
    c) Clustering
    d) Association Rules
16. __________ is used to determine sequential patterns in data.
    a) TSA
    b) Sequence discovery
    c) Clustering
    d) Association rules

17. The definition of KDD includes the keyword __________.
    a) Useful
    b) This
    c) DM
    d) All the above

18. In transformation ____________ is used to reduce the number of possible data values being considered.
    a) Data reduction
    b) Data interchange
    c) Errorneous of data
    d) Clearence of data
19. ____________ techniques are used to make the data easier to mine and more useful and to provide meaningful results.
    a) Preprocessing
    b) Selection
    c) Transformation
    d) Interpretation
20. __________ refers to the visual representation of data.
    a) GUI
    b) Interpretation
    c) Visualization
    d) Hybrid
21. _______ techniques include the box plot and scatter diagram.
    a) Graphical
    b) Geometric
    c) Icon-based
    d) Pixel-based
22. __________ is used to proceed from specific knowledge to more general information.
    a) Compression
    b) Induction
    c) Hybrid
    d) Pruning
23. ____________ occurs when the model does not fit future states.
    a) Overfitting
    b) Human interaction
    c) Outliers
    d) Integration
24. There are many data entries that do not fit nicely into derived model.
    a) Overfitting
    b) Human interaction
    c) Outliers
    d) Integration
25. IR stands for____________.
    a) Information reduction
    b) Information retrieval
    c) Information results
    d) Information relation
26. __________ is a software that is used to access the database.
    a) DBMS
    b) OLTP
    c) SQL
    d) CFMS

27. __________ data is said to be invalid or incorrect.
   a) Missing data
   b) Irrelevant data
   c) Noisy data
   d) Changing data

28. ROI stands for _______________.
   a) Return on investment
   b) Return on instruction
   c) Return on information
   d) Return on invalid data

29. The use of other attributes that increase the complexity and decrease in algorithm is called _____________.
   a) Dimensionality Curse
   b) Dimensionality reduction
   c) Dimensionality attribute
   d) Dimensionality

30. __________ techniques are targeted to such application as fraud detection, criminal suspects, prediction of terrorist.
   a) DM
   b) DB
   c) DBMS
   d) OLTP

31. ___________ access a database using a well defined query stated in language such as SQL.
   a) DBMS
   b) DBS
   c) KDD
   d) Database queries

32. A database is partitioned into disjoint grouping of similar tuples called __________.
   a) Clustering
   b) Classification
   c) Segmentation
   d) Generlization

33. __________ finds occurrences of a predefined pattern in data.
   a) Patterning
   b) Pattern recognition
   c) Patterning of data
   d) Pattern analysis

34. In KDD, the input to the process is known as _______ and the Output is ________.
   a) Informtion , data
   b) Field,record
   c) Record,field
   d) Data ,information

35. In KDD,obtaining the data from various DB, files,and other sources is called ______.
   a) Preprocessing
   b) Selection
   c) Tranformation
   d) Evaluation

36. Link analysis is otherwise called as_________.
    a) Association
    b) Association rule
    c) Affinity analysis
    d) All the above
37. Prediction application include ___________.
    a) Flooding
    b) Speech recognition
    c) Machine learning
    d) All the above
38. In regression, some type of error analysis is used to determine which function is____.
    a) Good
    b) Best
    c) Excellent
    d) Bad
39. Data mining is otherwise called as_______________.
    a) Data analysis
    b) Data discovery
    c) Deductive learning
    d) All the above
40. The rise of DBMS tool is________.
    a) 1960
    b) 1970
    c) 1980
    d) 1990
41. The metrics used include the traditional metrics of space and time based on _________.
    a) Complexity analysis
    b) Effectiveness
    c) Usefulness of data
    d) Scalability
42. ____________ data are noisy and have many missing attributes values.
    a) Real world
    b) Abstract
    c) Assumption
    d) Authorized
43. The use of ________ data is found in GIS data base .
    a) Missing
    b) Irrelevant
    c) Noisy
    d) Multimedia

44. A large DB can be viewed as using___________ to help uncover hidden information about the data.
    a) Search
    b) Compression

c) Approximation
d) Querying

45. Interfaces between technical experts and domain comes under______ issues.
a) Overfitting
b) Human interaction
c) Outlier
d) Application

46. The data Mining process can itself be vies a type of _____ underlying database.
a) Querying
b) Induction
c) Search
d) Processing

47. ___ requests may be treated as special,unusual or one time needs.
a) KDD
b) DM
c) DBMS
d) DB

48. ______ and___________ are effective tools to attack scalability problems.
a) Dimensionality & Parallelization
b) Sampling &Dimensionality
c) Effectiveness &Sampling
d) Sampling & Parallelization

49. Large data set is otherwise called as ________.
a) Massive datasets
b) High datasets
c) Noisy datasets
d) Irrelevent datasets

50. KDD process consists of _________ steps .
a) One
b) Three
c) Four
d) Five

UNIT-II

51. ________ models describe the relationship between I/O through algebraic equation.
    a) Parametric
    b) Non-parametric
    c) Static
    d) Dynamic
52. _______ may also be used to estimate error.
    a) Squared error
    b) Root mean error
    c) Mean Root square
    d) Mean squared error
53. _________ assumes that a linear relationship exists between the input data and the output data.
    a) Bivariate regression
    b) Correlation
    c) Multiple regression
    d) Linear regression
54. The _________ algorithm solves the estimation problem with incomplete data.
    a) Expectation maximization
    b) Expectation minimization
    c) Summarization-maximization
    d) Summarization minimization
55. Decision tree uses a _________ techniques.
    a) Greedy
    b) Divide & Conquer
    c) Shortest Path
    d) BFS
56. Null hypothesis and _______ hypothesis are two complementary hypothesis.
    a) Classical
    b) Testing
    c) Alternative
    d) None of the above
57. The BIAS of an estimator is the difference between ______ & _______values.
    a) Expected,actual
    b) Actual ,Expected
    c) Maximal,Minimal
    d) Minimal,Maximal
58. An __________ estimator is one whose BIAS is 0.
    a) Unbiased
    b) Rule biased
    c) Mean Root square
    d) Mean squared error

59. ________ is defined as the expected value of the squared difference between the estimate and the actual value.
    a) MSE
    b) RMS
    c) EM
    d) MLE
60. The________ may also be used to estimate error or another statistic to describe a distribution.
    a) RMS
    b) MLE
    c) EM
    d) MSE
61. _______ is a technique to estimate the likelihood of a property given the set of data as evidence or input.
    a) Point Estimation
    b) Models based on summarization
    c) Bayes theorem
    d) Hypothesis testing

62. In Box plot the Total range of the data value is divided into ________.
    a) Regions
    b) Quartiles
    c) Divisions
    d) Partitions
63. ________ measure is used instead of similarity measures.
    a) Distance
    b) Dissimilarity
    c) Both a,b
    d) None of the above
64. _________ relates the overlap to the average size of the two sets together.
    a) Dice
    b) Jaccard
    c) Cosine
    d) Overlap

65. ________ is used to measure the overlap of two sets as related to the whole set caused by their union.
    a) Dice
    b) Jaccard
    c) Cosine
    d) Overlap
66. ________ coefficient relates the overlap to the geometric average of the two sets.
    a) Dice
    b) Jaccard
    c) Cosine
    d) Overlap

67. The_________ metrics determines the degree to which the two sets overlap.
   a) Dice
   b) Jaccard
   c) Cosine
   d) Overlap

68. ________ is a predictive modeling technique used in classification ,clustering,etc.
   a) Neural networks
   b) Decision tree
   c) Genetic algorithm
   d) All the above

69. The neural networks can be viewed as a directed graph with __________ nodes.
   a) Two
   b) Three
   c) Four
   d) One

70. Internal nodes are also called as _________.
   a) Input
   b) Output
   c) Hidden
   d) Sink

71. In neural networks _______ activation function produces a linear output value based on the input.
   a) Threshold
   b) Step
   c) Linear
   d) Sigmoid

72. ______ is a bell shaped curve with output values in the range [0,1].
   a) Linear
   b) Guassian
   c) Hyperbolic
   d) Sigmoid

73. In neural network , ___________ is an S shaped curve with output values -1,1
   a) Sigmoid
   b) Linear
   c) Step
   d) Hyperbolic

74. The crossover technique generates new individual called________.
   a) Offspring
   b) Children
   c) Both a, b
   d) None of the above

75. _______ is used to determine the best individuals in a population.
   a) Crossover
   b) Mutation
   c) Fitness function
   d) All the above

76. The______________ operation randomly changes character in the offspring.
    a) Crossover
    b) Mutation
    c) Fitness function
    d) Both a,b

77. __________ is defined by precise algorithms that indicate how to combine the given set of individual to produce new ones.
    a) Production
    b) Reproduction
    c) Genetic algorithms
    d) Crossover

78. The activation function is also called as____________.
    a) Processing element function
    b) Squashing function
    c) Firing rule
    d) All the above

79. The subsections of the chromosomes are called___________.
    a) Cross over
    b) Genes
    c) Alleles
    d) Offspring

80. ________ is used to estimate error or to describe a distribution.
    a) RMS
    b) MSE
    c) SE
    d) Jackknife

81. ________ can be defined as a value proportional to actual probability with specific distribution.
    a) Likelihood
    b) Maximum Likelihood
    c) Estimation
    d) None of the above

82. In hypothesis testing O represents _____________.
    a) Outliers
    b) Observed data
    c) Output
    d) None of the above

83. One standard formula to measure linear correlation is the ____________.
    a) Correlation coefficient
    b) Classification
    c) Clustering
    d) Dissimilarity measures

84. _________ are often used instead o similarity measures.
    a) Distance
    b) Dissimilarity measure
    c) Both a,b
    d) None of the above

85. A variation of sigmoid function is called___________.
   a) Gaussian
   b) Hyperbolic
   c) Linear
   d) Threshold

86. Gaussian function is a ___________ shaped curve.
   a) S
   b) V
   c) Bell
   d) C

87. ________ is used to determine the best individuals in a population.
   a) Mutation
   b) Fitness function
   c) Crossover
   d) Starting set

88. One of the most important components of a genetic algorithm is__________.
   a) How to select individual
   b) How to select offspring
   c) How to select crossover
   d) How to select fitness

89. _________ coefficient is used to measure the overlap of two sets as related to whole set caused by their union.
   a) Dice
   b) Jaccard
   c) Cosine
   d) Overlap

90. _______ coefficient is used to relates the overlap to the average size of two sets together.
   a) Dice
   b) Jaccard
   c) Cosine
   d) Overlap

91. _______ coefficient relates the overlap to the geometric average of the two sets.
   a) Dice
   b) Jaccard
   c) Cosine
   d) Overlap

92. The_______ metric determines the degree to which the two sets overlap.
   a) Overlap
   b) Dice
   c) Cosine
   d) Jaccard

93. Rejection of null hypothesis causes another hypothesis called__________ hypothesis.
   a) Alternative
   b) Similarity measure
   c) Correlation
   d) Mutation

94. The input nodes exist in ___________ layer.
   a) Output
   b) Input
   c) Hidden
   d) All the above

95. Internal nodes is called _________ nodes.
   a) Input
   b) Output
   c) Hidden
   d) All the above

96. Artificial NNs can be classified based on the type of_____________.
   a) Connectivity
   b) Learning
   c) Both a, b
   d) None of the above

