|  |  |
| --- | --- |
| **S.**  **No.** | **Objective Questions (MCQ /True or False / Fill up with Choices )** |
| 1. | Which of the following is not an example of Social Media?   1. Twitter 2. **Google** 3. Insta 4. Youtube |
| 2. | By 2025, the volume of digital data will increase to   1. TB 2. YB 3. **ZB** 4. EB |
| 3. | For Drawing insights for Business what are need?   1. Collecting the data 2. Storing the data 3. Analysing the data 4. **All the above** |
| 4. | Does Facebook uses "Big Data " to perform the concept of Flashback? Is this True or False.   1. **TRUE** 2. FALSE |
| 5. | The Process of describing the data that is huge and complex to store and process is known as   1. Analytics 2. Data mining 3. **Big Data** 4. Data Warehouse |
| 6. | Data generated from online transactions is one of the example for volume of big data. Is this true or False.   1. **TRUE** 2. FALSE |
| 7. | Velocity is the speed at which the data is processed   1. TRUE 2. **FALSE** |
| 8. | have a structure but cannot be stored in a database.   1. Structured 2. **Semi-Structured** 3. Unstructured 4. None of these |
| 9. | refers to the ability to turn your data useful for business.   1. Velocity 2. Variety 3. **Value** 4. Volume |

|  |  |
| --- | --- |
| 10. | Value tells the trustworthiness of data in terms of quality and accuracy.   1. TRUE 2. **FALSE** |
| 11. | GFS consists of a Master and Chunk Servers   1. Single, Single 2. Multiple, Single 3. **Single, Multiple** 4. Multiple, Multiple |
| 12. | Files are divided into sized Chunks.   1. Static 2. Dynamic 3. **Fixed** 4. Variable |
| 13. | is an open source framework for storing data and running application on clusters of commodity hardware.   1. HDFS 2. **Hadoop** 3. MapReduce 4. Cloud |
| 14. | HDFS Stores how much data in each clusters that can be scaled at any time?   1. 32 2. 64   **c. 128**  d. 256 |
| 15. | Hadoop MapReduce allows you to perform distributed parallel processing on large volumes of data quickly and efficiently… is this MapReduce or Hadoop… i.e statement is True or False   1. **TRUE** 2. FALSE |
| 16. | Hortonworks was introduced by Cloudera and owned by Yahoo.   1. TRUE 2. **FALSE** |
| 17. | Hadoop YARN is used for Cluster Resource Management in Hadoop Ecosystem.   1. **TRUE** 2. FALSE |
| 18. | Google Introduced MapReduce Programming model in 2004.   1. **TRUE** 2. FALSE |
| 19. | phase sorts the data & creates logical clusters.   1. Reduce, YARN 2. MAP, YARN 3. REDUCE, MAP 4. **MAP, REDUCE** |

|  |  |
| --- | --- |
| 20. | There is only one operation between Mapping and Reducing is it True or False…   1. **TRUE** 2. FALSE |
| 21. | is factors considered before Adopting Big Data Technology.   1. **Validation** 2. Verification 3. Data 4. Design |
| 22. | for improving supply chain management to optimize stock management, replenishment, and forecasting;   1. Descriptive 2. Diagnostic 3. **Predictive** 4. Prescriptive |
| 23. | which among the following is not a Data mining and analytical applications?   1. profile matching 2. social network analysis 3. facial recognition 4. **Filtering** |
| 24. | as a result of data accessibility, data latency, data availability, or limits on bandwidth in relation to the size of inputs.   1. Computation-restricted throttling 2. Large data volumes 3. **Data throttling** 4. Benefits from data parallelization |
| 25. | As an example, an expectation of using a recommendation engine would be to increase same-customer sales by adding more items into the market basket.   1. Lowering costs 2. **Increasing revenues** 3. Increasing productivity 4. Reducing risk |
| 26. | Which storage subsystem can support massive data volumes of increasing size.   1. Extensibility 2. Fault tolerance 3. **Scalability** 4. High-speed I/O capacity |
| 27. | provides performance through distribution of data and fault tolerance through replication   1. **HDFS** 2. PIG 3. HIVE 4. HADOOP |

