**Department of Computer Science**

**Class: T. Y. B. Sc. (CS)**

**Semester: VI**

**Subject: Data Science**

**Sample Questions**

**Multiple Choice Questions**

1. Data that summarize all observations in a category are called \_\_\_\_\_\_\_\_\_\_ data.
	1. frequency
	2. summarized
	3. raw
	4. cleaned
2. What are the two types of data, with regard to Data Science?
	1. Traditional and Modern
	2. Small and Big data
	3. Qualitative and Quantitative
	4. Int and Float
3. Qualitative data is also called as data.
	1. Continuous
	2. Numerical
	3. Categorical
	4. Discrete
4. Quantitative data is also called as data.
	1. Ordinal
	2. Numerical
	3. Categorical
	4. Nominal
5. Student name is type data.
	1. Ordinal
	2. Nominal
	3. Binary
	4. Numerical
6. Result of a coin toss is type data.
	1. Numerical
	2. Ordinal
	3. Nominal
	4. Binary
7. Movie rating is type data.
	1. Ordinal
	2. Nominal
	3. Binary
	4. Continuous
8. Days in a year is type data.
	1. Binary
	2. Continuous
	3. Discrete
	4. Nominal
9. Weight of a product is type data.
	1. Continuous
	2. Discrete
	3. Nominal
	4. Ordinal
10. Interval and Ratio are types of data.
	1. Nominal
	2. Discrete
	3. Continuous
	4. Categorical
11. The data originally collected by the researcher through direct efforts, is called
	1. Primary data
	2. Secondary data
	3. Unknown data
	4. Processed data
12. Data published in books and journals is called
	1. Primary data
	2. Secondary data
	3. First hand data
	4. Processed data
13. Which of the following high-level languages is NOT a preferred language for Data Science?
	1. Java
	2. Python
	3. R
	4. Both R and Python
14. is the IDE for R language.
	1. RStudio
	2. PyCharm
	3. Numpy
	4. Scilab
15. Which of the following is another name for raw data?
	1. destination data
	2. eggy data
	3. secondary
	4. machine learning
16. EDA stands for
	1. Extensive Data Analysis
	2. Experimental Data Analysis
	3. Exploratory Data Analysis
	4. Educational Data Analysis
17. EDA is used to
	1. Detect outliers
	2. Visualize potential relationships between variables
	3. Implement Dimension reduction
	4. Both A and B
18. is a non-graphical method of EDA
19. Boxplot
20. Histogram
21. Median
22. Piechart
23. is a graphical method of EDA.
	1. Boxplot
	2. Interquartile range
	3. Mean
	4. Standard Deviation
24. is the process of creating visual representation of data to bring some meaningful insights.
	1. Data Curation
	2. Data Visualization
	3. Data Cleaning
	4. Data Transformation
25. The graph that represents the five-number summary.
	1. Histogram
	2. Scatter plot
	3. Box plot
	4. Dendogram
26. Which of the following cannot be determined from a boxplot?
	1. Minimum and Maximum values
	2. Inter-quartile range
	3. Median
	4. Mode
27. The summary statistics that finds the average value of a data set is called
	1. Median
	2. Mode
	3. Mean
	4. Quartile
28. is the exact middle value in a sorted data set.
	1. Median
	2. Mode
	3. Mean
	4. Variance
29. is a measure of a distribution’s asymmetry.
	1. IQR
	2. Standard Deviation
	3. Skewness
	4. Mean
30. measures the degree of the relationship between two random variables.
	1. Correlation
	2. Skewness
	3. Standard Deviation
	4. IQR
31. When two variables move together in the same direction, they will have Positive covariance
	1. Negative covariance
	2. Zero covariance
	3. Equal covariance
32. Excel files are examples of data.
	1. Structured
	2. Unstructured
	3. Semi- Structured
	4. Web data
33. Data generated from Twitter is example of data.
	1. Structured
	2. Unstructured
	3. Semi- Structured
	4. Tidy data
34. Non-digitized data is data.
	1. Structured
	2. Unstructured
	3. Semi- Structured
	4. Media data
35. is the organization, publication and presentation of data for reuse.
	1. Data Curation
	2. Data smoothing
	3. Data transformation
	4. Data wrangling
36. is a web-based version-control and collaboration platform for software developers.
	1. XML
	2. GitHub
	3. AWS
	4. HBase
37. In MongoDB, data is written in like format.
	1. JSON
	2. XML
	3. HTML
	4. C
38. data has each variable saved in its own column.
	1. Structured
	2. Unstructured
	3. Web
	4. Unorganized
39. is the on-going management of data through its lifecycle.
	1. Data transformation
	2. Data Curation
	3. Data visualization
	4. Data cleaning
40. is an example of a NoSQL.
	1. MySQL
	2. ORACLE
	3. MongoDB
	4. Excel
41. tidyr is a reframing of \_\_\_\_\_\_\_ designed to accompany the tidy data framework.
	1. reshape5
	2. dplyr
	3. reshape2
	4. applyr
42. A MongoDB database is composed of
	1. Documents
	2. Collections
	3. Cursors
	4. Tables
43. A Collection in MongoDB is composed of
	1. Documents
	2. Tables
	3. Rows
	4. Columns
44. Consider a collection book, which has the following fields: \_id, title, author, and price. Which of the following query retrieves ONLY the key named title?
	1. db.book.find({},{\_id:0, title:1})
	2. db.book.find({title:1})
	3. db.book.find({},{title:1})
	4. db.book.select({},{title:1})
45. MongoDB is a based database.
	1. Column
	2. Graph
	3. Document
	4. Relation
46. Which of the following statements is INCORRECT about MongoDB?
	1. Index can be created on any field.
	2. Database has a fixed schema.
	3. Data is stored in the form of JSON style documents.
	4. Database is composed of ‘collections’, which are made of ‘documents’.
47. Method used in MongoDB to query documents
	1. Select()
	2. Retrieve()
	3. Find()
	4. Insert()
48. Method used in MongoDB to create an index
	1. createIndex()
	2. makeIndex()
	3. addIndex()
	4. dropIndex()
49. Method used in MongoDB to delete a database