97. ________ occurs when the NNs is trained to fit one set to data.
   a) Outlier
   b) Noisy data
   c) Missing data
   d) Overfitting

98. To avoid overfitting _______ NNs are advisable.
   a) Larger
   b) Smaller
   c) Medium
   d) All the above

99. In sigmoid , c is a _____________.
   a) Change
   b) Constant
   c) Crossover
   d) Children

100. ____ is defined as the excepted value of the squared difference between the estimate and the actual value.
   a) MSE
   b) RMSE
   c) BIAS
   d) RMS

UNIT-III

101. Estimation and prediction may be viewed as types of _____________.

a) Clustering

b) Classification

c) Regression

d) Time Series

102. Classification performed by dividing the input space of potential database tuples into _______.

a) Regions

b) Class

c) Space

d) Sector

103. ________ values cause during both training and the classification process itself.

a) Data

b) Class

c) Predicate

d) Missing data

104. The performance of classification usually examined by evaluating the ___of the classification.

a) Accuracy

b) Contribution

c) Special value

d) Missing values

105. Classification true positives and false positives are calculated by the following curve.

a)MOC

b) NOC

c)ROC

d) COC

106. The __________ matrix illustrates the accuracy of the solution to a classification problem.

a) Confusion

b) Mutation

c) Crossover

d) Gaussian

107. ______ problems deal with estimation of an output value based on input values.

a) Prediction

b) Classification

c) Clustering

d) Regression

108. _____ is erroneous data.

a) OC
b) Regression
c) Noise
d) Linear model

109. Which are data values that are exceptions to the usual and expected data?

a) Outliers
b) Noise
c) Regression
d) Poor fit

110. The _______ classification can be viewed as both a descriptive and a predictive type of algorithm.

a) Naive
b) Bayes
c) Naive bayes
d) Prediction

111. The similarity (or) distance measures may be used to identify the _____ of different items in the database.

a) Likeness
b) Alikeness
c) Outliers
d) Centroid

112. A straightforward distance based approach assuming the each class Ci is represented by__.

a) Centroid
b) Outlier
c) Medoid
d) KNN

113. Expand : KNN

a) K Normal Neighbors
b) K Nearest Neighbor
c) K Normal Nextvalue
d) K Nearest Nest

114. The decision tree approach to classification is to divide search space into _______ regions.

a) Square
b) Triangular
c) Circular
d) Rectangular

115. In DT ,each internal node is labled with an _______.

a) Class

b) Attribute

c) Arc

d) Database

116. In DT, each leaf node labled with ____.

a) Class

b) Attribute

c) Arc

d) Link

117. The _____ technique to building a DT is based on information theory and attempts to minimize the expected number of comparisons.

a) CART

b) ID3

c) C.4.5

d) ROC

118. Neural networks are more robust than DTs because of the _________.

a) Arcs

b) Links

c) Weights

d) Classes

119. In NN, the normal approach used for processing is called________.

a) Activation function

b) Interconnections

c) Training data

d) Propagation

120. The NN starting state is modified based on feedback of its performance is referred to as__.

a) Supervised

b) Unsupervised

c) Both (a) and (b)

d) None of these

121. ________ learning can also be performed if the output is not known.

a) Supervised

b) Unsupervised

c) Neither (a) or (b)

d) Oral

122. The Mean Squared Error (MSE) is found by ________.

a) (yi-di)2/2

b) (yi+di)2/2

c) (di-yi)2/2

d) (di+yi)2/2

123. The ____ can be used to find a total error over all nodes in the network.

a) RDF

b) ROC

c) MSE

d) CMC

124. Which learning technique that adjusts weights in the NN by propagating weight changes backward from the sink to the source nodes?

a) Propagation

b) perceptrons

c) MSE

d) Back propagation

125. In radial basis function (RBF) central point value is ______.

a) 0

b) 1

c) +1

d) -1

126. The simplest Neural Network is called a ________.

a) Neuron

b) Gene

c) Perceptron

d) Single neuron

127. In rule-based algorithms, _____rules that cover all cases.

a) if-else

b) if-then

c) switch-case

d) nested if

128. The __________is used to predict a future classification value.

a) Genetic algorithm

b) Decision Tree

c) Rule-based Algorithm

d) Neural Network

129. Multiple independent approaches can be applied to a classification problem is referred to as ___.

a) CMC
b) RBF
c) ROC
d) DCS

130. In which technique the classifier that has the best accuracy in database sample?

a) CMC
b) RBF
c) DCS
d) ROC

131. OC stands for_______________.

a) Operating characteristics
b) Operating curve
c) Operating classifications
d) None of the above

132. Rule based classification algorithms generate ______ rules to perform the classifications.

a) If
b) Then
c) If-then
d) If - else

133. In OC curve , the horizontal axis has the percentage of _________Positives for a sample DB.

a) False
b) True
c) Either a, b
d) None of the above

134. In OC curve , the vertical axis has the percentage of _________Positives for a sample DB.

a) False
b) True
c) Either a, b
d) None of the above

135. The__________ approach is most useful in classification problem.

a) Incremental rule
b) Cluster
c) NN
d) Decision tree

136. _________ techniques use labeling of the items to assist In the classification process.

a) Intrinsic
b) Extrinsic
c) Overlapping
d) Numerical

137. A __________ curve shows the relationship between false positives and true positives.

a) BOC
b) ROS
c) ROC
d) BOS

138. Task of CART is_________.

a) Only regression
b) Only classification
c) Both a,b
d) None of the above

139. A variation of the complete link algorithm is called ___________ algorithm.

a) Nearest
b) Neighbour
c) Farthest Neighbour
d) All the above

140. K nearest neighbor is a classification scheme based on the use of_________.

a) Distance Measure
b) Similarity
c) Complete link
d) Average

141. A perceptron is a ___________ neuron with multiple inputs and one output.

a) single
b) Multiple
c) Double
d) None of the above

142. The classes that exist for a classification problem are indeed _________.

a) Equivalence classes
b) Variance classes
c) Mean classes
d) Median

143. The formula for straight line is_________.

a) Y=mx+b
b) y=mx
c) Y=M+b
d) Y=m

144. _________ are data values that are exception to the usual and expected data.

a) Outliers
b) Noise
c) Error
d) Overfit

145. ________ is an errorneous data.

a) Overfit
b) Outlier
c) Noise
d) Missing

146. ___________ problems deal with the estimation of output value based on input value.

a) Baysian classification
b) K nearest Neighbour
c) Regression
d) All the above

147. ____ problem can be thought of as estimating the formula for a straight line.

a) Regression
b) Linear regression
c) Bayesian classification
d) K nearest neighbour

148. Logistic regression uses ___________ technique.

a) Box plot
b) Logistic curve
c) Straight line
d) Logistic line

149. Decision tree is otherwise called as___________.

a) Classification tree
b) Regression tree
c) K nearest neighbor
d) Clustering tree

150. Data objects are described by a number of ________that capture the basic characteristics of an object.

a) Data sets
b) Elements
c) Record
d) Attribute

UNIT-IV

151. __________is similar to classification in that data are grouped.

a) Classification

b) Regression

c) Clustering

d) DT

152. One of the first domain in which clustering was used as __________taxonomy.

a) Biological

b) Zoological

c) Mathematical

d) Scientific

153. Cluster results are________.

a) Static

b) Realistic

c) Acoustic

d) Dynamic

154. _______ clustering , the algorithm creates only one set of clusters.

a) Dynamic

b) Hierarchical

c) Partitional

d) Static

155. With ______ clustering, a nested set of clusters to be created.

a) Partitional

b) Hierarchical

c) Dynamic

d) Static

156. In similarity measures, metric attributes satisfy the ______ inequality.

a) Rectangular

b) Triangular

c) Square

d) Circle

157. The ____ is the “middle” of the cluster it need not be actual point in the cluster.

a) Radius

b) Diameter

c) Centroid

d) Metoid

158. The cluster is represented by one centrally located object in the cluster called a______.

a) Centroid
b) Medoid
c) Radius
d) Diameter

159. The ____is the square root of the average mean squared distance from any point in the cluster to centroid.

a) Radius
b) Medoid
c) Diameter
d) Centroid

160. The _____is the square root of the average mean squared distance between all pairs of points in the cluster.

a) Radius
b) Medoid
c) Diameter
d) Centroid

161. Largest distance between an element in one cluster and an element in the other is_____.

a) Single Link
b) Complete Link
c) Average Link
d) Centroid

162. Smallest distance between an element in the cluster and an element in the other is____.

a) Centroid
b) Medoid
c) Complete link
d) Single link

163. _________ are sample points with values much different from those of the remaining set of data.

a) Centroid
b) Medoid
c) Outliers
d) Compression

164. In hierarchical clustering , a tree data structure is called ______.

a) Connected component
b) Dendrogram
c) Minimum spanning tree
d) Bond energy

165. The root in the dendrogram tree contains ____ clusters ,where all elements aretogether.

a) Four

b) Three

c) Two

d) One

166. The space complexity for hierarchical algorithms is_______.

a) O(n)

b) O(N+2)

c) O(n2)

d) O(2N)

167. A ________ component is a graph in which there exists a path between any two vertices.

a) Connected

b) Unconnected

c) Nested

d) Stylish

168. A _____ is a maximal graph in which there is an edge between vertices.

a) Connected graph

b) Clique

c) Candidates

d) Dendrogram

169. The ____are sample points with values much different from those of the remaining set of data.

a) Clusters

b) Outliers

c) Candidates

d) Mining

170. _____is the process of identifying outliers in a set of data.

a) Outlier detection

b) Outlier avoidance

c) Outlier collision

d) Outlier prediction

171. The outliers can be detected by well-known tests such as _________-.

a) Chi-square test

b) Random test

c) Discordancy test

d) Unit test

172. Clustering applications include plant and _____ classifications.

a) Medical

b) Biological

c) Zoological

d) Animal

173. ____ clustering , all items are initially placed in one cluster and clusters are repeat.

a) Random

b) Divisive

c) Nearest neighbour

d) Partitional

174. BEA stands for__.

a) Band Echo Algorithm

b) Bond Echo Algorithm

c) Balance Energy Algorithm

d) Bond Energy Algorithm

175. _________is an iterative clustering algorithm.

a) K-means

b) LARGE DB

c) KDD

d) BEA

176. The nearest neighbor algorithm uses ______technique.

a) Single link

b) Complete link

c) Average link

d) Centroid

177. The PAM algorithm also called _______algorithm.

a) K-means

b) K-medoids

c) K-centroid

d) K-radius

178. The time complexity of nearest neighbor algorithm is________.

a) O(n)

b) O(N+2)

c) O(n2)

d) O(2N)

179. In a distributed database, each resulting cluster is called a _______.

a) Horizontal Fragment

b) Vertical Fragment

c) Both(a) & (b)

d) None

180. In neural network, the number of input nodes is the same as the number of___.

a) Levels

b) Clusters

c) Points

d) Attributes

181. 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

182. __________ 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

183. ________ seeks to find groups of closely related observations so that observations that belong the same cluster are more similar to each other.

a) Association

b) Anomaly detection

c) Clustering

d) None

184. In web mining, _______ is used to find natural groupings of users, pages, etc.

a) Clustering
b) Associations
c) Sequential analysis
d) Classification

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

a) k-medoid

b) k-means

c) STIRR

d) ROCK

186. 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

187. Pick out a k-medoid algoithm.

a) DBSCAN

b) BIRCH

c) PAM

d) CURE

188. Pick out a hierarchical clustering algorithm.

a) DBSCAN

b) BIRCH

c) PAM

d) CURE

189. _______ is the process of identifying outliers in a set of data.

a) Outlier

b) Outlier detection

c) Segmentation

d) Processing

190. The space complexity of adjacency matrix is_______.

a) O(n)

b) O(kn)

c) O(n2)

d) None o the above

191. A variation of complete link algorithm is called the ______.

a) Farthest nearest neighbor

b) Nearest neighbor

c) Average

d) Single

192. A tree data structure called________ is used to illustrate the hierarchical clustering technique.

a) Dendogramming

b) Dendo

c) Dendogram

d) Dendograms

193. The term _______ indicates the ability of these NN to organize the nodes into clusters based on the similarity between them.

a) Competitive

b) Non-competitive

c) Self organizing

d) None of the above

194. CF stands for_________

a) Clustering Features

b) Clustering future

c) Classification Features

d) Classification Future

195. The space complexity for K-means is_______.

a) O(n)

b) O(kn)

c) n

d) O(n2)

196. The squared error algorithm has _______ type.

a) Hierarchical

b) Partitional

c) Mixed

d) Agglomeative.