|  |  |
| --- | --- |
| 28. | is a programming model for writing applications that can process Big Data in parallel on multiple nodes.   1. HDFS 2. **MAP REDUCE** 3. HADOOP 4. HIVE |
| 29. | takes the grouped key-value paired data as input and runs a Reducer function on each one of them.   1. MAPPER 2. **REDUCER** 3. COMBINER 4. PARTITIONER |
| 30. | is a type of local Reducer that groups similar data from the map phase into identifiable sets.   1. MAPPER 2. REDUCER 3. **COMBINER** 4. PARTITIONER |
| 31. | While Installing Hadoop how many xml files are edited and list them ?   1. **core-site.xml** 2. hdfs-site.xml 3. mapred.xml 4. yarn.xml |
| 32. | Write the code for core-site.xml ?  <?xml version="1.0" encoding="UTF-8"?>  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>  <configuration>  <property>  <name>hadoop.tmp.dir</name>  <value>D:\hadoop\temp</value>  </property>  <property>  <name>fs.default.name</name>  <value>hdfs://localhost:50071</value>  </property>  </configuration>  </?xml > |
| 33. | Write the code for hdfs-site.xml ? |

|  |  |
| --- | --- |
| **S.**  **No.** | **Objective Questions (MCQ /True or False / Fill up with Choices )** |
| 1. | Movie Recommendation systems are an example of   1. Classification 2. Clustering 3. Reinforcement Learning 4. Regression    1. 2 Only    2. 1 and 2    3. 1 and 3    4. **2 and 3** |
| 2. | Sentiment Analysis is an example of   1. Regression 2. Classification 3. Clustering 4 Reinforcement Learning    1. **1, 2 and 4**    2. 1 and 3    3. 1, 2 and 3    4. 1 and 2 |
| 3. | Can decision trees be used for performing clustering?   1. **True** 2. False |
| 4. | What is the minimum no. of variables/ features required to perform clustering? 1. 0  **2. 1**  3. 2  4. 3 |
| 5. | For two runs of K-Mean clustering is it expected to get same clustering results?   1. Yes 2. **No** |
| 6. | Which of the following can act as possible termination conditions in K-Means?   1. For a fixed number of iterations. 2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum. 3. Centroids do not change between successive iterations. 4.Terminate when RSS falls below a threshold.    1. 1, 3 and 4    2. 1, 2 and 3    3. 1, 2 and 4    4. **All of the above** |
| 7. | Which of the following algorithm is most sensitive to outliers?   1. **K-means clustering algorithm** 2. K-medians clustering algorithm 3. K-modes clustering algorithm 4. K-medoids clustering algorithm |
| 8. | After performing K-Means Clustering analysis on a dataset, you observed the following  dendrogram. Which of the following conclusion can be drawn from the dendrogram? |

|  |  |
| --- | --- |
|  | 1. There were 28 data points in clustering analysis 2. The best no. of clusters for the analyzed data points is 4 3. The proximity function used is Average-link clustering 4. **The above dendrogram interpretation is not possible for K-Means clustering analysis** |
| 9. | In the figure below, if you draw a horizontal line on y- axis for y=2. What will be the number of clusters formed?    1. 1  **2. 2**  3. 3  4. 4 |
| 10. | In which of the following cases will K-Means clustering fail to give good results?   1. Data points with outliers 2. Data points with different densities 3. Data points with round shapes 4. Data points with non-convex shapes    1. 1 and 2    2. 2 and 3    3. 2 and 4    4. **1, 2 and 4** |
| 11. | The discrete variables and continuous variables are two types of   1. Open end classification 2. Time series classification 3. Qualitative classification 4. **Quantitative classification** |