	1. db.deleteDatabase()
	2. db.dropDatabase()
	3. db.removeDatabase()
	4. db.cutDatabase()
50. A in MongoDB is analogous to a table in RDBMS.
	1. Collection
	2. Document
	3. Index
	4. Field
51. is a query language for selecting nodes from an XML document.
	1. JQuery
	2. SQL
	3. XPath
	4. HTML
52. Which of the following is INCORRECT about NoSQL?
	1. NoSQL stands for ‘Not Only SQL’ or ‘Not SQL’
	2. NoSQL databases are table based
	3. NoSQL are non-relational, compared to the relational SQL databases
	4. NoSQL databases are horizontally scalable or distributed
53. The NoSQL type which is useful for sparse datasets
	1. Graph stores
	2. Key value stores
	3. Document stores
	4. Column family stores
54. MongoDB is a type NoSQL.
	1. Key value store
	2. Document store
	3. Column family store
	4. Graph store
55. is an example of a database implementation that uses Graph store type NoSQL.
	1. Facebook
	2. Blog Entry
	3. Web page content management
	4. Google
56. is the security issue of gaining unauthorized access to a system.
	1. Phishing
	2. Hacking
	3. Spoofing
	4. encrypting
57. involves creating fake websites that look like legitimate business websites or emails to induce individuals to reveal personal information.
	1. Phishing
	2. Hacking
	3. Identity theft
	4. Virus
58. is the process of transforming data from one format to another.
	1. Data Mining
	2. Data wrangling
	3. Data visualization
	4. Data transformation
59. is a programming model for the processing of big data.
	1. MapReduce
	2. HBase
	3. GitHub
	4. MySQL
60. is a Hadoop based non-relational distributed database system.
	1. MySQL
	2. XPath
	3. MapReduce
	4. HBase
61. is an open source framework written in java that allows distributed processing of large datasets across clusters of computers.
	1. MySQL
	2. Hadoop
	3. ORACLE
	4. XPath
62. Amazon Web Services (AWS) is a service.
	1. cloud computing
	2. web development
	3. data mining
	4. Data analysis
63. is a Sequential action in Data curation.
	1. Dispose
	2. Appraise
	3. Reappraise
	4. Migrate
64. is an Occasional action in Data curation.
	1. Create
	2. Ingest
	3. Access
	4. Dispose
65. In data curation, is the action that transfers data to an archive.
	1. Appraise
	2. Ingest
	3. Transform
	4. Dispose
66. In data curation, is the action that evaluates and selects data for long-term preservation.
	1. Appraise
	2. Ingest
	3. Transform
	4. Create
67. learning does not have a response variable.
	1. Supervised
	2. Unsupervised
	3. Classification
	4. Regression
68. In learning, for each observation of predictor variables, there is an associated response variable.
	1. Supervised
	2. Unsupervised
	3. Reinforced
	4. deep learning
69. is an unsupervised learning method.
	1. Linear regression
	2. Logistic regression
	3. Cluster Analysis
	4. SVM
70. is a supervised learning method.
	1. K-means clustering
	2. Principal Component Analysis
	3. Hierarchical clustering
	4. SVM
71. Problems with a quantitative response are called
	1. Classification
	2. Regression
	3. Clustering
	4. Dimension reduction
72. Problems with a qualitative response are often called
	1. Classification
	2. Regression
	3. Time series
	4. Decision tree
73. are techniques used to reduce error by avoiding overfitting.
	1. Smoothing
	2. Aggregation
	3. Cross validation Regularization
	4. Cross validation
74. RSS stands for
	1. Repeated sum of scores
	2. Residual sum of squares
	3. Retrieved sum of scales
	4. Repeated sum of scales
75. When the model tries to learn or fit every noise data point in the data set, it results in .
	1. Overfitting
	2. Underfitting
	3. Smoothing
	4. Aggregation
76. Overfitting can be controlled by
	1. Regularization
	2. Increasing the size of training dataset
	3. Both A and B
	4. Decreasing the size of training dataset
77. causes underfitting.
	1. Bias
	2. Variance
	3. Regularization
	4. Normalization
78. causes overfitting.
	1. Bias
	2. Variance
	3. Regularization
	4. Normalization
79. is the error caused when a complicated problem is approximated to a simpler model.
	1. Bias
	2. Variance
	3. Regularization
	4. Normalization
80. is the amount that the estimate of the target function will change for a different training dataset.
	1. Bias
	2. Variance
	3. Regularization
	4. Normalization
81. A supervised machine learning algorithm should aim to have bias and variance.
	1. Low, high
	2. High, low
	3. Low, low
	4. High, high
82. Increasing the bias will
	1. Decrease
	2. Increase
	3. Not affect
	4. Nullify
83. Increasing the variance will
	1. Increase
	2. Decrease
	3. Nullify
	4. Not affect
84. AIC stands for
	1. Approximate Information Criterion
	2. Akaike Information Criterion
	3. Average Information Criterion
	4. Appropriate Information Criterion
85. BIC stands for
	1. Bayesian Information Criterion
	2. Basic Information Criterion
	3. Bayesian Integrity Criterion
	4. Baeson Integrity Criterion
86. A good machine learning model is the one that has AIC among all the models.
	1. Maximum
	2. Minimum
	3. Average
	4. Constant
87. is a technique used to estimate the likelihood of a model to predict future values.
	1. L1 regression
	2. L2 regression
	3. AIC
	4. L3 regression
88. is also known as Schwarz criterion.
	1. AIC
	2. BIC
	3. LASSO
	4. Ridge regression
89. Which of the following is a criterion for model selection?
	1. AIC
	2. BIC
	3. RSS
	4. Both A and B
90. The best fitting model will have a AIC and a BIC.
	1. Low, low
	2. High, high
	3. Low, high
	4. High, low
91. For k number of parameters and a likelihood L, the equation to estimate AIC is

