SUBJECT CODE: 305 **SUBJECT NAME: BIGDATA**

Unit 1-Introduction to Bigdata

1. Data in _____ bytes size is called Big Data.

A. Tera
B. Giga
C. Peta
D. Meta

Ans : C

2. How many V's of Big Data

A. 2
B. 3
C. 4
D. 5

Ans : D

3. Transaction data of the bank is?

A. structured data
B. unstructured datat
C. Both A and B
D. None of the above

Ans : A

4. In how many forms BigData could be found?

A. 2
B. 3

C. 4
D. 5

Ans : B

5. Which of the following are Benefits of Big Data Processing?

A. Businesses can utilize outside intelligence while taking decisions
B. Improved customer service
C. Better operational efficiency
D. All of the above

Ans : D

6. Which of the following are incorrect Big Data Technologies?

A. Apache Hadoop
B. Apache Spark
C. Apache Kafka
D. Apache Pytarch

Ans : D

7. The overall percentage of the world's total data has been created just within the past two years is ?

A. 80%
B. 85%
C. 90%
D. 95%

Ans : C

8. Apache Kafka is an open-source platform that was created by?

A. LinkedIn
B. Facebook
C. Google
D. IBM

Ans : A

9. What was Hadoop named after?

A. Creator Doug Cutting's favorite circus act
B. Cuttings high school rock band
C. The toy elephant of Cutting's son
D. A sound Cutting's laptop made during Hadoop development

Ans : C

10. What are the main components of Big Data?

A. MapReduce
B. HDFS
C. YARN
D. All of the above

Ans : D

11. All of the following accurately describe Hadoop, EXCEPT _____

A. Open-source
B. Real-time
C. Java-based
D. Distributed computing approach

Ans : B

12. _____ has the world's largest Hadoop cluster.

A. Apple
B. Datamatics
C. Facebook
D. None of the above

Ans : C

13. Facebook Tackles Big Data With _____ based on Hadoop.

A. Project Prism
B. Prism
C. Project Big
D. Project Data

Ans : A

14. _____ is general-purpose computing model and runtime system for distributed data analytics.

A. Mapreduce
B. Drill
C. Oozie
D. None of the above

Ans : A

15. The examination of large amounts of data to see what patterns or other useful information can be found is known as

A. Data examination
B. Information analysis
C. Big data analytics
D. Data analysis

Ans : C

16. Big data analysis does the following except?

A. Collects data
B. Spreads data
C. Organizes data
D. Analyzes data

Ans : D

17. What makes Big Data analysis difficult to optimize?

A. Big Data is not difficult to optimize
B. Both data and cost effective ways to mine data to make business sense out of it
C. The technology to mine data
D. None of the above

Ans : B

18. The new source of big data that will trigger a Big Data revolution in the years to come is?

A. Business transactions
B. Social media
C. Transactional data and sensor data
D. RDBMS

Ans : C

19. The unit of data that flows through a Flume agent is

A. Log
B. Row
C. Record
D. Event

Ans : D

20. Listed below are the three steps that are followed to deploy a Big Data Solution except

A. Data Processing
B. Data dissemination
C. Data Storage
D. Data Ingestion

Ans : B

21. Who popularized bigdata term?
A. John deere
B. John Mashey
C. johny Mashe
D. Jhon Mash
Ans : B

22. Numbers ,text, image, audio and video data is _____
A. Volume
B. Value
C. Varity
D. Variety
Ans : D

23. Real time data is _____.
A. Field
B. Primary Key
C. unique
D. record
Ans : C

24. _____ is the term that is used to describe data that is high volume , high velocity and /or high variety.
A. Analytics
B. Bigdata
C. Hadoop Data
D. Bigdata analytics
Ans : B

25. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?
A. Big data management and data mining
B. Data warehousing and business intelligence
C. Management of Hadoop clusters
D. Collecting and storing unstructured data
Ans : A

26. Point out the wrong statement.
A. Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data
B. Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data
C. The programming model, MapReduce, used by Hadoop is difficult to write and test
D. All of these
Ans : C