197. The time complexity for single link algorithm is__________.

a) O(kn2)

b) O(n)

c) O(kn)

d) O(n2)

198. The squared error clustering algorithm minimizes_______ .

a) Error

b) Squared error

c) Square

d) All the above

199. With _________ clustering the algorithm creates only one set of clusters.

a) Partitional

b) Hierarchical

c) Agglomerative

d) None of the above

200. ________ techniques use labeling of the items to assist in the classification process.

a) Intrinsic

b) Extrinsic

c) Both a,b

d) All the above

UNIT-V

201. The purchasing of one product when another product is purchased represents an__________.

a) Decision Tree

b) Association Rule

c) Classification

d) Clustering

202. The ______of an item is the percentage of transactions in which that item occurs.

a) Confidence

b) Support

c) Association rule

d) Itemset

203. The _____is called the number of scans of the database.

a) Support

b) Confidence

c) Strength

d) Both (b) & (c)

204. Potentially large item sets are called ____.

a) Support

b) Confidence

c) Candidates

d) Itemset

205. In association rule algorithm, the notation “P” indicates.

a) Confidence

b) Candidates

c) Partitions

d) Transactions

206. Any subset of a large itemset must be _____.

a) Small
b) Medium
c) Average
d) Large

207. The large itemsets are also said to be _______closure.

a) Upward

b) Middleware

c) Downward

d) None

208. Additional candidates are determined by applying the _________ border function.

a) Positive

b) Negative

c) Average

d) Medium

209. The Apriori algorithm shows the sample is performed using a support called __.

a) High

b) Low

c) Average

d) Smalls

210. The basic _________ reduces the number of database scans to two.

a) Divisive algorithm

b) Parallel algorithm

c) Partition algorithm

d) Sampling algorithm

211. The candidates are partitioned and counted separately at each processor is called____.

a) Data parallelism

b) Task parallelism

c) Candidates

d) Data reduction

212. One data parallelism algorithm is the _____.

a) MSE

b) FIS

c) DDA

d) CDA

213. One task parallelism algorithm is called _____.

a) CDA

b) MSE

c) DDA

d) BCD

214. An algorithm all rules that satisfy a given support and confidence level is called____.

a) Target

b) Type

c) Data type

d) Data source

215. The most common data structure used to store the candidates itemsets and their counts is a_______.

a) Binary tree

b) B-tree

c) Balanced tree

d) Hash tree

216. Which technique is used to improve on the performance of an algorithm given distribution Or amount of main memory?

a) Architecture

b) Optimization

c) Parallelism

d) Itemset

217. A leaf node in the hash tree contains_____.

a) Attributes

b) Itemset

c) Candidates

d) Data

218. One incremental approach,______is based on the Apriori algorithm.

a) CDA

b) DDA

c) fast update

d) slow update

219. A variation of generalized rules are _____ association rules.

a) Multiple-level

b) Hierarchical-level

c) Multi-level

d) Hybrid-level

220. A ______ association rule is one that involves categorical and quantitative data.

a) Categorical

b) Qualitative

c) Quantitative

d) Spanning

221. MIS stands for _______.

a) Medium item support

b) Maximum item support

c) Minimum item support

d) Medium item scale

222. A __rule is defined as a set of itemsets that are correlated.

a) Correlation

b) Co-efficient

c) MIS

d) Modification

223. Correlation(A=>B)= ________?

a) P(A,B) / P(A)P(B)

b) (b)P(A) / (P(A) P(B)

c) P(B) / P(A) P(B)

d) P(A) P(B) / P(A) – P(B)

224. Conviction has a value of ____ if A and B are not related.

a) 0

b) 1

c) 2

d) ∞

225. Which one is not an association rule algorithm?

a) Apriori

b) CDA

c) DDA

d) PAM

226. _______ algorithms may be able to adapt better to limited main memory.

a) Divisive

b) Sampling

c) Partitioning

d) Distributed

227. During the _______ scan, additional candidates are generated and counted.

a) First

b) Second

c) Third

d) Fourth

228. Chi-squared statistic is denoted by the _______symbol.

a) X2

b) E[X]

c) 2X

d) X3

229. ___ are used to show the relationships between data items.

a) Clustering

b) Regression

c) Association rules

d) Classification

230. The most two important property of an association rules are _____.

a) Support, confidence

b) Itemset, data

c) Neuron, gene

d) Lift, interest

231. A __________ is defined as a set of itemsets that are correlated.

a) Correlation rule

b) Association rule

c) Conviction

d) Probability of correlation

232. Confidence or strength are indicated by ___________.

a) ©

b) ®

c) €

d) α

233. In association rule l stands for__________.

a) Large item sets in L

b) Set of large item set

c) Both a,b

d) None of the above

234. ________is the most well known association rule algorithm and is used in most commercial products.

a) Apriori algorithm

b) Partition algorithm

c) Distributed algorithm

d) Pincer-search algorithm

235. 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

236. 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.

237. 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

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

a) Upward

b) Downward

c) Breadthwise

d) Both upward and downward

239. 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

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

a) Candidate generation

b) Itemset generation

c) Pruning

d) Partitioning

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

a) Candidate generation

b) Itemset generation

c) Pruning

d) Partitioning

242. The A Priori algorithm is a ___________.

a) top-down search

b) breadth first search

c) depth first search

d) bottom-up search

243. A priori algorithm is otherwise called as __________.

a) width-wise algorithm

b) level-wise algorithm

c) pincer-search algorithm

d) FP growth algorithm

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

a) Consequent

b) Onset

c) Antecedent

d) Precedent

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

a) Consequent

b) Onset

c) Antecedent

d) Precedent

246. 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

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

a) Confidence

b) Support

c) Support count

d) None Of the above

248. ______ are effective tools to attack the scalability problem.

a) Sampling

b) Parallelization

c) Both A & B

d) None of the above

249. Discovery of cross-sales opportunities is called ________________.

a) Segmentation

b) Visualization

c) Correction

d) Association

250. In web mining, _________ is used to know which URLs tend to be requested together.

a) Clustering

b) Associations

c) Sequential analysis

d) Classification

ANSWER KEY

UNIT-I

1	A	2	C	3	C	4	A	5	C	6	A	7	B	8	A	9	C	10	A
11	D	12	D	13	A	14	A	15	D	16	B	17	A	18	A	19	C	20	C
21	B	22	B	23	A	24	C	25	B	26	A	27	C	28	A	29	A	30	A
31	D	32	C	33	B	34	D	35	B	36	D	37	D	38	B	39	D	40	B
41	A	42	A	43	D	44	C	45	B	46	A	47	A	48	D	49	A	50	D

UNIT-II

51	**A**	**52**	**B**	**53**	**D**	**54**	**A**	**55**	**B**	**56**	**C**	**57**	**A**	**58**	**A**	**59**	**A**	**60**	**A**
61	C	62	B	63	C	64	A	65	B	66	C	67	D	68	B	69	B	70	C
71	C	72	B	73	A	74	C	75	C	76	B	77	B	78	C	79	B	80	A
81	A	82	B	83	A	84	C	85	B	86	C	87	B	88	A	89	B	90	A
91	C	92	A	93	A	94	B	95	C	96	C	97	D	98	B	99	B	100	A

UNIT-III

101	**B**	**102**	**A**	**103**	**D**	**104**	**A**	**105**	**C**	**106**	**A**	**107**	**D**	**108**	**C**	**109**	**A**	**110**	**C**
111	B	112	A	113	B	114	D	115	B	116	C	117	B	118	C	119	D	120	A
121	B	122	A	123	C	124	D	125	A	126	C	127	B	128	D	129	A	130	C
131	A	132	C	133	A	134	C	135	B	136	B	137	C	138	C	139	C	140	A
141	A	142	A	143	A	144	A	145	C	146	C	147	B	148	B	149	A	150	D

UNIT-IV

151	**C**	**152**	**A**	**153**	**D**	**154**	**C**	**155**	**B**	**156**	**B**	**157**	**C**	**158**	**B**	**159**	**A**	**160**	**C**
161	B	162	D	163	C	164	B	165	D	166	C	167	A	168	B	169	B	170	A
171	C	172	D	173	B	174	D	175	A	176	A	177	B	178	C	179	B	180	D
181	C	182	B	183	C	184	A	185	B	186	A	187	C	188	A	189	B	190	C
191	A	192	C	193	C	194	A	195	A	196	B	197	A	198	B	199	A	200	B

UNIT-V

201	B	202	B	203	D	204	C	205	C	206	D	207	C	208	B	209	D	210	C
211	B	212	D	213	C	214	A	215	D	216	B	217	C	218	C	219	A	220	C
221	C	222	A	223	A	224	B	225	D	226	A	227	B	228	A	229	C	230	A
231	A	232	D	233	A	234	A	235	A	236	A	237	B	238	A	239	B	240	C
241	A	242	B	243	A	244	C	245	A	246	A	247	C	248	C	249	D	250	B

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. What does the leaf node in decision tree indicates

A : sub tree

B : class label

C : testing node

D : condition

Q.no 2. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 3. the negative tuples that were correctly labeled by the classifier

A : False positives(FP)

B : True positives(TP)

C : True negatives (TN)

D : False negatives(FN)

Q.no 4. Removing duplicate records is a process called

A : recovery

B : data cleaning

C : data cleansing

D : data pruning

Q.no 5. For Apriori algorithm, what is the first phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 6. Multi-class classification makes the assumption that each sample is assigned to

A : one and only one label

B : many labels

C : one or many labels

D : no label

Q.no 7. Multilevel association rules can be mined efficiently using

A : Support

B : Confidence

C : Support count

D : Concept Hierarchies under support-confidence framework

Q.no 8. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 9. Self-training is the simplest form of

A : supervised classification

B : semi-supervised classification

C : unsupervised classification

D : regression

Q.no 10. Which of the following is direct application of frequent itemset mining?

A : Social Network Analysis

B : Market Basket Analysis

C : Outlier Detection

D : Intrusion Detection

Q.no 11. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 12. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 13. The Synonym for data mining is

A : Data warehouse

B : Knowledge discovery in database

C : ETL

D : Business Intelligemce

Q.no 14. Which of the following are methods for supervised classification?

