# DATA MINING QUESTIONS BANK

| Sr. No. | Questions |
|---|---|
| 1 | The earliest step in the data mining process is usually?<br>A. Visualization<br>**B. Preprocessing**<br>C. Modelling<br>D. Deployment |
| 2 | If a record data matrix has reduced number of rows after a transformation, the transformation has performed:<br>**A. Data Sampling**<br>B. Dimensionality Reduction<br>C. Noise Cleaning<br>D. Discretization |
| 3 | Friendship structure of users in a social networking site can be considered as an example of:<br>**A. Record data**<br>B. Ordered data<br>C. Graph data<br>D. None of the above |
| 4 | A decision tree can be used to build models for:<br>A. Regression problems<br>B. Classification problems<br>**C. Both of the above**<br>D. None of the above |
| 5 | If a decision tree is expressed as a set of logical rules, then:<br>A. the internal nodes in a branch are connected by AND and the branches by AND.<br>B. the internal nodes in a branch are connected by OR and the branches by OR.<br>C. **the internal nodes in a branch are connected by AND and the branches by OR.**<br>D. the internal nodes in a branch are connected by OR and the branches by AND. |
| 6 | In data mining, how many categories of functions are included?<br>A. 5<br>B. 4<br>**C. 2**<br>D. 3 |
| 7 | A relational database logically stores data as:<br>A. Files<br>**B. Tables**<br>C. Trees<br>D. Arrays |

| | |
|---|---|
| 8 | A committee of 3 person is to be formed from 3 men and 2 women. In how many ways can the committee be formed if we have to include at least one women?<br>A. 9<br>B. 1<br>C. 3<br>**D. 6** |
| 9 | Mode of the set of numbers {5,6,7,10,11} is?<br>A. 1.5<br>B. 2.5<br>**C. No mode**<br>D. 4.0 |
| 10 | One card is drawn randomly from a well shuffled deck of 52 cards. What is the probability that the card drawn is not an ace?<br>A. 1/13<br>**B. 12/13**<br>C. 1/4<br>D. 1/8 |
| 11 | What is the time complexity to find an element x in an array? The array is sorted in decreasing order.<br>**a. O(logn)**<br>b. O(n)<br>c. O(n2)<br>d. O(n3) |
| 12 | Which one of the following can be defined as the data object which does not comply with the general behavior (or the model of available data)?<br>A. Evaluation Analysis<br>**B. Outliner Analysis**<br>C. Classification<br>D. Prediction |
| 13 | Entropy value of _____ represents that the data sample has a 50-50 split belonging to two categories:<br>A. **1**<br>B. 0<br>C. 0.5<br>D. None of the above |
| 14 | Which of the following is the correct advantage of the Update-Driven Approach?<br>A. This approach provides high performance.<br>B. The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.<br>**C. Both A and B**<br>D. None of the above |

| | |
|---|---|
| 15 | Which one of the clustering technique needs the merging approach?<br>A. Partitioned<br>B. Naïve Bayes<br>**C. Hierarchical**<br>D. Both A and C |
| 16 | Which one of the following can be considered as the correct application of the data mining?<br>A. Fraud detection<br>B. Corporate Analysis & Risk management<br>C. Management and market analysis<br>**D. All of the above** |
| 17 | Consider a binary classification problem with two classes C1 and C2. Class labels of ten other training set instances sorted in increasing order of their distance to an instance x is as follows: {C1, C2, C1, C2, C2, C2, C1, C2, C1, C2}. How will a K=5 nearest neighbor classifier classify x?<br>A. There will be a tie<br>B. C1<br>**C. C2**<br>D. Not enough information to classify |
| 18 | Issues with Euclidean measure are: (1 mark)<br>A. High dimensional data.<br>B. Can produce counter-intuitive results.<br>C. Shrinking density – sparsification effect<br>**D. All of the above.** |
| 19 | Artificial neural networks can be used for:<br>A. Pattern Recognition<br>B. Classification<br>C. Clustering<br>**D. All of the above** |
| 20 | A perceptron can correctly classify instances into two classes where the classes are:<br>A. Overlapping<br>B. **Linearly separable**<br>C. Non-linearly separable<br>D. None of the above |
| 21 | The logic function that cannot be implemented by a perceptron having two inputs is?<br>A. AND<br>B. OR<br>C. NOR<br>**D. XOR** |
| 22 | A neuron with 3 inputs has the weight vector [0.2 -0.1 0.1]^T and a bias θ = 0. If the input vector is X = [0.2 0.4 0.2]^T , then the total input to the neuron is: |