27. All of the following accurately describe Hadoop, EXCEPT _____
A. Open-source
B. Real-time
C. Java-based
D. Distributed computing approach
Ans: B

28. _____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.
A.MapReduce
B.Mahout
C.Oozie
D.All of the mentioned
Ans: A

29. _____ has the world's largest Hadoop cluster.
A. Apple
B. Datamatics
C. Facebook
D. None of the mentioned
Ans:C

30. Facebook Tackles Big Data With _____ based on Hadoop.
A. 'Project Prism'
B. 'Prism'
C. 'Project Big'
D. 'Project Data'

Ans:A

31. Data science is the process of diverse set of data through ?

A. organizing data
B. processing data
C. analysing data
D. All of the above

Ans : D

32. The modern conception of data science as an independent discipline is sometimes attributed to?

A. William S.
B. John McCarthy
C. Arthur Samuel
D. Satoshi Nakamoto

Ans : A

33. Which of the following language is used in Data science?

A. C
B. C++
C. R
D. Ruby

Ans : C

34. Which of the following is false?

A. Subsetting can be used to select and exclude variables and observations
B. Raw data should be processed only one time.
C. Merging concerns combining datasets on the same observations to produce a result with more variables
D. None Of the above

Ans : B

35. What is the work of Data Architect?

A. utilize large data sets to gather information that meets their company's needs
B. work with businesses to determine the best usage of the information yielded from data
C. build data solutions that are optimized for performance and design applications
D. All of the above

Ans : C

36. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics
B. Machine Learning / Deep Learning
C. Data Wrangling
D. All of the above

Ans : D

37. Which of the following are correct component for data science?

A. Data Engineering
B. Advanced Computing
C. Domain expertise
D. All of the above

Ans : D

38. Which of the following is not a part of data science process?

A. Discovery
B. Model Planning
C. Communication Building
D. Operationalize

Ans : C

39. Which of the following are the Data Sources in data science?

A. Structured
B. Unstructured
C. Both A and B
D. None Of the above

Ans : C

40. Which of the following is not a application for data science?

A. Recommendation Systems
B. Image & Speech Recognition
C. Online Price Comparison
D. Privacy Checker

Ans : D

41. Point out the correct statement.

A. Raw data is original source of data
B. Preprocessed data is original source of data
C. Raw data is the data obtained after processing steps
D. None of the above

Ans : A

42. Which of the following is one of the key data science skills?

A. Statistics
B. Machine Learning
C. Data Visualization
D. All of the above

Ans : D

43. Which of the following is a key characteristic of a hacker?

A. Afraid to say they don't know the answer
B. Willing to find answers on their own
C. Not Willing to find answers on their own
D. All of the above

Ans : B

44. Raw data should be processed only one time.

A. True
B. False
C. Can be true or false
D. Can not say

Ans : B

45. Which of the following is the common goal of statistical modelling?

A. Inference
B. Summarizing
C. Subsetting
D. None of the above

Ans : A

46. Causal analysis is commonly applied to census data.

A. True
B. False
C. Can be true or false
D. Can not say

Ans : B

47. Which of the following model is usually a gold standard for data analysis?

A. Inferential
B. Descriptive
C. Causal
D. All of the above

Ans : C

48. Which of the following is a revision control system?

A. Git
B. Numpy
C. Scipy
D. Slidify

Ans : A

49. Which of the following step is performed by data scientist after acquiring the data?

A. Data Cleaning
B. Data Integration
C. Data Replication
D. All of the above

Ans : A

50. Which of the following focuses on the discovery of (previously) unknown properties on the data?

A. Data mining
B. BigData
C. Data wrangling
D. Machine Learning

Ans : A

Unit 2-Introduction to Data science

1. Which of the following can be used to create sub–samples using a maximum dissimilarity approach?
A. minDissim
B. maxDissim
C. inmaxDissim
D. All of the Mentioned
Ans :B

2. Which of the following can be used to impute data sets based only on information in the training set?
A. postprocess
B. preProcess
C. process
D. All of the Mentioned
Ans :B