A : Decision tree

B : K-Means

C : Hierarchical

D : Apriori

Q.no 15. These are the intermediate servers that stand in between a relational back-end server and client front-end tools

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 16. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 17. What are two steps of tree pruning work?

A : Pessimistic pruning and Optimistic pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : None of the options

Q.no 18. A data cube is defined by

A : Dimensions

B : Facts

C : Dimensions and Facts

D : Dimensions or Facts

Q.no 19. For Apriori algorithm, what is the second phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 20. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 21. Lazy learner classification approach is

A : learner waits until the last minute before constructing model to classify

B : a given training data constructs a model first and then uses it to classify

C : the network is constructed by human experts

D : None of the options

Q.no 22. Cross validation involves

A : testing the machine on all possible ways by substituting the original sample into training set

B : testing the machine on all possible ways by dividing the original sample into training and validation sets.

C : testing the machine with only validation sets

D : testing the machine on only testing datasets.

Q.no 23. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 24. Data independence means

A : Data is defined separately and not included in programs

B : Programs are not dependent on the physical attributes of the data

C : Programs are not dependent on the logiical attributes of the data

D : Programs are not dependent on the physical attributes as well as logical attributes of the data

Q.no 25. Which of the following is a predictive model?

A : Clustering

B : Regression

C : Summarization

D : Association rules

Q.no 26. The data cubes are generally

A : 1 Dimensional

B : 2 Dimensional

C : 3 Dimensional

D : n-Dimensional

Q.no 27. Identify the example of sequence data

A : weather forecast

B : data matrix

C : market basket data

D : genomic data

Q.no 28. The frequent-item-header-table consists of number fields

A : Only one

B : Two

C : Three

D : Four

Q.no 29. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 30. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 31. When do you consider an association rule interesting?

A : If it only satisfies minimum support

B : If it only satisfies minimum confidence

C : If it satisfies both minimum support and minimum confidence

D : There are other measures to check interesting rules

Q.no 32. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 33. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 34. Which of the following probabilities are used in the Bayes theorem.

A : P(Ci|X)

B : P(Ci)

C : P(X|Ci)

D : P(X)

Q.no 35. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 36. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 37. Handwritten digit recognition classifying an image of a handwritten number into a digit from 0 to 9 is example of

A : Multiclassification

B : Multi-label classification

C : Imbalanced classification

D : Binary Classification

Q.no 38. What type of data do you need for a chi-square test?

A : Categorical

B : Ordinal

C : Interval

D : Scales

Q.no 39. For a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is not true in such a case?

A : Imbalaced problems should not be measured using Accuracy metric.

B : Accuracy metric is not a good idea for imbalanced class problems.

C : Precision and recall metrics aren’t good for imbalanced class problems.

D : Precision and recall metrics are good for imbalanced class problems.

Q.no 40. Which of the following property typically does not hold for similarity measures between two objects ?

A : Symmetry

B : Definiteness

C : Triangle inequality

D : Transitive

Q.no 41. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 42. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 43. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 44. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 45. Which of the following operation is requird to calculate cosine similarity?

A : Vector dot product

B : Exponent

C : Modulus

D : Percentage

Q.no 46. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 47. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 48. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 49. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 50. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 51. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 52. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 53. Which operation data warehouse requires ?

A : Initial loading of data

B : Transaction processing

C : Recovery

D : Concurrency control mechanisms

Q.no 54. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 55. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 56. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 57. What does a Pearson's product-moment allow you to identify?

A : Whether there is a relationship between variables

B : Whether there is a significant effect and interaction of independent variables

C : Whether there is a significant difference between variables

D : Whether there is a significant effect and interaction of dependent variables

Q.no 58. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 59. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 60. These server performs the faster computation

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Answer for Question No 1. is b

Answer for Question No 2. is b

Answer for Question No 3. is c

Answer for Question No 4. is b

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is d

Answer for Question No 8. is b

Answer for Question No 9. is b

Answer for Question No 10. is b

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is a

Answer for Question No 17. is b

Answer for Question No 18. is c

Answer for Question No 19. is a

Answer for Question No 20. is a

Answer for Question No 21. is a

Answer for Question No 22. is c

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is b

Answer for Question No 26. is d

Answer for Question No 27. is d

Answer for Question No 28. is b

Answer for Question No 29. is c

Answer for Question No 30. is a

Answer for Question No 31. is c

Answer for Question No 32. is a

Answer for Question No 33. is c

Answer for Question No 34. is a

Answer for Question No 35. is b

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is a

Answer for Question No 39. is c

Answer for Question No 40. is c

Answer for Question No 41. is a

Answer for Question No 42. is d

Answer for Question No 43. is b

Answer for Question No 44. is c

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is c

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is d

Answer for Question No 51. is d

Answer for Question No 52. is d

Answer for Question No 53. is a

Answer for Question No 54. is b

Answer for Question No 55. is d

Answer for Question No 56. is a

Answer for Question No 57. is a

Answer for Question No 58. is a

Answer for Question No 59. is a

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Postpruning is

A : Removing branches from fully grown tree

B : Stop constructing tree if this would result in the measure falling below a threshold

C : construting a new tree

D : Flow-Chart

Q.no 2. If two documents are similar, then what is the measure of angle between two documents?

A : 30

B : 60

C : 90

D : 0

Q.no 3. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 4. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 5. These are the intermediate servers that stand in between a relational back-end server and client front-end tools

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 6. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 7. Which of the following is not a type of constraints?

A : Data constraints

B : Rule constraints

C : Knowledge type constraints

D : Time constraints

Q.no 8. Baysian classification in based on

A : probability for the hypothesis

B : Support

C : tree induction

D : Trees

Q.no 9. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 10. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 11. What is an alternative form of Euclidean distance?

A : L1 norm

B : L2 norm

C : Lmax norm

D : L norm

Q.no 12. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 13. What are closed frequent itemsets?

A : A closed itemset

B : A frequent itemset

C : An itemset which is both closed and frequent

D : Not frequent itemset

Q.no 14. A decision tree is also known as

A : general tree

B : binary tree

C : prediction tree

D : None of the options

Q.no 15. cross-validation and bootstrap methods are common techniques for assessing

A : accuracy

B : Precision

C : recall

D : performance

Q.no 16. A multidimensional data model is typically organized around a central theme which is represented by

A : Dimension table

B : Fact table

C : Dimension table and Fact table

D : Dimension table or Fact table

Q.no 17. The problem of agents to learn from the environment by their interactions with dynamic environment is done in

A : Reinforcement learning

B : Multi-label classification

C : Binary Classification

D : Multiclassification

Q.no 18. Entropy is a measure of

A : impurity of an attribute

B : Purity of an attribute

C : Weight of an attribute

D : Class of an attribute

Q.no 19. the negative tuples that were correctly labeled by the classifier

A : False positives(FP)

B : True positives(TP)

C : True negatives (TN)

D : False negatives(FN)

Q.no 20. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 21. What is another name of data matrix?

A : Single mode

B : Two mode

C : Multi mode

D : Large mode

Q.no 22. Which of the following is a predictive model?

A : Clustering

B : Regression

C : Summarization

D : Association rules

Q.no 23. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 24. Data independence means

A : Data is defined separately and not included in programs

B : Programs are not dependent on the physical attributes of the data

C : Programs are not dependent on the logiical attributes of the data

D : Programs are not dependent on the physical attributes as well as logical attributes of the data

Q.no 25. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 26. If first object X and Y coordinates are 3 and 5 respectively and second object X and Y coordinates are 10 and 3 respectively, then what is Manhattan disstance between these two objects?

A : 8

B : 13

C : 9

D : 10

Q.no 27. Number of records are comparatively more in

A : OLAP

B : OLTP

C : Same in OLAP and OLTP

D : Can not compare

Q.no 28. Which of the following operations are used to calculate proximity measures for ordinal attribute?

A : Replacement and discretization

B : Replacement and characterizarion

C : Replacement and normalization

D : Normalization and discretization

Q.no 29. Which of the following is necessary operation to calculate dissimilarity between ordinal attributes?

A : Replacement of ordinal categories

B : Correlation coefficient

C : Discretization

D : Randomization

Q.no 30. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 31. In a decision tree each leaf node represents

A : Test conditions

B : Class labels

C : Attribute values

D : Decision

Q.no 32. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 33. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 34. The property of Apriori algorithm is

A : All nonempty subsets of a frequent itemsets must also be frequent

B : All empty subsets of a frequent itemsets must also be frequent

C : All nonempty subsets of a frequent itemsets must be not frequent

D : All nonempty subsets of a frequent itemsets can frequent or not frequent

Q.no 35. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 36. The probability of a hypothesis before the presentation of evidence is called as

A : Apriori probability

B : subjective probability

C : posterior probability

D : conditional probability

Q.no 37. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 38. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 39. The accuracy of a classifier on a given test set is the percentage of

A : test set tuples that are correctly classified by the classifier

B : test set tuples that are incorrectly classified by the classifier

C : test set tuples that are incorrectly misclassified by the classifier

D : test set tuples that are not classified by the classifier

Q.no 40. Which of the following is measure of document similarity?

A : Cosine dissimilarity

B : Sine similarity

C : Sine dissimilarity

D : Cosine similarity

Q.no 41. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 42. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 43. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 44. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 45. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 46. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 47. A concept hierarchy that is a total or partial order among attributes in a database schema is called

A : Mixed hierarchy

B : Total hierarchy

C : Schema hierarchy

D : Concept generalization

Q.no 48. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 49. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 50. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 51. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 52. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 53. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 54. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 55. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 56. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 57. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 58. A data normalization technique for real-valued attributes that divides each numerical value by the same power of 10.

A : min-max normalization

B : z-score normalization

C : decimal scaling

D : decimal smoothing

Q.no 59. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 60. Holdout method, Cross-validation and Bootstrap methods are techniques to estimate

A : Precision

B : Classifier performance

C : Recall

D : F-measure

Answer for Question No 1. is a

Answer for Question No 2. is d

Answer for Question No 3. is c

Answer for Question No 4. is d

Answer for Question No 5. is a

Answer for Question No 6. is b

Answer for Question No 7. is d

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is c

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is c

Answer for Question No 14. is c

Answer for Question No 15. is a

Answer for Question No 16. is b

Answer for Question No 17. is a

Answer for Question No 18. is a

Answer for Question No 19. is c

Answer for Question No 20. is d

Answer for Question No 21. is b

Answer for Question No 22. is b

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is c

Answer for Question No 27. is b

Answer for Question No 28. is c

Answer for Question No 29. is a

Answer for Question No 30. is c

Answer for Question No 31. is b

Answer for Question No 32. is c

Answer for Question No 33. is d

Answer for Question No 34. is a

Answer for Question No 35. is b

Answer for Question No 36. is a

Answer for Question No 37. is b

Answer for Question No 38. is b

Answer for Question No 39. is a

Answer for Question No 40. is d

Answer for Question No 41. is d

Answer for Question No 42. is b

Answer for Question No 43. is a

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is c

Answer for Question No 47. is c

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is a

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is d

Answer for Question No 55. is b

Answer for Question No 56. is d

Answer for Question No 57. is b

Answer for Question No 58. is c

Answer for Question No 59. is a

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 2. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 3. cross-validation and bootstrap methods are common techniques for assessing

A : accuracy

B : Precision

C : recall

D : performance

Q.no 4. The task of building decision model from labeled training data is called as

A : Supervised Learning

B : Unsupervised Learning

C : Reinforcement Learning

D : Structure Learning

Q.no 5. A multidimensional data model is typically organized around a central theme which is represented by

A : Dimension table

B : Fact table

C : Dimension table and Fact table

D : Dimension table or Fact table

Q.no 6. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 7. What is association rule mining?