|  |  |
|---|---|
|  | A. 0.2<br>**B. 0.02**<br>C. 0.4<br>D. 0.10 |
| 23 | Overfitting is expected when we observe that?<br>A.With training iterations, error on training set as well as test set decreases<br>**B. With training iterations, error on training set decreases but test set increases**<br>C. With training iterations, error on training set as well as test set increases<br>D. With training iterations, training set as well as test set error remains constant |
| 24 | The leaves of a dendrogram in hierarchical clustering represent?<br>A. **Individual data points**<br>B. Clusters of multiple data points<br>C. Distances between data points<br>D. Cluster membership value of the data points |
| 25 | The classification or mapping of a class using a predefined class or group is called:<br>a. Data Sub Structure<br>b. Data Set<br>c. **Data Discrimination**<br>d. Data Characterisation |
| 26 | Which one of the following refers to the binary attribute?<br>**A. This takes only two values. In general, these values will be 0 and 1, and they can be coded as one bit**<br>B. The natural environment of a certain species<br>C. Systems that can be used without knowledge of internal operations<br>D. All of the above |
| 27 | The class under study in Data Characterization is known as:<br>a. Final Class<br>b. **Target Class**<br>c. InitialClass<br>d. Study Class |
| 28 | Which one of the following refers to querying the unstructured textual data?<br>A. Information access<br>B. Information update<br>**C. Information retrieval**<br>D. Information manipulation |
| 29 | Which of the following also used as the first step in the knowledge discovery process?<br>A. Data selection<br>**B. Data cleaning**<br>C. Data transformation |

| | |
|---|---|
| | D. Data integration |
| 30 | Regression is used in:<br>A. **predictive data mining**<br>B. exploratory data mining<br>C. descriptive data mining<br>D. explanative data mining |
| 31 | The output of a regression algorithm is usually a:<br>A. **real variable**<br>B. integer variable<br>C. character variable<br>D. string variable |
| 32 | Which of the following refers to the steps of the knowledge discovery process, in which the several data sources are combined?<br>A. Data selection<br>B. Data cleaning<br>C. Data transformation<br>**D. Data integration** |
| 33 | Which one of the following refers to the Black Box?<br>**A. It can be referred as the system that can be used without the knowledge of the internal operations**<br>B. It referrers the natural environment of the specific species<br>C. It takes only two values at most that are 0 and 1<br>D. All of the above |
| 34 | The issue of Pattern evaluation comes under which of these?<br>a. Performance Issues<br>b. Diverse Data Types Issues<br>c. **User Interaction and Mining Methodology Issues**<br>d. None of the above |
| 35 | Which one of the following issues must be considered before investing in data mining?<br>A. Compatibility<br>B. Functionality<br>C. Vendor consideration<br>**D. All of the above** |
| 36 | The term "DMQL" stands for _____<br>A. Data Marts Query Language<br>B. DBMiner Query Language<br>**C. Data Mining Query Language**<br>D. None of the above |
| 37 | In the regression equation Y = 21 - 3X, the slope is<br>A. 21<br>B. -21<br>C. 3 |

| | |
|---|---|
| | D. **-3** |
| 38 | Functions of Data Mining are<br>A. Association and correctional analysis classification<br>B. Prediction and characterization<br>C. Cluster analysis and evolution analysis<br>**D. All of the above** |
| 39 | In certain cases, it is not clear what kind of pattern need to find, data mining should_________:<br>a) Try to perform all possible tasks<br>b) Perform both predictive and descriptive task<br>**c) It may allow interaction with the user so that he can guide the mining process**<br>d) All of the above |
| 40 | Which of the following correctly refers to the term "Data Independence"?<br>a) It means that the programs are not dependent on the logical attributes<br>b) It refers to that data that is defined separately, not included in the program<br>**c) It means that the programs are totally dependent on the physical attributes of data**<br>d) Both A and C |
| 41 | Margin of a hyperplane is defined as:<br>A. The angle it makes with the axes<br>B. The intercept it makes on the axes<br>**C. Perpendicular distance from its closest point**<br>D. Perpendicular distance from origin |
| 42 | In a hard margin support vector machine:<br>**A. No training instances lie inside the margin**<br>B. All the training instances lie inside the margin<br>C. Only few training instances lie inside the margin<br>D. None of the above |
| 43 | The Lagrange multipliers corresponding to the support vectors have a value:<br>A. equal to zero<br>B. less than zero<br>**C. greater than zero**<br>D. can take on any value |
| 44 | The SVM's are less effective when:<br>A. The data is linearly separable<br>B. The data is clean and ready to use<br>**C. The data is noisy and contains overlapping points**<br>D. None of the above |
| 45 | The dual optimization problem in SVM design is solved using:<br>A. Linear programming<br>**B. Quadratic programming**<br>C. Dynamic programming |