3. Which of the following model model include a backwards elimination feature selection routine?
A. MCV
B. MARS
C. MCRS
D. All of the Mentioned
Ans :B

4. Which of the following is a categorical outcome?
A. RMSE
B. RSquared
C. Accuracy
D. All of the Mentioned
Ans :C

5. Which of the following function provides unsupervised prediction ?
A. cl_forecast
B. cl_nowcast
C. cl_precast
D. None of the Mentioned
Ans :D

6. What is the work of Data Architect?

A. utilize large data sets to gather information that meets their company's needs
B. work with businesses to determine the best usage of the information yielded from data
C. build data solutions that are optimized for performance and design applications
D. All of the above

Ans : C

7. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics
B. Machine Learning / Deep Learning
C. Data Wrangling
D. All of the above

Ans : D

8. Which of the following are correct component for data science?

A. Data Engineering
B. Advanced Computing
C. Domain expertise
D. All of the above

Ans : D

9. Which of the following is not a part of data science process?

A. Discovery
B. Model Planning
C. Communication Building
D. Operationalize

Ans : C

10. Which of the following are the Data Sources in data science?

A. Structured
B. Unstructured
C. Both A and B
D. None Of the above

Ans : C

11. What is true about Machine Learning?

A. Machine Learning (ML) is that field of computer science
B. ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method.
C. The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.
D. All of the above

Ans : D

12. ML is a field of AI consisting of learning algorithms that?

A. Improve their performance
B. At executing some task
C. Over time with experience
D. All of the above

Ans : D

13. $p \rightarrow 0q$ is not a?

A. hack clause
B. horn clause
C. structural clause
D. system clause

Ans : B

14. The action _____ of a robot arm specify to Place block A on block B.

A. STACK(A,B)
B. LIST(A,B)
C. QUEUE(A,B)
D. ARRAY(A,B)

Ans : A

15. A_____ begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written.

A. bottow-up parser
B. top parser
C. top-down parser
D. bottom parser

Ans : C

16. A model of language consists of the categories which does not include _____.

A. System Unit
B. structural units.
C. data units
D. empirical units

Ans : B

17. Different learning methods does not include?

A. Introduction
B. Analogy
C. Deduction
D. Memorization

Ans : A

18. The model will be trained with data in one single batch is known as ?

A. Batch learning
B. Offline learning
C. Both A and B
D. None of the above

Ans : C

19. Which of the following are ML methods?

A. based on human supervision
B. supervised Learning
C. semi-reinforcement Learning
D. All of the above

Ans : A

20. In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called ?

A. mini-batches
B. optimizedparameters
C. hyperparameters
D. superparameters

Ans : C

21. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

A. Decision Tree
B. Regression
C. Classification
D. Random Forest

Ans : D

22. To find the minimum or the maximum of a function, we set the gradient to zero because:

A. The value of the gradient at extrema of a function is always zero
B. Depends on the type of problem
C. Both A and B
D. None of the above

Ans : A

23. Which of the following is a disadvantage of decision trees?

A. Factor analysis
B. Decision trees are robust to outliers
C. Decision trees are prone to be overfit
D. None of the above

Ans : C

24. How do you handle missing or corrupted data in a dataset?

A. Drop missing rows or columns
B. Replace missing values with mean/median/mode
C. Assign a unique category to missing values
D. All of the above

Ans : D

25. When performing regression or classification, which of the following is the correct way to preprocess the data?

A. Normalize the data -> PCA -> training
B. PCA -> normalize PCA output -> training
C. Normalize the data -> PCA -> normalize PCA output -> training
D. None of the above

Ans : A

26. Which of the following statements about regularization is not correct?

A. Using too large a value of lambda can cause your hypothesis to underfit the data.
B. Using too large a value of lambda can cause your hypothesis to overfit the data
C. Using a very large value of lambda cannot hurt the performance of your hypothesis.
D. None of the above

Ans : D

27. Which of the following techniques can not be used for normalization in text mining?

A. Stemming
B. Lemmatization
C. Stop Word Removal
D. None of the above

Ans : C

28. 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 nonconvex shapes

A. 1 and 2
B. 2 and 3
C. 1 and 3
D. All of the above

Ans : D

29. Which of the following is a reasonable way to select the number of principal components "k"?

A. Choose k to be the smallest value so that at least 99% of the varinace is retained.
B. Choose k to be 99% of m (k = 0.99*m, rounded to the nearest integer).
C. Choose k to be the largest value so that 99% of the variance is retained.
D. Use the elbow method.