A : Using association to find correlation rules

B : Same as frequent itemset mining

C : Finding of strong association rules using frequent itemsets

D : Finding of frequent itemset from large database

Q.no 8. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 9. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 10. OLAP database design is

A : Application-oriented

B : Object-oriented

C : Goal-oriented

D : Subject-oriented

Q.no 11. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 12. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 13. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 14. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 15. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 16. Data used to build a data mining model.

A : Validation Data

B : Training Data

C : Testing Data

D : Hidden Data

Q.no 17. The problem of agents to learn from the environment by their interactions with dynamic environment is done in

A : Reinforcement learning

B : Multi-label classification

C : Binary Classification

D : Multiclassification

Q.no 18. accuracy is used to measure

A : classifier’s true abilities

B : classifier’s analytic abilities

C : classifier’s decision abilities

D : classifier’s predictive abilities

Q.no 19. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 20. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 21. What is uniform support in multilevel association rule minig?

A : Use of minimum support

B : Use of minimum support and confidence

C : Use of same minimum threshold at each abstraction level

D : Use of minimum support and support count

Q.no 22. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 23. Which of the following operation is correct about supremum distance?

A : It gives maximum difference between any attribute of the objects

B : It gives minimum difference between any attribute of the objects

C : It gives maximum difference between fisrt attribute of the objects

D : It gives minimum difference between fisrt attribute of the objects

Q.no 24. Frequent patterns generated from association can be used for classification is called

A : Naïve Bays

B : Associative Classification

C : Preditctive Mining

D : Decision Tree

Q.no 25. Holdout and random subsampling are common techniques for assessing

A : K-Fold validation

B : cross validation

C : accuracy

D : sampling

Q.no 26. Which statement is true about the decision tree attribute selection process

A : A categorical attribute may appear in a tree node several times but a numeric attribute may appear at most once.

B : A numeric attribute may appear in several tree nodes but a categorical attribute may appear at most once.

C : Both numeric and categorical attributes may appear in several tree nodes.

D : Numeric and categorical attributes may appear in at most one tree node.

Q.no 27. Which of the following is not correct use of cross validation?

A : Selecting variables to include in a model

B : Comparing predictors

C : Selecting parameters in prediction function

D : classification

Q.no 28. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 29. When do you consider an association rule interesting?

A : If it only satisfies minimum support

B : If it only satisfies minimum confidence

C : If it satisfies both minimum support and minimum confidence

D : There are other measures to check interesting rules

Q.no 30. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 31. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 32. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 33. Which is the wrong combination.

A : True negative=correctly indentified

B : False negative=incorrectly identified

C : False positive=correctly identified

D : True positive=correctly identified

Q.no 34. The data cubes are generally

A : 1 Dimensional

B : 2 Dimensional

C : 3 Dimensional

D : n-Dimensional

Q.no 35. A nearest neighbor approach is best used

A : with large-sized datasets.

B : when irrelevant attributes have been removed from the data.

C : when a generalized model of the data is desireable.

D : when an explanation of what has been found is of primary importance.

Q.no 36. The confusion matrix is a useful tool for analyzing

A : Regression

B : Classification

C : Sampling

D : Cross validation

Q.no 37. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 38. What type of data do you need for a chi-square test?

A : Categorical

B : Ordinal

C : Interval

D : Scales

Q.no 39. Sensitivity is also referred to as

A : misclassification rate

B : true negative rate

C : True positive rate

D : correctness

Q.no 40. Number of records are comparatively more in

A : OLAP

B : OLTP

C : Same in OLAP and OLTP

D : Can not compare

Q.no 41. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 42. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 43. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 44. Consider three itemsets V1={tomato, potato,onion}, V2={tomato,potato}, V3={tomato}. Which of the following statement is correct?

A : support(V1) is greater than support (V2)

B : support(V3) is greater than support (V2)

C : support(V1) is greater than support(V3)

D : support(V2) is greater than support(V3)

Q.no 45. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 46. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 47. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 48. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 49. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 50. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 51. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 52. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 53. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 54. What does a Pearson's product-moment allow you to identify?

A : Whether there is a relationship between variables

B : Whether there is a significant effect and interaction of independent variables

C : Whether there is a significant difference between variables

D : Whether there is a significant effect and interaction of dependent variables

Q.no 55. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 56. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 57. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 58. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 59. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 60. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Answer for Question No 1. is c

Answer for Question No 2. is d

Answer for Question No 3. is a

Answer for Question No 4. is a

Answer for Question No 5. is b

Answer for Question No 6. is d

Answer for Question No 7. is c

Answer for Question No 8. is b

Answer for Question No 9. is c

Answer for Question No 10. is d

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is a

Answer for Question No 14. is a

Answer for Question No 15. is c

Answer for Question No 16. is b

Answer for Question No 17. is a

Answer for Question No 18. is d

Answer for Question No 19. is a

Answer for Question No 20. is c

Answer for Question No 21. is c

Answer for Question No 22. is a

Answer for Question No 23. is a

Answer for Question No 24. is b

Answer for Question No 25. is c

Answer for Question No 26. is b

Answer for Question No 27. is d

Answer for Question No 28. is c

Answer for Question No 29. is c

Answer for Question No 30. is b

Answer for Question No 31. is b

Answer for Question No 32. is b

Answer for Question No 33. is c

Answer for Question No 34. is d

Answer for Question No 35. is b

Answer for Question No 36. is b

Answer for Question No 37. is a

Answer for Question No 38. is a

Answer for Question No 39. is c

Answer for Question No 40. is b

Answer for Question No 41. is b

Answer for Question No 42. is d

Answer for Question No 43. is c

Answer for Question No 44. is b

Answer for Question No 45. is d

Answer for Question No 46. is b

Answer for Question No 47. is c

Answer for Question No 48. is d

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is d

Answer for Question No 52. is a

Answer for Question No 53. is a

Answer for Question No 54. is a

Answer for Question No 55. is d

Answer for Question No 56. is d

Answer for Question No 57. is b

Answer for Question No 58. is a

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 2. In Data Characterization, class under study is called as?

A : Study Class

B : Intial Class

C : Target Class

D : Final Class

Q.no 3. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 4. the negative tuples that were correctly labeled by the classifier

A : False positives(FP)

B : True positives(TP)

C : True negatives (TN)

D : False negatives(FN)

Q.no 5. A person trained to interact with a human expert in order to capture their knowledge.

A : knowledge programmer

B : knowledge developer

C : knowledge engineer

D : knowledge extractor

Q.no 6. Removing duplicate records is a process called

A : recovery

B : data cleaning

C : data cleansing

D : data pruning

Q.no 7. Self-training is the simplest form of

A : supervised classification

B : semi-supervised classification

C : unsupervised classification

D : regression

Q.no 8. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 9. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 10. The task of building decision model from labeled training data is called as

A : Supervised Learning

B : Unsupervised Learning

C : Reinforcement Learning

D : Structure Learning

Q.no 11. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 12. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 13. A decision tree is also known as

A : general tree

B : binary tree

C : prediction tree

D : None of the options

Q.no 14. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 15. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 16. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 17. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 18. The example of knowledge type constraints in constraint based mining is

A : Association or Correlation

B : Rule templates

C : Task relevant data

D : Threshold measures

Q.no 19. Which technique finds the frequent itemsets in just two database scans?

A : Partitioning

B : Sampling

C : Hashing

D : Dynamic itemset counting

Q.no 20. A data matrix in which attributes are of the same type and asymmetric is called

A : Pattern matrix

B : Sparse data matrix

C : Document term matrix

D : Normal matrix

Q.no 21. Specificity is also referred to as

A : true negative rate

B : correctness

C : misclassification rate

D : True positive rate

Q.no 22. If first object X and Y coordinates are 3 and 5 respectively and second object X and Y coordinates are 10 and 3 respectively, then what is Manhattan disstance between these two objects?

A : 8

B : 13

C : 9

D : 10

Q.no 23. Which of the following property typically does not hold for similarity measures between two objects ?

A : Symmetry

B : Definiteness

C : Triangle inequality

D : Transitive

Q.no 24. The property of Apriori algorithm is

A : All nonempty subsets of a frequent itemsets must also be frequent

B : All empty subsets of a frequent itemsets must also be frequent

C : All nonempty subsets of a frequent itemsets must be not frequent

D : All nonempty subsets of a frequent itemsets can frequent or not frequent

Q.no 25. One of the most well known software used for classification is

A : Java

B : C4.5

C : Oracle

D : C++

Q.no 26. This supervised learning technique can process both numeric and categorical input attributes.

A : linear regression

B : Bayes classifier

C : logistic regression

D : backpropagation learning

Q.no 27. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 28. K-fold Cross Validation envisages

A : partitioning of the original sample into one sample.

B : partitioning of the original sample into ‘k’ equal sized sub-samples.

C : partitioning of the original sample into ‘k’ unequal sized sub-samples.

D : partitioning of the original sample into ‘k’ random samples.

Q.no 29. The fact table contains

A : The names of the facts

B : Keys to each of the related dimension tables

C : Facts and keys

D : Facts or keys

Q.no 30. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 31. Which of the following operation is correct about supremum distance?

A : It gives maximum difference between any attribute of the objects

B : It gives minimum difference between any attribute of the objects

C : It gives maximum difference between fisrt attribute of the objects

D : It gives minimum difference between fisrt attribute of the objects

Q.no 32. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 33. Sensitivity is also referred to as

A : misclassification rate

B : true negative rate

C : True positive rate

D : correctness

Q.no 34. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 35. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 36. Which of the following sequence is used to calculate proximity measures for ordinal attribute?

A : Replacement discretization and distance measure

B : Replacement characterizarion and distance measure

C : Normalization discretization and distance measure

D : Replacement normalization and distance measure

**Q.no 37. For a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is not true in such a case?**

A : Imbalaced problems should not be measured using Accuracy metric.

B : Accuracy metric is not a good idea for imbalanced class problems.

C : Precision and recall metrics aren’t good for imbalanced class problems.

D : Precision and recall metrics are good for imbalanced class problems.

Q.no 38. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 39. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 40. What is another name of data matrix?

A : Single mode

B : Two mode

C : Multi mode

D : Large mode

Q.no 41. Holdout method, Cross-validation and Bootstrap methods are techniques to estimate

A : Precision

B : Classifier performance

C : Recall

D : F-measure

Q.no 42. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 43. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 44. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 45. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 46. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 47. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 48. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 49. These server performs the faster computation

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 50. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 51. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 52. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 53. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 54. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 55. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 56. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 57. Consider three itemsets V1={tomato, potato,onion}, V2={tomato,potato}, V3={tomato}. Which of the following statement is correct?