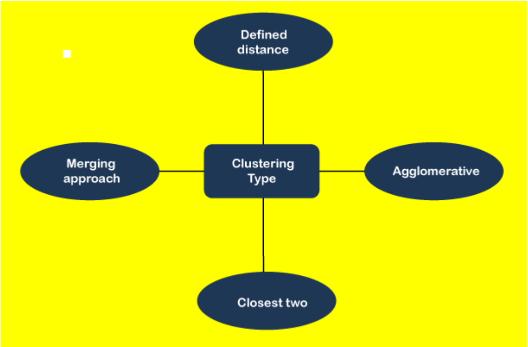
| | |
|---|---|
| | D. Integer programming |
| 46 | Which of the following is an essential process in which the intelligent methods are applied to extract data patterns?<br>A. Warehousing<br>**B. Data Mining**<br>C. Text Mining<br>D. Data Selection |
| 47 | What is KDD in data mining?<br>**A. Knowledge Discovery Database**<br>B. Knowledge Discovery Data<br>C. Knowledge Data definition<br>D. Knowledge data house |
| 48 | What are the functions of Data Mining?<br>A. Association and correctional analysis classification<br>B. Prediction and characterization<br>C. Cluster analysis and Evolution analysis<br>**D. All of the above** |
| 49 | Which one of the following can be considered as the final output of the hierarchal type of clustering?<br>**A. A tree which displays how the close thing are to each other**<br>B. Assignment of each point to clusters<br>C. Finalize estimation of cluster centroids<br>D. None of the above |
| 50 | Which of the following statements about hierarchal clustering is incorrect?<br>**A. The hierarchal clustering can primarily be used for the aim of exploration**<br>B. The hierarchal clustering should not be primarily used for the aim of exploration<br>C. Both A and B<br>D. None of the above |
| 51 | Which of the following refers to the problem of finding abstracted patterns (or structures) in the unlabeled data?<br>a) Supervised learning<br>**b) Unsupervised learning**<br>c) Hybrid learning<br>d) Reinforcement learning |
| 52 | Identify the term used to define the task of inferring a model from labeled training data.<br>**a) Supervised learning**<br>b) Unsupervised learning<br>c) Hybrid learning<br>d) Reinforcement learning |
| 53 | “Handling the rational and complex types of data” comes under the |

| | |
|---|---|
| | __________ category.<br>**A. Diverse Data Type**<br>B. Performance issues<br>C. Both (a) and (b)<br>D. Neither (a) nor (b) |
| 54 | Identify the incorrect option among the following which is not involved in data mining.<br>a) Data exploration<br>b) Knowledge extraction<br>**c) Data Abstraction**<br>d) Data archaeology |
| 55 | Which of the following focuses on the discovery of (previously) unknown properties on the data?<br>**a) Data mining**<br>b) Big Data<br>c) Data wrangling<br>d) Machine Learning |
| 56 | Which of the following can be considered as the correct process of Data Mining?<br>**a) Infrastructure, Exploration, Analysis, Interpretation, Exploitation**<br>b) Exploration, Infrastructure, Analysis, Interpretation, Exploitation<br>c) Exploration, Infrastructure, Interpretation, Analysis, Exploitation<br>d) Exploration, Infrastructure, Analysis, Exploitation, Interpretation |
| 57 | Which of the following is the correct advantage of the Update-Driven Approach?<br>A. This approach provides high performance.<br>B. The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.<br>**C. Both A and B**<br>D. None of the above |
| 58 | For what purpose, the analysis tools pre-compute the summaries of the huge amount of data?<br>a) In order to maintain consistency<br>b) For authentication<br>c) For data access<br>**d) To obtain the queries response** |
| 59 | Which of the following statements is incorrect about hierarchical clustering?<br>**a)The hierarchal type of clustering is also known as the HCA**<br>b) The choice of an appropriate metric can influence the shape of the cluster<br>c) In general, the splits and merges both are determined in a greedy manner<br>d) All of the above |