Ans : A

30. What is a sentence parser typically used for?

A. It is used to parse sentences to check if they are utf-8 compliant.
B. It is used to parse sentences to derive their most likely syntax tree structures.
C. It is used to parse sentences to assign POS tags to all tokens.
D. It is used to check if sentences can be parsed into meaningful tokens.

Ans : B

31. Data Analysis is a process of?

A. inspecting data
B. cleaning data
C. transforming data
D. All of the above

Ans : D

32. Which of the following is not a major data analysis approaches?

A. Data Mining
B. Predictive Intelligence
C. Business Intelligence
D. Text Analytics

Ans : B

33. How many main statistical methodologies are used in data analysis?

A. 2
B. 3
C. 4
D. 5

Ans : A

34. In descriptive statistics, data from the entire population or a sample is summarized with ?

A. integer descriptors
B. floating descriptors
C. numerical descriptors
D. decimal descriptors

Ans : C

35. Data Analysis is defined by the statistician?

A. William S.
B. Hans Peter Luhn
C. Gregory Piatetsky-Shapiro
D. John Tukey

Ans : D

36. Which of the following is true about hypothesis testing?

A. answering yes/no questions about the data
B. estimating numerical characteristics of the data
C. describing associations within the data
D. modeling relationships within the data

Ans : A

37. The goal of business intelligence is to allow easy interpretation of large volumes of data to identify new opportunities.

A. TRUE
B. FALSE
C. Can be true or false
D. Can not say

Ans : A

38. The branch of statistics which deals with development of particular statistical methods is classified as

A. industry statistics
B. economic statistics
C. applied statistics
D. applied statistics

Ans : D

39. Which of the following is true about regression analysis?

A. answering yes/no questions about the data
B. estimating numerical characteristics of the data
C. modeling relationships within the data
D. describing associations within the data

Ans : C

40. Text Analytics, also referred to as Text Mining?

A. TRUE
B. FALSE
C. Can be true or false
D. Can not say

Ans : A

41. What is true about Data Visualization?

A. Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts.
B. Data Visualization helps users in analyzing a large amount of data in a simpler way.
C. Data Visualization makes complex data more accessible, understandable, and usable.
D. All of the above

Ans : D

42. Data can be visualized using?

A. graphs
B. charts
C. maps
D. All of the above

Ans : D

43. Data visualization is also an element of the broader _____.

A. deliver presentation architecture
B. data presentation architecture
C. dataset presentation architecture
D. data process architecture

Ans : B

44. Which method shows hierarchical data in a nested format?

A. Treemaps
B. Scatter plots
C. Population pyramids
D. Area charts

Ans : A

45. Which is used to inference for 1 proportion using normal approx?

A. fisher.test()
B. chisq.test()

C. Lm.test()
D. prop.test()
Ans : D

46. Which is used to find the factor congruence coefficients?

A. factor.mosaicplot
B. factor.xyplot
C. factor.congruence
D. factor.cumsum
Ans : C

47. Which of the following is tool for checking normality?

A. qqline()
B. qline()
C. anova()
D. lm()
Ans : A

48. Which of the following is false?

A. data visualization include the ability to absorb information quickly
B. Data visualization is another form of visual art
C. Data visualization decrease the insights and take solwer decisions
D. None Of the above
Ans : C

49. Common use cases for data visualization include?

A. Politics
B. Sales and marketing
C. Healthcare
D. All of the above
Ans : D

50. Which of the following plots are often used for checking randomness in time series?
A. Autocausation
B. Autorank
C. Autocorrelation
D. None of the above
Ans : C

Unit 3-Introduction to Machine learning

1. All of the following accurately describe Hadoop, EXCEPT _____

A. Open-source
B. Real-time
C. Java-based
D. Distributed computing approach

Ans : B

2. _____ has the world's largest Hadoop cluster.