A : support(V1) is greater than support (V2)

B : support(V3) is greater than support (V2)

C : support(V1) is greater than support(V3)

D : support(V2) is greater than support(V3)

Q.no 58. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 59. A concept hierarchy that is a total or partial order among attributes in a database schema is called

A : Mixed hierarchy

B : Total hierarchy

C : Schema hierarchy

D : Concept generalization

Q.no 60. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Answer for Question No 1. is d

Answer for Question No 2. is c

Answer for Question No 3. is b

Answer for Question No 4. is c

Answer for Question No 5. is c

Answer for Question No 6. is b

Answer for Question No 7. is b

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is a

Answer for Question No 11. is d

Answer for Question No 12. is b

Answer for Question No 13. is c

Answer for Question No 14. is c

Answer for Question No 15. is b

Answer for Question No 16. is c

Answer for Question No 17. is c

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is b

Answer for Question No 21. is a

Answer for Question No 22. is c

Answer for Question No 23. is c

Answer for Question No 24. is a

Answer for Question No 25. is b

Answer for Question No 26. is b

Answer for Question No 27. is a

Answer for Question No 28. is d

Answer for Question No 29. is c

Answer for Question No 30. is c

Answer for Question No 31. is a

Answer for Question No 32. is d

Answer for Question No 33. is c

Answer for Question No 34. is d

Answer for Question No 35. is b

Answer for Question No 36. is d

Answer for Question No 37. is c

Answer for Question No 38. is b

Answer for Question No 39. is b

Answer for Question No 40. is b

Answer for Question No 41. is b

Answer for Question No 42. is a

Answer for Question No 43. is b

Answer for Question No 44. is a

Answer for Question No 45. is b

Answer for Question No 46. is d

Answer for Question No 47. is c

Answer for Question No 48. is b

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is a

Answer for Question No 53. is d

Answer for Question No 54. is d

Answer for Question No 55. is d

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is d

Answer for Question No 59. is c

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. The problem of agents to learn from the environment by their interactions with dynamic environment is done in

A : Reinforcement learning

B : Multi-label classification

C : Binary Classification

D : Multiclassification

Q.no 2. Baysian classification in based on

A : probability for the hypothesis

B : Support

C : tree induction

D : Trees

Q.no 3. Which of the following is correct about Proximity measures?

A : Similarity

B : Dissimilarity

C : Similarity as well as Dissimilarity

D : Neither similarity nor dissimilarity

Q.no 4. For Apriori algorithm, what is the second phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 5. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 6. The most widely used metrics and tools to assess a classification model are:

A : Conusion Matrix

B : Support

C : Entropy

D : Probability

Q.no 7. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 8. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 9. Multilevel association rules can be mined efficiently using

A : Support

B : Confidence

C : Support count

D : Concept Hierarchies under support-confidence framework

Q.no 10. Which of the following is not a type of constraints?

A : Data constraints

B : Rule constraints

C : Knowledge type constraints

D : Time constraints

Q.no 11. Data matrix is also called as

A : Object by object structure

B : Object by attribute structure

C : Attribute by attribute structure

D : Attribute by object structure

Q.no 12. Each dimension is represented by only one table. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 13. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 14. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 15. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 16. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 17. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 18. cross-validation and bootstrap methods are common techniques for assessing

A : accuracy

B : Precision

C : recall

D : performance

Q.no 19. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 20. OLAP database design is

A : Application-oriented

B : Object-oriented

C : Goal-oriented

D : Subject-oriented

Q.no 21. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 22. This supervised learning technique can process both numeric and categorical input attributes.

A : linear regression

B : Bayes classifier

C : logistic regression

D : backpropagation learning

Q.no 23. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 24. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 25. A frequent pattern tree is a tree structure consisting of

A : A frequent-item-node

B : An item-prefix-tree

C : A frequent-item-header table

D : both B and C

Q.no 26. Learning with a complete system in mind with reference to interactions among
the systems and subsystems with proper understanding of systemic boundaries is

A : Multi-label classification

B : Reinforcement learning

C : Systemic learning

D : Machine Learning

Q.no 27. Handwritten digit recognition classifying an image of a handwritten number into a digit from 0 to 9 is example of

A : Multiclassification

B : Multi-label classification

C : Imbalanced classification

D : Binary Classification

Q.no 28. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 29. The frequent-item-header-table consists of number fields

A : Only one

B : Two

C : Three

D : Four

Q.no 30. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 31. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 32. What is another name of data matrix?

A : Single mode

B : Two mode

C : Multi mode

D : Large mode

Q.no 33. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 34. In a decision tree each leaf node represents

A : Test conditions

B : Class labels

C : Attribute values

D : Decision

Q.no 35. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 36. If first object X and Y coordinates are 3 and 5 respectively and second object X and Y coordinates are 10 and 3 respectively, then what is Manhattan disstance between these two objects?

A : 8

B : 13

C : 9

D : 10

Q.no 37. one-versus-one(OVO) and one-versus-all (OVA) classification involves

A : more than two classes

B : Only two classes

C : Only one class

D : No class

Q.no 38. K-fold Cross Validation envisages

A : partitioning of the original sample into one sample.

B : partitioning of the original sample into ‘k’ equal sized sub-samples.

C : partitioning of the original sample into ‘k’ unequal sized sub-samples.

D : partitioning of the original sample into ‘k’ random samples.

Q.no 39. One of the most well known software used for classification is

A : Java

B : C4.5

C : Oracle

D : C++

Q.no 40. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 41. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 42. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 43. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 44. Which of the following operation is requird to calculate cosine similarity?

A : Vector dot product

B : Exponent

C : Modulus

D : Percentage

Q.no 45. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 46. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 47. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 48. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 49. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 50. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 51. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 52. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 53. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 54. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 55. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 56. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 57. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 58. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 59. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 60. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is c

Answer for Question No 4. is a

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is c

Answer for Question No 8. is d

Answer for Question No 9. is d

Answer for Question No 10. is d

Answer for Question No 11. is b

Answer for Question No 12. is a

Answer for Question No 13. is d

Answer for Question No 14. is b

Answer for Question No 15. is c

Answer for Question No 16. is b

Answer for Question No 17. is d

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is d

Answer for Question No 21. is d

Answer for Question No 22. is b

Answer for Question No 23. is d

Answer for Question No 24. is c

Answer for Question No 25. is d

Answer for Question No 26. is c

Answer for Question No 27. is a

Answer for Question No 28. is a

Answer for Question No 29. is b

Answer for Question No 30. is a

Answer for Question No 31. is d

Answer for Question No 32. is b

Answer for Question No 33. is b

Answer for Question No 34. is b

Answer for Question No 35. is a

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is c

Answer for Question No 41. is d

Answer for Question No 42. is a

Answer for Question No 43. is d

Answer for Question No 44. is a

Answer for Question No 45. is b

Answer for Question No 46. is a

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is c

Answer for Question No 52. is d

Answer for Question No 53. is a

Answer for Question No 54. is a

Answer for Question No 55. is d

Answer for Question No 56. is b

Answer for Question No 57. is d

Answer for Question No 58. is b

Answer for Question No 59. is a

Answer for Question No 60. is a

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 2. Data mining is best described as the process of

A : identifying patterns in data

B : deducing relationships in data

C : representing data

D : simulating trends in data

Q.no 3. These are the intermediate servers that stand in between a relational back-end server and client front-end tools

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 4. What does the leaf node in decision tree indicates

A : sub tree

B : class label

C : testing node

D : condition

Q.no 5. A multidimensional data model is typically organized around a central theme which is represented by

A : Dimension table

B : Fact table

C : Dimension table and Fact table

D : Dimension table or Fact table

Q.no 6. Data used to build a data mining model.

A : Validation Data

B : Training Data

C : Testing Data

D : Hidden Data

Q.no 7. The Synonym for data mining is

A : Data warehouse

B : Knowledge discovery in database

C : ETL

D : Business Intelligemce

Q.no 8. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 9. Cotraining is one form of

A : sampling

B : Reinforcement learning

C : unsupervised classification

D : semi-supervised classification

Q.no 10. What is C4.5 is used to build

A : Decision tree

B : Regression Analysis

C : Induction

D : Association Rules

Q.no 11. Training process that generates tree is called as

A : Pruning

B : Rule generation

C : Induction

D : spliiting

Q.no 12. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 13. Which of the following is not frequent pattern?

A : Itemsets

B : Subsequences

C : Substructures

D : Associations

Q.no 14. What is an alternative form of Euclidean distance?

A : L1 norm

B : L2 norm

C : Lmax norm

D : L norm

Q.no 15. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 16. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 17. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 18. Which of the following are methods for supervised classification?

A : Decision tree

B : K-Means

C : Hierarchical

D : Apriori

Q.no 19. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 20. Removing duplicate records is a process called

A : recovery

B : data cleaning

C : data cleansing

D : data pruning

Q.no 21. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 22. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 23. If x and y are two objects of nominal attribute with COMP and IT values respectively, then what is the similarity between these two objects?

A : Zero

B : Infinity

C : Two

D : One

Q.no 24. The accuracy of a classifier on a given test set is the percentage of

A : test set tuples that are correctly classified by the classifier

B : test set tuples that are incorrectly classified by the classifier

C : test set tuples that are incorrectly misclassified by the classifier

D : test set tuples that are not classified by the classifier

Q.no 25. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 26. What is uniform support in multilevel association rule minig?

A : Use of minimum support

B : Use of minimum support and confidence

C : Use of same minimum threshold at each abstraction level

D : Use of minimum support and support count

Q.no 27. Which of the following is not correct use of cross validation?

A : Selecting variables to include in a model

B : Comparing predictors

C : Selecting parameters in prediction function

D : classification

Q.no 28. The frequent-item-header-table consists of number fields

A : Only one

B : Two

C : Three

D : Four

Q.no 29. Which of these distributions is used for a testing hypothesis?

A : Normal Distribution

B : Chi-Squared Distribution

C : Gamma Distribution

D : Poisson Distribution

Q.no 30. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 31. Joins will be needed to execute the query. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 32. Which of the following sequence is used to calculate proximity measures for ordinal attribute?

A : Replacement discretization and distance measure

B : Replacement characterizarion and distance measure

C : Normalization discretization and distance measure

D : Replacement normalization and distance measure

Q.no 33. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 34. Which statement is true about the KNN algorithm?

A : All attribute values must be categorical

B : The output attribute must be cateogrical.

C : Attribute values may be either categorical or numeric.

D : All attributes must be numeric.

Q.no 35. The correlation coefficient is used to determine:

A : A specific value of the y-variable given a specific value of the x-variable

B : A specific value of the x-variable given a specific value of the y-variable

C : The strength of the relationship between the x and y variables

D : None of these

Q.no 36. What type of data do you need for a chi-square test?

A : Categorical

B : Ordinal

C : Interval

D : Scales

Q.no 37. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 38. Which of the following is measure of document similarity?