|  |  |
| --- | --- |
| 12. | Bayesian classifiers is   1. **A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.** 2. Any mechanism employed by a learning system to constrain the search space of a hypothesis 3. 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. 4. None of these |
| 13. | Classification accuracy is   1. A subdivision of a set of examples into a number of classes 2. **Measure of the accuracy, of the classification of a concept that is given by a certain theory** 3. The task of assigning a classification to a set of examples 4. None of these |
| 14. | Classification task referred to   1. A subdivision of a set of examples into a number of classes 2. A measure of the accuracy, of the classification of a concept that is given by a certain theory 3. **The task of assigning a classification to a set of examples** 4. None of these |
| 15. | Euclidean distance measure is   1. A stage of the KDD process in which new data is added to the existing selection. 2. 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 3. The distance between two points as calculated using the Pythagoras theorem 4. **None of these** |
| 16. | is good at handle missing data and support both the kind of attributes ( i.e Categorial and Continuous attributes )  a. ID3.  **b. C4.5.**   1. CART. 2. Naïve Bayes. |
| 17. | Decision trees use , in that they always choose the option that seems the best available at that moment.   1. **Greedy Algorithms.** 2. Divide and Conquer. 3. Backtracking. 4. Shortest Path Method. |
| 18. | Decision trees cannot handle categorical attributes with many distinct values, such as country codes for telephone numbers.   1. TRUE 2. **FALSE** |
| 19. | are easy to implement and can execute efficiently even without |

|  |  |
| --- | --- |
|  | prior knowledge of the data, they are among the most popular algorithms for classifying text documents.   1. ID3 2. **Naïve Bayes classifiers** 3. CART 4. None of these. |
| 20. | High entropy means that the partitions in classification are   1. Pure 2. **Not pure** 3. Useful 4. Useless |
| 21. | Which of the following statements about Naive Bayes is incorrect?   1. Attributes are equally important. 2. **Attributes are statistically dependent of one another given the class value.** 3. Attributes are statistically independent of one another given the class value. 4. Attributes can be nominal or numeric |
| 22. | The maximum value for entropy depends on the number of classes so if we have 8 Classes what will be the max entropy.   1. Max Entropy is 1 2. Max Entropy is 2 3. **Max Entropy is 3** 4. Max Entropy is 4 |
| 23. | John flies frequently and likes to upgrade his seat to first class. He has determined that if he checks in for his flight at least two hours early, the probability that he will get an upgrade is 0.75; otherwise, the probability that he will get an upgrade is 0.35. With his busy schedule, he checks in at least two hours before his flight only 40% of the time. Suppose John did not receive an upgrade on his most recent attempt. What is the probability that he did not arrive two hours early?  a. 0.892  **b. 0.796**  c. 0.685  d. 0.999 |
| 24. | Point out the wrong statement.   1. **k-nearest neighbor is same as k-means** 2. k-means clustering is a method of vector quantization 3. k-means clustering aims to partition n observations into k clusters 4. none of the mentioned |
| 25. | Consider the following example “How we can divide set of articles such that those articles have the same theme (we do not know the theme of the articles ahead of time) " is this:   1. **Clustering** 2. Classification |
|  | 1. Regression 2. None of These |
| 26. | Can we use K Mean Clustering to identify the objects in video?   1. **Yes** 2. No |
| 27. | Clustering techniques are in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.   1. **Unsupervised** 2. Supervised 3. Reinforcement 4. Neural network |