| | |
|---|---|
| 60 | Which one of the following can be considered as the final output of the hierarchal type of clustering?<br>a)**A tree which displays how the close thing are to each other**<br>b) Assignment of each point to clusters<br>c) Finalize estimation of cluster centroids<br>d) None of the above |
| 61 | Which of the following can be considered as the classification or mapping of a set or class with some predefined group or classes?<br>a) Data set<br>b) Data Characterization<br>c) Data Sub Structure<br>d) **Data Discrimination** |
| 62 | Which one of the following correctly refers to the Class study in the data cauterization?<br>a) Final class<br>b) Study class<br>**c) Target class**<br>d) Both A and C |
| 63 | The issues like "handling the rational and complex types of data" comes under which of the following categories?<br>**a) Diverse Data Type**<br>b) Mining methodology and user interaction Issues<br>c) Performance issues<br>d) All of the above |
| 64 | Which one of the following refers to the Black Box?<br>**a) It can be referred as the system that can be used without the knowledge of the internal operations**<br>b) It referrers the natural environment of the specific species<br>c) It takes only two values at most that are 0 and 1<br>d) All of the above |
| 65 | Which of the following is generally used by the E-R model to represent the weak entities?<br>a) Diamond<br>**b) Doubly outlined rectangle**<br>c) Dotted rectangle<br>d) Both B & C |
| 66 | In certain cases, it is not clear what kind of pattern need to find, data mining should_________:<br>a) Try to perform all possible tasks<br>b) Perform both predictive and descriptive task<br>**c) It may allow interaction with the user so that he can guide the mining process**<br>d) All of the above |

| | |
|---|---|
| 67 | Which one of the following issues must be considered before investing in data mining?<br>a) Compatibility<br>b) Functionality<br>c) Vendor consideration<br>**d) All of the above** |
| 68 | Which of the following correctly refers to the term "Data Independence"?<br>a) It means that the programs are not dependent on the logical attributes<br>b) It refers to that data that is defined separately, not included in the program<br>c) It means that the programs are totally dependent on the physical attributes of data<br>**d) Both A and C** |
| 69 | Which of the following can be considered as the drawback of the query-Driven approach in data warehousing?<br>a) This approach is expensive for queries that require aggregations<br>b) This approach is expensive insufficient, and very frequent queries<br>c) This approach requires a very complex integration and filtering process<br>**d) All of the above** |
| 70 | Which of the following refers to the steps of the knowledge discovery process, in which the several data sources are combined?<br>a) Data selection<br>b) Data cleaning<br>c) Data transformation<br>**d) Data integration** |
| 71 | Which of the following is also used as the first step in the knowledge discovery process?<br>a) Data selection<br>**b) Data cleaning**<br>c) Data transformation<br>d) Data integration |
| 72 | Which one of the following can be considered as the correct application of data mining?<br>a) Fraud detection<br>b) Corporate Analysis & Risk management<br>c) Management and market analysis<br>**d) All of the above** |
| 73 | Which of the following correctly refers to the data selection?<br>a) A subject-oriented integrated time-variant non-volatile collection of data in support of management<br>b) The actual discovery phase of a knowledge discovery process<br>**c) The stage of selecting the right data for a KDD process**<br>d) All of the above |
| 74 | Which one of the following correctly defines the term cluster? |

| | |
|---|---|
| | **a) Group of similar objects that differ significantly from other objects**<br>b) Symbolic representation of facts or ideas from which information can potentially be extracted<br>c) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm<br>d) All of the above |
| 75 | The issues like efficiency, scalability of data mining algorithms comes under_______<br>**a) Performance issues**<br>b) Diverse data type issues<br>c) Mining methodology and user interaction<br>d) All of the above |
| 76 | Which one of the following statements is not correct about the data cleaning?<br>a) It refers to the process of data cleaning<br>b) It refers to the transformation of wrong data into correct data<br>c) It refers to correcting inconsistent data<br>**d) All of the above** |
| 77 | The analysis performed to uncover the interesting statistical correlation between associated -attributes value pairs is known as the _______.<br>a) Mining of association<br>**b) Mining of correlation**<br>c) Mining of clusters<br>d) All of the above |
| 78 | “Hybrid” is defined as<br>**A. Combining different types of method or information**<br>B. Information base filled with the knowledge of an expert<br>C. The design of learning algorithms which are lined along the theory of evolution<br>D. None of the above |
| 79 | In data mining, how many categories of functions are included?<br>a) 5<br>b) 4<br>**c) 2**<br>d) 3 |
| 80 | Which of the following statements is true about the classification?<br>a)It is a measure of accuracy<br>**b) It is a subdivision of a set**<br>c) It is the task of assigning a classification<br>d) None of the above |
| 81 | The self-organizing maps can also be considered as the instance of _________ type of learning.<br>a)Supervised learning<br>**b) Unsupervised learning** |