A : Cosine dissimilarity

B : Sine similarity

C : Sine dissimilarity

D : Cosine similarity

Q.no 39. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 40. In multilevel association rules, which strategy is employed

A : Top-down

B : Recursive

C : Bottom-up

D : Divide and conquer

Q.no 41. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 42. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 43. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 44. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 45. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 46. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 47. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 48. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 49. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 50. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 51. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 52. Holdout method, Cross-validation and Bootstrap methods are techniques to estimate

A : Precision

B : Classifier performance

C : Recall

D : F-measure

Q.no 53. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 54. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 55. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 56. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 57. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 58. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 59. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 60. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Answer for Question No 1. is c

Answer for Question No 2. is a

Answer for Question No 3. is a

Answer for Question No 4. is b

Answer for Question No 5. is b

Answer for Question No 6. is b

Answer for Question No 7. is b

Answer for Question No 8. is a

Answer for Question No 9. is d

Answer for Question No 10. is a

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is d

Answer for Question No 14. is b

Answer for Question No 15. is a

Answer for Question No 16. is a

Answer for Question No 17. is b

Answer for Question No 18. is a

Answer for Question No 19. is c

Answer for Question No 20. is b

Answer for Question No 21. is c

Answer for Question No 22. is d

Answer for Question No 23. is a

Answer for Question No 24. is a

Answer for Question No 25. is a

Answer for Question No 26. is c

Answer for Question No 27. is d

Answer for Question No 28. is b

Answer for Question No 29. is b

Answer for Question No 30. is a

Answer for Question No 31. is b

Answer for Question No 32. is d

Answer for Question No 33. is b

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is a

Answer for Question No 37. is b

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is a

Answer for Question No 41. is c

Answer for Question No 42. is a

Answer for Question No 43. is d

Answer for Question No 44. is d

Answer for Question No 45. is d

Answer for Question No 46. is b

Answer for Question No 47. is d

Answer for Question No 48. is d

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is b

Answer for Question No 53. is c

Answer for Question No 54. is a

Answer for Question No 55. is c

Answer for Question No 56. is a

Answer for Question No 57. is d

Answer for Question No 58. is c

Answer for Question No 59. is d

Answer for Question No 60. is d

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. The Synonym for data mining is

A : Data warehouse

B : Knowledge discovery in database

C : ETL

D : Business Intelligemce

Q.no 2. The example of knowledge type constraints in constraint based mining is

A : Association or Correlation

B : Rule templates

C : Task relevant data

D : Threshold measures

Q.no 3. If two documents are similar, then what is the measure of angle between two documents?

A : 30

B : 60

C : 90

D : 0

Q.no 4. The most widely used metrics and tools to assess a classification model are:

A : Conusion Matrix

B : Support

C : Entropy

D : Probability

Q.no 5. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 6. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 7. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 8. Cotraining is one form of

A : sampling

B : Reinforcement learning

C : unsupervised classification

D : semi-supervised classification

Q.no 9. Which is the keyword that distinguishes data warehouses from other data repository systems ?

A : Subject-oriented

B : Object-oriented

C : Client server

D : Time-invariant

Q.no 10. Self-training is the simplest form of

A : supervised classification

B : semi-supervised classification

C : unsupervised classification

D : regression

Q.no 11. Which of the following is correct about Proximity measures?

A : Similarity

B : Dissimilarity

C : Similarity as well as Dissimilarity

D : Neither similarity nor dissimilarity

Q.no 12. For Apriori algorithm, what is the first phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 13. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 14. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 15. What is C4.5 is used to build

A : Decision tree

B : Regression Analysis

C : Induction

D : Association Rules

Q.no 16. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 17. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 18. An automatic car driver and business intelligent systems are examples of

A : Regression

B : Classification

C : Machine Learning

D : Reinforcement learning

Q.no 19. Which of the following is direct application of frequent itemset mining?

A : Social Network Analysis

B : Market Basket Analysis

C : Outlier Detection

D : Intrusion Detection

Q.no 20. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 21. The Microsoft SQL Server 2000 is the example of

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 22. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 23. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 24. What is uniform support in multilevel association rule minig?

A : Use of minimum support

B : Use of minimum support and confidence

C : Use of same minimum threshold at each abstraction level

D : Use of minimum support and support count

Q.no 25. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 26. Where does the bayes rule used?

A : Solving queries

B : Increasing complexity

C : Decreasing complexity

D : Answering probabilistic query

Q.no 27. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 28. If x and y are two objects of nominal attribute with COMP and IT values respectively, then what is the similarity between these two objects?

A : Zero

B : Infinity

C : Two

D : One

Q.no 29. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 30. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 31. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 32. In a decision tree each leaf node represents

A : Test conditions

B : Class labels

C : Attribute values

D : Decision

Q.no 33. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 34. To improve the accuracy of multiclass classification we can use

A : cross validation

B : sampling

C : Error-detecting codes

D : Error-correcting codes

Q.no 35. Cross validation involves

A : testing the machine on all possible ways by substituting the original sample into training set

B : testing the machine on all possible ways by dividing the original sample into training and validation sets.

C : testing the machine with only validation sets

D : testing the machine on only testing datasets.

Q.no 36. OLAP Summarization means

A : Consolidated

B : Primitive

C : Highly detailed

D : Recent data

Q.no 37. Identify the example of sequence data

A : weather forecast

B : data matrix

C : market basket data

D : genomic data

Q.no 38. Which of the following is necessary operation to calculate dissimilarity between ordinal attributes?

A : Replacement of ordinal categories

B : Correlation coefficient

C : Discretization

D : Randomization

Q.no 39. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 40. Which of the following probabilities are used in the Bayes theorem.

A : P(Ci|X)

B : P(Ci)

C : P(X|Ci)

D : P(X)

Q.no 41. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 42. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 43. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 44. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 45. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 46. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 47. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 48. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 49. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 50. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 51. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 52. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 53. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 54. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 55. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 56. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 57. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 58. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 59. A concept hierarchy that is a total or partial order among attributes in a database schema is called

A : Mixed hierarchy

B : Total hierarchy

C : Schema hierarchy

D : Concept generalization

Q.no 60. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Answer for Question No 1. is b

Answer for Question No 2. is a

Answer for Question No 3. is d

Answer for Question No 4. is a

Answer for Question No 5. is b

Answer for Question No 6. is d

Answer for Question No 7. is d

Answer for Question No 8. is d

Answer for Question No 9. is a

Answer for Question No 10. is b

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is c

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is b

Answer for Question No 17. is c

Answer for Question No 18. is d

Answer for Question No 19. is b

Answer for Question No 20. is a

Answer for Question No 21. is c

Answer for Question No 22. is c

Answer for Question No 23. is d

Answer for Question No 24. is c

Answer for Question No 25. is b

Answer for Question No 26. is d

Answer for Question No 27. is b

Answer for Question No 28. is a

Answer for Question No 29. is d

Answer for Question No 30. is d

Answer for Question No 31. is b

Answer for Question No 32. is b

Answer for Question No 33. is a

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is a

Answer for Question No 37. is d

Answer for Question No 38. is a

Answer for Question No 39. is c

Answer for Question No 40. is a

Answer for Question No 41. is c

Answer for Question No 42. is d

Answer for Question No 43. is b

Answer for Question No 44. is d

Answer for Question No 45. is c

Answer for Question No 46. is a

Answer for Question No 47. is a

Answer for Question No 48. is a

Answer for Question No 49. is a

Answer for Question No 50. is d

Answer for Question No 51. is a

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is b

Answer for Question No 55. is c

Answer for Question No 56. is b

Answer for Question No 57. is c

Answer for Question No 58. is a

Answer for Question No 59. is c

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 2. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 3. What is C4.5 is used to build

A : Decision tree

B : Regression Analysis

C : Induction

D : Association Rules

Q.no 4. Which of the following is not frequent pattern?

A : Itemsets

B : Subsequences

C : Substructures

D : Associations

Q.no 5. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 6. A data cube is defined by

A : Dimensions

B : Facts

C : Dimensions and Facts

D : Dimensions or Facts

Q.no 7. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 8. Which of the following is the data mining tool?

A : Borland C

B : Weka

C : Borland C++

D : Visual C

Q.no 9. Cotraining is one form of

A : sampling

B : Reinforcement learning

C : unsupervised classification

D : semi-supervised classification

Q.no 10. Each dimension is represented by only one table. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 11. What are two steps of tree pruning work?

A : Pessimistic pruning and Optimistic pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : None of the options

Q.no 12. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 13. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 14. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 15. to evaluate a classifier’s quality we use

A : confusion matrix

B : error detection code

C : error correction code

D : classifier

Q.no 16. accuracy is used to measure

A : classifier’s true abilities

B : classifier’s analytic abilities

C : classifier’s decision abilities

D : classifier’s predictive abilities

Q.no 17. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 18. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 19. What are closed frequent itemsets?

A : A closed itemset

B : A frequent itemset

C : An itemset which is both closed and frequent

D : Not frequent itemset

Q.no 20. In Data Characterization, class under study is called as?

A : Study Class

B : Intial Class

C : Target Class

D : Final Class

Q.no 21. A nearest neighbor approach is best used

A : with large-sized datasets.

B : when irrelevant attributes have been removed from the data.

C : when a generalized model of the data is desireable.

D : when an explanation of what has been found is of primary importance.

Q.no 22. Lazy learner classification approach is

A : learner waits until the last minute before constructing model to classify

B : a given training data constructs a model first and then uses it to classify

C : the network is constructed by human experts

D : None of the options

Q.no 23. Which of the following probabilities are used in the Bayes theorem.

A : P(Ci|X)

B : P(Ci)

C : P(X|Ci)

D : P(X)

Q.no 24. A frequent pattern tree is a tree structure consisting of

A : A frequent-item-node

B : An item-prefix-tree

C : A frequent-item-header table

D : both B and C

Q.no 25. Holdout and random subsampling are common techniques for assessing

A : K-Fold validation

B : cross validation

C : accuracy

D : sampling

Q.no 26. Specificity is also referred to as

A : true negative rate

B : correctness

C : misclassification rate

D : True positive rate

Q.no 27. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 28. To improve the accuracy of multiclass classification we can use

A : cross validation

B : sampling

C : Error-detecting codes

D : Error-correcting codes

Q.no 29. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 30. one-versus-one(OVO) and one-versus-all (OVA) classification involves

A : more than two classes

B : Only two classes

C : Only one class

D : No class

Q.no 31. OLAP Summarization means

A : Consolidated

B : Primitive

C : Highly detailed

D : Recent data

Q.no 32. When you use cross validation in machine learning, it means

A : you verify how accurate your model is on multiple and different subsets of data.

B : you verify how accurate your model is on same dataset.

C : you verify how accurate your model is on new dataset.

D : you verify how accurate your model on unknown dataset

Q.no 33. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 34. Which of the following operations are used to calculate proximity measures for ordinal attribute?

A : Replacement and discretization

B : Replacement and characterizarion

C : Replacement and normalization

D : Normalization and discretization

Q.no 35. In Apriori algorithm, for generating e. g. 5 itemsets, we use

A : Frequent 5 itemsets

B : Frequent 3 itemsets

C : Frequent 4 itemsets

D : Frequent 6 itemsets

Q.no 36. Which of the following is a predictive model?

A : Clustering

B : Regression

C : Summarization

D : Association rules

Q.no 37. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 38. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 39. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 40. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 41. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 42. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 43. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 44. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 45. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 46. Which operation data warehouse requires ?

A : Initial loading of data

B : Transaction processing

C : Recovery

D : Concurrency control mechanisms

Q.no 47. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 48. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 49. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 50. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 51. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 52. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 53. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 54. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 55. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 56. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 57. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 58. A data normalization technique for real-valued attributes that divides each numerical value by the same power of 10.

