	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