| | |
|---|---|
| | c) Missing data imputation<br>d) Both A & C |
| 82 | Which one of the clustering techniques needs the merging approach?<br>a)Partitioned<br>b) Naïve Bayes<br>**c) Hierarchical**<br>d) Both A and C |
| 83 | Which one of the following statements about the K-means clustering is incorrect?<br>a)The goal of the k-means clustering is to partition (n) observation into (k) clusters<br>b) K-means clustering can be defined as the method of quantization<br>**c) The nearest neighbor is the same as the K-means**<br>d) All of the above |
| 84 | Which one of the following can be considered as the final output of the hierarchal type of clustering?<br>**a)A tree which displays how the close thing are to each other**<br>b) Assignment of each point to clusters<br>c) Finalize estimation of cluster centroids<br>d) None of the above |
| 85 | In the following given diagram, which type of clustering is used?<br> Defined distance<br>Merging approach<br>Clustering Type<br>Agglomerative<br>Closest two<br>**A. Hierarchal**<br>B. Naive Bayes<br>C. Partitional<br>D. None of the above |
| 86 | Which of the following statements about hierarchical clustering is incorrect?<br>**a)The hierarchical clustering can primarily be used for the aim of exploration**<br>b) The hierarchical clustering should not be primarily used for the aim of exploration<br>c) Both A and B<br>d) None of the above |

| | |
|---|---|
| 87 | Which of the following correctly refers to the term "Data Independence"?<br>a) It means that the programs are not dependent on the logical attributes<br>b) It refers to that data that is defined separately, not included in the program<br>c) It means that the programs are totally dependent on the physical attributes of data<br>**d) Both A and C** |
| 88 | Which of the following is a good alternative to the star schema?<br>a) snow flake schema<br>b) star schema<br>c) star snow flake schema<br>**d) fact constellation** |
| 89 | Which of the following is true for Classification?<br>**a) A subdivision of a set**<br>b) A measure of the accuracy<br>c) The task of assigning a classification<br>d) All of these |
| 90 | What is noise?<br>a) component of a network<br>**b) context of KDD and data mining**<br>c) aspects of a data warehouse<br>d) None of these |
| 91 | Which of the following forms of data mining assigns records to one of a predefined set of classes?<br>(A). Classification<br>**(B). Clustering**<br>(C). Both A and B<br>(D). None |
| 92 | According to storks' population size, find the total number of babies from the following example of predicting the number of babies.<br>(A). feature<br>**(B). outcome**<br>(C). attribute<br>(D). observation |
| 93 | Which of the following is not belong to data mining?<br>(A). Knowledge extraction<br>**(B). Data transformation**<br>(C). Data exploration<br>(D). Data archaeology |
| 94 | Which of the following terms is used as a synonym for data mining?<br>**(A). knowledge discovery in databases**<br>(B). data warehousing<br>(C). regression analysis<br>(D). parallel processing in databases |

| | |
|---|---|
| 95 | Which of the following forms of data mining assigns records to one of a predefined set of classes?<br>(A). Classification<br>**(B). Clustering**<br>(C). Both A and B<br>(D). None |
| 96 | The analysis performed to uncover the interesting statistical correlation between associated -attributes value pairs are known as the _______.<br>A. Mining of association<br>**B. Mining of correlation**<br>C. Mining of clusters<br>D. All of the above |
| 97 | Name of a person, can be considered as an attribute of type?<br>**a) Nominal**<br>b) Ordinal<br>c) Interval<br>d) Ratio |
| 98 | If a record data matrix has reduced number of rows after a transformation, the transformation has performed:<br>**a) Data Sampling**<br>b) Dimensionality Reduction<br>c) Noise Cleaning<br>d) Discretization |
| 99 | Based on the results in (9), confidence of association rules {b,d}->{e} and {e}->{b,d} are:<br>**a) 0.8, 1**<br>b) 1, 0.8<br>c) 0.25, 1<br>d) 1, 0.25 |
| 100 | Why is the snowflake schema applied?<br>A. Transformation<br>B. Aggregation<br>**C. Normalization**<br>D. Generalization |

**NOTE:** The Option in Bold is the Correct Answer for each Question.