A. Apple
B. Datamatics
C. Facebook
D. None of the above

Ans : C

3. Facebook Tackles Big Data With _____ based on Hadoop.

A. Project Prism
B. Prism
C. Project Big
D. Project Data

Ans : A

4. _____ is general-purpose computing model and runtime system for distributed data analytics.

A. Mapreduce
B. Drill
C. Oozie
D. None of the above

Ans : A

5. The examination of large amounts of data to see what patterns or other useful information can be found is known as

A. Data examination
B. Information analysis
C. Big data analytics
D. Data analysis

Ans : C

6. Big data analysis does the following except?

A. Collects data
B. Spreads data
C. Organizes data
D. Analyzes data

Ans : D

7. What makes Big Data analysis difficult to optimize?

A. Big Data is not difficult to optimize
B. Both data and cost effective ways to mine data to make business sense out of it
C. The technology to mine data
D. None of the above

Ans : B

8. The new source of big data that will trigger a Big Data revolution in the years to come is?

A. Business transactions
B. Social media
C. Transactional data and sensor data
D. RDBMS

Ans : C

9. The unit of data that flows through a Flume agent is

A. Log
B. Row
C. Record
D. Event

Ans : D

10. Listed below are the three steps that are followed to deploy a Big Data Solution except

A. Data Processing
B. Data dissemination
C. Data Storage
D. Data Ingestion

Ans : B

11. What is true about Machine Learning?

A. Machine Learning (ML) is that field of computer science
B. ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method.
C. The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.
D. All of the above

Ans : D

12. ML is a field of AI consisting of learning algorithms that?

A. Improve their performance
B. At executing some task
C. Over time with experience
D. All of the above

Ans : D

13. p → 0q is not a?

A. hack clause
B. horn clause
C. structural clause
D. system clause

Ans : B

14. The action _____ of a robot arm specify to Place block A on block B.

A. STACK(A,B)
B. LIST(A,B)
C. QUEUE(A,B)
D. ARRAY(A,B)

Ans : A

15. A_____ begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written.

A. bottow-up parser
B. top parser
C. top-down parser
D. bottom parser

Ans : C

16. A model of language consists of the categories which does not include _____.

A. System Unit
B. structural units.
C. data units
D. empirical units

Ans : B

17. Different learning methods does not include?

A. Introduction
B. Analogy

C. Deduction
D. Memorization
Ans : A

18. The model will be trained with data in one single batch is known as ?

A. Batch learning
B. Offline learning
C. Both A and B
D. None of the above
Ans : C

19. Which of the following are ML methods?

A. based on human supervision
B. supervised Learning
C. semi-reinforcement Learning
D. All of the above
Ans : A

20. In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called ?

A. mini-batches
B. optimizedparameters
C. hyperparameters
D. superparameters
Ans : C

21. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

A. Decision Tree
B. Regression
C. Classification
D. Random Forest
Ans : D

22. To find the minimum or the maximum of a function, we set the gradient to zero because:

A. The value of the gradient at extrema of a function is always zero
B. Depends on the type of problem

C. Both A and B
D. None of the above
Ans : A

23. Which of the following is a disadvantage of decision trees?

A. Factor analysis
B. Decision trees are robust to outliers
C. Decision trees are prone to be overfit
D. None of the above
Ans : C

24. How do you handle missing or corrupted data in a dataset?

A. Drop missing rows or columns
B. Replace missing values with mean/median/mode
C. Assign a unique category to missing values
D. All of the above
Ans : D

25. When performing regression or classification, which of the following is the correct way to preprocess the data?

A. Normalize the data -> PCA -> training
B. PCA -> normalize PCA output -> training
C. Normalize the data -> PCA -> normalize PCA output -> training
D. None of the above
Ans : A

26. Which of the following statements about regularization is not correct?

A. Using too large a value of lambda can cause your hypothesis to underfit the data.
B. Using too large a value of lambda can cause your hypothesis to overfit the data
C. Using a very large value of lambda cannot hurt the performance of your hypothesis.
D. None of the above
Ans : D