a. -2\*ln(L) + 2\*ln(k)

 b. -2\*ln(L) + 2\*k

c. -2\*ln(L) - 2\*k

d. -2\*In(L) \* 2-k

1. For k number of parameters, N number of observations, and a likelihood L, the equation to estimate BIC is
	1. -2\*ln(L) + ln(N)\*k
	2. -2\*ln(L) + ln(k)\*N
	3. -2\*ln(L) - ln(N)\*k
	4. -2\*ln(L) - ln(N)-k
2. refers to the techniques for evaluating the results of a machine learning model for an independent data set.
	1. Regularization
	2. Normalization
	3. Cross Validation
	4. generalization
3. Which one of the following is a Cross Validation method?
	1. Validation set approach
	2. LOOCV
	3. K-fold validation
	4. Non Approc
4. The cross validation method in which the available data set is randomly divided into training set and validation set, is called
	1. Validation set approach
	2. LOOCV
	3. K-fold validation
	4. Non Approc
5. The cross validation method in which a single observation is kept as validation set, is called
	1. Validation set approach
	2. LOOCV
	3. K-fold validation
	4. Non Approc
6. The cross validation method in which the available data set is randomly divided into k groups, one of which is used as the validation set, is called
	1. Validation set approach
	2. LOOCV
	3. K-fold validation
	4. Non Approch
7. L1 and L2 regression are techniques.
	1. Validation
	2. Regularization
	3. Approximation
	4. Ensemble
8. Regularization techniques add a term to the loss function.
	1. Penalty
	2. Overfitting
	3. Underfitting
	4. Below lifting
9. The loss function in Regularization, is given by
	1. Residual Sum of Squares(RSS)
	2. Within Cluster Sum of Squares(WCSS)
	3. Gini Index
	4. Residual Sum
10. L1 Regularization is also called
	1. Ridge regression
	2. Lasso regression
	3. Logistic regression
	4. Classification
11. L2 Regularization is also called
	1. Ridge regression
	2. Lasso regression
	3. Logistic regression
	4. Classification
12. The key difference between L1 and L2 regularization techniques is
	1. Loss function
	2. Penalty term
	3. Both
	4. No difference
13. The penalty term in Lasso regression is
	1. Absolute value of magnitude of coefficients
	2. Square of magnitude of coefficients
	3. Product of magnitude of coefficients
	4. Absolute value of coefficients