|  |  |
| --- | --- |
| **S.**  **No.** | **Objective Questions (MCQ /True or False / Fill up with Choices )** |
| 1. | metric is examined to determine a reasonably optimal value of k.   1. Mean Square Error 2. **Within Sum of Squares (WSS)** 3. Speed 4. None of These |
| 2. | If an itemset is considered frequent, then any subset of the frequent itemset must also be frequent.   1. Apriori Property 2. Downward Closure Property 3. Either 1 or 2 4. **Both 1 & 2** |
| 3. | if {bread,eggs,milk} has a support of 0.15 and {bread,eggs} also has a support of 0.15, the confidence of rule {bread,eggs}→{milk} is  1. 0  2**. 1**  3. 2  4. 3 |
| 4. | Confidence is a measure of how X and Y are really related rather than coincidentally happening together.   1. True 2. **False** |
| 5. | A high-confidence rule can sometimes be misleading because confidence does not consider support of the itemset in the rule consequent. Is This True ?   1. **Yes** 2. No |
| 6. | recommend items based on similarity measures between users and/or items.   1. Content Based Systems 2. Hybrid System 3. **Collaborative Filtering Systems** 4. None of These |
| 7. | There are major Classification of Collaborative Filtering Mechanisms 1. 1  **2. 2**  3. 3  4. None of These |
| 8. | Movie Recommendation to peoples is an example of   1. User Based Recommendation 2. **Item Based Recommendation** 3. Knowledge Based Recommendation 4. Content Based Recommendation |
| 9. | recommenders rely on an explicitly defined set of recommendation rules.   1. Constraint Based 2. **Case Based** 3. Content Based 4. User Based |
| 10. | Parallelized hybrid recommender systems operate dependently of one another and produce separate recommendation lists.   1. True 2. **False** |
| 11. | Association rules are sometimes referred to as   1. ***market basket analysis*** 2. Itemset Filtering 3. Frequent Itemset Analysis 4. None of these. |
| 12. | if 80% of all transactions contain itemset {bread}, then the support of {bread} is 0.8. Similarly, if 60% of all transactions contain itemset {bread,butter}, then the support of  {bread,butter} is  a. 0.4  b. 0.5  **c. 0.6**  d. 0.7 |
| 13. | Lift is defined as the measure of certainty or trustworthiness associated with each discovered rule.   1. TRUE 2. **FALSE** |
| 14. | is able to identify trustworthy rules, but it cannot tell whether a rule is coincidental.   1. Lift 2. **Confidence** 3. Support 4. Leverage |
| 15. | recommend items based on similarity measures between users and/or items. The items recommended to a user are those preferred by similar users.   1. **Collaborative Filtering System** 2. Content Based Recommendation 3. Knowledge Based Recommendation 4. Hybrid Approaches |
| 16. | Pure collaborative approaches take a matrix of given user–item ratings as the only input and typically produce output. Is it Pure Collaborative?   1. **Yes** 2. No |
| 17. | With respect to the determination of the set of similar users, one common measure used in  recommender systems is |
|  | 1. Cosine Similarity Measure 2. **Pearson’s correlation coefficient.** 3. Mean Squared Error Method 4. None of these. |
| 18. | Large-scale e-commerce sites, often implement a different technique,  which is more apt for offline preprocessing and thus allows for the computation of recommendations in real time even for a very large rating matrix.   1. **Item-Based Recommendation** 2. User-Based Recommendation 3. Content-Based Recommendation 4. None of these |
| 19. | Here are two very short texts to compare and find the cosine similarity measure?   1. Julie loves me more than Linda loves me 2. Jane likes me more than Julie loves me a. 0.6   b. 0.7  **c. 0.8**  d. 0.9 |
| 20. | is based on the availability of item descriptions and a profile that assigns importance to these characteristics.   1. Item-Based Recommendation 2. User-Based Recommendation 3. **Content-Based Recommendation.** 4. None of these |
| 21. | Consider the features of a movie which are not relevant to a recommendation system.   1. The set of actors of the movie. 2. The Director 3. The Year in which the movie was made 4. **The Budget of the movie.** |
| 22. | A has been implemented, for similarity based retrieval under nearest neighbors.   1. **k-nearest-neighbor method (kNN)** 2. Conventional Neural Network (CNN) 3. Bayes Theorem 4. Naïve Bayes Classifier |
| 23. | Case-based recommenders focus on the retrieval of similar items on the basis of different types of similarity measures   1. **TRUE** 2. FALSE |
| 24. | In recommendation approaches, items are retrieved using similarity measures that describe to which extent item properties match some given user’s requirements.   1. Item-Based 2. **Case-Based** |
|  | 1. Content-Based 2. User-Based |
| 25. | are based on a sequenced order of techniques, in which each succeeding recommender only refines the recommendations of its predecessor.   1. Weighted Hybrids 2. Mixed Hybrids 3. **Cascade Hybrids** 4. Switching Hybrids |
| 26. | require an oracle that decides which recommender should be used in a specific situation, depending on the user profile and/or the quality of recommendation   1. Weighted Hybrids 2. Mixed Hybrids 3. Cascade Hybrids 4. **Switching Hybrids** |