27. Which of the following techniques can not be used for normalization in text mining?

A. Stemming
B. Lemmatization

C. Stop Word Removal
D. None of the above

Ans : C

28. 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 nonconvex shapes

A. 1 and 2
B. 2 and 3
C. 1 and 3
D. All of the above

Ans : D

29. Which of the following is a reasonable way to select the number of principal components "k"?

A. Choose k to be the smallest value so that at least 99% of the varinace is retained.
B. Choose k to be 99% of m (k = 0.99*m, rounded to the nearest integer).
C. Choose k to be the largest value so that 99% of the variance is retained.
D. Use the elbow method.

Ans : A

30. What is a sentence parser typically used for?

A. It is used to parse sentences to check if they are utf-8 compliant.
B. It is used to parse sentences to derive their most likely syntax tree structures.
C. It is used to parse sentences to assign POS tags to all tokens.
D. It is used to check if sentences can be parsed into meaningful tokens.

Ans : B

31. Data science is the process of diverse set of data through ?

A. organizing data
B. processing data
C. analysing data
D. All of the above

Ans : D

32. The modern conception of data science as an independent discipline is sometimes attributed to?

A. William S.
B. John McCarthy
C. Arthur Samuel
D. Satoshi Nakamoto

Ans : A

33. Which of the following language is used in Data science?

A. C
B. C++
C. R
D. Ruby

Ans : C

34. Which of the following is false?

A. Subsetting can be used to select and exclude variables and observations
B. Raw data should be processed only one time.
C. Merging concerns combining datasets on the same observations to produce a

result with more variables
D. None Of the above

Ans : B

35. What is the work of Data Architect?

A. utilize large data sets to gather information that meets their company's needs
B. work with businesses to determine the best usage of the information yielded from data
C. build data solutions that are optimized for performance and design applications
D. All of the above

Ans : C

36. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics
B. Machine Learning / Deep Learning
C. Data Wrangling
D. All of the above

Ans : D

37. Which of the following are correct component for data science?

A. Data Engineering
B. Advanced Computing
C. Domain expertise
D. All of the above

Ans : D

38. Which of the following is not a part of data science process?

A. Discovery
B. Model Planning

C. Communication Building
D. Operationalize

Ans : C

39. Which of the following are the Data Sources in data science?

A. Structured
B. Unstructured
C. Both A and B
D. None Of the above

Ans : C

40. Which of the following is not a application for data science?

A. Recommendation Systems
B. Image & Speech Recognition
C. Online Price Comparison
D. Privacy Checker

Ans : D

41. Point out the correct statement.

A. Raw data is original source of data
B. Preprocessed data is original source of data
C. Raw data is the data obtained after processing steps
D. None of the above

Ans : A

42. Which of the following is one of the key data science skills?

A. Statistics
B. Machine Learning
C. Data Visualization
D. All of the above

Ans : D

43. Which of the following is a key characteristic of a hacker?

A. Afraid to say they don't know the answer
B. Willing to find answers on their own

C. Not Willing to find answers on their own
D. All of the above
Ans : B

44. Raw data should be processed only one time.

A. True
B. False
C. Can be true or false
D. Can not say
Ans : B

45. Which of the following is the common goal of statistical modelling?

A. Inference
B. Summarizing
C. Subsetting
D. None of the above
Ans : A

46. Causal analysis is commonly applied to census data.

A. True
B. False
C. Can be true or false
D. Can not say
Ans : B

47. Which of the following model is usually a gold standard for data analysis?

A. Inferential
B. Descriptive
C. Causal
D. All of the above
Ans : C

48. Which of the following is a revision control system?

A. Git
B. Numpy
C. Scipy
D. Slidify

Ans : A

49. Which of the following step is performed by data scientist after acquiring the data?

A. Data Cleaning
B. Data Integration
C. Data Replication
D. All of the above

Ans : A

50. Which of the following focuses on the discovery of (previously) unknown properties on the data?

A. Data mining
B. BigData
C. Data wrangling
D. Machine Learning

Ans : A

Unit 4-Data Analytics with R/Weka Machine Learning

1. Which of the following can be used to create sub–samples using a maximum dissimilarity approach?
A. minDissim
B. maxDissim
C. inmaxDissim
D. All of the Mentioned
Ans :B