A : min-max normalization

B : z-score normalization

C : decimal scaling

D : decimal smoothing

Q.no 59. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 60. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Answer for Question No 1. is a

Answer for Question No 2. is d

Answer for Question No 3. is a

Answer for Question No 4. is d

Answer for Question No 5. is b

Answer for Question No 6. is c

Answer for Question No 7. is d

Answer for Question No 8. is b

Answer for Question No 9. is d

Answer for Question No 10. is a

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is d

Answer for Question No 17. is c

Answer for Question No 18. is c

Answer for Question No 19. is c

Answer for Question No 20. is c

Answer for Question No 21. is b

Answer for Question No 22. is a

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is a

Answer for Question No 27. is b

Answer for Question No 28. is d

Answer for Question No 29. is d

Answer for Question No 30. is a

Answer for Question No 31. is a

Answer for Question No 32. is a

Answer for Question No 33. is a

Answer for Question No 34. is c

Answer for Question No 35. is c

Answer for Question No 36. is b

Answer for Question No 37. is b

Answer for Question No 38. is b

Answer for Question No 39. is c

Answer for Question No 40. is a

Answer for Question No 41. is a

Answer for Question No 42. is d

Answer for Question No 43. is c

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is d

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is c

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is a

Answer for Question No 58. is c

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. For Apriori algorithm, what is the second phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 2. Which of these is not a frequent pattern mining algorithm?

A : Decision trees

B : Eclat

C : FP growth

D : Apriori

Q.no 3. Which of the following is not a type of constraints?

A : Data constraints

B : Rule constraints

C : Knowledge type constraints

D : Time constraints

Q.no 4. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 5. If two documents are similar, then what is the measure of angle between two documents?

A : 30

B : 60

C : 90

D : 0

Q.no 6. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 7. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 8. The fact is also called as

A : Dimension

B : Key

C : Schema

D : Measure

Q.no 9. The most widely used metrics and tools to assess a classification model are:

A : Conusion Matrix

B : Support

C : Entropy

D : Probability

Q.no 10. A person trained to interact with a human expert in order to capture their knowledge.

A : knowledge programmer

B : knowledge developer

C : knowledge engineer

D : knowledge extractor

Q.no 11. Training process that generates tree is called as

A : Pruning

B : Rule generation

C : Induction

D : spliiting

Q.no 12. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 13. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 14. to evaluate a classifier’s quality we use

A : confusion matrix

B : error detection code

C : error correction code

D : classifier

Q.no 15. For Apriori algorithm, what is the first phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 16. The example of knowledge type constraints in constraint based mining is

A : Association or Correlation

B : Rule templates

C : Task relevant data

D : Threshold measures

Q.no 17. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 18. A data cube is defined by

A : Dimensions

B : Facts

C : Dimensions and Facts

D : Dimensions or Facts

Q.no 19. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 20. What are two steps of tree pruning work?

A : Pessimistic pruning and Optimistic pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : None of the options

Q.no 21. The Microsoft SQL Server 2000 is the example of

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 22. The property of Apriori algorithm is

A : All nonempty subsets of a frequent itemsets must also be frequent

B : All empty subsets of a frequent itemsets must also be frequent

C : All nonempty subsets of a frequent itemsets must be not frequent

D : All nonempty subsets of a frequent itemsets can frequent or not frequent

Q.no 23. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 24. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 25. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 26. Sensitivity is also referred to as

A : misclassification rate

B : true negative rate

C : True positive rate

D : correctness

Q.no 27. In Apriori algorithm, for generating e. g. 5 itemsets, we use

A : Frequent 5 itemsets

B : Frequent 3 itemsets

C : Frequent 4 itemsets

D : Frequent 6 itemsets

Q.no 28. Handwritten digit recognition classifying an image of a handwritten number into a digit from 0 to 9 is example of

A : Multiclassification

B : Multi-label classification

C : Imbalanced classification

D : Binary Classification

Q.no 29. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 30. Specificity is also referred to as

A : true negative rate

B : correctness

C : misclassification rate

D : True positive rate

Q.no 31. To improve the accuracy of multiclass classification we can use

A : cross validation

B : sampling

C : Error-detecting codes

D : Error-correcting codes

Q.no 32. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 33. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 34. For a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is not true in such a case?

A : Imbalaced problems should not be measured using Accuracy metric.

B : Accuracy metric is not a good idea for imbalanced class problems.

C : Precision and recall metrics aren’t good for imbalanced class problems.

D : Precision and recall metrics are good for imbalanced class problems.

Q.no 35. one-versus-one(OVO) and one-versus-all (OVA) classification involves

A : more than two classes

B : Only two classes

C : Only one class

D : No class

Q.no 36. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 37. OLAP Summarization means

A : Consolidated

B : Primitive

C : Highly detailed

D : Recent data

Q.no 38. A frequent pattern tree is a tree structure consisting of

A : A frequent-item-node

B : An item-prefix-tree

C : A frequent-item-header table

D : both B and C

Q.no 39. The confusion matrix is a useful tool for analyzing

A : Regression

B : Classification

C : Sampling

D : Cross validation

Q.no 40. Cross validation involves

A : testing the machine on all possible ways by substituting the original sample into training set

B : testing the machine on all possible ways by dividing the original sample into training and validation sets.

C : testing the machine with only validation sets

D : testing the machine on only testing datasets.

Q.no 41. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 42. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 43. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 44. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 45. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 46. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 47. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 48. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 49. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 50. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 51. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 52. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 53. A data normalization technique for real-valued attributes that divides each numerical value by the same power of 10.

A : min-max normalization

B : z-score normalization

C : decimal scaling

D : decimal smoothing

Q.no 54. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 55. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 56. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 57. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 58. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 59. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 60. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is d

Answer for Question No 4. is d

Answer for Question No 5. is d

Answer for Question No 6. is b

Answer for Question No 7. is c

Answer for Question No 8. is d

Answer for Question No 9. is a

Answer for Question No 10. is c

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is c

Answer for Question No 16. is a

Answer for Question No 17. is d

Answer for Question No 18. is c

Answer for Question No 19. is a

Answer for Question No 20. is b

Answer for Question No 21. is c

Answer for Question No 22. is a

Answer for Question No 23. is c

Answer for Question No 24. is a

Answer for Question No 25. is d

Answer for Question No 26. is c

Answer for Question No 27. is c

Answer for Question No 28. is a

Answer for Question No 29. is a

Answer for Question No 30. is a

Answer for Question No 31. is d

Answer for Question No 32. is b

Answer for Question No 33. is c

Answer for Question No 34. is c

Answer for Question No 35. is a

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is c

Answer for Question No 41. is d

Answer for Question No 42. is c

Answer for Question No 43. is a

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is b

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is d

Answer for Question No 50. is a

Answer for Question No 51. is d

Answer for Question No 52. is c

Answer for Question No 53. is c

Answer for Question No 54. is b

Answer for Question No 55. is d

Answer for Question No 56. is c

Answer for Question No 57. is d

Answer for Question No 58. is c

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 2. OLAP database design is

A : Application-oriented

B : Object-oriented

C : Goal-oriented

D : Subject-oriented

Q.no 3. Multilevel association rules can be mined efficiently using

A : Support

B : Confidence

C : Support count

D : Concept Hierarchies under support-confidence framework

Q.no 4. accuracy is used to measure

A : classifier’s true abilities

B : classifier’s analytic abilities

C : classifier’s decision abilities

D : classifier’s predictive abilities

Q.no 5. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 6. The task of building decision model from labeled training data is called as

A : Supervised Learning

B : Unsupervised Learning

C : Reinforcement Learning

D : Structure Learning

Q.no 7. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 8. Multi-class classification makes the assumption that each sample is assigned to

A : one and only one label

B : many labels

C : one or many labels

D : no label

Q.no 9. Which of these is not a frequent pattern mining algorithm?

A : Decision trees

B : Eclat

C : FP growth

D : Apriori

Q.no 10. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 11. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 12. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 13. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 14. Each dimension is represented by only one table. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 15. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 16. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 17. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 18. Which of the following is the data mining tool?

A : Borland C

B : Weka

C : Borland C++

D : Visual C

Q.no 19. A decision tree is also known as

A : general tree

B : binary tree

C : prediction tree

D : None of the options

Q.no 20. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 21. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 22. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 23. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 24. Which of the following sequence is used to calculate proximity measures for ordinal attribute?

A : Replacement discretization and distance measure

B : Replacement characterizarion and distance measure

C : Normalization discretization and distance measure

D : Replacement normalization and distance measure

Q.no 25. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 26. Which of the following is not correct use of cross validation?

A : Selecting variables to include in a model

B : Comparing predictors

C : Selecting parameters in prediction function

D : classification

Q.no 27. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 28. The fact table contains

A : The names of the facts

B : Keys to each of the related dimension tables

C : Facts and keys

D : Facts or keys

Q.no 29. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 30. The accuracy of a classifier on a given test set is the percentage of

A : test set tuples that are correctly classified by the classifier

B : test set tuples that are incorrectly classified by the classifier

C : test set tuples that are incorrectly misclassified by the classifier

D : test set tuples that are not classified by the classifier

Q.no 31. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 32. The confusion matrix is a useful tool for analyzing

A : Regression

B : Classification

C : Sampling

D : Cross validation

Q.no 33. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 34. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 35. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 36. One of the most well known software used for classification is

A : Java

B : C4.5

C : Oracle

D : C++

Q.no 37. Identify the example of sequence data

A : weather forecast

B : data matrix

C : market basket data

D : genomic data

Q.no 38. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 39. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 40. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 41. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 42. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 43. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 44. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 45. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 46. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 47. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 48. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 49. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 50. What does a Pearson's product-moment allow you to identify?

A : Whether there is a relationship between variables

B : Whether there is a significant effect and interaction of independent variables

C : Whether there is a significant difference between variables

D : Whether there is a significant effect and interaction of dependent variables

Q.no 51. Consider three itemsets V1={tomato, potato,onion}, V2={tomato,potato}, V3={tomato}. Which of the following statement is correct?

A : support(V1) is greater than support (V2)

B : support(V3) is greater than support (V2)

C : support(V1) is greater than support(V3)

D : support(V2) is greater than support(V3)

Q.no 52. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 53. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 54. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 55. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 56. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 57. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 58. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 59. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 60. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Answer for Question No 1. is b

Answer for Question No 2. is d

Answer for Question No 3. is d

Answer for Question No 4. is d

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is a

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is d

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is d

Answer for Question No 14. is a

Answer for Question No 15. is b

Answer for Question No 16. is d

Answer for Question No 17. is c

Answer for Question No 18. is b

Answer for Question No 19. is c

Answer for Question No 20. is a

Answer for Question No 21. is a

Answer for Question No 22. is a

Answer for Question No 23. is d

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is d

Answer for Question No 27. is c

Answer for Question No 28. is c

Answer for Question No 29. is d

Answer for Question No 30. is a

Answer for Question No 31. is b

Answer for Question No 32. is b

Answer for Question No 33. is b

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is b

Answer for Question No 37. is d

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is b

Answer for Question No 41. is d

Answer for Question No 42. is b

Answer for Question No 43. is d

Answer for Question No 44. is a

Answer for Question No 45. is c

Answer for Question No 46. is b

Answer for Question No 47. is d

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is c

Answer for Question No 53. is c

Answer for Question No 54. is b

Answer for Question No 55. is b

Answer for Question No 56. is a

Answer for Question No 57. is d

Answer for Question No 58. is a

Answer for Question No 59. is a

Answer for Question No 60. is d