2. Which of the following can be used to impute data sets based only on information in the training set?
A. postprocess
B. preProcess
C. process
D. All of the Mentioned
Ans :B

3. Which of the following model model include a backwards elimination feature selection routine?
A. MCV
B. MARS
C. MCRS
D. All of the Mentioned
Ans :B

4. Which of the following is a categorical outcome?
A. RMSE
B. RSquared
C. Accuracy
D. All of the Mentioned
Ans :C

5. Which of the following function provides unsupervised prediction ?
A. cl_forecast
B. cl_nowcast
C. cl_precast
D. None of the Mentioned
Ans :D

6. What is the work of Data Architect?

A. utilize large data sets to gather information that meets their company's needs
B. work with businesses to determine the best usage of the information yielded from data
C. build data solutions that are optimized for performance and design applications
D. All of the above

Ans : C

7. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics
B. Machine Learning / Deep Learning
C. Data Wrangling
D. All of the above

Ans : D

8. Which of the following are correct component for data science?

A. Data Engineering
B. Advanced Computing
C. Domain expertise
D. All of the above

Ans : D

9. Which of the following is not a part of data science process?

A. Discovery
B. Model Planning
C. Communication Building
D. Operationalize

Ans : C

10. Which of the following are the Data Sources in data science?

A. Structured
B. Unstructured
C. Both A and B
D. None Of the above

Ans : C

11. What is true about Machine Learning?

A. Machine Learning (ML) is that field of computer science
B. ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method.
C. The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.
D. All of the above

Ans : D

12. ML is a field of AI consisting of learning algorithms that?

A. Improve their performance
B. At executing some task
C. Over time with experience
D. All of the above

Ans : D

13. $p \rightarrow 0q$ is not a?

A. hack clause
B. horn clause
C. structural clause
D. system clause

Ans : B

14. The action _____ of a robot arm specify to Place block A on block B.

A. STACK(A,B)
B. LIST(A,B)
C. QUEUE(A,B)
D. ARRAY(A,B)

Ans : A

15. A_____ begins by hypothesizing a sentence (the symbol S) and successively predicting lower level constituents until individual preterminal symbols are written.

A. bottow-up parser
B. top parser
C. top-down parser
D. bottom parser

Ans : C

16. Big data analysis does the following except?

A. Collects data
B. Spreads data
C. Organizes data
D. Analyzes data

Ans : D

17. What makes Big Data analysis difficult to optimize?

A. Big Data is not difficult to optimize
B. Both data and cost effective ways to mine data to make business sense out of it
C. The technology to mine data
D. None of the above

Ans : B

18. The new source of big data that will trigger a Big Data revolution in the years to come is?

A. Business transactions
B. Social media

C. Transactional data and sensor data
D. RDBMS
Ans : C

19. The unit of data that flows through a Flume agent is

A. Log
B. Row
C. Record
D. Event
Ans : D

20. Listed below are the three steps that are followed to deploy a Big Data Solution except

A. Data Processing
B. Data dissemination
C. Data Storage
D. Data Ingestion
Ans : B

21. Who popularized bigdata term?
A. John deere
B. John Mashey
C. johny Mashe
D. Jhon Mash
Ans : B

22. Numbers ,text, image, audio and video data is ____
A. Volume
B. Value
C. Varity
D. Variety
Ans : D

23. Real time data is _____.
A. Field
B. Primary Key
C. unique
D. record
Ans : C

24. _____ is the term that is used to describe data that is high volume , high velocity and /or high variety.
A. Analytics
B. Bigdata
C. Hadoop Data
D. Bigdata analytics
Ans : B

25. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?
A. Big data management and data mining
B. Data warehousing and business intelligence
C. Management of Hadoop clusters
D. Collecting and storing unstructured data
Ans : A

26. Point out the wrong statement.
A. Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data
B. Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data
C. The programming model, MapReduce, used by Hadoop is difficult to write and test
D. All of these
Ans : C

27. All of the following accurately describe Hadoop, EXCEPT _____
A. Open-source

B. Real-time
C. Java-based
D. Distributed computing approach
Ans: B

28. _____ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.
A.MapReduce
B.Mahout
C.Oozie
D.All of the mentioned
Ans: A

29. _____ has the world's largest Hadoop cluster.
A. Apple
B. Datamatics
C. Facebook
D. None of the mentioned
Ans:C

30. Facebook Tackles Big Data With _____ based on Hadoop.
A. 'Project Prism'
B. 'Prism'
C. 'Project Big'
D. 'Project Data'

Ans:A

31. Data science is the process of diverse set of data through ?

A. organizing data
B. processing data
C. analysing data
D. All of the above

Ans : D

32. The modern conception of data science as an independent discipline is sometimes attributed to?

A. William S.
B. John McCarthy
C. Arthur Samuel
D. Satoshi Nakamoto

Ans : A

33. Which of the following language is used in Data science?

A. C
B. C++
C. R
D. Ruby

Ans : C

34. Which of the following is false?

A. Subsetting can be used to select and exclude variables and observations
B. Raw data should be processed only one time.
C. Merging concerns combining datasets on the same observations to produce a result with more variables
D. None Of the above

Ans : B

35. What is the work of Data Architect?

A. utilize large data sets to gather information that meets their company's needs
B. work with businesses to determine the best usage of the information yielded from data
C. build data solutions that are optimized for performance and design applications
D. All of the above

Ans : C

35. What is the work of Data Architect?

A. utilize large data sets to gather information that meets their company's needs
B. work with businesses to determine the best usage of the information yielded from data
C. build data solutions that are optimized for performance and design applications
D. All of the above

Ans : C

36. Which of the following is correct skills for a Data Scientist?

A. Probability & Statistics
B. Machine Learning / Deep Learning
C. Data Wrangling
D. All of the above

Ans : D

37. Which of the following are correct component for data science?

A. Data Engineering
B. Advanced Computing
C. Domain expertise
D. All of the above

Ans : D

38. Which of the following is not a part of data science process?

A. Discovery
B. Model Planning
C. Communication Building
D. Operationalize

Ans : C

39. Which of the following are the Data Sources in data science?

A. Structured
B. Unstructured
C. Both A and B
D. None Of the above

Ans : C

40. Which of the following is not a application for data science?

A. Recommendation Systems
B. Image & Speech Recognition
C. Online Price Comparison
D. Privacy Checker

Ans : D

41. Point out the correct statement.

A. Raw data is original source of data
B. Preprocessed data is original source of data
C. Raw data is the data obtained after processing steps
D. None of the above

Ans : A

42. Which of the following is one of the key data science skills?

A. Statistics
B. Machine Learning
C. Data Visualization
D. All of the above

Ans : D

43. Which of the following is a key characteristic of a hacker?

A. Afraid to say they don't know the answer
B. Willing to find answers on their own
C. Not Willing to find answers on their own
D. All of the above

Ans : B

44. Raw data should be processed only one time.

A. True
B. False
C. Can be true or false
D. Can not say

Ans : B

45. Which of the following is the common goal of statistical modelling?

A. Inference
B. Summarizing
C. Subsetting
D. None of the above

Ans : A

46. _____ Programming language is dialect of S.

A. B
B. C
C. R
D. None of the above

Ans : C

47. Which of the following model is usually a gold standard for data analysis?

A. Inferential
B. Descriptive
C. Causal
D. All of the above

Ans : C

48. File containing R scripts end with extension _____.

A. .R
B. .S
C. .bigdata
D. All of the above

Ans : A

49. Which of the following is a subset of machine learning?

A. Numpy
B. SciPy
C. Deep Learning
D. All of the above

Ans : C

50. How many layers Deep learning algorithms are constructed?

A. 2
B. 3
C. 4
D. 5

Ans : B